Situated and Semantics-Aware Mixed Reality

Dissertation
zur Erlangung des Doktorgrades
- Dr. rer. nat -

der Fakultät für Informatik
der Universität Duisburg-Essen

vorgelegt von
Mohamed Kari

betreut durch
Prof. Dr. Reinhard Schütte
Lehrstuhl für Wirtschaftsinformatik und
integrierte Informationssysteme

Essen, Mai 2024

Betreuer:	Prof. Dr. Reinhard Schütte *Lehrstuhl für Wirtschaftsinformatik und integrierte Informationssyteme*
Gutachter:	Prof. Dr. Reinhard Schütte Prof. Dr. Volker Gruhn Prof. Dr. Stefan Eicker
Tag der mündlichen Prüfung:	19. August 2024

DOI: 10.17185/duepublico/82367
URN: urn:nbn:de:hbz:465-20240826-091044-7

Abstract

Computers moved from the basement onto the user's desk into their pocket and around their wrist. Not only did they physically converge to the user's space of action, but sensors such as CMOS, GPS, NFC, and LiDAR also began to provide access to the environment, thus situating them in the user's direct reality. However, the user and their computer exhibit fundamentally different perceptual processes, resulting in substantially different internal representations of reality.

Augmented and Mixed Reality systems, as the most advanced situated computing devices of today, are mostly centered around geometric representations of the world around them, such as planes, point clouds, or meshes. This way of thinking has proven useful in tackling classical problems including tracking or visual coherence. However, little attention has been directed toward a semantic and user-oriented understanding of the world and the specific situational parameters a user is facing. This is despite the fact that humans not only characterize their environment by geometries, but they also consider objects, the relationships between objects, their meanings, and the affordance of objects. Furthermore, humans observe other humans, how they interact with objects and other people, all together stimulating intent, desire, and behavior. The resulting gap between the human-perceived and the machine-perceived world impedes the computer's potential to seamlessly integrate with the user's reality. Instead, the computer continues to exist as a separate device, in its own reality, reliant on user input to align its functionality to the user's objectives.

This dissertation on Situated and Semantics-Aware Mixed Reality aims to get closer to Augmented and Mixed Reality experiences, applications, and interactions that address this gap and seamlessly blend with a user's space and mind. It aims to contribute to the integration of interactions and experiences with the physicality of the world based on creating an understanding of both the user and their environment in dynamic situations across space and time.

Method-wise, the research presented in this dissertation is concerned with the design, implementation, and evaluation of novel, distributed, semantics-aware, spatial, interactive systems, and, to this end, the application of state-of-the-art computer vision for scene reconstruction and understanding, together enabling Situated and Semantics-Aware Mixed Reality.

Zusammenfassung

Computer sind vom Keller auf den Schreibtisch der Nutzer gelangt, von dort in ihre Tasche und bis hin um ihr Handgelenk. Sie haben sich jedoch nicht nur physisch praktischen Einsatzbereichen angenähert, vielmehr haben Sensoren wie CMOS, GPS, NFC, und LiDAR auch informationellen Zugang zur Umgebung ermöglicht, und sind somit auch im unmittelbar relevanten Realitätsauschnitt des Nutzers situiert. Jedoch weisen ein menschlicher Nutzer und ein digitaler Rechner fundamentale Unterschiede in ihrem Wahrnehmungsprozess auf, die in unterschiedlichen internen Repräsentationen des Realweltauschnitts resultieren.

Augmented-Reality- und Mixed-Reality-Systeme als die fortschrittlichsten Instanzen dessen, was in dieser Dissertation als Situated-Computing-Paradigma verstanden wird, sind vorrangig auf geometrische Repräsentationen der Welt ausgerichtet, etwa in Form von Ebenen, Punktwolken, oder Polygonnetzen. Diese konventionelle Betrachtungsweise hat sich als nützlich erwiesen um klassische Probleme der Disziplin zu bearbeiten, etwa zur Eigenbewegungsverfolgung oder der Herstellung visueller Kohärenz in Mixed-Reality-Systemen. Ein semantisches oder nutzerorientiertes Verständnis der Welt und deren situativer Parameter, denen sich ein Nutzer gegenübersieht, haben weit weniger Aufmerksamkeit in der Disziplin erfahren. Dies steht jedoch der Tatsache entgegen, dass Menschen ihre Umgebung nicht mittels Geometrien charakterisieren, sondern vielmehr mittels Objekten, der Beziehung zwischen Objekten, deren Bedeutungen und ihrer funktionalen Verwendbarkeit. Sie erkennen andere Menschen, wie andere Menschen mit Objekten und wieder anderen Menschen interagieren, und reagieren mit Gedanken, Absichten, und Verhalten. Die resultierende Diskrepanz zwischen der menschlich wahrgenommenen und der maschinell wahrgenommenen Welt mindert das Potenzial von Computern zur nahtlosen Integration in die Realität des Nutzers. Stattdessen verbleibt der Rechner in konzeptuell isolierter Existenz und Abhängigkeit expliziter Nutzereingaben, um seine Funktionalität mit den Zielen des Nutzers in Einklang zu bringen.

Die vorliegende Dissertation zu *Situated and Semantics-Aware Mixed Reality* zielt darauf ab, diese Diskrepanz mittels Augmented- und Mixed-Reality-Experiences, -Anwendungen, und -Interaktionen, die sich nahtlos in Raum und Geist des Nutzers einbetten, zu adressieren. Es wird angestrebt einen Beitrag zur Integration von Experiences und Interaktionen mit der Physikalität der nutzerumgebenden Welt zu leisten, indem ein maschinelles, gleichwohl semantisch verankertes Verständnis vom Nutzer und der Umgebung in räumlichen und zeitlichen Facetten erlangt wird.

Methodisch ist die dieser Dissertation zugrundeliegende Forschung im Entwurf, der Implementierung und der Evaluation neuartiger, verteilter, semantisch informierter, räumlicher, interaktiver Systeme konstituiert, wobei auf der Computer-Vision-Forschung zur Szenenrekonstruktion und -erfassung aufgebaut wird, um zusammengenommen der Vision von Situated and Semantics-Aware Mixed Reality näherzukommen.

Publications

Parts of the research presented in this dissertation have appeared in the following publications:

Kari, M., Schütte, R., and Sodhi, R. Scene Responsiveness for Visuotactile Illusions in Mixed Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023), Best Paper Honorable Mention Award, pp. 1–15

Kari, M., and Holz, C. HandyCast: Phone-based Bimanual Input for Virtual Reality in Mobile and Space-Constrained Settings via Pose-and-Touch Transfer. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–15

Kari, M., Grosse-Puppendahl, T., Jagaciak, A., Bethge, D., Schütte, R., and Holz, C. SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology* (2021), Best Paper Award, pp. 118–133

Kari, M., Grosse-Puppendahl, T., Coelho, L. F., Fender, A. R., Bethge, D., Schütte, R., and Holz, C. TransforMR: Pose-Aware Object Substitution for Composing Alternate Mixed Realities. In *Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality* (2021), pp. 69–79

Patents

Parts of the research underlying this dissertation have been filed or granted as patents:

Kari, M., and Sodhi, R. Techniques and graphics-processing aspects for enabling scene responsiveness in mixed-reality environments, including by using situated digital twins, and systems and methods of use thereof, 2023. USPTO, US 63/494,449, filed and under review

Kari, M. Method for supporting a user of a motor vehicle, 2022. DPMA, DE10 2022 126 876.4

Kari, M., and Dreharov, N. Method for providing an evaluation system, 2022. DPMA, DE10 2022 120 936.9

Kari, M., and Kerl, P. Computer-implemented method for transmitting information about a headlight of a motor vehicle with multiple light sources, 2022. DPMA, DE10 2022 113 682.5

Kari, M., and Bethge, D. Method and device for monitoring objects, 2022. DPMA, DE10 2022 110 349.8

Kari, M., and Bethge, D. Method for operating an automated maneuvering system of a vehicle, 2022. DPMA, DE10 2022 105 245.1, filed and under review

Kari, M., and Isele, S. T. Vehicle and control device and procedures for operating the vehicle, 2022. DPMA, DE10 2022 105 155.2

Kari, M., and Bethge, D. Devices and methods for joint route guidance, 2021. DPMA, DE10 2021 130 939.5

Kari, M., and Thum, M. Method, system and computer program product for analyzing time series datasets of an entity, 2021. DPMA, DE10 2021 130 938.7

Kari, M., and Bethge, D. Procedure for determining spare parts requirements, 2021. DPMA, DE10 2021 129 094.5

Kari, M., Grosse-Puppendahl, T., Jagaciak, A., and Bethge, D. Method and system for scene-synchronous selection and playback of audio sequences for a motor vehicle, 2021. DPMA, DE10 2021 110 268.5

Acknowledgments

I thank my mother **Elke Kari** for your unconditional support in every regard, enabling me to focus on the pursuit of any goal I might have, for providing better paper reviews than a peer reviewer even before the paper is written, and for having great video ideas, and my father **Mohamed Kari** for always believing in me, for being an unshakable source of calmness, and for staying close despite any distance. Thank you.

I thank **Reinhard Schütte** for being the mentor and *Doktorvater* everyone deserves, inspiring me in every discussion, encouraging me to take the leap, helping me to think further without stopping before reaching questions about the universe or its atoms, believing in me, being lenient with me despite the many reasons not to, animating me to be dissatisfied, valuing quality, creativity, and understanding over quantity and reproduction, shaping my professional, scientific, and personal ideals, making my development your utmost priority, teaching me to focus on the horizon, giving me the impression I can understand the world and am able to contribute a few things to it, and sometimes giving me the highly satisfactory feeling we know a thing or two others have missed.

I thank **Felix Weber** for being a great friend and discussion partner, not only but also during long car rides from Berlin, Hamburg, Amsterdam, Münster, or Moers. I thank **Sarah Seufert**, **Tobias Wulfert**, **Michael Harr**, **Mareen Wienand**, **Dustin Syfuß**, **Hendrik Obertreis**, **Michel Muschkiet**, **Clemens Brackmann**, **Nadine Weiss** and **Thomas Kern** for covering my back countless times.

I thank **Volker Gruhn** for taking the time. I also thank **Stefan Eicker** for your time, **Erik Heimann** for your dedication to helping others and your support, and **Pete Schuler**, **Barbara Schiller**, and **Peter M. Schuler** for introducing me to the academic world.

Regarding my Porsche time

I thank **Katerina Kourti** for the joy you brought to our team. You are dearly missed.

I thank **David Bethge** for being the best PhD peer one can wish for. Thank you for conspiring together, and inspiring me, for taking every idea–as crazy as it might seem–seriously, and for helping me grow into the person that might achieve some of them.

I thank **Jochen Gross** for your impregnable trust, providing an unheard-of degree of autonomy, making my personal advancement your personal objective, and being the paragon of leadership.

I am grateful to **Luis Coelho** for being the best co-author I ever had. I thank you and **Philipp Hallgarten**, **Philipp Wolters**, **Constantin Patsch**, **Daniel Cantz**, **Jason Roy**, **Finn Weiler**, **Satiya Murugaboopathy**, and **Alexander Jagaciak** for making Emerging Tech Research what it was, growing from students to peers and teachers.

I thank **Tobias Grosse-Puppendahl** for founding Emerging Tech Research and making me a part of it, introducing me to the HCI community, valuing research excellence, providing autonomy,

and teaching me the importance of a good story. I thank **Johannes Immel** and **Tim Laudahn** for establishing a great team spirit.

I thank **Claudia Feiner** for being a friend and hero, demonstrating what fighting and winning an uphill battle means.

I thank **Simon Isele**, **Michael Herrmann**, and the many other PhD peers from the Porsche PhD community for a great time together, and the great help during the SoundsRide study. I thank **Frank-Steffen Walliser** and **Mattias Ulbrich** for fostering a company culture that enables creative and risky research and taking an interest in it.

I thank **Moritz Rabe** for being a great friend and colleague, bringing me not only to intellectual but also to phyiscal heights. I thank **Martin Mayer**, **Andrada Tatu**, and **Helge Silberhorn** for a great preparatory time for the PhD and for helping me become a well-rounded Porscheaner. I thank **Tetiana Aymelek**, **Marco Wiedner**, and **Maximilian Thum** for being the colleagues the company and I needed. Not sure many people get to work on the future of Porsche models as a side project. I thank **Matthias Zimmer** for being the best patent assessor the world has seen, and to the patent attorneys doing much of the grassroots work.

Regarding my ETH Zürich time

I thank **Christian Holz** for a transformative time in your wonderful group of extraordinary talent density, taking me in at the right phase, introducing me to many tools of the trade, helping me become a better computer scientist, redefining my understanding of impactful technical HCI research, correct capitalization in English, and the attribute "world-class", and inspiring me to think about my long-term research agenda.

I thank **Max Möbus** for the friendship, the always open door, surprising Tuesday energy, and wonderful discussions, **Tiffany Luong** for the couch spot and balancing coffee breaks, and **Paul Streli** for being a friend and night owl. I thank **Andreas Fender** for being a friend, being the most creative mind I know, conducting the kind of HCI research that the field needs most, inspiring me, and helping me get things right. I thank **Jiaxi Jiang**, **Manuel Meier**, **Björn Braun**, and **Yi Fei Cheng** for being brilliant colleagues. I am infinitely grateful to **Thomas Roberts**. Thanks for the wonderful time, may it be in Zurich in STD, CNB, places around Hardturm, the Ale House, or the Züri Bistro, at CHI in Hamburg, or at UIST in San Francisco.

I thank **Philipp Herholz** for being a friend and the most interesting person I know, always available to talk about anything, and for the always open door (I know what you'll say to this one). I thank **Luca Cavalli** for being a friend and **Iro Armeni** for providing me with options, for the coffee and for the talk opportunity. I thank **Marille Hahne** and **Jill Scott** for letting me live in the most beautiful place in Zurich and getting me out of a jam a few times.

Regarding my Apple time

I thank **Tom Runia** and **Vittorio Megaro** for moonshot ideas and down-to-earth guidance on how to execute them, getting closer to the vision step by step. Thank you for being great leaders, great scientists, great engineers, and great doers, trusting me and helping me become a bit more the researcher I aspire to be.

I thank **Marcel Germann** for being a great leader who always finds a way.

I thank **Evan Ntavelis** for being the peer I needed, and **Mohammad Mahdi Johari** for being a great colleague.

I thank **Fabio Maninchedda**, **Sebi Martin**, **Peter Kaufmann**, **Brian Amberg**, and the nearly countless other colleagues for being a team that leaves no doubt why Apple products are so good.

Regarding my Meta time

I thank **Raj Sodhi** for a magical time, for setting me up for a life-changing experience, for the incredible autonomy, for the high idea density in every discussion, the support, going for crazy ideas, and redefining what an internship can be.

I thank **Roko Parać** for your great support and your friendship. I thank **Rishi Hasra** for the many table tennis rounds, may it be Firefly, Origin, Dexter, Bellevue, MPK, Burlingame, Vancouver, Sausalito, or San Francisco skyscrapers. I thank **Michael Nebeling** for great discussions, may it be at IHOP or Mahoney's. I thank **Raejoon Jung** and **Roman Raedle** for being great listeners and supporters. I thank **Ran Tan** and **Ryo Takahashi** for being great official desk mates, and **Dimitri Bouche** for being a great inofficial deskmate.

I thank **Eric Whitmire, Neil Weiss, Purnendu, Jason Bou Kheir, Ran Tan, Jacqui Fashimpaur, Alec Pierce, Amy Karlson, Andre Levi, Evan Strasnic, Nathan Godwin, Felix Izarra, Sebastian Freitag, Tianyi Wang, Roger Boldu, Tanya Yonker, Zhenhong Hu, Joseph Faller**, and **Austin Lee** for your feedback on the Scene Responsiveness project and motivating me to keep going. I thank **Mike Miller** and **Alexander Stein** for your great help.

I thank **Hrvoje Benko** and **Wolf Kienzle** for providing a highly innovative research environment that can shape the future. Thank you for supporting me and helping me contribute to that future.

I thank **Michael Abrash** for your openness, some unforgettable quotes, great feedback, and for creating a globally unique place with special people, together defining the future.

I thank **Sean Keller**, **Andrew "Boz" Bosworth**, **Gabe Aul** for your feedback, catapulting my motivation the next level.

I also thank all the other nearly countless talents in Meta that make this place and my time with you so extraordinary.

Beyond

I thank **Parastoo Abtahi** for helping me take the next step.

I thank **Ruofei Du** for the discussions and for inspiring a lot of my work. I also thank the nearly countlessly many great fellow scholars in the human-computer interaction and computer vision community on whose foundations this work is built on.

I thank **Lukas Hilleheim** for introducing me to the world of computer science, teaching me my first programming language, and thus enabling everything I am doing now. I thank you and **Jan-Niklas Schnalke**, **Tobias Klöppner**, and **Julian Spee** for being great friends with countless care-free evenings playing Wizard.

I thank **Philipp Gabrys**, **Dave Mölenkamp**, **Kevin Siebert**, **Maurice Rothe**, **Dennis Zelecnik**, **Christopher Belusa**, and **Lukas Bücking** for being friends and providing a reliable reference "in the homeland". I thank **Wolf Thomsen** for stimulating discussions, from caching in CPU registers to corporate structures.

I thank **Daniel Wagner**, **Bettina Lüscher**, **Daniel Cremers**, **Philipp Häusser**, **Juliane Kronen**, **Avery Wang**, **Vera Kostiuk Busch**, **Hanna Shvindina**, **Rainer Storb**, **Manuel Bewarder**, and **Gloria Mark** for helping me practice asking the right questions.

I thank **Elisabeth Radziejowska** for teaching me life and keyboard lessons, the latter equipping me with the ability to just focus on something else, for ten minutes or a day–who knows when playing.

Contents

List of Figures

List of Tables

Acronyms

3D	3-dimensional
AI	Artificial Intelligence
AR	Augmented Reality
DOF	Degrees of Freedom
ECG	Electrocardiography/Electrocardiogram
EMG	Electromyography
FOV	Field of View
GPS	Global Positioning System
HCI	Human-Computer Interaction
HMD	Head-Mounted Display
IK	Inverse Kinematics
IMU	Inertial Measurement Unit
IR	Infrared
LiDAR	Light Detection and Ranging
LLM	Large Language Model
ML	Machine Learning
MR	Mixed Reality
PPD	Pixels per Degree
PPG	Photoplethysmogram
RGB	Red-Green-Blue
SAR	Spatial Augmented Reality
SLAM	Simultaneous Localization and Mapping
UI	User Interface
UWB	Ultra-Wideband
VIO	Visual-Inertial Odometry
VR	Virtual Reality
XR	Extended Reality
xR	"Anything" Reality

1 Exposition

1.1 From Mainframe Computing to Situated Computing

Figure 1.1: Evolution of commercially relevant computing platforms. Computers moved ⓐ from the basement ⓑ onto the user's desk and ⓒ into their pockets, ⓓ around their wrists and ⓔ in front of their eyes. Note that each computing platform in this evolutionary process complements the previously existing computing platforms, rather than replacing them. Also note the specialization relationship between the platforms: mobile computers are a specialization of personal computers, wearable computers are special mobile computers, and spatial computers are special wearable computers.

Computers moved from the basement onto the user's desk and into their pockets, around their wrists and in front of their eyes (Figure 1.1). Miniaturization, portability, and connectivity have allowed computers to physically converge to the user's space of action. Rather than forcing the user to change their location, the user gained the ability to interact with their computing devices *in loco*.

Simultaneously, devices started to be equipped with systems that enabled sensing the user and their environment (Figure 1.2). While the first mainframe computers in the 1950s, such as the UNIVAC I [79], only offered a keyboard to "sense" the user without any sensory interface to the environment at all, personal computers such as the Macintosh added a user-facing web camera (in the form of the Connectix QuickCam in 1994 [365]), a capacitive touch sensor (in the form of the Cirque GlidePoint trackpad in 1994 [265, p. 139]), and a microphone (in the form of the PlainTalk microphone in 1993 [326]). Mobile computers such as the iPhone [8] further included a structured light scanner, a proximity sensor, multiple ambient light sensors, an inertial measurement unit (IMU), a compass, a barometer, a light detection and ranging (LiDAR) scanner,

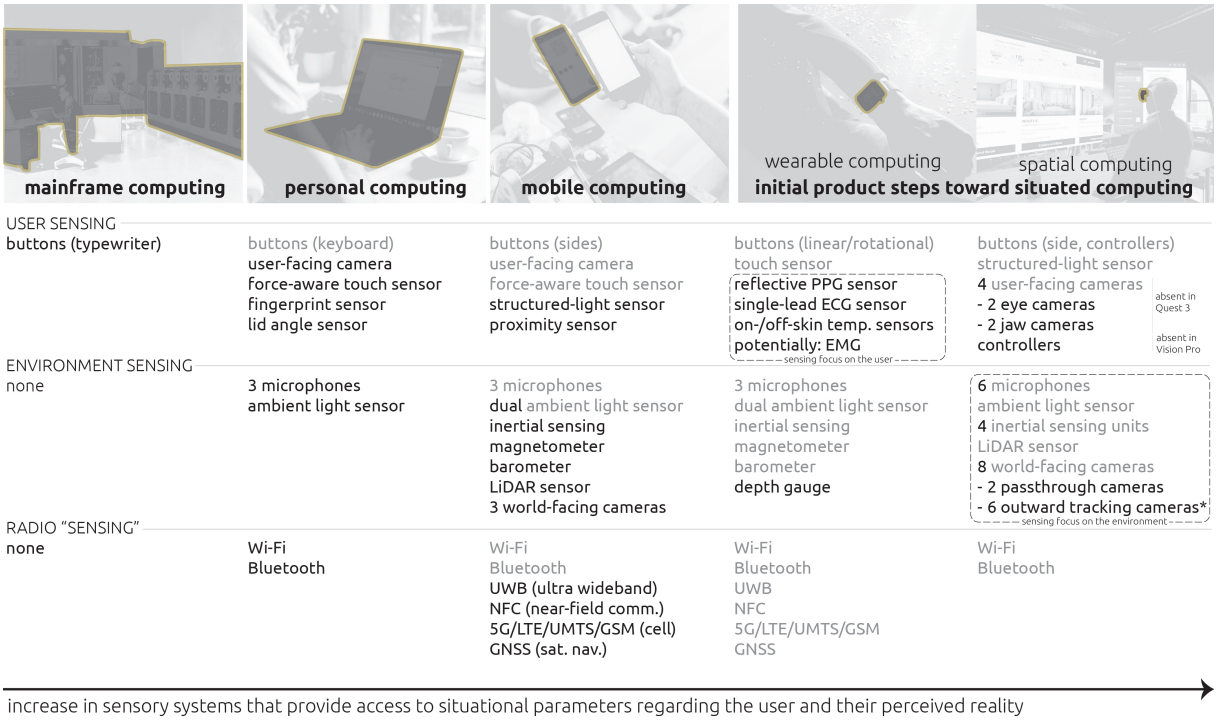


Figure 1.2: Evolution of sensors in commercially relevant computing platforms.

additional microphones, an array of world-facing cameras, and a variety of radio modules such as a global positioning system (GPS) for global and an ultra-wideband (UWB) module for local positioning in their package. Wearable computers such as the Apple Watch [6] moreover added physiological sensory systems for photoplethysmograms (PPG) and electrocardiograms (ECG) for sensing cardiovascular activity in the user. Research not only envisions the integration of additional sensors for electromyography (EMG) [70] but the advancement from wearable computing to body-integrated computing [241], where brain implants [337] are only the most conspicuous–and maybe even least original–example.

The progression across these computing platforms is characterized by increasing integration of sensing technologies that make the computing device situationally aware of the user and the user's direct reality, even envisioned in an "always-on" fashion. The device is not only at the user's location, it is part of the user's situation respecting situational parameters beyond the location. In turn, this offers users the ability to interact with the device *in situ*.

Mixed Reality (MR) headsets, both in the research and product landscape, offer to further advance the state-of-the-art of situated computing.

On the input side, they offer remarkable *sensing capabilities*, i.e., capabilities to transduce physical phenomena to electrical signals: By boosting the number of forward-facing cameras, e.g., to 6 cameras in the case of Apple Vision Pro [9], featuring a LiDAR scanner, and packaging multiple IMUs [148], MR headsets have significantly elevated capabilities of obtaining interpretable high-resolution data about the local physical 3-dimensional (3D) environment. Algorithms for visual-inertial odometry (VIO) [93; 94] and visual-inertial simultaneous localization and mapping (SLAM) on this high-resolution input cannot only provide view poses with 6 degrees of freedom (DOF) at high accuracy but also dense representations of the environment.

On the output side, MR headsets offer comprehensive *rendering capabilities* for visual and audi-

tive signals, i.e., capabilities of transducing electrical signals to physical phenomena: Taking the example of Apple Vision Pro, each display in front of the eyes features as many as 3660x3200 pixels at 40 pixels per degree (PPD) of field of view [146], thereby yielding an increase of an order of magnitude since the 1990s [263] on the path toward retinal resolution at 60 PPD [388].

These advanced sensing and rendering capabilities take a large stride toward the vision of Augmented Reality (AR), canonically defined as the visual real-time integration of virtual 3D objects into a physical 3D environment [17]. Estimated view poses enable view-dependent, world-locked rendering for the spatially registered virtual scene. Reconstructed 3D geometries increase visual coherency between the virtual and the physical scene for the plausible rendering of virtual occlusions, collisions, and lighting behaviors. Both together enable an estimation of the warp field required to compute the perspectively corrected passthrough-background layer for the displays from the RGB (red-green-blue) camera input. This geometrical awareness has motivated their denomination as "spatial computing devices"–also before the term was popularized by vendors [355, p. 1].

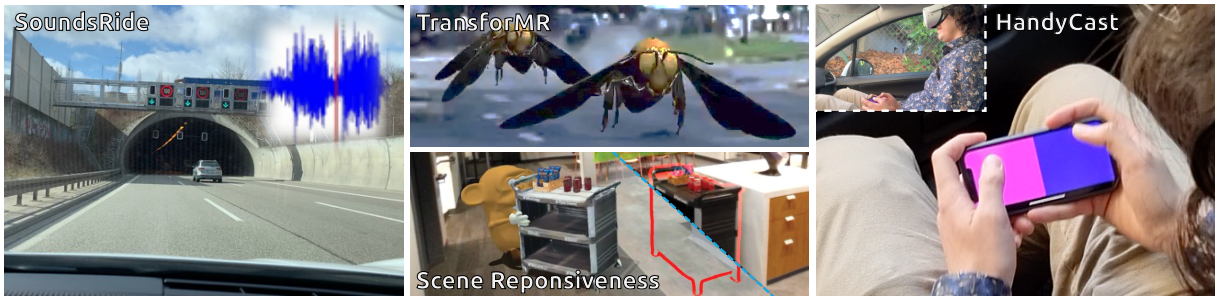
However, when complemented with technologies that also allow to *interpret* the user and their environment, MR devices far exceed the potential of a "view-adaptive stereoscopic passthrough display". On the one hand, inward-facing infra-red IR cameras for eye tracking and face tracking, structured-light sensing for facial reconstruction, downward cameras for hand and body pose tracking, and additional body-worn or body-integrated sensors promise to provide data for the reconstruction of biomechanical, physiological, and even mental properties of the user. On the other hand, advances in computational hardware as well as in artificial intelligence (AI) and machine learning (ML) allow to dramatically increase the opportunities for processing the signals they sense. Therefore, while these headsets of today materialize impressive AR or "spatial computing" devices, they motivate a broader vision of situated computing.

1.2 Beyond the 3rd Dimension: Mixed Reality from a Situated Computing Perspective

From our perspective of situated computing, MR comprises systems that sense physical properties of the user and their environment to render virtual signals in a direct relationship to inferred situational properties. This perspective pushes the traditional characteristics, often associated with such systems, into the background.

First, it is *modality-agnostic* and pushes the visual property into the background. Instead, it emphasizes the targeted real-time integration of any kind of signal into the physically pre-existing entirety of user-perceivable phenomena in whatever ways are meaningful and needed in the specific situation. While such an integration can and often will be visual, alternative mixes between virtual and physical signals are also captured. For example, this understanding allows to argue that MR can be equally, if not more, useful to members of the blind and visually impaired community [387] than to humans with full vision.

Second, it is *user-perception-oriented* and pushes the augmentation property into the background. A mechanism for noise-canceling headphones that blocks all ambient sounds and voices from a soundscape, and only passes through the extracted voice of a select subject to the headphone driver [336] would be perceived as a signal removal procedure from the perspective of the user. A similar approach, where the focal speech is also translated in real-time, would be perceived as a signal alteration rather than an augmentation. This understanding includes modulation or transduction of physical signals in any purposeful way, rather than only focusing on



ⓐ beyond the visual sense ⓑ beyond augmentation ⓒ beyond standard form factors

Figure 1.3: Mixed Reality beyond adding a third dimension. From a situated computing perspective, Mixed Reality offers to provide virtual signals beyond traditional visual augmentations in a spatially registered 3D user interface. ⓐ In the SoundsRide system, presented in this dissertation, virtual content is situated in the user's perceivable reality beyond the visual sense. ⓑ With the dissertation's contributions of TransforMR (top) and Scene Responsiveness (bottom), virtual content is not only overlayed as in classical AR but physical objects are removed for subsequent situated substitution or manipulation. ⓒ In the HandyCast system, we replace traditional dedicated controllers for mixed reality with a smartphone for situated on-the-go usage in physically space-constrained environments.

augmentations. It also motivates the use of the MR term over the AR term in this dissertation.

Third, it is *functionally oriented* and pushes the form factor into the background. Rather than coupling the notion of MR to the headset form factor, it couples it more generally to input and output devices that provide insights into a user's situation and allow to place signals therein. For example, on the input side, while headsets of today employ optical hand tracking with headset-integrated cameras, research [121] and products envision optical hand tracker systems worn beyond the head [180]. To realize more subtle applications, much research has been also dedicated to input sensing without cameras [192; 355], e.g., by instrumenting the wrist, the whole hand, or individual fingers.

Figure 1.3 gives an insight into the contributions of this dissertation from this situated-computing perspective, emphasizing the situational aspects of the user interaction. At the same time, and most importantly, this perspective moves the need for a semantic interpretation of situational parameters into the foreground.

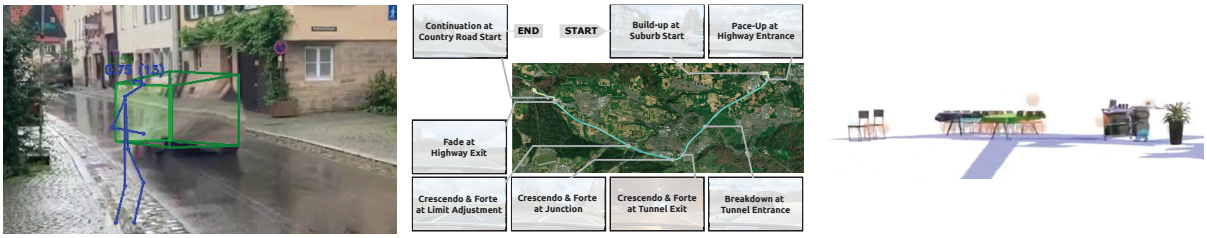
1.3 The Need for Semantic Awareness in Situated Mixed Reality

Today's commercially available MR systems are mostly centered around geometrical representations of the surrounding environment, such as point clouds or polygonal surface meshes. However, human users not only characterize the world around them by geometries, but instead reason about objects, the relationships between objects, their meanings, and the affordance of objects and surfaces [103]. They observe other humans, how these other humans interact with objects and even again with other humans, all together stimulating impressions, thought, intent, and behavior [143, p. 16].

This misalignment between the human-perceived and the machine-perceived world inhibits the purposeful linkage between virtuality and physicality. The computer continues to exist in a disconnected reality, reliant on user input to align its functionality to the user's objectives.

The remedy lies in the *semantic awareness* of the MR system.

Questions of semantics have been discussed in philosophy, linguistics, psychology, and com-



ⓐ awareness of subjects and objects ⓑ awareness of sound affordances ⓒ awareness of actor-object affordance

Figure 1.4: Semantic awareness in Mixed Reality. ⓐ TransforMR is aware of subjects and objects, ⓑ SoundsRide is aware of sound affordances, and ⓒ the Scene Responsiveness subsystems are aware of objects actor-object affordance.

puter science with its subfields of artificial intelligence, natural language processing, knowledge representation, computer vision, software engineering, data engineering, information systems engineering, formal language design, and human-computer interaction. In our practically oriented understanding, semantic awareness of an MR system refers to its property of representing reality with concepts that correspond with the concepts of a presumed mental model of reality of a user. This intention of correspondence between the machine's digital model and the user's mental model aims to enable the purposeful rendering of virtual signals that are meaningful to the user with their situationally specific perception and interpretation of physicality. Figure 1.4 provides an overview of the representations underlying the semantic awareness of the systems presented in this dissertation.

Many of the sensors of an MR system are exposed to the same or similar physical phenomena to those that the user is exposed to. This gives rise to the distant hope that "the right software" could computationally reproduce a model of reality in real-time that corresponds with the user's model of reality in wide parts. While advances in the methods of artificial intelligence, particularly in machine learning, computer vision, and visual language modeling, renew this hope, today, these methods remain task-specific tools, far from exhibiting human-like "world understanding". Therefore, AI can play a pivotal role in extracting and contributing specific situational information in real-time, while the preceding design of the system, the design of underlying conceptual models in anticipation of the user's mental models for the intended extent of situations, and the scope of AI integration remains in the responsibility and deliberation of the system creators such as designers, engineers, and researchers.

1.4 Contributions and Structure

This dissertation aims to get closer to the vision of MR experiences, interactions, and applications that seamlessly blend in with the user's space and mind. To this end, its *overarching contribution* lies in the invention, realization, and study of systems with the following properties:

1. exhibiting and leveraging **large-scale yet high-resolution semantic awareness** of the environment
2. enabling engaging and **situated mixed reality experiences beyond single-object augmentations** that
   (a) are derived from and deeply integrated with the physical scene semantics,
   (b) maintain the user's perception of select aspects of physicality, and
   (c) adapt to the environment or the user,
3. operable on general-purpose hardware **without requiring specialized devices** on the user or in the environment.

In **Chapter 2**, we unfold our *framework contribution* of situated and semantics-aware MR in more detail and position it against the background of related work. In particular, we present our higher-order reality model as a conceptual framework to discuss situated and semantics-aware MR systems.

In each of the subsequent chapters, we detail the design, implementation, and user-study-based evaluation of situated and semantics-aware MR systems with their underlying *technical, conceptual, and empirical contributions*. In the remainder of this section, we give an overview of each system, and relate them to the properties of semantic awareness, situatedness, and general-purpose hardware usage as outlined above.

In **Chapter 3**, we present SoundsRide with its *awareness of sound affordances* for car drives, spanning tens of kilometers in scale, yet audio-augmented with scene-adaptive music that is synchronized at sub-second precision. In SoundsRide, meaningful events in the music, such as beat drops or song transitions, are *spatially and temporally situated* with meaningful events along the route, such as highway entrances or tunnel exits. Situational changes, resulting from dynamics in user behavior or environmental conditions, are handled with new techniques for predictive and corrective mix generation. Our system prototype runs *without specialized hardware* in a client-server setup involving a real-time music mixing server on a laptop computer and a GPS-based localization and prediction client app on a smartphone. In a technical evaluation and in a user evaluation with 8 participants, we gained insights into the system's perceptual qualities, in particular its synchronicity, affordance noticeability, and mix artifact perceptibility.

In the two subsequent chapters, we draw inspiration from the metaphysical ideas of *parallel worlds* and *possible worlds* [312].

In **Chapter 4**, we present TransforMR with its *awareness of object semantics and their 3D poses as well as humans and their articulated 18-keypoint body poses* for the automatic composition of parallel worlds, well interpretable by users. Through TransforMR's notion of pose-aware object substitution, *semantic simulacra are situated* in the physical scene, taking the place, function, and semantic context of physical counterpart objects. Despite a comprehensive and compute-intensive real-time computer vision and graphics pipeline, that performs 3D object detection, 3D body pose estimation, 2D semantic instance segmentation, 2D inpainting, and 3D rendering, we can achieve interactive frame rates by means of cloud offloading and our proposed approach for pipeline parallelism. The user requires just a mobile smartphone and network access *without any dedicated hardware* to experience parallel worlds in open-ended, unprepared, and previously unseen environments.

In **Chapter 5**, we further push the conceptual and technical limits of scene integration in MR with Scene Responsiveness, enabling to experience other possible worlds that are deeply integrated with and arguably indistinguishable from physical reality. Scene Responsiveness leverages *comprehensive awareness of geometric, photometric, and semantic scene information*, all together represented in exhaustive and integrated digital space and object twins, up to the level of object-specific interaction affordances with their respective affordance features: Scene Responsiveness not only incorporates the knowledge that there is a chair, but also where it is, how it is oriented, at which height the seating surface is, whether it has armrests and where they are, which structural components afford grasping, how the hand is oriented when grasping, how much it weighs, and how its weight impacts the way a human would move it, e.g., carry it in front of the body or drag it behind. Scene Responsiveness makes use of this awareness to significantly deepen the illusion of MR by facilitating physically and semantically convincing in-situ manipulations of objects. The RealityToggle subsystem provides *situated space twins*

and *situated object twins* as needed, based on its object virtualization concept that enables changing the "reality states of objects", from physical to hidden to virtualized. The Spielberg subsystem provides *situated characters* and *situated avatars*. Rather than simply rendering a floating 3D head avatar into the physical scene, or positioning a standing avatar on the floor, situated characters and situated avatars are meaningfully integrated into the physical scene, e.g., they seem to sit on a physical couch, push physical buttons, or open physical doors. Spielberg also enables adaptivity to dynamic user behavior through different techniques that ensure visuotactile consistency for end-to-end illusionary experiences across sensing modalities. Scene Responsiveness requires *no specialized hardware* beyond a consumer-grade MR headset during the experience, and a LiDAR-enabled smartphone and a computer beforehand as part of the TwinBuilder subsystem.

In **Chapter 6**, we present HandyCast, shifting the emphasis from semantics-aware and *situated output* to *situated input*. In the analogy to early computing and today's desktop computing, where users must enter a dedicated room or space to use their computer (e.g., their desk), many MR and, in particular, many VR applications still require the user to enter a dedicated empty area to use their VR headset. This empty area ensures the user's unobstructed maneuverability when walking or even when just reaching out. However, such a spaciousness requirement is incongruous with in-situ interactions, i.e., interactions in any situation of the user's discretion. For example, many out-of-home situations, such as commuting, waiting, or being in an office only afford seated interaction with very limited space. HandyCast enables *situated input* by introducing space-efficient pose-and-touch transfer functions. It retrofits a smartphone as a 6 DoF + 2x2 DoF controller and is therefore operable *without dedicated hardware*.

Chapter 7 concludes this disseration with two perspective on the future. First, we consider implications derived from the observations made as part of this dissertation along three distinct phases. Finally, against the backdrop of the contributions of this dissertation, we revisit the vision of situated and semantics-aware MR systems.

2

Background and Related Work

2.1 Preliminary Epistemological Remarks

2.1.1 Balancing the Fine Line of Scientific Justification

Among the classical descriptive, explicative, and constructive objectives of science [381], the guiding scientific objective of this dissertation lies in the construction. However, in contrast to construction-oriented research that investigates, for example, operational details of the fabrication process of an electric motor, this specific dissertation already touches in its title as well as in its premise the simultaneously basic, yet complex concepts of reality and even realit*ies*, worlds, universes, physicality, perception, semantics, understanding, and the like.

Generally, research contributions tend to either be positioned against the backdrop of an existing, in itself more or less coherent set of terminological and logical presuppositions[1] (e.g., the electric-motor dissertation example), or, alternatively, revisit the primary and secondary sources inspiring an assumed formed consensus to then consolidate them differently under its own purpose and perspective.

The first approach seems misdirected for this dissertation. We object to the notion of "importing" any single, explicitly pre-existing set of terminological and logical presuppositions. As will be shown in the respective chapters, some of the concepts, contributed as part of this dissertation, aim to break with conventional terminology. For example, the Scene Responsiveness work presented in Chapter 5 introduces the notion of "toggling the reality state of an object", thus blurring the lines between the concepts classically subsumed in the fields of Virtual Reality and Augmented Reality.

However, the second approach of aiming to revisit the primary and secondary sources regarding the aforementioned terms of reality, worlds, universes, physicality, etc. seems like a fatuous endeavor. The reason lies not only in the sheer scale of the philosophical academic corpus.

[1] Also cf. the ontological, epistemological, and methodological basis decided upon within an observed discipline's paradigm as discussed by Zelewski [381]

Instead, even when focusing on just the subset of maybe the most seminal sources in the respective fields, one is faced with a plethora of problems such as insurmountable conflicts between prevailing perspectives, potentially as drastic as theoretical incommensurability [298, p. 165], the need to consider the "conceptual phylogenesis" of these primary sources over time as they evolve into different schools of thought, the need to consult secondary resources as a reference for understanding the primary sources, paired with the problem that the secondary sources in themselves are colored with a subjective scientific stance. Of course, more generally, this is the non-lamentable "scientist's fate" but the challenge to provide a consolidated understanding of nature as part of a construction-oriented research undertaking is considerable.

In light of these considerations, we follow Albert's position [2, p. 18] that the lack of the "absolute foundation"[2] is a logical necessity as a result of the Münchhausen trilemma. Consequently, we argue that the value of this work, especially against the backdrop of its construction-oriented ambition, lies not in an ill-fated pursuit of providing an absolute foundation, but is instead rather found in providing plausible, purposeful, and coherent views of statements and concepts that inform useful constructions. In addition, we argue that our pragmatic orientation allows for the freedom to draw inspiration from different scientific movements that might inspire a useful artifact without fully committing to a single school of thought. For example, we argue that it is permissible to draw inspiration from the metaphysical notion of *possible worlds* to draw inspirations for some of our ideas and terms, even without reviewing the full metaphysical discourse or providing a consolidated view thereof, as long as our interpretation allows consistent integration into our own structure of propositions, even if these individual inspirations are otherwise to be interpreted in their larger scientific framework.

2.1.2 Epistemological Positions

In this dissertation, epistemological positions are of twofold importance, namely on a meta-level and on an object level. Regarding the meta-level, epistemology provides a backdrop against which this scientific *text* and the principles that govern its writing and mode of knowledge acquisition can be interpreted. Regarding the object level, epistemology provides an entry point for thinking about the principles that govern the actual user experience when using the technical system implementations that underlie this research text. To clarify the latter point, take a user employing a VR headset to explore a virtual environment featuring virtual objects. As we will describe in more detail throughout this chapter, creating plausible *sense data*–in a more intuitive understanding of that wording–is at the core of a purposeful VR illusion. At the same time, sense data–as a technical term–has been proposed, discussed, and criticized epistemologically by philosophers such as Russell [286, chapter 1], Wittgenstein [364, proposition 486], Ogden [256, p. 49], or Wright [366]. A recent technophilosophical perspective on VR has been presented by Chalmers [50, pp. 148-237].

For both the meta-level and the object level, this dissertation will most often find itself in the vicinity of *moderate constructivism* [298, p. 163; 296], refusing both the notion of naive realism as well as radical constructivism. It assumes the existence of physicality beyond the individual yet acknowledges–and, in fact, when it comes to our system implementations, leverages–the far-reaching implications of individual perception and cognition. At the same time, even though certainly irreconcilable when all implications are considered together, it appreciates ideas found in logical positivism as represented by Wittgenstein [363; 14], critical realism [298, p. 163],

[2] Albert's "Letztbegründung" [2, p. xvi]

critical rationalism [296, p. 221] as represented by Popper, empiricism as represented by Quine [334], cognitivism as it is underlying many works in the Human-Computer Interaction community [122], and as evident from this very sentence, therefore also sees value in Feyerabend's "Anything goes" [89, p. 14].

2.2 Related Concepts in Literature

2.2.1 Overview

While this dissertation presents its own conceptualization of *Situated and Semantics-Aware Mixed Reality*, many of its underlying ideas have been the subject of decade-spanning research in the field of human-computer interaction and related disciplines within and beyond computer science. Some of the ideas in the field are advocated by individual researchers, while other ideas have evolved into wider, more or less well-delineated umbrella terms. This background chapter will take a wider perspective while more concrete aspects are discussed in dedicated sections of the subsequent chapters as needed.

2.2.2 Virtual Reality

Virtual Reality (VR) systems have been defined by their ability "to immerse a user in a synthetically generated environment" [264]. Immersion is understood as the sense of being there [263] and generally pursued by means of a stereoscopic, head-mounted display that creates the illusion of a 3D surrounding world [279]. Other principles such as projector-equipped dedicated rooms have also been proposed [69]. Since its inception in the 1960s [319], producing a visual signal that provides a convincing illusion for VR took a paramount role in VR research [40]. At the same time, even early VR prototypes outlined the simulation of auditory and haptic signals by means of binaural auditory displays [90] and haptic displays [41], from gloves to body suits [301].

2.2.3 Augmented Reality

Canonically, Augmented Reality (AR) systems [91; 83; 347; 135] have been defined by their ability to visually integrate virtual 3D objects into a 3D real environment in real-time [17; 16; 29; 31]. Other frequently found phrasings define AR as the ability to superimpose or overlay computer-generated imagery on the real world [30; 276] by adding "virtual information to a user's sensory perceptions" [84]. "[T]he primacy of the physical world" [347] is generally emphasized as the demarcation line to VR [83; 31]. The combined scene is presented to the user by means of a display. Such displays include spatial displays, handheld displays, and head-worn displays [16; 29, p. 140; 31, p. 72].

Spatial displays, discussed in the field of Spatial Augmented Reality (SAR) [31], present virtual information by projecting light directly onto physical surfaces in the environment using projectors. For convincing augmentations that are coherently integrated into physical space, the projection system fundamentally needs to account for the user's head position for view dependency, and for the non-planarity of physical surfaces [31, p. 88].

Handheld displays, nowadays most often in the form of smartphones, employ a video-passthrough approach, where an outward-facing camera captures a video feed, which is augmented by the AR

software in real-time, and then displayed on the inward-facing device screen [31, p. 88]. This approach allows the display to act as a "magic lens" [28] that can reveal the AR scene depending on the pose-tracked display's position and orientation. However, generally, this also results in the dual-view problem [68] where the AR scene is rendered and displayed from the perspective of the camera, rather than the perspective of the user's eyes. Approaches have been presented [68; 234] that enable user-perspective rendering on handheld devices, however, the fundamental limitation that only a fraction of the user's field of view can be augmented with a small display held at arm's length remains.

Head-Mounted Displays (HMDs) or head-worn displays generally feature two optics-and-display elements, positioned in front of each eye, to leverage stereoscopic vision in the human visual system. Each display composes a view that visually merges the physical scene with a virtual scene. Different display principles can be distinguished [84]. *Optical seethrough displays* feature a light engine, positioned in the housing of the HMD, that emits light of a 2D raster image corresponding to the virtual scene [76]. This light is coupled into a waveguide, i.e., a more or less transparent material in front of each eye, that guides the light to predefined extraction points, arranged in a grating, where the light can exit the waveguide toward the user's respective eye. The waveguide's transparency results in a combination of the artificially created light rays and the ambient light rays [75]. The most prominent representative of this principle is Microsoft HoloLens 2, however, it has also been employed by manufacturers such as Magic Leap, WaveOptics, Dispelix, or DigiLens [76]. *Video passthrough displays*, while researched for decades [84], have only found their way into popular, comprehensive products more recently, e.g., with the launch of the Meta Quest Pro, the Meta Quest 3, the Varjo XR-3, and the Apple Vision Pro. In a naive approach, an eye-specific camera streams each captured frame to the corresponding eye-specific display. However, such a naive approach disregards the visual displacement between the user's eyes and the cameras. The potentially relatively substantial difference between the user's anatomical inter-pupil distance and the distance between the headset's camera positions, as well as the forward-shifted camera position as a result of the device package, leads to perceptual inconsistencies [112, p. 8]. Therefore, modern products with video passthrough displays employ perspective correction through means of reprojection [52; 368]. Apart from the objective of minimizing reprojection-caused visual artifacts, such a processing pipeline also introduces high demands on computational efficiency to maintain the natural perception of the physical scene. The classical VR quality objective of minimizing motion-to-photon latency is complemented by the objective of minimizing photon-to-photon latency, i.e., the duration from light entering the camera to the corresponding signal exiting the display [111]. A major goal, pursued by research in MR [392], lies in the increase of brightness, field-of-view, resolution while reducing weight. One could summarize it as follows: *more nits, more degs, more pixels, but less atoms.* Past and current *smartglasses*, such as Google Glass, Amazon Echo Frames, or Ray-Ban Meta glasses, are canonically not subsumed under the AR definition as they lack the corresponding pose tracking and display components for view-dependent rendering. The declination of view-dependent rendering in smartglasses allows for substantially more compact form factors, esp. when featuring no or just a monocular small-FOV display, yet significantly reduces integrability of concepts that blend virtuality and physicality. Other close-to-eye technologies such as retinal projectors [339] or AR contact lenses [356] play a minor role as of now.

2.2.4 Mixed Reality

The Mixed Reality (MR) term has maybe been used with the broadest conceptual scope among the three discussed terms of AR, VR, and MR [309]. Sometimes, definitions of MR seem to refer to the same terminological extension as AR definitions; more recently, MR has been used to specifically refer to *video passthrough headsets*; at other times, it is understood as an umbrella term comprising AR and VR; and then again it has been used to refer to systems that are distinct variations of AR or VR. Authors who discern MR as a self-sufficient "phenomenon", next to AR and VR have introduced new umbrella terms such Extended Reality (XR) [133] and Cross-Reality [12], and then other authors again have surrendered and introduced the xR symbol meaning something like "Anything" Reality.

A well-recognized terminological perspective on MR has been presented by Milgram and Kishino [229] with subsequent elaborations [230; 227; 228] on the seminal paper. At the core, they present the *virtuality continuum* (shortly later renamed in a paper on AR featuring the same authors to the *reality-virtuality continuum* [230]) that stretches from devices that display the real environment, e.g., a conventional video camera, to devices that display a virtual environment, e.g., VR HMDs. They position MR displays in between these extrema, thereby encompassing augmented reality displays and augmented virtuality displays, arguing the characteristic property of MR lies in presenting virtual-world objects and real-world objects together within a single display.

The presented continuum and the subsequently discussed implications offer a range of merits. First, it rightly points out the seemingly simple, yet practically non-trivial problem that arises when distinguishing between objects that are virtual and those that are not [229, p. 1324]. Even though an image of a physical object is not of the same physicality as the physical object itself, the image nonetheless is in some sense physical. Second, the authors rightly point out that a discriminating factor in different MR systems lies not only in the proportion of pixels pertaining to virtual or physical objects on the display but also in its *extent of world knowledge*. For example, a naive video camera would have no knowledge regarding the contents of the frames it captures and displays [229, p. 1326]. In contrast, world knowledge would be considered high if an AR application had information on the geometry, location, and orientation of an object relative to the camera. Skarbez et al. [307] similarly draw on the notion of world awareness.

However, we argue that some of the presented ideas suffer from shortcomings from today's perspective:

First, a more purposeful name for the continuum seems to be *physicality*-virtuality continuum rather than *reality*-virtuality continuum. On the one hand, already the VR term suggests the orthogonality between reality and virtuality, which, otherwise, would be oxymoronic. Of course, one could argue that the VR term is the actual misnomer, and the continuum employs the more appropriate mental model. More importantly, however, on the other hand, the naming of *reality-virtuality* continuum presupposes that something virtual cannot be real. As will be outlined in the subsequent section in more detail, we argue that reality can be interpreted as a subjective construction, irrespective of its material qualities. Similar criticism has also been expressed in literature [50, p. 236].

Second, it seems questionable that VR HMDs are explicitly excluded. For example, most VR HMDs have no means of canceling out natural sound sources in the physical scene. Similarly, the ground on which the user walks continues to be physical ground, and in many VR applications, the user remains in danger of colliding with physical objects. Therefore, even the most advanced VR systems of today, and most probably, also those in the foreseeable future, provide, in fact, a

blend of virtual and physical signals to the user's senses, rather than exclusively those of a virtual environment. As a result, this dissertation maintains that it is more useful to include VR under the MR term. Similar criticism has also been discussed in literature [307]. Furthermore, from this terminological perspective, it is also more purposeful to consider VR applications a special case of MR applications, rather than the other way around as proposed in the first sentence of the seminal paper. This view is also corroborated by the fact that a video-passthrough headset that can run AR applications can also run VR applications, but not all VR headsets can run AR applications as they lack outward-facing cameras.

Third, we argue that the presented continuum only insufficiently conveys the degrees of freedom with respect to environment modeling. In particular, the authors of the continuum fundamentally and explicitly argue that the extent of world knowledge is perfect in VR whereas it is incomplete in AR. We oppose this view and instead argue that, for any MR system–and following our previous point of criticism, this inevitably not only includes AR but also VR systems–, it is necessary to maintain a representation of the physical scene, and simultaneously, a representation of the virtual scene. On the one hand, the virtual scene representation is, by definition, perfectly known by the system in both AR and VR. On the other hand, all MR systems, both AR and VR systems, require a more or less detailed and accurate representation of the physical world. VR systems make use of a physical representation, at least for tracking, and research has investigated many different concepts of physicality-aware VR, either to reappropriate haptics of the physical scene for the virtual experience [305], or, on the contrary, to avoid collisions between the user and the physical scene, e.g., through the procedural generation of virtually corresponding obstacles [56]. AR systems make use of a physical representation, e.g., to inform the augmentation process that manages the virtual scene.

Fourth, the user and user interfaces in general are not discussed. The authors deliberately position the framework for its use in systematizing MR displays. However, as a result, the MR definition given in the paper is reduced merely to display-related criteria. Instead, we argue that MR is not only a display technique, but also emphasizes the computer's role as an intimate, personalized device, able to sense and model various user-oriented aspects. This includes creating models in order to solve the strongly structured problems of head, face, eye, hand, and full-body tracking and reconstruction, as well as creating models for weakly structured problems, e.g., to estimate attention focus, cognitive load, intent, or affect.

Apart from this seminal perspective, many other frameworks on MR, some with a more general scope, reasoning about a wide class of applications, and, others with a more specific scope, have been presented in literature. Speicher et al. [309] present a conceptual framework of MR, spanning five dimensions to purposefully distinguish between MR experiences: number of environments, number of users, level of immersion, level of virtuality, and degree of interaction. Skarbez et al. [307] present a framework spanning three dimensions: the extent of world knowledge, immersion, and coherence. Mann [216] argues that the definition of MR excludes some modulating operations, such as modifying or diminishing reality, and therefore introduces the notion of mediated reality. However, despite finding acknowledgment in the community, the notion of mediated reality itself has not gained widespread adoption, at least not when compared to the original MR term. Instead, many authors, us included, follow the conception that said examples of diminished and modified reality can still be understood as instances of MR applications. More problem-specific related work will be discussed throughout the dissertation as needed.

2.3 Situated and Semantics-Aware Mixed Reality

2.3.1 Overview

In the previous section, we have outlined common understandings, found in literature, that provide mental models to make sense of technological developments of the past and the future. While many of these understandings have been useful in comprehending the inceptive idea associated with the MR term for decades, some of them begin falling short of pinpointing the conceptual characteristics and technological visions that are explicitly and implicitly associated with MR today.

To develop a perspective that enables understanding current and future developments, we put two assumptions to record. First, the assumption of the *progression of situated computing*, and second, the assumption that it remains useful to *uphold the distinction between AR and VR*. We elaborate on both assumptions in the following.

First, we assume the *progression of situated computing*. Why is it that countless, some of which of the most talented, academics around the world spend significant parts of their lifetime on MR and its related subjects, and some of the most economically successful technology companies of the world invest tens of billions of dollars into its development and research? While the answer to these questions is surely not monocausal, we argue that MR is as much a technology to render 3D objects in the scene as is the mobile phone a technology to make phone calls. To make sense of the larger developments in MR and its implications, we fundamentally consider MR devices *versatile sensor platforms*, providing an unprecedented opportunity to gain an understanding of the user and their reality. To hyperbolize this point–and drawing technology companies darker than they might be–MR platforms would be of high value to companies even if the device would not have any means of rendering any signal, neither auditive nor visual, just because understanding a user's reality is of (pecuniary) value. Such a "product" would be a non-seller, of course. However, a versatile sensor platform coupled with *versatile rendering capabilities* to provide auditive, visual, and potentially haptic signals, meets a demand, as many users, justifiably, see high value in a computer that has a deep understanding of them and their reality, and can situate its experiences and applications therein. Against this backdrop, it stands to reason that the trajectory of MR is to be seen in the light of the progression toward situated computing.

Second, we assume the need of *upholding the distinction between AR and VR*. There seems to be a trend of identifying a convergence between AR and VR, resulting in newer uniformist ideas and terms such as Cross-Reality, Extended Reality, and even xR ("Anything" Reality). While these conceptions are undoubtedly useful for certain hardware considerations and applications, we argue that some of the broader developments in the MR field can be better interpreted when upholding the distinction between AR and VR. On the one hand, recent artifacts such as the *FluidReality* [300] haptic glove aim to decouple the user's sensory system from the physical scene. On the other hand, artifacts such as systems helping people with low vision to navigate stairs [387] are on the other side of the scale, by enhancing their user's awareness of the physical scene. As a result of this user-oriented perspective, we believe it remains useful to maintain the differentiation between AR and VR.

Taking both assumptions together, we conceive our understanding of *Situated and Semantics-Aware Mixed Reality* systems by first outlining the underlying vision, and then providing a definition. It is best understood in its distinction from VR. The grand vision of VR has been

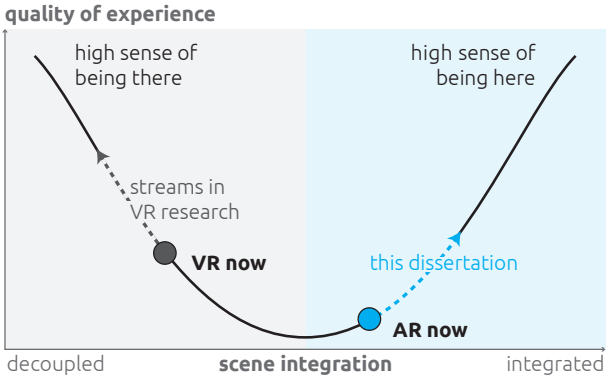


Figure 2.1: Motivation of situated and semantics-aware MR experiences. This dissertation is motivated by the desire to get closer to MR experiences that seamlessly blend with the user's space and mind. We pursue this vision by contributing concepts and technologies that enable to situate virtual content under semantic awareness in the physical scene, thereby deepening scene integration and increasing the quality of experience.

tangibly formulated in the pursuit of the *ultimate display* [318] or the *sensory Turing test* (inspired by the "visual Turing test of VR" [392] and similar [151], which in turn is inspired by the Turing test of AI [329]). Both formulations capture the objective to produce artificial signals representing a virtual scene that are indistinguishable from signals of a natural source pertaining to the physical scene. As the technical ability to decouple virtuality from physicality increases, the user's sensation of being present in the virtual environment also increases, thus presumably improving the quality of the experience. As a result, prominent research streams in VR aim to advance the perceived realism of artificially producible haptic, visual, and auditive signals. Against this backdrop, the historic path of VR research and products can be well-understood. In contrast, AR has lacked a comparably tangible idea or catchphrase that can clarify the vision beyond more diffuse fantastical fragments from Cameron's Terminator or Spielberg's Minority Report. This dissertation argues that the vision of AR is, in a certain sense, opposed to the vision of VR. Rather than being in pursuit of decoupling the virtual scene from the physical scene, AR pursues to increase the integration between the virtual and the physical scene. That is, rather than creating the VR illusion that the user is really there [308], it can be argued that AR is concerned with creating the illusion that virtual content is really here. To improve this illusion, AR requires a representation of physicality that is in line with the user's understanding of physicality, to then situate virtual content therein. We argue that the AR system's representation of physical reality is not exhausted in building a geometric representation, but also requires reconstructing a photometric and constructing a semantic representation. Figure 2.1 illustrates this ambition.

Guided by this vision, we define situated and semantics-aware MR systems as technical systems that 1) exhibit awareness of the user, 2) represent the physical environment semantically, and 3) leverage both to provide user experiences, semantically situated in physicality and meaningful to the user. Our understanding emancipates itself from constraints such as visual modalities, augmentations, or the form factor. Coming back to the smartglass example of before, in contrast with the canonical AR view, a smartglass device *is subsumed* under our definition of situated and semantics-aware MR if, for example, a built-in camera provides the system with situational awareness, thereby enabling a speech synthesis module to provide scene-adaptive output, even if this device does not even feature a visual display. However, a smartglass device *is not subsumed* under our definition if it simply provides the user with push notifications, even if those push notifications would be visually displayed. Instead, our understanding is merely attached to providing a physically situated experience, leveraging semantic environment awareness and user awareness.

In the remainder of this section, we detail our terminology with respect to reality, physicality,

and virtuality on the one hand, and situatedness, scene awareness, and semantic awareness on the other hand.

2.3.2 *Perspective Relativity* of Reality, Physicality, and Virtuality

Overview

In an intuitive understanding, world, reality, and physicality are often used synonymously. However, Wittgenstein, from his logical positivist stance, states that "[t]he world is the totality of facts, not of things" [363, proposition 1.1], that "facts in logical space are the world" [363, proposition 1.13], and "[t]he total reality is the world" [363, proposition 2.063]. Popper, critical realist, distinguishes "first, the world of physical objects or of physical states; the world of states of consciousness or of mental states [...]; and thirdly, the world of objective contents of thoughts, especially of scientific and poetic thoughts and of works of art" [268, p. 1]. Frege considers a similar differentiation between a realm of things ("Reiche der Dinge"), the realm of imaginations ("Reich der Vorstellungen") referring to thoughts in the mind of individuals, and a third realm referring to thoughts that exist independent of a specific individual [96, pp. 42-44]. Among many others, Davidson distinguishes between subjective, intersubjective, and objective propositional knowledge [72, p. xiiv]. These examples already reveal the more intricate relationship that can be assigned between terms such as the world, things, reality, and physicality. In this dissertation, we also do not follow a synonymous usage of world, reality, and physicality as this would reduce basic terms of virtual reality and augmented reality in our discipline ad absurdum. Instead, we maintain the differentiation between objective reality and subjective reality while simultaneously conceding that any thought or language artifact, even when *aiming* to deal with objective reality is of subjective nature as individuals lack "unfiltered" access to objective reality or means to reconstruct it objectively.

To clarify our understanding of virtuality, physicality, and reality, and demonstrate that the terms are fundamentally conditioned on subjective perspective, we outline the following thought experiment. It conceptually resorts to a passthrough mixed-reality headset as it is offered in different products of today.

Perspective Relativity as a Consequence of Constructivism–A Thought Experiment

Step 1 Imagine two users standing in a room with a red table. The table features certain wavelength reflection properties that justify its denomination as red. The first user perceives the table as red. The second user suffers from a congenital red-green color blindness and perceives the table as brownish. In this first step, one can distinguish between a single instance of *objective reality* and two different instances of *subjective reality*. For some purposes, it would be useful to refer to these different instances with three different denominations. For other purposes, it would be sufficient to refer to a single physicality with a single physical table, verbally omitting the distortion that takes place "by the human condition".

Step 2 Now each user puts on a passthrough mixed-reality headset that features outward-facing RGB cameras and inward-facing RGB displays, thereby providing a more or less faithful reproduction of the previous visual sensory signals that reach the users' eyes. For the system

engineers of the headset that are concerned with the algorithms to transform the multi-camera input into a perspectively correct passthrough view to be rendered, it will be useful to distinguish between physicality and the virtual reproduction of physicality that is displayed to each eye display. For the users, however, it might remain useful to refer to a single "physicality" and a single "physical" table, even though what their cognition operates on has been processed and distorted multiple times now until it becomes accessible to reasoning.

Step 3 Next, a third person, taking the role of an experimenter, in the room turns on "blue mode" on both user's headsets via the press of a button. This blue mode changes all red pixels to blue pixels in each of the users' headsets. When asked which color "the table" has, users might distinguish between the virtual color of the table and the physical color of the table.

Step 4 Then, with the headset on their head, both users step through a door into the next room with a table. It appears to be blue for both. The users have no way of deciding whether the physical table is "physically blue" or not. It could be assumed that users might refer to the table as "the blue table" when asked for a description. As it fulfills its purpose exactly as well as any other table they have used so far in their life experience, we might assume users will certainly refer to the table as "real".

Step 5 Now, assume the experimenter asks the users, one after the other, to take a seat on the "physical chair". Only once the users had followed the instructions, did the experimenter reveal the "physical chair" was also altered in color from black to gray. However, from the perspective of the user, up to the experimenter's revelation, the chair was as physical as any other chair to the users because they never even doubted its qualities.

Step 6 Next, a dog enters the room through the door. Maybe, the users will think this is a physical dog and try to pet it only to find out they grasp into thin air, calling it a virtual animal from now on. A few moments later, a unicorn follows the dog through the door. Both users will immediately know that this unicorn is virtual. For the users, the "reality state" of the dog was ambiguous, but from the perspective of the system, the dog and the unicorn always took exactly the same role.

Step 7 Finally, assume, the headset could not only visually render the object, but–by means of a built-in science-fiction-like 3D printer–in real-time make a table appear in the middle of the room out of thin air. However, the printer can only produce exactly one object at a single time, and it must liquidate the produced object before it can create a new one. While there can be no doubt that, the user does not differentiate between the tables anymore, the system engineers must continue to distinguish between physical and virtual objects to maintain their ability to liquidate the produced object in order to produce the next one. However, if the experimenter uses the science-fiction printer to create a perfect replica of the marriage rings of users, suddenly, the users will also maintain the differentiation between the "real ring" and the "fake ring".

Implications

These considerations offer multiple insights.

First, system engineers might find it useful to distinguish between "physicality" in the sense of "objective reality", the input processed by the headset (e.g., "input signals"), the representation inside the headset (the technical "system's reality"), the output of the headset ("output signals"), the sensed input to the users' eyes, and the mental representation that a user will form inside their mind (the "user's reality").

Second, users might distinguish between virtual and physical objects if it makes a difference, otherwise the distinction is of little value.

Third, from the user's perspective, physicality or virtuality can be perceived as equally real, and physicality is conceptually different from reality. We refer to reality as either the input or the output of the sensory, perceptual, and cognitive process. Objective reality is at the beginning of this process, and subjective reality stands at the output of that process. MR relies on some existence in objective reality, i.e., requires a means to influence the sensory, perceptual, and cognitive process.

Fourth, there is nothing "essentialist" that distinguishes physical from virtual objects. Instead, the adjectives "physical" and "virtual" are subjective assignments, following the purpose of the subject that performs the assignment based on subjectively assumed qualities. So, what the system engineer refers to as virtual (e.g., rendering a virtual dog), could be called physical by the user if convincing enough. From the system engineer's perspective, the goal can but does not necessarily lie in creating the illusion that something is perceived as physical by the user–anticipating what users perceive as physical or virtual. This perspective is also found in empiricist views such as the one by Quine: "Physical objects are conceptually imported into the situation as convenient intermediaries [...], comparable, epistemologically, to the gods of Homer [...] The myth of physical objects is epistemologically superior to most in that has proved more efficacious than other myths" [334, sec. VI]. Similarly, Winograd [361, p. 97] singles out the naive realism underlying the concept of situation semantics by Barwise and Perry [23] which "takes for granted the existence of specific objects and properties independent of language" [362, p. 69].

Fifth, an object can exist in different *representational stages* until it evolves into its "final" representation in the user's mind. As system creators, this provokes the need for a more differentiated terminology to refer to these representational stages, motivating the next considerations.

2.3.3 A *Higher-Order Reality Model* for Situated and Semantics-Aware MR

2.3.3.1 First-Order and Higher-Order Reality

To meaningfully distinguish between physically pre-existing signals, and artificially generated physical signals, produced by the MR system, we introduce the *Higher-Order Reality Model*. This model provides a terminology for signals sensed by the user, based on the strength of *physicality-representational correspondence*.

We define the *user-sensed scene* as the entirety of signals, transduced by the user's sensory system. This definition captures Goldman's idea [104, p. 198] that a human's perceptual process takes transducer activity as input while expanding the frequently found concept of scenes as the entirety of visual signals [278; 27] to the entirety of all potential signal types.

Our higher-order reality model establishes that the user perceives the scene as a mix of first-order, second-order, third-order, and fourth-order signals with the following distinction:

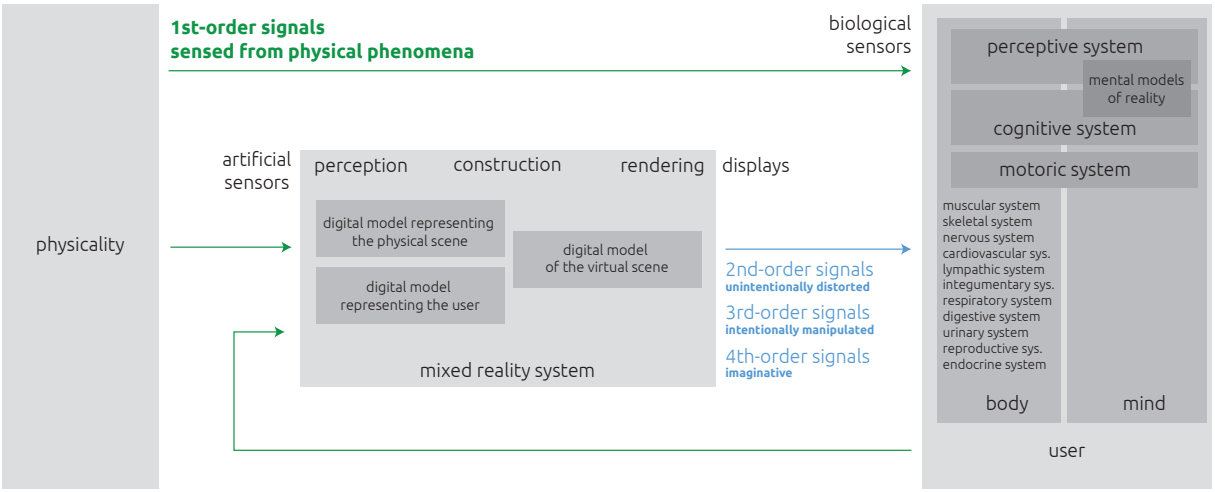


Figure 2.2: Perceiving physicality and virtuality in terms of the *Higher-Order Reality Model*. The user perceives the scene as a mix of first-order, second-order, third-order, and fourth-order signals. Human organ system adopted from Wakim et al. [340].

First-order signals (signals of "physical reality of first order") pertain to physical phenomena in objective reality and are sensed by the user without ever passing through the MR system. Subjective interpretation of first-order signals yields *first-order objects*. They have the highest representational correspondence with physicality. First-order signals can also be algorithmically interpreted by the MR system, and–assuming ideal conditions where the MR algorithm encodes an *intersubjective* interpretation scheme–yields comparable first-order object references.

Second-order signals (signals of "unintentionally distorted reality") pertain to physical (i.e., first-order) objects that are processed by the MR system and passed through with the least possible manipulation or distortion. They have a strong correspondence in physicality, yet might be distorted before reaching the user's sensory system due to technical imperfections. In video-passthrough MR, may it be on HMDs (such as found in our Scene Responsiveness implementation) or handheld devices (such as found in our TransforMR implementation), and in audio-passthrough MR, this particularly refers to the potentially artifact-polluted passthrough layer, as captured and processed by the outward-facing camera or microphone system. In optical-seethrough AR systems, this refers to light rays that passively propagate through the waveguide material from the outside.

Third-order signals (signals of "intentionally manipulated reality") pertain to physical objects that are processed by the MR system and, on their way, are intentionally manipulated for application purposes before being presented to the user. While the manipulation process can be slight or heavy, it maintains physicality-representational correspondence. In video-passthrough MR, this refers for example to the process of superimposing wireframes, contours, or x-ray visualizations. In audio-passthrough MR, this might refer to a voice of different a person in the room that is either artificially increased in volume while dimming down all other voices and noises, or even to a synthetically generated voice that translates the other person's speech from German to English in real-time. The distinguishing criterion between second-order and third-order signals lies in the intentionality of manipulations: Second-order signals are altered unintentionally, while third-order signals are altered intentionally.

Fourth-order signals (signals of "imaginative reality") pertain to fully imagined objects that are generated by the MR system without any representational correspondence in the physical scene.

In video-passthrough MR, this refers to classical AR augmentations for example to a unicorn or an exploding bomb in a gaming application, a locally visualized avatar of a remote interlocutor in a telepresence application, or a shoe that a user considers buying in an e-commerce application. This also refers to graphical user interface elements, such as menus, widgets, text annotations, arrows, waypoints, or similar.

We refer to the collection of first-order objects as the "physical scene". We refer to the collection of third- and fourth-order objects as "virtual scene in the narrower sense". We refer to the collection of second-, third-, and fourth-order objects as "virtual scene in the broader sense". We refer to the combination of the physical and virtual scene, i.e., encompassing the blend of signals of all orders, as the "user-sensed scene". We refer to any logical unit that is purposeful to consider from the perspective of the user as "object" (in the broader sense). We refer to any logical unit that is purposeful to consider from the perspective of the user, however excluding the background, as "objects in the narrower sense". Often, the above distinction is provided implicitly through textual context.

Note that just because a fourth-order signal does not exhibit physicality-representational correspondence, it can, and often should, be situated in the landscape of sensed first-order, rendered second-order, and rendered third-order signals with *semantic correspondence*. We will discuss the notion of semantics after establishing the insight that MR systems must, by definition, convert from first-order to higher-order signals.

2.3.3.2 Inevitability of First-Order Signal Conversion

By definition, any signal emitted by the MR system and sensed by the user cannot be a first-order signal but must be a second-order, third-order, or fourth-order signal. Furthermore, any first-order signal sensed by the MR system must be either converted into a second-order, a third-order, or a fourth-order signal, or it must be eliminated. This insight results from the consideration of the following four mutually exclusive, collectively exhausive alternatives:

First, the MR system turns the first-order signal into a second-order signal by passing it through without intentional modification.

Second, the MR system turns the first-order signal into a third-order signal by manipulating it intentionally.

Third, generating a fourth-order signal can be lossless or lossy. In audio MR, for example, adding an audio signal in a different region of the audio spectrum than existing first-order audio signals is a lossless operation. However, in video passthrough MR, rendering an imaginative dog removes the user's ability to sense the background occluded by the dog, and therefore is a lossy operation. An equivalent phrasing would the rendered dog substitutes parts of the first-order signal pertaining to the background. While losing background information is lossy in information-theoretic terms, it is generally not perceived as a *loss in meaning* from the subjective interpretation of the user. However, substituting first-order signals pertaining to a "first-order dog", existing in physicality, with background information could be very much considered a *loss in meaning*. Therefore, when generating fourth-order signals, we conceptually distinguish between object augmentation and object removal operations, even though the object removal effect has been achieved via signal generation, e.g., superimposing an object with its estimated background color in video-passthrough AR, or generating anti-noise in audio AR.

Fourth, while humans, broadly speaking, cannot *not* see with their eyes open so that the "default"

of vision is seeing *some*thing, in the audio realm, the default is not hearing anything. Therefore, in audio-passthrough MR, first-order signals, can, in fact, be eliminated without the need to estimate "neutral" information for which to generate a fourth-order signal.

2.3.4 Semantic Awareness in Situated Mixed Reality Systems

2.3.4.1 Overview

With our *Higher-Order Reality Model*, described in the previous section, we provided a differentiated terminology for MR experiences, based on *physicality-representational correspondence* for second- and third-order objects, and we already hinted at the idea of *semantic correspondence* for fourth-order objects. In the following, we detail the idea of representations, semantics and semantic awareness against the backdrop of literature.

Questions of semantics and meaning arise in various scientific fields such as philosophy with its subfields of epistemology [209; 143; 45; 364; 87; 88], ontology and meta-ontology [297, p. 10], metaphysics [273, p. 35], and logic [95; 285; 287; 315], in linguistics with its subfield of semiotics [256], in psychology [27; 278] with its subfields of psychology of discourse, and cognitive psychology, and in computer science with its subfields of artificial intelligence [380, p. 193], natural language processing [360], knowledge representation, computer vision, software, data [314], and information systems engineering, formal language design [64; 97], and human-computer interaction. Within the field of human-computer interaction, considerations on semantics stretch from user-oriented software for language [360] and user interface (UI) design [4; 259] via multimodal systems such as Bolt's "Put-That-There" demo [32] to MR [59].

In the following, we first discuss semantics from the perspective of linguistics, scene perception, and cognition, as all of these fields are concerned with the question of how humans and even machines can deal with representations of reality. Afterward, we synthesize a unifying understanding of semantic awareness in situated MR systems.

2.3.4.2 Semantics in Linguistics, Perception, and Cognition

Semantics from a linguistic perspective is commonly differentiated into the theory of reference and the theory of meaning, as presented, for example, by Bar-Hillel [20] and Quine [334]. According to Bar-Hillel [20, p. 236], Carnap [44] implicitly follows the same distinction with his differentiation between intensionality and extensionality. As discussed by Chomsky [63], the concept of a theory of meaning proves problematic in a variety of aspects, which maybe serve as an explanation for the dominance of the theory of reference in literature, also observed by Bar-Hillel [20]. According to the theory of reference, a word refers to a referent, e.g., an entity in subjective or inter-subjective reality. In an easy case, the referent is a definite entity in intersubjective reality [360], e.g., "the current US president". However, the case becomes more difficult when the reference cannot be resolved an unambiguous referent, e.g., when employing an indefinite reference such as "a president" or when referring to a conceptual entity [360, p. 288]. For example, Locke [209, Chapter III] points out, both "horse" and "unicorn" are words, free of inconsistencies and are equally "certain, steady and permanent" complex abstract ideas.

A further common distinction refers to the semantic and pragmatic properties of linguistic artifacts. However, the idea of separating semantics as uncontextualized meaning and pragmatics as its contextualized counterpart is controversial as seen by discussion from Winograd [360, p.

272] or Chomsky [65], if one declines the existence of a context-free symbol. It is therefore disregarded in this dissertation.

Framing our previously presented higher-order reality model in these terms, our second-order objects (unintentionally distorted representations of a first-order object) and third-order objects (intentionally manipulated representations of a first-order object) are definite references to entities in extra-subjective reality. As discussed in detail before, fourth-order signals have no representational correspondence in the physical scene, but can still be semantically related to a first-order object. Then, however, the first-order object becomes the referent, i.e., the object that is of meaning to the user, and the MR systems must gain an awareness of this same referent, augmenting the fourth-order object in relationship to this first-order referent.

Semantics from a scene perception perspective are usually concerned with visual scenes. Regarding *human* scene perception, Rensink [278] distinguishes the sketch as low-level representations, object structure and object dynamics, scene layouts or maps, and scene gists as mid-level representations, and the scene scheme as a high-level representation during a human's perceptual process. Similarly, Biederman defines semantics as the schema of the scene which is the "overall internal representation of the scene that integrates the scene's entities and relations" [27].

Regarding *machine* scene perception, semantics are frequently used as an adjective as in semantic segmentation [304], semantic SLAM [288], semantic correspondence [118], semantic editing [253], and similar. While the term is mostly used in an intuitive manner, it generally refers to an approach that incorporates class labels or instance labels for each processed 2D pixel or 3D point, where each class label is one of a predetermined set of class labels. These technically motivated approaches create well-structured formal problems that abstract away the human cognitive perspective.

In contrast, **semantics from a cognitive perspective** are emphasized by Winograd [360, p. 265] who considers (his take on) semantics the study of understanding how language interacts with cognitive functioning, and underlines the importance of concepts such as representations, knowledge, and models to discuss the "set of symbol structure‘ in a physical symbol system such as a computer or nervous system. Furthermore, he discusses the notion of a "language of thought" that includes abstractions such as physical and institutional objects, perceptual properties, events, abstract categorizations, "complex conceptual objects, properties, and events built up out of descriptions couched in terms of other mental entities", linguistic objects, and "hypothesized versions of the entities in the symbol systems of other people" [362, p. 266].

Winograd [360, p. 288] also provides a nexus from *semantics* in the above-described referential sense to *mental models*: Accordingly, a mental model in an individual comprises conceptual entities. The conceptual entities in two individuals may correspond with each other, or not. Correspondence between the conceptual entities can even be established if the individuals do not have fully identical or compatible mental models.

The importance of mental models has also been discussed in human-computer interaction, for example, distinguishing the conceptual model of a target system, the user's mental model of that system, and the scientist's conceptualization of the mental model [251, p. 7]. In HCI, mental models have been defined as "users' own mental representation of their interactions with devices" [311].

2.3.4.3 A Unifying Perspective on Semantics in Situated MR Systems

Considering the above, we define semantics as *the property of an MR system to represent reality with concepts that correspond with concepts in a presumed mental model of reality of a user.* Our definition takes into account the representational nature of semantics, the subjectivity that underlies it, the HCI-driven necessity to take the user into account, and the engineering-oriented perspective to anticipate users' mental models. It is not static and allows for the fact that the mental models of users are also influenced by the system's underlying representational model. Furthermore, it is largely compatible with other works on semantics in MR [59; 336; 245; 226; 307; 135].

At the same time, it has two main implications: First, due to their first-person authority [72], users have "privileged access" to their mental models, so that neither a technical MR system nor a designer, an engineer, not even a scientist, can faithfully reconstruct the mental models a user has. Instead, the challenge lies in making assumptions about concepts and representations that potentially correspond with concepts pertaining to the user's mental models. Second, different users have different mental models. The challenge lies in anticipating the most useful concepts for the addressed user group.

That means there is no *universal semantic representation* of reality. However, this is not a concession, specific to the development of MR systems. Instead, they are the raison d'être of design and research in HCI, including this dissertation.

2.3.5 Semantic Awareness for Higher-Order Reality Generation

In the previous paragraphs, we outlined that the MR system must always perform a conversion of a first-order signal to either a second-order, third-order, or fourth-order signal or–in the audio-passthrough case–choose to eliminate it completely.

We also defined semantic awareness as *the property of an MR system to represent reality with concepts that correspond with concepts in a presumed mental model of reality of a user.*

Semantic awareness of the physical scene is a necessary prerequisite for establishing *semantic correspondence* of a virtual scene with the physical scene. In the following, we discuss how MR systems turn first-order signals into higher-order signals by making use of semantic awareness.

To turn first-order signals into second-order signals, i.e., passing through sensed first-order signals to the user without intended modifications, the MR system, generally, semantic awareness of the physical scene is not required. Instead, first-order signals are sensed and forward while being agnostic of the interpretation the user will conceive. This is not to say that it could not be possibly helpful to be aware of the nature of the first-order signals. For example, by being aware of the fact that input signals stem from a starlit sky during the night (e.g., for an astronomy education AR app that augments the names of planets and star constellations), the governing parameters for computational photography could be adjusted accordingly. However, just passing through the real-time sky video frames to the display, does not require detailed semantic information about the starscape.

To turn first-order signals into third-order signals, i.e., sensing the first-order signals, turning them into an application-specific representation, and then visualizing the application-specific representation depends on a semantic awareness, i.e., an understanding of what the user conceives as first-order objects. For example, to visualize an X-ray view according to the magic lens

metaphor, the MR system requires a digital representation of the first-order object of interest to the user. It must also be able to identify the virtual object in the entirety of first-order signals sensed–may it be by the user of AR fiducial markers or dynamic computer vision models–and then provide corresponding pose information for rendering the digital third-order object. On the other hand, to eliminate distracting voices during a cocktail party in an audio smartglasses app, the smartglasses need to disaggregate the aggregated audio signal into multiple audio tracks–one per voice plus additional tracks for background music and other noises–, identify the voice of interest, potentially using lip-sync detection on the camera stream and then emit a real-time copy thereof to intensify it.

As stated before, even though fourth-order objects have no physicality-representational correspondence, they can, and often will or even should be, in a semantic correspondence with the first-order objects. For example, a text annotation that indicates the price tag for a product in a stationary retail AR application is in a strong semantic correspondence with the first-order object, i.e., the retail product. To display a virtual monster that enters the scene through a window requires knowledge of the windows in the physical scene. In contrast, a shoe that the user considers buying which is floating in mid-air has practically no semantic correspondence to the physical scene. Similarly, the much-heralded virtual HMD screens that shall replace a physical desktop monitor, and can be positioned everywhere in any size, are also fourth-order objects without physicality-representational nor semantic correspondence in the physical scene.

These examples illustrate a scale of different levels to which the various third-order and fourth-order objects–that is, X-Ray views, price tags, monsters, voices, panels, or e-commerce products–can be in semantic correspondence with the scene. We also refer to this semantic correspondence as physical-virtual scene integration and distinguish the following level: First-order and second-order can be in coexistence, in coherence, or in situatedness with third-order and fourth-order signals. As shown in Figure 2.3, higher levels of scene integration depend on all levels below.

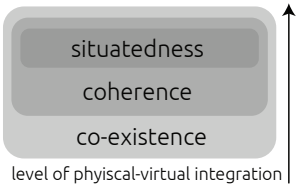


Figure 2.3: Co-existence, coherence, and situatedness. Generally, we distinguish between three levels of physical-virtual scene integration, namely co-existence, coherence, and situatedness. This dissertation is concerned with two aspects. First, it investigates new ways of creating scene situatedness. Second, it proposes a new level of scene integration, complementing the existing levels of scene co-existence, scene coherence, and scene situatedness which we name *Scene Responsiveness* (see chapter 5).

Thus, in this dissertation, situatedness is defined on third-order (intentionally manipulated evolution of the physical prototype) and fourth-order (imaginative without a physical prototype) objects and understood as featuring a semantic correspondence to first-order (directly stemming from the physical prototype) and second-order (unintentionally distorted evolution of the physical prototype) signals.

More abstractly, our understanding of situatedness is in line with White's [354] definition of situated visualization as "visual representation of data presented in its spatial and semantic context", a concept that has also been taken up explicitly and implicitly by others [38; 359; 92; 292; 154]. While not defined explicitly, this understanding of a semantic relationship as being definitional to the situatedness term is well-suitable with the implicit usage of the term in early AR works, e.g., the notion of *situated documentaries* by Hoellerer, Feiner and Pavlik [134], the notion of *situated information spaces* by Fitzmaurice [91], the notion of *situated communication* by Rekimoto and Ayatsuka [275], or the notion of *situated media* by Güven, Feiner, and Oda [115].

Availing of this terminology, we can concisely describe the contributions, presented in this thesis.

In our SoundsRide project, we situate fourth-order audio signals, being aware of the semantic concept of sound affordances in the physical scene and maintaining corresponding representations thereof. In our TransforMR project, we situate third-order visual objects, leveraging semantic awareness of humans and vehicle counterpart objects in the physical scene to eliminate them and substitute them with corresponding functionally corresponding counterpart objects. In our Scene Responsiveness project, for a wide range of objects, we dynamically swap between second-order and third-order status and also render a fourth-order character that can interact with these second-order and third-order objects, leveraging knowledge about object poses and affordances. In our HandyCast project, we enable situated input in a full fourth-order environment that is situated in that it avoids haptic collisions.

	1st order	2nd order	3rd order	4th order
SoundsRide	sound affordances			synchronized music mix
TransforMR	out-of-viewport correspondence	semantic context	semantic simulacra	
Scene Responsiveness	consistent haptics	semantic context	virtualized object twins	characters and avatars
HandyCast	*avoiding* 1st-order haptic collisions			spacious virtual environment

Table 2.1: Reality orders of central concepts in this dissertation's contributions.

3

SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality

Music is a central instrument in video gaming to attune a player's attention to the current atmosphere and increase their immersion in the game. We transfer the idea of scene-adaptive music to car drives and propose SoundsRide, an in-car audio augmented reality system that mixes music in real-time synchronized with sound affordances along the ride. After exploring the design space of affordance-synchronized music, we design SoundsRide to temporally and spatially align high-contrast events on the route, e. g., highway entrances or tunnel exits, with high-contrast events in music, e. g., song transitions or beat drops, for any recorded and annotated GPS trajectory by a three-step procedure. In real-time, SoundsRide 1) estimates temporal distances to events on the route, 2) fuses these novel estimates with previous estimates in a cost-aware music-mixing plan, and 3) if necessary, re-computes an updated mix to be propagated to the audio output. To minimize user-noticeable updates to the mix, SoundsRide fuses new distance information with a filtering procedure that chooses the best updating strategy given the last music-mixing plan, the novel distance estimations, and the system parameterization. We technically evaluate SoundsRide and conduct a user evaluation with 8 participants to gain insights into how users perceive SoundsRide in terms of mixing, affordances, and synchronicity. We find that SoundsRide can create captivating music experiences and positively as well as negatively influence subjectively perceived driving safety, depending on the mix and user.

3.1 Introduction

Music has become an integral part of a driver's in-car experience. It is powerful in influencing the driver's mental state, being able to provoke negative as well as positive effects on driving experience and performance through mechanisms of arousal and distraction or concentration in the driver [348; 82; 39; 333; 390; 252]. More pragmatically, music has been instrumented purposefully to assist in tasks such as keeping a certain speed [43] or navigating towards a specified destination [3].

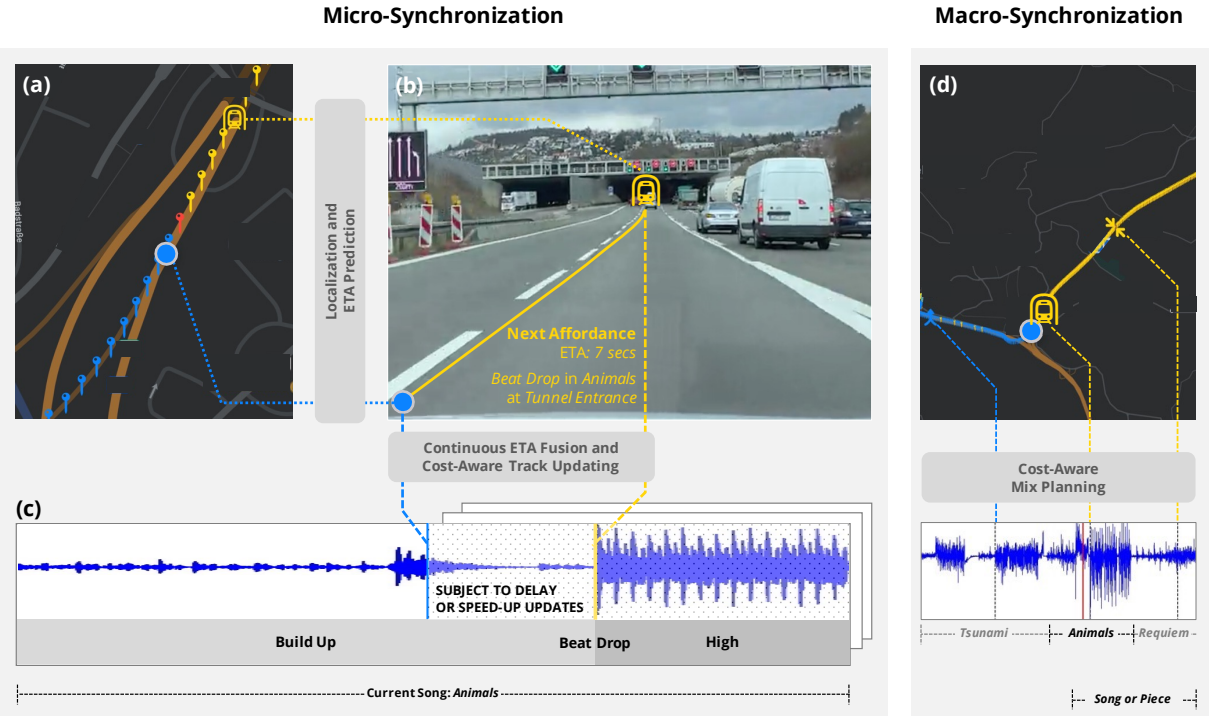


Figure 3.1: SoundsRide is an in-car audio augmented reality system that mixes music in real-time synchronized with sound affordances along the ride. (a) SoundsRide continuously predicts the Estimated Time to Arrival (ETA) to the next affordances (b) to temporally and spatially align high-contrast events on the ride – such as a tunnel entrance – (c) with high-contrast events in the music – such as a beat drop – by speeding up or delaying a mix event through track updates if necessary. (d) On a macro-level, SoundsRide makes use of affordance ETAs to plan a cost-aware mix for the entire ride by deciding on the song sequence and when to transition between songs.

To enhance a user's music experience, previous work has explored adapting music to the driver's context based on their internal state, behavior, or environment. However, music adaption to the user's internal state [80] is challenging due to the complex nature of mental state and the difficulties in estimating and influencing it purposefully. Approaches that adapt music to a user's behavior [81; 195] do not capture the potential to increase situational awareness for the driving task. Previous approaches that adapt music to a user's environment typically serve as mere decision support systems [139; 36; 61; 290; 19] without enabling novel forms of user experiences.

In this chapter, we propose *SoundsRide*, an in-car audio augmented reality system that mixes music in real-time synchronized with the events surrounding the driver's ride. To categorize such events, we introduce a series of *sound affordances* that describe such momentous and well-noticeable events in the environment. For example, the exit of a tunnel with its stark switch in illumination and acoustics affords a strong musical event such as a beat drop or a crescendo that builds up until the instant of the vehicle exiting the tunnel. We argue that aligning such high-contrast events along the ride with high-contrast events in the music can create captivating experiences shaped by memorable moments along the ride and not yet achievable by users on their own without such technical enablement.

The core technical contribution of SoundsRide lies in its capability of synchronizing music to affordances in real-time by creating a mix from a song database so that high-contrast intra-song or inter-song events in the mixed audio signal are temporally and spatially aligned with the sound affordances. Precisely scheduling such auditory signals in response to a driving car is

challenging and comprises three subproblems: 1) localizing sound affordances on the route and predicting an estimated time to arrival ("affordance ETAs"), 2) determining a mixing plan that minimizes undesirable effects such as silences between songs, and 3) continuously integrating updated affordance ETAs without causing disruptions in the mixed audio signal.

In the approach section, we first describe the design alternatives of affordance-synchronized music and then detail SoundsRide's integration of geo-referenced affordances to enable ETA predictions, our heuristics-based scheduling method for affordance-synchronized music mixes, as well as our recursive filtering technique for updating the audio signal to updated ETA predictions.

In our qualitative evaluation, eight participants drove a route with SoundsRide and reported on their impression of the ride. Participants generally commented positively on the drive and seven pointed out their excitement of experiencing music that adapts to the scene. The results also showed that SoundsRide helped participants become more aware of their environment while driving, as they freely listed half of all affordances without explicitly having been instructed about them; they remembered 18 % more affordances when explicitly asked about them, and recognized another quarter of affordances when watching a video replay of their ride.

In our quantitative evaluation of SoundsRide's performance, we examined the level of affordance synchronicity achieved based on the rides recorded during the participant evaluation. SoundsRide ensured synchronicity within $\pm 1.1s$ in 47.7 % of cases and with no more than 1 noticeable update to the audio signal within 15 seconds of an affordance location. In 77.7 % of cases, SoundsRide is able to ensure synchronicity with a maximum misalignment of $1.9s$ and a maximum number of 2 updates to the audio track.

Contributions

In this chapter, we make the following contributions:

1. a novel end-to-end approach for creating suspenseful and affordance-synchronized in-car music experiences based on temporal distance estimation for reinforcing high-contrast events along a ride with high-contrast events in music,

2. a design space of affordance-based in-car audio augmented reality for music experiences,

3. a cost-based method for deriving affordance-synchronized mix plans featuring inter- and intra-song mix events with song snippets from an annotated song databases, and

4. a recursive filter algorithm for incorporating continuously updated ETA information while minimizing obtrusive audio track manipulations.

Combining these contributions in a real-time system, SoundsRide brings dynamic and environment-aware audio augmented reality to everyday car rides, supporting the joy of driving and drawing the driver's attention to the periphery. We provide our implementation of SoundsRide and accompanying assets to the community for future work[1].

[1] `https://github.com/MohamedKari/soundsride`

3.2 Background and Related Work

SoundsRide's idea of affordance-synchronized music mixing builds on concepts found in context-adaptive music, music-to-video alignment, procedural game music, and in-transit audio augmented reality.

3.2.1 Context-Adaptive Music

Approaches that aim to adapt music to a user's internal state draw on user-focusing context variables such as emotion, mood, or fatigue [19; 80; 382; 139; 117]. These context-adaptive approaches often operate on difficult-to-test assumptions about the complex mental dynamics and interdependent processes in humans that govern music listening preferences in a specific real-world situation.

Within the field of approaches that aim to adapt music to a user's environment, location-aware music recommendation [19; 36; 290; 61; 195; 129; 128] takes a particularly prominent role. However, the main objective of these music recommender systems typically consists of improving or simplifying a user's *decision making* in terms of music selection by drawing on typically coarse-grained location classes as decision parameters.

Approaches that aim to adapt music to a user's externally observable behavior, such as pace [81] or driving style [19] reflect and might therefore reinforce a user's activity, however do not expose interactivity with the environment around the user.

In contrast to location-aware music recommendation systems, SoundsRide does not aim at supporting decision making but rather at enabling scene-synchronized music mixing based on accurate estimation of temporal distance to sound affordance moments for captivating and suspenseful experiences. In contrast to behavior-adaptive music playback, SoundsRide is bound to sound affordances in the environment, therefore featuring a novel user-environment interaction pattern. In contrast to affect-adaptive music playback, beyond the fundamental hypothesis that scene-synchronized music is interesting to users, SoundsRide does not make implicit assumptions about the complex interrelationship between affective state and music.

3.2.2 Automatic Music-to-Video Alignment

UnderScore by Rubin et al. [283] automatically derives a musical underlay for an audio story, constrained by emphasis points in speech. Rubin et al. [282] build on the idea of UnderScore, however focus on emotions as a key constraint under which alignment takes place. Sato et al. [289] present a system that automatically arranges a soundtrack so that it fits climaxes in a video. Wang et al. [345] present a system that synthesizes background music after visually classifying intervals in a given video by emotions. Frid et al. [98] propose a system that allows MIDI-based synthesis of music, similar to a provided reference song, in order to acoustically underlay a video.

While these systems are not location-aware, they are abstractly similar in that they align music with externally specified moments in time. However, they are very different in that they operate on all video data available, while SoundsRide operates in real-time and under uncertainty, thus requiring not only a prediction but also a correction procedure for incorporating updated affordance ETAs into the signal output.

3.2.3 Procedural Game Music

To increase immersion [42; 272], many games feature a rich soundscape, ranging from simple player-controlled sounds such as footsteps or collecting coins to atmospheric and narrative-supporting game scores that are bound to certain trigger points in the game's visual space. Creating these soundscapes is often based on so-called procedural or non-linear music composition that allows to add in, subtract, transpose, or swap layers of instrumentation or parameterize a predefined musical sequence in terms of jumps, repeats, or loops [66; 350]. The landmark system iMuse by Land and McConnell [191] from 1991 has heralded non-linear music in gaming by enabling seamless transitions, triggered by switching the gameplay levels or certain events in the gameplay and based on decision points in the score that allow branching to one or the other sequence [266].

SoundsRide's problem statement is similar to such game score engines as it temporally synchronizes music with trigger points, however is different in that it aligns *intra-song events in common songs* with these trigger points rather than branching from a dedicated score. As a consequence, to schedule a song ahead, SoundsRide predicts the time to arrival to the next trigger point, which due to the fewer degrees of freedom in driving, can be more deterministically planned than a player's gaming interactions.

3.2.4 In-Transit Audio AR

We refer to the common notion of Audio AR as superimposing an audio signal on top of the real world, as the user moves within it [220; 24; 188]. On a basic level, common navigation systems with speech output such as Google Maps[2] or Waze[3] can be interpreted as audio augmented reality applications that supply context-bound information to drivers, pedestrians, or cyclists on the auditory channel. Systems like GyPSy Guide[4] provide audio commentary playback for in-car usage along a path of specified locations to provide a location-bound virtual tour guide.

On a more sophisticated level, HindSight [294] employs a sonification of objects detected in real time in continuously streamed 360° video to increase an in-transit user's awareness of vehicles in the surroundings. Audible Panorama by Huang et al. [140] augments a panorama picture with an audio signal, generated by first detecting objects in the image and then mixing sounds corresponding to the object class.

Both aforementioned systems are conceptually distinct from the music-to-video alignment task because they take a much more high-resolution understanding of reality as a basis – using machine-learned computer vision algorithms – in contrast to operating on a series of non-further qualified video frames. Even though Audible Panorama is not focused on in-transit usage, both the ideas in HindSight and Audible Panorama can be interpreted as "affordance-oriented" in that they assign certain artificial sounds to real-world observations.

However, SoundsRide is different from both these systems in that 1) it focuses on music instead non-musical sounds, 2) must plan ahead and update plans under location-awareness in real-time to temporally align mix events with environment events, and 3) opens up interesting user-controlled interaction patterns between the user, the system, and the environment.

[2] `https://maps.google.com/`
[3] `https://waze.com/`
[4] https://gypsyguide.com/

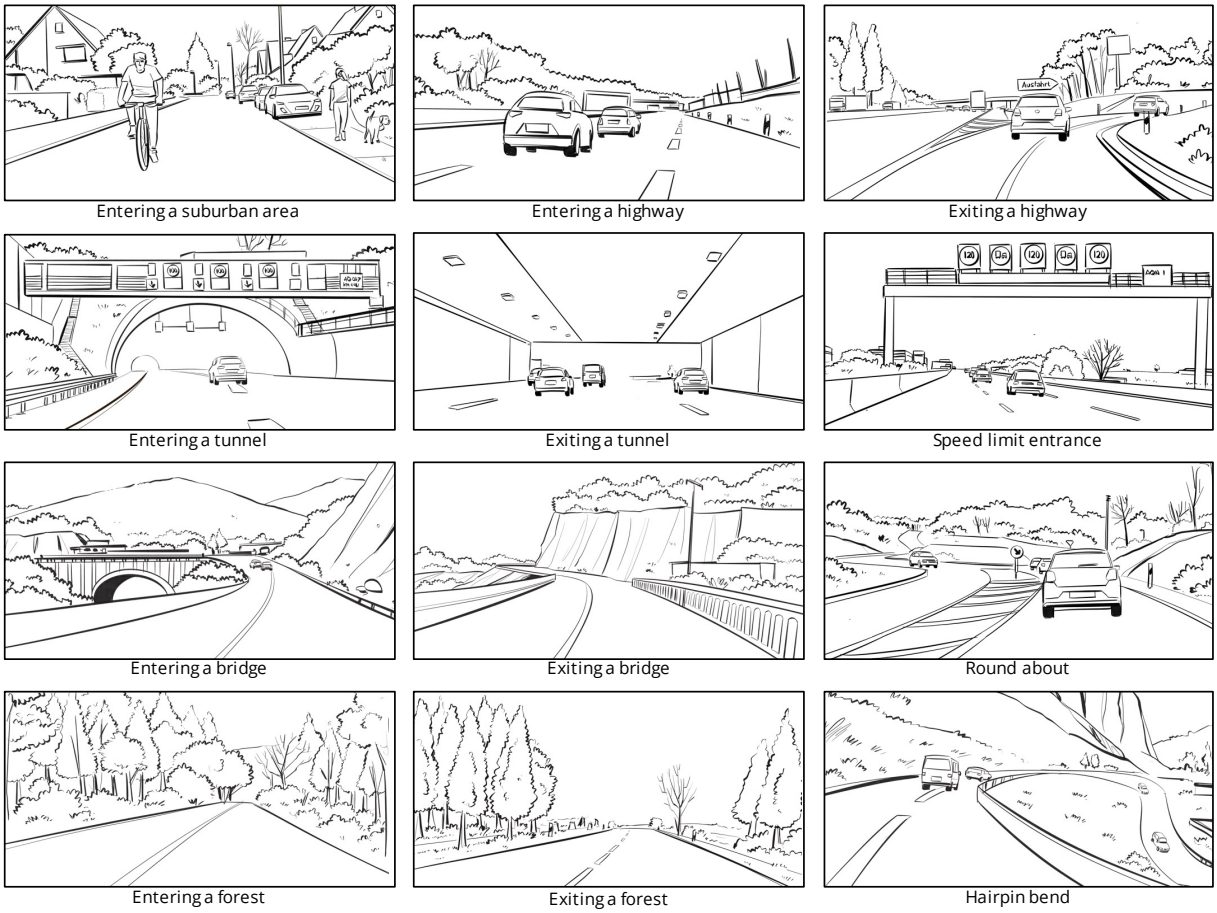


Figure 3.2: Examples for location-bound affordance situations. SoundsRide's design allows annotating any of these affordance situations on the route and to temporally and spatially align events in the music to them.

3.3 Affordance-Based Music

3.3.1 Sound Affordances: Features in the environment that lend themselves to acoustic events

Affordances are action possibilities [156]. We define *sound affordances* as momentous and well-noticeable events in the environment, characterized by a salient contrast in some user-perceivable aspect and offering the possibility to assign a certain musical event. These contrastive aspects are not limited to visual contrasts – such as entering or exiting a tree-lined alley, but might alternatively or additively refer to a contrast in *g*-force – e. g., when passing the crest of a hill, a contrast in sound perception – e. g., when entering or exiting a tunnel, a contrast in the traffic guidances – e. g., specification or revocation of a speed limit, or a contrast in the overall traffic flow – e. g., when entering or exiting a highway. In our understanding, sound affordances are constituted by an *affordance situation* that provides the opportunity for an *affordance action*.

Sound affordance situations can be either bound to the environment or bound to the user. User-bound affordances are endogenous and can either focus on the users' internal state or their observable behavior. On the other hand, environment-bound affordances are exogenous and comprise locations and traffic situations. While variables such as weather or daytime can serve

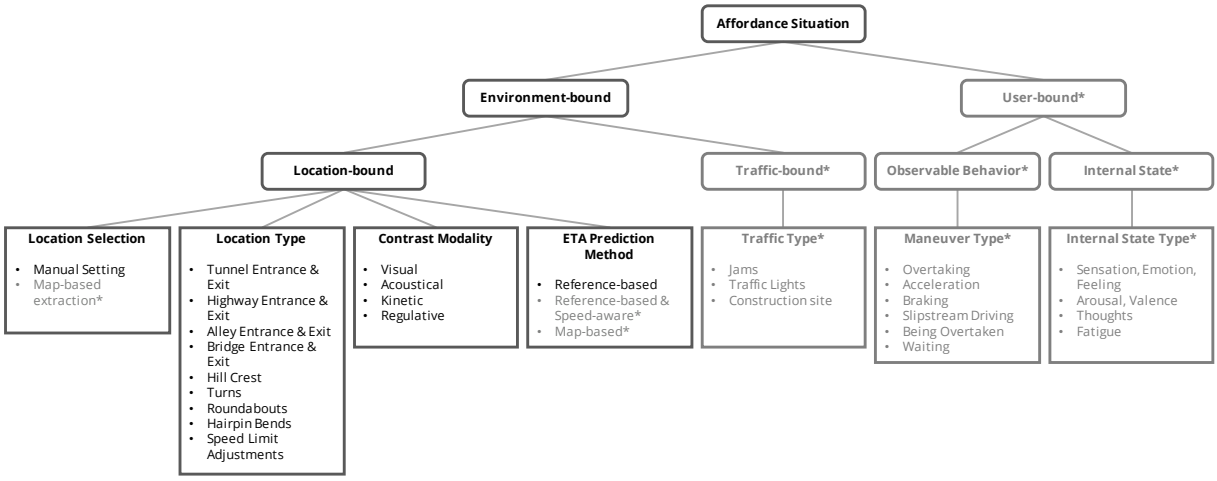


Figure 3.3: Taxonomy of affordance situations in affordance-based music. Design choices for SoundsRide are shown in black (without asterisk). SoundsRide is focused on location-bound affordance situations that are annotated by the user with a mobile app on a self-recorded GPS trajectory. This enables users to label any location as an affordance situation of a certain type and underline any contrast modality with music. The same GPS trajectory used for annotation is also used for affordance ETA prediction.

as more general context variables, they do not represent contrastive events in a ride that can be annexed for events in music. Within the space of location-bound affordance situations, relevant dimensions for positioning a location-bound affordance situation include the type of the location, the modality under which the contrast unfolds, the way it can be determined as an affordance, and the method that can be used to estimate the temporal distance to it. Figure 3.2 shows a set of examples for location-bound affordance situations.

A *sound affordance action* is taken by the affordance-handling system to create a musical audio signal from information about affordance situations. Context-aware music recommender systems such as [128] or [129] aim to reflect the *mood* in a driving segment with a length of minutes or tens of minutes in the music. We call this objective *macro-synchronicity*. Alignment in affordance-based music, first and foremost, aims at *micro-synchronicity*. Micro-synchronicity is characterized by temporal alignment of *events* in the music mix and estimated events in the environment in a subsecond-to-seconds time horizon. Temporal alignment will ensure spatial alignment given accurate ETA predictions. While micro-synchronicity can entail macro-synchronicity with a suitable song selection per affordance, the inverse is generally not true. Conceptually, concerning sound affordance actions, we distinguish between music playback, music mixing, and music generation. These different modes of affordance actions differ in terms of the degrees of freedom they can control to align music with affordance situations. In *music playback*, the song sequence is the only decision variable. This mode of operation is employed in context-aware music recommender systems and can only aim at macro-synchronicity. In contrast, music generation and music mixing can aim at micro-synchronicity. In *music generation*, MIDI-based techniques as found in procedural game music or music live-coding techniques are used to arrange elementary sounds or more complex instrument loops to create a novel piece of music, thus offering the maximum degree of freedom. In *music mixing*, decision variables include song-structure-aware fade-in and fade-out of songs, deliberate repetition of segments such as beats, bars, or parts, sound effects, frequency equalization or filtering, stretching, panning and balancing, transition effects, looping, etc. To ensure micro-synchronicity, a system needs to react to unexpected changes in ETA predictions by delaying or speeding-up a planned event in the

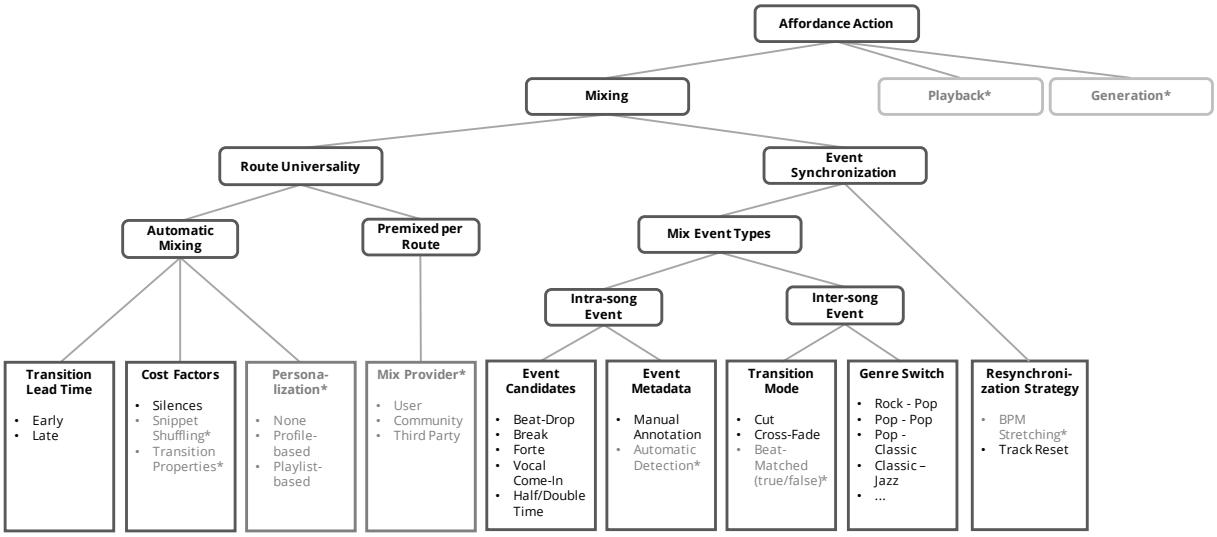


Figure 3.4: Taxonomy of affordance actions in affordance-based music. Design choices for SoundsRide are shown in black (without asterisk). SoundsRide is focused on mixing music so that intra-song or inter-song events are temporally and spatially synchronized with affordance situations. Synchronization takes place on a micro-level by track resetting alignment as well as on a macro-level by affordance-oriented song selection. It allows automatic mixing for any user-recorded and user-annotated GPS trajectory as well as aligning a provided route-specific mix for the current ride.

mix using a *resynchronization strategy* such as BPM stretching or track resetting. Events in the mix to be aligned are *inter-song transitions* or *intra-song features* that are either recognized automatically or annotated manually.

3.3.2 Affordance-Synchronized Music Mixing in SoundsRide

Situation-wise, location-bound affordances are particularly interesting for SoundsRide as we assume that 1) they are intuitively comprehensible to users, 2) are more robustly measurable than internal state, 3) at least partially capture the spirit of user-bound affordances implicitly, assuming that location influences the user's behavior and internal state, and 4) might increase the situational awareness beneficially for driving as opposed to user-bound affordances. Figure 3.3 summarizes our taxonomy of affordance situations and the position of SoundsRide.

Action-wise, for SoundsRide, we choose a music mixing approach that allows automatic mixing as well as aligning provided route-specific mixes for the current ride. We design SoundsRide for intra-song as well as inter-song events and use track resetting for delaying as well as speeding up planned events. Figure 3.4 summarizes our taxonomy of affordances and the position of SoundsRide.

In order to mix music based on location-bound sound affordances, SoundsRide tackles three problems: 1) Affordance ETA Prediction, 2) Mix Planning, and 3) Innovative Information Fusion.

3.3.2.1 Affordance ETA Prediction

The problem of estimating the temporal distance to a certain location in the route ahead is commonly solved in navigation systems such as Google Maps or Waze. However, the navigational use case of these systems typically tolerates a deviation of a couple of seconds. More precisely,

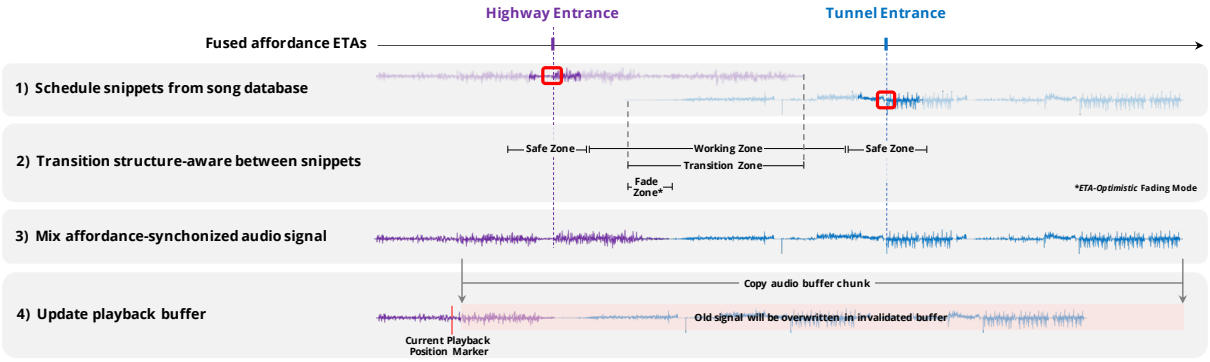


Figure 3.5: Whenever a resynchronization is triggered by the information fusion, SoundsRides overwrites the played-back audio buffer with an updated audio signal. This audio signal is produced by 1) aligning events in or transitions between songs snippets from a song database on an audio track along a given set of affordance ETAs, 2) fading these songs snippets in at some point in time, and fading them out at some point later in time. Fading times are based on the annotated structure of the song and the overlaps between the snippets. 3) ETA-aware positioning ensures alignment of song events in the mix with affordances in the environment. 4) Future chunks of the played-back audio buffer are then updated in-place.

navigational use cases do not require an accuracy exact to the second quite a time ahead of the actual event. However, to begin a music snippet so that an event in the music tens of seconds later is aligned with an upcoming sound affordance requires such an estimation - at least if last-second manipulations are to be avoided. We call this special case *micro-temporal estimation*.

For the purposes of our application, we approach the problem of micro-temporal estimation by allowing the user to first record a reference GPS trajectory, i. e., a path of GPS points. After recording, the user can mark points as affordance locations by tapping the respective pin (see Figure 3.1) and toggling through the offered affordance types. For the actual SoundsRide session, the user is re-localized against the reference GPS trajectory and the estimated time to arrival (ETA) to each of the succeeding sound affordances is easily computed as an aggregate over the temporal distance between pins. Figure 3.2 shows types of affordance situations that users might annotate on a route of their choosing, e. g., their commute.

3.3.2.2 Mix Planning

Given a specification of affordance ETAs from the step above, SoundsRide creates a mix plan with scheduled mix events. Figure 3.5 shows how a mix plan is generated from the affordance ETAs by aligning events in a song snippet (i. e., intra-song events) or transitions between songs (i. e., an inter-song event) with the ETAs, then transitioning between song snippets, and finally mixing the audio signal to be written to the played-back audio buffer. The mix event type is determined by a predefined mapping from affordance situations to mix event types. SoundsRide looks up songs in a database of annotated songs that contain events of the needed event type. From all matching song events, SoundsRide chooses the song that incurs the least cost. Cost is incurred if a song is so short that it ends before the next song can fade-in, hence resulting in silence. Overlaps between songs are eliminated by applying a set of cross-fading rules, making sure that cross-fades only take place beyond a safe zone around the aligned mix event. This step features a configurable parameter *ETA accuracy optimism* that determines how early or late the next song is faded in before the next affordance situation. Fading in early means that non-

negligible updates to the next affordance ETA will become more likely, whereas fading in late, e. g., 10 seconds before the next affordance, means that build-up time towards this affordance is reduced, thus possibly undesirably surprising the user. Finally, the mix plan is mixed to an audio signal. We employ a window-based approach that produces the audio signal only up to a defined time horizon in the future to avoid wasting computation time on a signal that will be overridden by another update anyways.

3.3.2.3 Innovative Information Fusion

After playback of a mixed audio signal has started, affordance ETAs continue to be updated at the GPS sample rate, typically revealing discrepancies between the novel and the previous ETA. Generally, as geographic distance to the affordance shrinks, the risk of error in the novel ETA shrinks and its trustworthiness increases. In order to incorporate innovative affordance ETA predictions, we developed a recursive filtering algorithm that continuously fuses the most recent affordance ETA predictions with previous predictions, and determines whether a resynchronization needs to be triggered.

This resynchronization decision is based on the 1) current timestamp, 2) the previously planned ETA of the next affordance, and 3) its novel ETA. We compute a set of descriptive metrics to estimate the current state of the world – that is the environment and the vehicle within that environment – and derive an audio updating strategy.

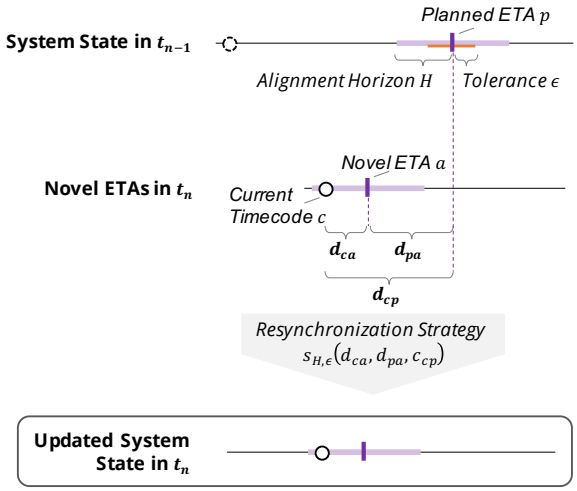


Figure 3.6: Time-to-arrival estimations are continuously updated by the localization module. To avoid continuous experience-degrading resynchronization updates to the audio signal, SoundsRide must carefully trade off updates against temporal misalignments between the events in music and in the environment. We implement this trade-off in a recursive filtering procedure that continuously fuses novel ETA predictions with former predictions to decide whether a resynchronization needs to be performed.

Figure 3.6 shows the overall approach. Algorithm 1 gives the in-depth procedure of assessing the state of the system given new ETAs. We *temporize*, i. e., we do not update to new affordance ETAs if there is a deviation between planned and innovative ETA, however it is so far in the future that any update made now is likely to be updated again anyways. This avoids unnecessary but experience-degrading updates. We *neglect* any deviation between planned and innovative ETA if it falls below a specified tolerance threshold. We *delay* the next mix event by updating the state of the affordance ETA predictions if the innovative affordance ETA is farther away than the planned affordance ETA. Analogously, we *accelerate* if the innovation suggests an earlier than expected event time. We *endure* a misalignment if the affordance is still ahead according to the innovation but should have been passed according to the plan. Only if a threshold is exceeded, we *redispatch* the song snippet that features the musical event. In summary, the updating behavior is controlled by the configurable parameters *misalignment tolerance* and the *alignment horizon*.

Algorithm 1: SoundsRide's ETA fusion algorithm for resynchronization

Input: current timestamp c, novel ETA timestamp a, previously planned ETA timestamp p, alignment horizon H, misalignment threshold ε
Output: Updating strategy s for the current timestep

```
s ← undefined
d_cp ← p − c, d_ca ← a − c, d_pa ← a − p
if d_ca ≥ 0 and d_cp ≥ 0 then
    if d_ca > H and d_cp > H then
        // beyond hot zones of novel and planned ETA
        s ← Temporize
    else if d_ca ≤ H and d_cp ≤ H then
        // within hot zone of novel and planned ETA
        if abs(d_pa) < ε then s ← NeglectMisalignment
        else if d_pa ≥ ε then s ← Delay
        else if d_pa ≤ −ε then s ← Accelerate
    else if d_ca ≥ H and d_cp ≤ H then
        // within hot zone of planned ETA
        if abs(d_pa) ≤ ε then s ← NeglectMisalignment
        else s ← Delay
    else if d_ca ≤ H and d_cp ≥ H then
        // within hot zone of novel ETA
        if abs(d_pa) ≤ ε then s ← NeglectMisalignment
        else s ← Accelerate
else
    if d_pa > ε then s ← RedispatchMissedAffordance
    else s ← EndureMissedAffordance
return s
```

3.4 Implementation

We built SoundsRide in a client-server setup. The *SoundsRide server* is responsible for managing the mixing plan, fusing new affordance ETA information with the mixing plan, computing the mixed audio signal, and forwarding the signal to the vehicle's audio output system via Bluetooth or latency-free AUX. It is implemented in Python. The *SoundsRide client* is responsible for localization using GPS and affordance ETA estimation. It is implemented in Swift for iOS. Inter-process communication is realized through remote procedure calls over HTTP/2 and WiFi using gRPC[5]. Mixing, playback from an audio buffer, real-time visualization, and server communication are all running in different threads on the server. Source code for both the client and the server software is available on GitHub. We chose this prototypical setup to enable rapid implementation on a full-fledged Python server while being able to either consume the GPS information from a common smartphone or from an experimentally tapped vehicle navigation system. For a production deployment, we could imagine both a purely Python-based version running on a vehicle-integrated and Linux-based computation unit, or running as a purely mobile-device-based app, or even running in such a distributed fashion, however delivering the final audio signal over cellular network.

[5] https://grpc.io/

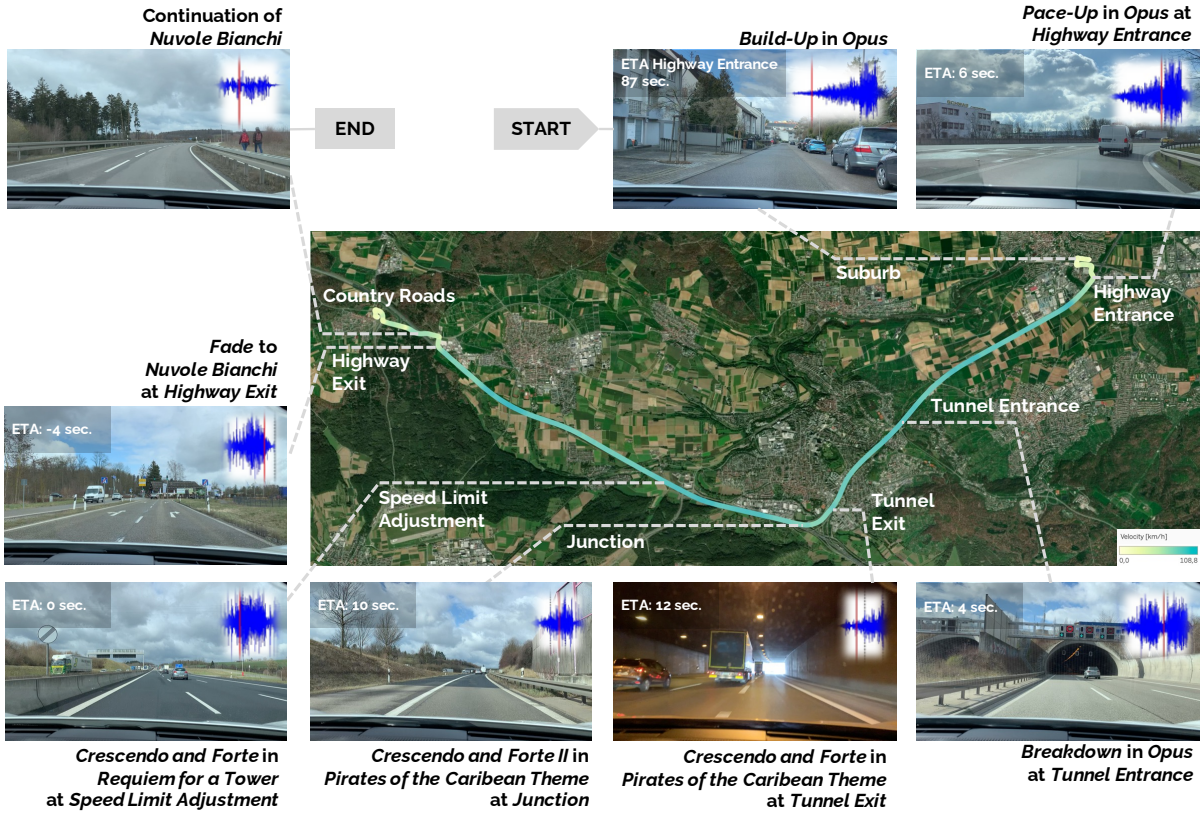


Figure 3.7: Route and Affordances for our User Evaluation in Southwest Direction. The red marker in the audio signal indicates the current position in the audio track. We selected a test route of 15.9 km length that takes about 12 to 15 minutes on average at average traffic with 4 affordances in northeast direction and 6 affordances in southwest direction shown here. We chose an electric vehicle from our institution. We recorded reference GPS trajectories in both directions. We employed SoundsRide's automatic mixing mode during the studies, but for comparability reasons, fixed the sequence of songs. Misalignment tolerance was set to 1000ms and the alignment horizon to 15s.

3.5 Exploratory User Evaluation

Participants and Apparatus To gain insights into how users perceive SoundsRide and which potential for immersion it offers, we conducted a user evaluation ($n = 8, n_{female} = 1, \mu_{age} = 32.9, \sigma_{age} = 8.0, \mu_{km\_pa\_driven} = 18000, \sigma_{km\_pa\_driven} = 11263$). We ran SoundsRide on a common smartphone (in our setup, an Apple iPhone 11 Pro) for GPS localization at 1 Hz and a common laptop (in our setup, a MacBook Pro A2141) for mixing. To measure user reactions to our system, we also tapped gas pedal position signals sent through the vehicle's communication bus system. We recorded the environment with a front-facing camera as well as the sound within the car with a microphone. Figure 3.7 shows the southwest direction of the test route and its sound affordances. We invited the participants one by one over two days to a location close to the route's start point, starting at approx. 9 am and ending at approx. 21 pm, thus covering different daytime conditions. In one half of the cases, we started with the southwest direction, in the other half, we started the same route in the northeast direction.

Rationale We choose a field setup over a simulator for two reasons. First, we want to examine user impressions of SoundsRide's real-time capabilities rather than the idea under

ideal conditions. Second, as SoundsRide does not assist in a driving task but aims at evoking subjective impressions, we give weight to the fidelity of the evaluation and want to ensure full visual, physical and auditory contrasts, e. g., at the tunnel entrance affordance. On the other hand, a simulation might entail questions on the real-world transferability.

Following Qin et al. [272] for immersion in digital gaming, control, concentration and comprehension serve as explanatory factors for immersion. In particular, comprehension is a necessary condition for immersion. Therefore, first, we wanted to understand how much sense SoundsRide makes to users intuitively, i. e., without exactly knowing its functionality. Once we have investigated this necessary condition, we evaluate which potential for immersion it offers to users.

Therefore, after a baseline part, the subsequent procedure was composed of two main segments, namely 1) investigating perceptibility, and 2) investigating the potential for immersion. The rationale for conducting separate studies to investigate perceptibility and potential for immersion is twofold:

1. By not previously telling participants what will happen, we do not expose them to confirmation bias when investigating how comprehensible the system is to users and what impression it evokes in them
2. By separating both study aspects, we can more reliably isolate insights concerning technical design aspects from insights concerning conceptual design aspects. If users were to undertake only an immersion study, in the case of a participant showing no indications of immersion, it is unclear whether this is *because of not* perceiving the system or *despite* perceiving it.

3.5.1 Study Segment 1: Investigating Perceptibility

3.5.1.1 Procedure

To investigate the system's perceptibility, each participant was shortly briefed with respect to the abstract capability of the system – that is that the system is "able to align certain events in the music with certain events during the ride". They were *not* informed about the concrete types of sound affordances supported. Further, we asked participants to "think out loud" and state their assumptions about what they believe the system is doing. We navigated participants by verbal instructions. After arriving at the end of the test route, we ask a cascade of questions to understand the system's percepbility to users, asking for free recall of affordances, then cued recall to video recall or no qualification as a music-environment-aligned affordance. The caption in Table 3.1 elaborates on the details of the procedure.

3.5.1.2 Results and Discussion

Table 3.1 shows the data collected during the study and aggregations thereof. 67.5 % of all affordances were either recalled freely or recalled after a cue question. 25.0 % of all affordances were not recognized by participants during the ride, however were qualified as such when reviewing the ride on the recorded video.

Both P1 and P4 did not remember a music synchronization at the highway entrance, but expressed puzzlement when looking at the recording and asked themselves why they did not notice

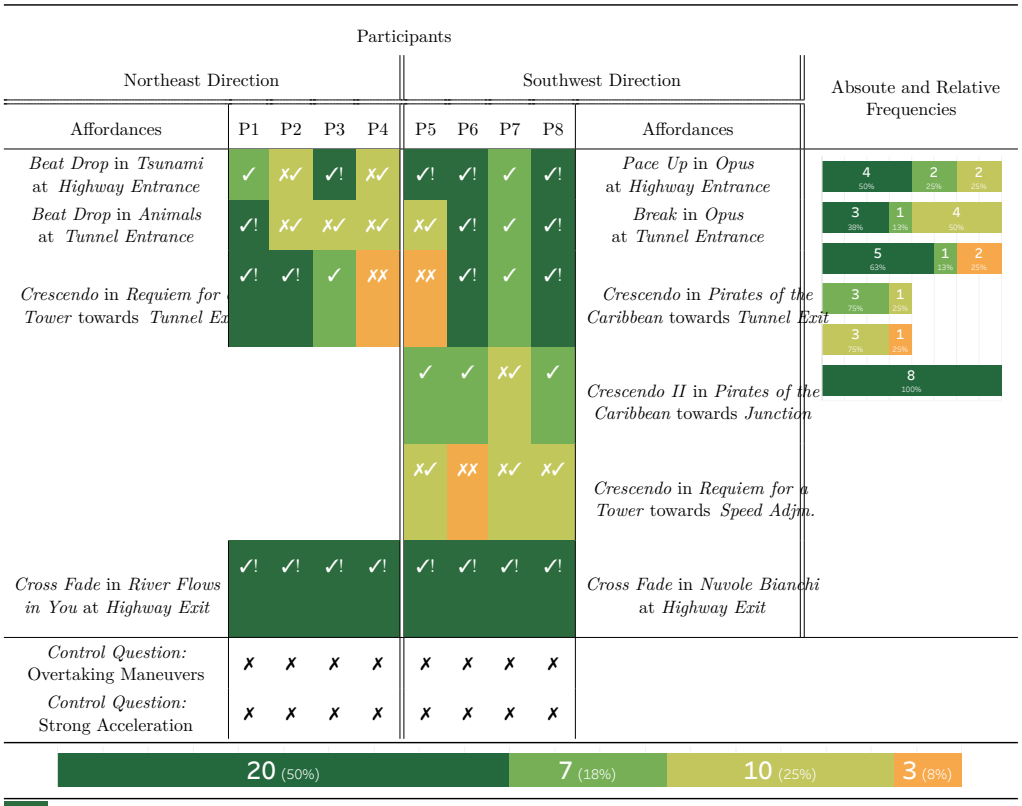


Table 3.1: After arriving at the end of the test route, we first asked the participant to list the most memorable events from the ride. Then, for each non-mentioned affordance as well as for two control items (i. e., environment events SoundsRide does not take into account, namely overtaking maneuvers and strong acceleration), we asked participants directly whether they perceived it (e. g., "From what you remember, did you perceive it as if a beat drop was aligned with the tunnel entrance?"). If answered negatively, we showed the participant the front-facing video recording of the ride including sound, and asked again shortly after the sound affordance took place (e. g., "Looking at the video recording, do you perceive it as if a beat drop was aligned with the tunnel entrance?"), to understand whether the reason for not remembering was not perceiving it *just in that moment* while during, e. g., due to high cognitive load, or whether the reason lied in generally not appreciating the affordance, e.ġ, because of a too large time delta between affordance trigger and affordance action or too much subjective indifference towards the contrast in music or environment.

the well-aligned and lucid beat drop in Tsunami. Similarly, P8 did not remember music synchronization with the speed limit adjustment, however suddenly remembered that they were very focused on overtaking a "white Tesla". In the later study segment on immersion, P8 also stated that they felt the high-energy music during the highway entrance was not supporting their mindfulness. Possibly, P1 and P4 blocked out the music to direct their attention to merge on the highway. *From this we conclude* that perceptibility must not be an overarching objective of the system. On the contrary, the system should fade into the background during critical segments such as highway entrances and only become noticeable again once the segment is finished. This can be realized by either moving the affordance location forward, selecting the music accordingly or both.

While P3, P4, and P7, missed two affordances each, P1 and P6 did not miss any. This does not correlate to the data on driving experience in age or km driven per year. However, as 6 out of 10 misses happened at the NE tunnel entrance and SW speed adjustment, both not visible from far as the former is hidden behind the curve and the latter is only a small road sign, we hypothesize that users miss affordances without sufficient time for anticipating or following the build-up in music. As inter-song events do not benefit from anticipation and are more distinct, they might be better suited for such locations.

In 7.5 % of affordances, participants rejected the notion that the music was adapting synchronously to the environment also after looking at the recording. In the case of P5 approaching the tunnel exit, a loss of the GPS signal inside the tunnel did result in a misalignment between music and tunnel exit of approx. 5 seconds, thus failing to keep music and environment in sync. In the case of P4 approaching the tunnel exit, the participant stated that the change in music at the tunnel exit was not enough to make a difference. Similarly, in the case of P6 approaching the speed limit adjustment, the same audio track was rejected for being too monotonous. However, the tunnel exit with the same orchestral piece was remembered by other participants 3 times. *From this we conclude* that the subjective feeling of energy a build-up or climax evokes is a significant determinant in the experience of SoundsRide.

All participants could freely recall the fade from orchestral music to piano music towards the highway exit. The speed limit adjustment was not recalled freely once. Surprisingly, *all* of the control item questions were correctly rejected by participants, strengthening our confidence in the validity of the answers overall.

Overall, we conclude that SoundsRide is well comprehensible to users but subjectivity of musical perception and mental load reduce perceptibility, first, indicating the need for customization of the song selection per user, and second, underlining the necessity to analyze the implications of SoundsRide for driving safety.

3.5.2 Study Segment 2: Investigating the Potential for Immersion

In the subsequent *Potential for Immersion study*, we wanted to understand whether the system allows users to immerse deeper in the music or in the environment than without SoundsRide. More specifically, we want to understand the experience of 1) synchronicity, 2) affordances, 3) mixing, 3) the effect on driving safety and 5) overall immersion.

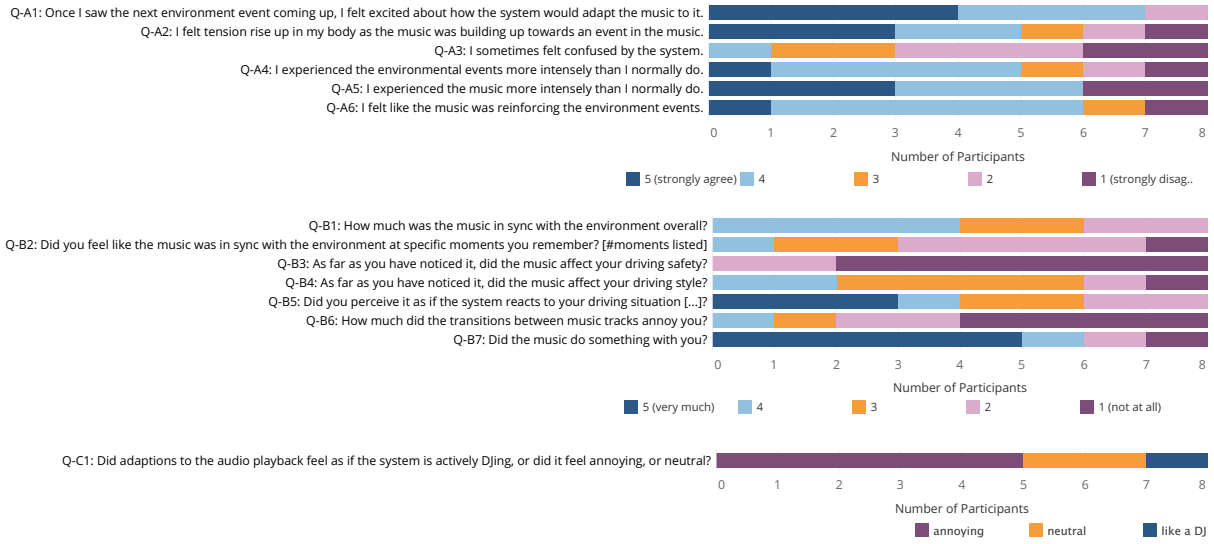


Figure 3.8: Participant's responses to our immersion questionnaire. After arriving at the end of the test route in study segment 2 on immersion, we employed a questionnaire to understand users' experiences in terms of affordances, mixing, and synchronicity, to understand the subjectively perceived effect on driving safety, and to understand the overall potential for immersion.

3.5.2.1 Procedure

We disclosed the affordances (using the word "event") taken into account by SoundsRide to align the music for the participants and asked them to drive in the opposite direction of the route for the perceptibility study. We announced the next affordance 30 to 60 seconds before it took place. Once again, we asked participants to "think out loud" and describe what they are thinking or feeling. Again, after arriving at the end of the test route, we employed a questionnaire.

3.5.2.2 Results and Discussion

Experience of Synchronicity Regarding *micro-synchronicity*, we found that participants judged temporal misalignments subjectively. While P4 negatively commented on the forte at the speed limit adjustment being off by approx. 1.5 seconds, P2 commented positively at the same affordance and offset. 4 out of 8 participants stated that they were very much (1x) or much (3x) disappointed when the system would miss an expected affordance by more than a second. Two participants stated the question was not applicable given they could not remember such situations. Two participants expressed little to no disappointment, indicating little emotional attachment. Also, we found that affordance positioning is user-dependent in some parts. While all participants gave positive feedback concerning the fade to the piano piece at the highway exit, P5 added that it was taking place too early in their mind as they drive the curve diverging from the highway at a higher speed than is adequate for the calming effect of the piano. Regarding *macro-synchronicity*, P4, P7, and P8 noted that the continuation of the orchestral "Requiem for a Tower" lost its fit to the environment continuously after the respective event (speed limit adjustment or tunnel exit respectively) had passed. In particular, P8 said that the continuation of the song would naturally lead to music events that are decoupled from the environment, even though they would not perceive this as detrimental to the experience. P1, P2, P3, P4, and P7 noted that overall synchronicity was impaired for a short duration at a traffic light where the

music was already building up to a beat drop while still waiting at the red traffic light.

Experience of Affordances *Situation-wise*, P5 and P6 dismissed the notion that a tunnel is an event during the ride worth synchronizing the music for. P6 reported that only the highway exit with the synchronized fade to piano music captured their attention while none of the other songs or pieces "did anything to me". However, all other participants deemed the affordance situations fitting and interesting. *Action-wise*, all participants favored the cross-fade to a piano piece at the highway exit. However, P6 stated, except for the piano piece, no piece or song was in the domain of their music preference and thus interesting to them. As a result, while they would auditorily recognize environment-triggered developments in the music, they did have the interest to actively follow it. P5 stated that driving is a "mechanical task that needs to be done" and not more, and hence "they block out the environment as far as it is not needed for driving safely". Therefore, the desire to listen to a certain song or is not a function of the ride or the location, but of their current mood. Except for P5 and P6, participants also liked the beat-drop in Animals (northeast direction) and the break in Opus (southwest direction) towards the tunnel entrance, however the beat-drop evoked more vivid feedback. Participants also expressed liking towards the orchestral crescendo at the tunnel exit. The continuous progression in tempo in Opus and the pace-up at the highway entrance was generally well-appreciated, however – as described in the results on study segment 1 – P8 expressed that they would have preferred music that enables them to better concentrate on the merge.

Experience of Mixing 7 out of 8 participants stated that the transitions from one song to the next, and the concomitant less-than-usual length of the song did not invoke stress, while one participant stated to experience light stress. However, 5 out of 8 participants reported that adaptions of an already playing song were annoying. These were the participants, who also noted that the traffic light was a source of asynchronicity, namely between a build-up in music and an unvaried environment. In particular, P2 described it "as being in a loop where the music gets faster, then gets slower, then gets faster and slower again while nothing actually happens". The described system behavior is a result of the system resetting the track to avoid a premature event in the music. However, this phenomenon generally did not lead to confusion, considering only 1 out of 8 participants agreed they felt confused by the system. Participants did not comment or call out on audio track modifications that were taking place very shortly, approx. 2 seconds, before the affordance situation took place. From this, we conclude that improving affordance ETAs on a larger scale are of first priority, e. g., by taking traffic lights into account during planning, while avoiding modifications in the last one or two seconds are of subordinate priority.

Effect on Driving Safety It is of crucial importance that SoundsRide does not impair driving safety through mechanisms of adverse incentives or distraction. 2 out of 8 participants responded with a value of 2 on a scale from 1 (not at all) to 5 (very much) to the question "As far as you have noticed, how much did the music affect your driving safety". One of these two participants has indicated the same in the questionnaire precursory to the ride to the same question, however in the general context of in-car music. This means SoundsRide did not negatively impact the participant compared to a baseline, but also did not manage to eliminate this by guiding focus towards the environment. Further, P8 said that the pace-up on the highway entrance was opposite to their need for deep focus. 6 out of 8 participants responded with a value of 1. P7 stated that they would normally drive too fast after driving off the highway,

however due to the switch to piano music, they'd prefer "just coasting along". P2 stated that the music helped them concentrate by directing their focus on the environment and prepare for the next driving maneuver.

From these answers, we conceive that SoundsRide can *negatively influence* safety by incentivizing users to increase velocity with increasing energy in music, e. g., at the highway entrance, or by distracting them from the driving situation. On the other hand, we conceive that SoundsRide can *positively influence* safety by incentivizing users to decrease velocity with decreasing energy in music, e. g., slowing down according to calm piano music at the highway exit, or by increasing environment awareness, e. g., so that users slow down early enough before the highway exit as the song starts fading out. *Overall*, we derive the hypothesis that SoundsRide has both the potential to slightly increase as well as slightly decrease driving safety. As a consequence, further research needs to investigate which properties in the situation and user prejudice or contribute to safety and how to eliminate or foster these effects, e.g., through careful positioning of affordances before or behind precarious points on the route.

Overall Immersion Especially building on Brown and Cairns [42] as well as Qin et al. [272], Georgiou and Kyze [102] propose constructs to be measured for immersion as well a questionnaire to operationalize these constructs in a study. Three consecutive levels of immersion are distinguished: engagement, engrossment, and total immersion. Classically, Level 1, engagement, represents the basic level and is characterized by the will to interact with a system. Level 2, engrossment, is characterized by an emotional attachment. Level 3, total immersion, is characterized by presence, which of course cannot be a goal for driver-directed applications. While Georgiou and Kyze's questionnaire is designed for evaluating see-through AR games, we took it as as a starting point for our in-car audio AR evaluation questionnaire. *Considering the aggregated data*, 7 out of 8 participants agreed, some strongly, that they were excited about how the music would adapt to the next environment event (i. e., the affordance situation), once they saw it coming up. 5 out of 8 agreed, some strongly, that they felt the tension of a build-up in music towards an event physically. 6 out of 8 agreed, one strongly, that the music was reinforcing the environment events. 6 out of 8 agreed, some strongly, that they experienced the music more intensely than they normally do. 5 out of 8 participants agreed, some strongly, that they experienced the affordance situations more intensely than normally. Of course, attention bias due to the study setup mandates to interpret responses to both former questions cautiously. *Considering the individual data*, as described above, P6 stated multiple times that the system did not at all capture their attention or emotion because the music did not mean anything to them. However, the participant was open to retrying the system should it support custom music selections. On the contrary, P5 rejected the premise of environment-induced music overall. On the other hand, P3 stated they "felt like being in a car advertisement". P2 stated that they "didn't know how spectacular merging on a highway could feel". P1 stated that the system put them "in a spirit of optimism and anticipation for the highway". On a more general note, the participants' statements reveal that the quality of the experience is subject to a range of context variables. Concerning *traffic*, P3 stated that the energetic build-up in music at the highway entrance evoked the "urge to accelerate", but the "traffic wouldn't allow it". Concerning *aesthics of the scene*, P7 stated that they felt as if they were "experiencing something very special, but the grey trucks bring me down to earth again". Concerning *weather*, P7 expressed the wish that "the rain should reflect in music to convey a sense of melancholy". Concerning *daytime*, P4 said that the "gloomy music [of Requiem for a Tower] fits the mood". Concerning *vehicle type*, P3 indicated that the "electric engine works well with the peaceful piano at slow down" behind the

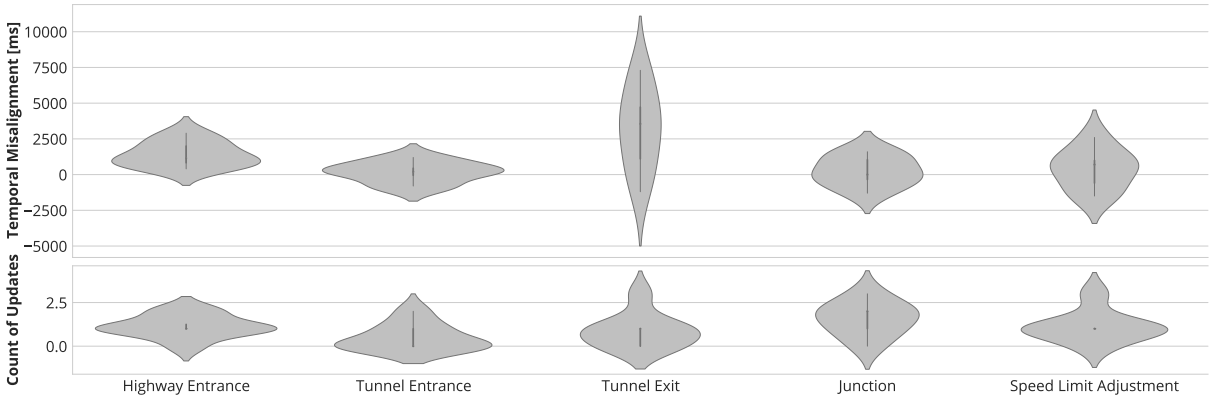


Figure 3.9: Distribution of temporal misalignment and number of updates per affordance across 8 rides, each with a misalignment tolerance of 1000 ms and an alignment horizon of 15 s. In 77.7 % of cases (73.7 % without the highway exit affordance), SoundsRide is able to ensure synchronicity with a maximum misalignment of ±1.9s and a maximum number of 2 updates to the audio track.

highway exit, whereas P6 noted that "a combustion engine fits the emotional experience better".

Overall, from the aggregate statistics and the individual responses, we conclude that SoundsRide does offer the potential for immersive experiences through emotional attachment, assuming a user is not rejecting the system's fundamental premise of affordance-based music. However, we note that profile-based or personalized music could also capture the attention of users whose music preferences were not covered in the selection, and that SoundsRide's experience is dependent on a variety of context variables which the system could potentially account for in the future.

3.6 Technical Evaluation

Since SoundsRide focuses on affordance synchronization, the main metric to evaluate its overall technical performance is given through temporal misalignment, i. e., the temporal distance between the mix event and the environment event. However, as temporal misalignment can be traded-off against re-synchronization updates to the audio signal, we also consider the number of audio signal updates before (speed-up or delay update) and after the affordance situation (redispatch).

3.6.1 Procedure

We recorded the rides during the exploratory user evaluation with a front-facing camera and incl. sound. Due to the nature of the setup, these rides feature different users, driving styles, and environment conditions, in particular rush hours in the morning and the afternoon, and free roads in the evening. We annotated the videos using a video editing software for all 48 affordances with the temporal misalignment and the number of audio track per affordance.

3.6.2 Results

Figure 3.9 shows the distribution of the temporal misalignment ($\mu = 942\,ms$, $\sigma = 1813\,ms$) and the number of update per affordance across all 8 rides.

Except for the tunnel exit affordance, all events were in a range of no more than four seconds. However, in approx. half of the cases (47.7 %, including the tunnel exit affordance) SoundsRide is able to ensure synchronicity within a time horizon of $\pm 1.1s$ and with no more than 1 track update in the 15 seconds before or after the affordance location is passed. In 77.7 % of cases, SoundsRide is able to ensure synchronicity with a maximum misalignment of $\pm 1.9s$ and a maximum number of 2 updates to the audio track. All rides were parameterized with a misalignment tolerance of $1000ms$ and an alignment horizon of $15s$.

The striking elongated distribution at the tunnel exit results from the decreased accuracy of the GPS signal in the tunnel. The number of updates is higher at average at the highway entrance and at the junction than at the tunnel entrance, as the exact timing of these affordance situations is more influenced by traffic and thus more difficult to predict from the reference trajectory. At the highway exit, we deliberately employed a long-running cross-fade of 10 seconds aligned with the exit lane branching from the highway without a point-exact location. In all rides, the affordance action was taken on the exit lane, thus meeting this condition.

3.7 Limitations and Future Work

Mixing The most frequent criticism of participants in the evaluation concerned the track resetting procedure. While we employ a user-parametrizable filtering procedure to minimize the number of noticeable mix modifications given new ETA information, updates that do take place are typically noticeable to users. By analyzing a song, e. g., detecting its BPM and segmenting it in individual bars and phases, music-theoretically more sound updating procedures to achieve temporal alignment can be imagined, e. g., bar or phase repetitions. In particular, employing BPM stretching in an acoustically safe range while simultaneously correcting for pitch could induce a very interesting, possibly unnoticeable, audio effect. That is, pushing the gas pedal or the brake pedal will indirectly lead to speeding up or slowing down the music in an effort of the system to reach the aligned event in the music earlier or later than originally planned.

Synchronicity Macro-synchronicity can be increased by manually annotating traffic lights or inferring them from rests in the GPS trajectory during route recording. Then, the system can ensure to only transition to the next song featuring the next musical event, once the traffic light has been passed, thus further reducing track resets from unexpected delays or speed-ups. To improve micro-synchronicity in general, reference-based ETA prediction could be enriched with speed awareness and acceleration awareness. To improve micro-synchronicity at tunnel exits specifically, by integrating SoundsRide with a vehicular information system, the GPS localization from a smartphone could be fused with odometric information from the vehicle's wheel speed sensor, compensating the loss of the GPS signal in the tunnel.

Affordance Situations We have designed SoundsRide for full flexibility on location-bound affordance situations. By first recording a reference route, users, community members, or third-party providers can freely annotate location-bound affordances. At the same time, the reference route is taken as a basis for predicting affordance ETAs. However, extracting location-bound affordances such as highways, tunnels, hill crests, etc. from and predicting their ETAs with mapping services such as OpenStreetMap could enable SoundsRide for one-step usage without manual annotation efforts just by entering the route to be driven. Going even further, continuous

automatic inference of the most probable next route segment the user will drive, even without explicit entry of a destination, might allow determining the next affordance on-the-fly.

Affordance Actions SoundsRide works best with songs featuring high contrastiveness and mixability. Therefore, EDM, orchestral, jazz-only, and piano-only music works particularly well. Nu metal, pop, rock, and pop-rock also work well in terms of contrasts, however, mixing runs the risk of pulling apart vocals when skipping or rewinding for resynchronization and fading. To make SoundsRide more general, future work can explore three strategies. First, by expanding the current structural awareness to vocal awareness, i. e., knowledge of timecodes of verses, the system could jump to dedicated markers, thereby keeping vocals together. Second, by further investigating how to improve micro-temporal estimation, the number of resynchronization updates and thus the risk of cutting vocals is reduced. Third, the adoption of BPM stretching as presented above could avoid jumps due to resynchronization entirely. In order to feature songs of different genres in the same mix, research into the automation of music-theoretically informed harmonic song transitions [335] gains in importance. While it is already perfectly possible to create a custom song database, this requires annotating each song with intra-song events and defining the mapping between sound affordance situations and the respective intra-song event. However, adopting approaches of structural segmentation [249] to detect segment boundaries and classes in songs could allow users to simply hand-over a playlist to the system before the next ride, thus conveniently enabling personalization and likely increasing the potential for immersion. Additionally, including further affordance actions such as balancing the music from front to back speakers when entering a tunnel, or panning the music from left to right when entering a highway based on spatial audio, or even employing other modalities such as actuated seats, could enrich SoundsRide's experience and increase its potential for immersion.

Adaption to Automated Driving The trend towards driving automation affects a variety of the system's aspects. For example, driving scenario detection could enable dynamic affordances that react on-the-fly to triggers in traffic, e. g., allowing to inject energetic music into the mix when overtaking or being overtaken. Visualizing affordance situations in the driving scene display could enable possibly interactive monitoring. Advanced depth-sensing might allow correcting ETAs on short term based on the environment scans, e. g., when detecting a forest boundary. Overall, ETA accuracy might benefit from automated driving due to improved predictability in the driving behavior. On the other hand, automation might impair the UX as the driver possibly starts shutting out the environment. Therefore, future work might also explore novel interaction patterns, e. g., where a user co-creates the mix plan by scheduling affordances on short-term notice and on-the-fly based on music choices offered by the system. Also, we imagine the system might enable audio-based in-car games, e. g. passengers guessing the current scene blindfolded based on music only.

Domain Adaption Finally, we see potential to transfer SoundsRide to other means of locomotion in general and bicycles in particular, e. g., in order to synchronize energetic music with uphill segments on the ride, or synchronize the transition from bicycle trails to public roads.

3.8 Conclusion

We presented SoundsRide, an in-car audio augmented reality system that synchronizes high-contrast events in music with high-contrast events in the environment. Our core technical contribution lies in predictive real-time music mixing, enabled by a novel approach comprising affordance ETA prediction, mix planning, and innovative information fusion.

After positioning SoundsRide in the design space of affordance-based music, we describe our technical approach and its implementation. Given the estimated temporal distance to location-bound affordances along the ride, we heuristically schedule a cost-aware music mix with intra-song and inter-song events that are temporally aligned with the affordances. Then, by continuously piping updated temporal affordances distances through a recursive filter, we determine any necessary updates to the audio signal while trading off manipulations of the audio signal against misalignments between music and environment.

On the one hand, using SoundsRide's automatic mixing mode, users are enabled to align events in any set of annotated songs with configured affordance situations along their ride, thus making it applicable for everyday rides. On the other hand, using a predefined mix possibly offered by a regional provider such as a national park service or a local tourism association, we also envision SoundsRide to further increase immersion for particularly captivating or scenic routes.

In our quantitative evaluation of SoundsRide's performance, we find that SoundsRide can convincingly ensure synchronicity of affordance situations in the environment and affordance actions in the music. In our qualitative evaluation, we find that SoundsRide's affordances as well as its mixing and synchronicity properties are well-received, thus enabling suspenseful and engrossing experiences, and that driving safety can be either slightly increased or decreased, depending on the mix and user.

4

TransforMR: Pose-Aware Object Substitution for Composing Alternate Mixed Realities

Figure 4.1: TransforMR is a video see-through mixed reality system for handheld devices that performs 3D pose-aware object substitution to create meaningful mixed reality scenes, enabling applications such as alternate mixed realities or real-time virtual character animation in context. (a) From just the monocular color camera of a mobile device, (b) TransforMR performs instance segmentation, (c) object removal, and (d) 3D pose estimation to substitute objects with pose awareness (top right).

In real-time and for previously unseen and unprepared real-world environments, TransforMR composes mixed reality scenes so that virtual objects assume behavioral and environment-contextual properties of replaced real-world objects. This yields meaningful, coherent, and human-interpretable scenes, not yet demonstrated by today's augmentation techniques. TransforMR creates these experiences through our novel pose-aware object substitution method building on different 3D object pose estimators, instance segmentation, video inpainting, and pose-

aware object rendering. TransforMR is designed for use in the real-world, supporting the substitution of humans and vehicles in everyday scenes, and runs on mobile devices using just their monocular RGB camera feed as input. We evaluated TransforMR with eight participants in an uncontrolled city environment employing different transformation themes. Applications of TransforMR include real-time character animation analogous to motion capturing in professional film making, however without the need for preparation of either the scene or the actor, as well as narrative-driven experiences that allow users to explore fictional parallel universes in mixed reality.

4.1 Introduction

Continuous advances in *geometric* scene understanding have contributed to the *physical* coherence of virtual objects in mixed reality scenes, for example through improvements in mesh reconstruction [374], occlusion shading [78], visual-inertial odometry [341; 178], or light source estimation [352]. Research on these topics is increasingly dedicated to extracting semantic information from the real-world scene [184; 110] to enable novel mixed reality (MR) experiences [226] or context-aware interactions between virtual-world characters and the real-world environment [323].

However, both *semantic* scene understanding and *functional*—rather than physical—reasoning [385] remain hard problems. Creating an alternate reality in MR from scratch that augments a real-world city scene is a considerable challenge. Take the example of SciFi-like hover cars that pace down the streets: The mixed reality system first needs to perform scene understanding, including recognizing lanes and the driving direction. To make virtual hover cars halt at a crossing zone while virtual robot pedestrians leave a real-world store entry, cross in front of the hover cars, and wait at a bus station, the system would then need to detect the crossing zones, store entries, bus stations, and sidewalks. To create such novel mixed reality experiences from scratch requires a level of scene understanding that draws on significant advances in machine perception, such as spatial scene decomposition and conceptual reasoning.

In this chapter, we propose *TransforMR*, a mixed reality system for theme-guided scene transformation through *pose-aware object substitution*. TransforMR is capable of creating such meaningful mixed-reality scenes as in the example, showing and letting the user interact inside alternate mixed realities that are situated in the real-world context. TransforMR accomplishes this by repurposing existing physical objects in the scene as proxy objects that transfer their semantics in their respective environment to virtual objects. In the scenes created by TransforMR, users may attribute behavioral and environment-contextual properties of replaced real-world objects to the virtual objects. This creates semantically consistent and more plausible interactions compared to virtually augmented objects that do not inherit real-world object context and merely co-exist in the real-world surroundings.

Our system transforms visual recordings on-the-fly and is independent of a specific environment, therefore also applicable in previously unseen scenes and locations. TransforMR processes the feed of a monocular RGB camera to derive a virtual scene through a pipeline of perception, transformation, and construction. In the *perception* step, we integrate deep learning models that run on a multi-GPU-accelerated back-end, and therefore offload all processing from the mobile device. Our back-end system analyzes the streamed-in video through semantic 2D instance segmentation [127] as well as 3D human keypoint [260; 223] and 6 degrees-of-freedom vehicle pose [207] estimation. For *transformation*, TransforMR logically maps recognized objects to virtual

objects according to a selected theme. In the *construction* step, TransforMR first removes the physical objects using 2D segmentation information and real-time video inpainting [182]. Lastly, TransforMR derives the final scene from projecting the theme-specific objects into the scene using the 3D pose information. Our use of inpainting allows transformed objects to occupy less display space than the removed objects, reconstructing the background where needed.

Figure 4.1 shows TransforMR in action. Here, a user is exploring the transformation of reality by looking through the tablet while freely walking through the real world. TransforMR substituted all pedestrians and vehicles from the scene with semantically corresponding objects from the "Animals" theme. As depicted in the figure, the transformed objects are shown in the context of all physical surroundings, allowing the user to maintain their frame of reference for safe navigation (e.g., when climbing stairs).

To the best of our knowledge, TransforMR is the first system with the capability of 3D pose-aware object substitution in unbounded, unprepared, and unseen environments with visually complex scenes. We enable this unprecedented live mixed reality experience with the sole requirement of a single RGB camera, making our system suitable for broad applicability in lower-end phones or tablets and high-end devices alike.

Contributions

Taken together, we make the following contributions in this chapter:

- Pose-aware object substitution as a novel technique to creating meaningful, theme-based alternate mixed reality scenes with virtual objects that assume behavioral and environment-contextual properties of replaced real-world objects,
- A camera-to-display system architecture, implementation, and design rationale for "TransforMR", a distributed mixed reality system that adapts, unifies, and integrates a series of deep learning-based 2D and 3D scene perception architectures as well as video inpainting for operating in-the-wild in real-time in unseen environments on commodity mobile devices,
- A parallization architecture employing three-step pipeline parallelism and three-fold task parallelism to achieve near-real-time operation of the integrated computer vision models at approximately 15 frames per second,
- An evaluation and discussion of the qualitative and technical aspects of TransforMR,
- Applications of TransforMR that comprise real-time character animation in real-world context and narrative-driven, consumptive experiences of alternate mixed reality scenes.

4.2 Background and Related Work

TransforMR composes the virtual scene with the context provided by the real scene. Previous work that uses real-world scene context for virtual-scene composition includes geometry- and depth-aware AR, superposition-based AR, and physicality-aware VR. As pose-aware object substitution features an object removal procedure, we consider diminished reality as well as as a

pipeline of diminished and augmented reality as related areas. Given the transformative character of TransforMR, we consider 3D scene reconstruction and transformation, as well as visual transformation as related research.

4.2.1 Context-Aware Mixed Reality

Geometry-aware AR and *Depth-aware AR* as implemented in Apple's ARKit[1] and Google's ARCore[2] enable applications to additively render new objects into a real scene while respecting its the geometrical context, providing capabilities for collision detection between real and virtual objects (e.g., virtual rain drops hitting the real ground, or virtual balls bouncing off the real walls), occlusion shading (i.e., partially covering virtual objects by real objects), or depth-of-field effects (e.g., bokeh or fog), using depth-from-motion approaches or depth sensors [78; 331]. Moreover, Nuernberger et al. [254] explore a concept for aligning virtual objects with edges in the real world. However, semantically meaningful augmentations are still difficult to achieve, in particular in an automated fashion without user guidance such as anchor setting, since they require the system to not only have an accurate geometric representation, but also a purposeful semantic representation beforehand. Furthermore, they do not feature object replacement or removal procedures.

Superposition-based AR is a widely established approach for anchoring a virtual object with a concealed real-world object, that is seen for the first time (e.g., a face) or has been incorporated into a set of reference objects. Examples include Annexing Reality by Hettiarachchi and Wigdor [131], Snapchat Lenses[3] and Apple Animojis[4]. This, of course, leads to the restriction that rendered objects must fully cover the real objects by being of similar shape and larger size. Especially when replacing multiple objects in a scene that are close to each other, this enlargement constraint cannot be satisfied without unnatural overlapping effects. In MediaPipe, Lugaresi et al. [212] extract object poses from shoes and chairs and superimpose virtual objects based on the pose information.

Physicality-Aware VR is concerned with enabling a virtual-reality experience that allows roaming the virtual environment while avoiding physical obstacles through redirected walking. Yang et al. [373] present DreamWalker, a system that - given a real-world destination – guides the user through a pre-authored virtual environment while avoiding physical static and moving obstacles. Cheng et al. [56] present VRoamer, a system that procedurally generates VR environments on-the-fly, constrained by the perceived physical obstacles. While the aforementioned work focuses on avoiding the interaction mismatches between the real and virtual environment with respect to walking, the subfield of tangible AR deals with reducing interaction mismatches with respect to touching, e.g., using physical proxy objects [305; 295; 242].

4.2.2 Dimished Reality

Diminished Reality aims at removing objects from a scene [113; 179; 245; 221; 177; 237; 130]. However, none of these systems simultaneously satisfy the three imposed constraints of (1) dimishing with only a real-time stream of monocular RGB information, (2) dimishing moving objects, and (3) dimishing all instances of a certain object class. More importantly however,

[1] ARKit:https://developer.apple.com/documentation/arkit
[2] ARCore:https://developers.google.com/ar/discover
[3] https://www.engadget.com/2020-02-20-snapchat-ground-lenses-floor-is-lava.html
[4] https://www.apple.com/newsroom/2019/12/clips-now-features-memoji-and-animoji-new-stickers-and-more/

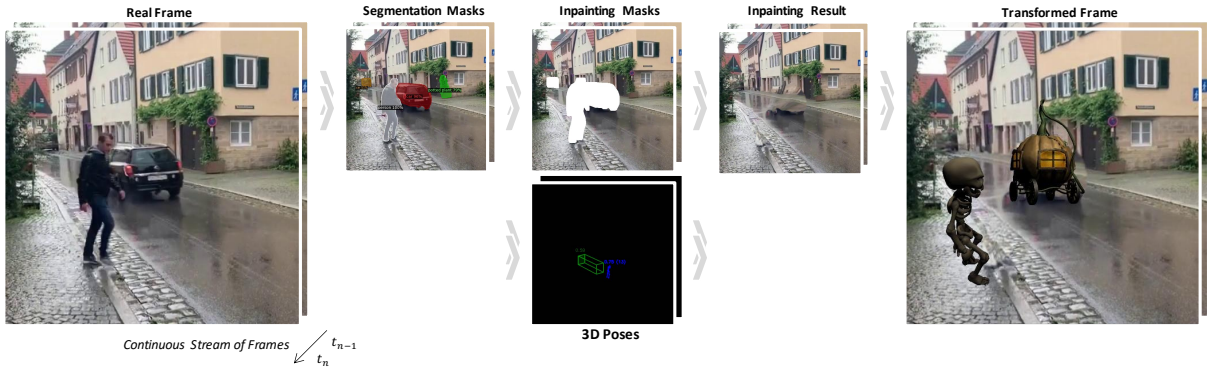


Figure 4.2: An example of continuously creating a Halloween-themed alternate reality through pose-aware object substitution, based on object removal through deep-learning-based instance segmentation and video inpainting, and as well as 3D pose estimations for humans and vehicles. Please also refer to the video figure.

diminished reality is not at all concerned with the simultaneous estimation of 3D poses or rendering replacement objects instead.

Piping Diminished Reality and Augmented Reality While our notion of Pose-Aware Object Substitution requires both diminishing real objects from and placing virtual objects in the scene, a naïve pipeline of separately applying a Diminished Reality system and then applying a geometry-aware AR system is fundamentally insufficient to achieve the envisioned result for TransforMR. Specifically, such a pipeline would not be able to achieve semantical coherence between the virtual objects and the environment. This inability results from the lack of semantic information or 3D pose information in either of the systems. For example, such a DR/AR pipeline could not create scenes in which vehicle-like objects move along real driving lanes, because the concept of a "lane" is not known to the AR system.

4.2.3 3D Scene Reconstruction and Transformation

Litany et al. [205] present an approach for semantics-invariant scene transformation based on point clouds in rooms. Izadinia et al. [147] present a system for transforming a single RGB image of a furnished room into a corresponding composition of CAD models, drawing on a database of such models. Their system is based on multiple applications of convolutional networks for object detection, scene segmentation into "ceiling", "right wall", "middle wall", etc. to derive room geometry, and estimating the objects' feature vectors for similarity measurements against the database. Finally, they apply a render-and-match approach to refine 3D poses. Avetisyan et al. [13] pursue the same objective, however rely on joint layout and object estimation. Shapira and Freedman [299] present Reality Skins, a system for generating virtual environments based on a 3D scan of a room. These setups are incompatible with our goal of allowing untethered open-world on-the-fly applications.

4.2.4 Visual Transformation

Non-photorealistic 2D rendering ranges from traditional convolutions to neural video style transfer [53; 141; 284] to create cartoon, night vision, art, or similar effects. However, by design, these approaches generally modify texture without the possibility to perform transformations such as replacing vehicles with animals.

Video-to-video translation has been used for input-conditioned creation of photorealistic videos. Thies et al. [327] present Face2Face for real-time facial reenactment. Wang et al. [344] present Few-Shot vid2vid for facial or body reenactment or converting semantic maps or human pose models to image sequences.

4.2.5 Summary

While we have identified *technically similar* or *conceptually similar* work in the preceding paragraphs, we argue that all of it is *functionally different* in that it does not aim at providing a real-time, on-the-fly user experience through transforming a real scene into a semantically transformed, yet isomorphic scene, which preserves the correspondence between real and virtual objects. These functional differences technically manifest in fundamentally different system architectures that for example do not comprise components for semantic mapping, object replacement or removal using temporal and spatial information, have less computation needs and do not need investigate offloading to multiple backend GPUs, nor deal with pipeline parallelism.

4.3 TransforMR: Design & Architecture

In this chapter, we propose a novel method for composing meaningful mixed reality scenes by transforming a real-world scene into an alternate reality through *pose-aware object substitution*. Figure 4.2 shows our proposed substitution procedure comprising object detection and pose estimation, object removal, object mapping, and pose-aware object rendering. In the following, we consider our design objectives, describe our system architecture for pose-aware object substitution, and describe its technical implementation.

4.3.1 Design Objectives

TransforMR builds on recent advances in computer vision to realize pose-aware object substitution under a set of design objectives that enable in-the-wild application:

- *Environment Independence.* We want TransforMR to operate on previously unseen scenes without prior preparation of the environment. This means our system cannot rely on on-site-installed camera systems known from room-scale experiences.
- *Handheld Display Rendering.* To allow users to comprehend the correspondences between the virtual and real objects, we display transformed scenes on a handheld display. This assorts well with the objective of environment independence, as handheld displays are, generally speaking, more broadly applicable in public spaces than head-mounted displays.
- *Mobile Device Compatibility.* As we envision broad applicability of TransforMR by enabling users to employ their own mobile device, we impose the constraint of compatibility with common smartphones or devices without the need for additional on-person hardware. As a consequence, perception must rely on a monocular RGB camera only and does not include time-of-flight sensor information.
- *Real-Time Execution Ability.* Being restricted to one monocular RGB camera only entails the requirement of compute-intensive machine vision methods for object and pose detection, and object removal. This conflicts with the limited hardware capacities present in

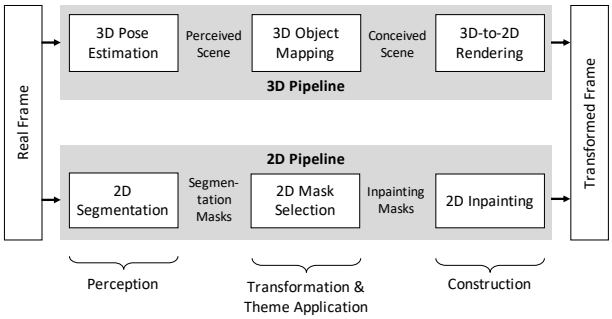


Figure 4.3: Component diagram of the TransforMR system implementation. TransforMR runs 2D and 3D pipelines in parallel with both pipelines performing perception, transformation, and construction steps. The 2D pipeline comprises instance segmentation and non-look-ahead video inpainting in image space. The 3D pipeline estimates object poses in 3D camera space, and renders objects at the same position with the same pose. Mapping is guided by themes through class-specific mapping instructions.

mobile devices. Nonetheless, we subject our system to real-time execution, that is we abstain from a post-capture AR approach and instead aim at processing frames in a real-time fashion.

4.3.2 Pose-Aware Object Substitution Architecture

Figure 4.3 gives an overview of our pose-aware object subsitution system. The overall system input is given by a real-time sequence of monocular RGB frames representing observations from the real environment. The overall system output is a sequence of RGB frames showing the transformed scene. As illustrated in Fig. 4.4, TransforMR performs a series of perception, transformation, and construction operations as described in the following.

Figure 4.4: Examples for our 3D-pose-aware object substitution approach. 3D pose estimation is performed for vehicles (top) and humans (bottom). Vehicles poses are estimated as oriented 3D bounding boxes. Human poses are estimated as 3D joint keypoints.

4.3.2.1 Perception: 3D Pose Estimation

We intend to render virtual objects into the conceived scene with the same pose as the physical object being replaced. Therefore, it is necessary to estimate the 3D poses of those physical objects first. While certain problems such as instance segmentation and 2D bounding box detection can be solved with general-purpose models trained on high-diversity datasets, 3D pose estimation algorithms are predominantly designed for specific purposes.

For detecting 3D poses – more specifically, 6 degrees-of-freedom pose – of vehicles in traffic scenes [240; 138; 189; 207], we employ SMOKE (Single-Stage Monocular 3D Object Detection via Keypoint Estimation) by Liu et al. [207]. Relying on a CenterNet-like network architecture with deep layer aggregation and deformable convolutions [377; 71; 389] for feature extraction, SMOKE directly regresses location and orientation parameters from a single monocular RGB frame without an intermediary step of inferring 2D object proposals first. Since SMOKE operates

on single frames and ignores the temporal dimension, we employ a distance-based tracking filter on the estimated centroid in 3D space to infer cross-temporal object identity.

For detecting 3D poses – more specifically, 18 different keypoints in 3D space – of humans [370; 277; 222; 261; 223], we employ *Lightweight Human Pose Estimation* by [261] which is based on a previously presented architecture [223], however, modified in order to decrease inference duration by using a MobileNet-like [137] feature extractor. It is noted that Android's ARCore doesn't support human pose estimation and Apple's ARKit supports 3D human pose estimation just for a single person and on recent chipsets [330] only. Lightweight Human Pose Estimation can estimate the poses of multiple persons simultaneously. As with the 3D vehicle pose estimation, a distance-based tracking is applied to infer object identity across a sequence of frames. Attached to the feature extractor are 2D and 3D keypoint detection stages.

Both pose estimation models are encapsulated as independent modules with the same abstract interface of consuming a single frame and thereupon returning a list of pose detections. State from processing previous frames is managed internally by each module. While models should predict pose information accurately relative to the camera, we adjust for different relative scales across the models by maintaining a model-specific scaling factor. Figure 4.4 visualizes the 3D awareness in image space, achieved by the pose estimation procedures described.

4.3.2.2 Perception: 2D Segmentation

For the object removal procedure, we rely on hole inpainting. We determine the holes to be inpainted through instance segmentation the Mask R-CNN algorithm [127]. Each instance segmentation mask corresponds to a detected instance and represents a bitmap the size of the input image which indicates the presence or absence of a pixel belonging to the respective instance. We use the *detectron2* implementation with a ResNet-50/FPN feature extractor.

4.3.2.3 Transformation: Theme-Guided Semantic Mapping

Transformed scenes are a function of the object-reduced input frame, the current object *detections* estimated by the system, and the *theme* selected by the user. A detection comprises estimations of the class, 3D pose information, and possibly additional information of a real-world object. The theme comprises *class-specific mapping instructions*. Each mapping instruction is scoped to a detection class supported by the perception module. It indicates which virtual object models can be rendered in lieu of the real object. A single class, e.g., *car*, can be mapped to different virtual object models, thus producing diverse transformations. In order to retain the mapping between a physical object in a frame and its previous mapping in preceding frames, a substitution state storing tracking IDs of detected objects and the corresponding virtual instances is managed. Figure 4.5 shows the themes we have prepared for use in TransforMR.

With these themes, users can employ TransforMR to either create their own narratives or to interactively consume provided narratives, e.g., created by a narrative provider in a certain context such as a museum or zoo. We discuss these narratives in Section 4.5.

4.3.2.4 Construction: Video-Inpainting-based Object Removal

We accomplish the goal of object removal through real-time video inpainting where the inpainting mask in each frame is filled by estimating the globally and locally most plausible pixel values.

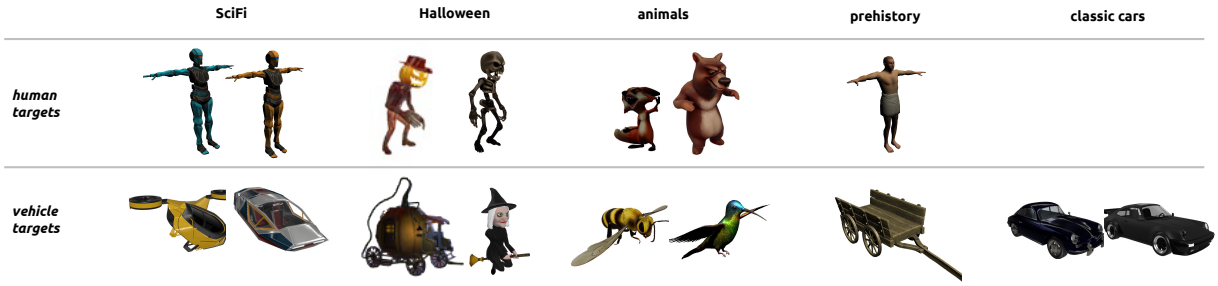


Figure 4.5: Overview of the specified SciFi, Halloween, Animals, Prehistory, and Classic Cars themes. The different themes feature 3D vehicular and humanoid object models.

Figure 4.6: Exemplaric comparison of the two alternative real-time video inpainting methods, integrated in TransforMR. (a) Based on the monocular RGB frame, (b) TransforMR runs 2D instance segmentation to produce the inpainting bitmap. (c) With our adaptions in VINet, inpainting can operate at approximately 4 FPS without lag at a single VINet model instance and approximately 7 FPS with two load-balanced VINet model instances. VINet yields visually coherent inpainting for large masks. (d) With our adaptions in LGTSM, a chunk size of 4 frames, and downsampling to a width and height of 200 px, can operate at approximately 22 FPS, however at the cost of visual coherence for larger holes.

Inpainting masks are derived from the instance segmentation bitmaps estimated as described above.

Since classical methods of image inpainting generally yield implausible results for larger holes or lead to inconsistent or flickering inpainting across consecutive video frames, we turn to learning-based video-inpainting methods [369; 182; 257; 194; 243; 384; 181; 383; 51; 322; 101]. Generally relying on optical-flow estimation to reconstruct the path of a pixel value through the temporal dimension, information is propagated from previous or future frames into the region to be filled. Filtering out methods which, by design, expect knowledge of all frames in advance, or methods with uncompetitive frame rates that are therefore unfit for our real-time objective, we integrated VINet [182; 181] and inpainting based on Learnable Gated Temporal Shift Modules (LGTSM) [51] as alternative methods into our system architecture.

While VINet is originally designed to peek five frames into the future, we adapted the inference logic so that no look-ahead is performed, thus reducing system lag. Further, while LGTSM-based inpainting originally chunks up a video into smaller batches, and operates on these smaller batches, we adapted the inference logic to feed-back inpainted frames into the next input chunk with an all-intact inpainting mask, thus yielding a frame-by-frame inpainting method suitable for real-time application.

VINet has an inference latency of approximately 250 ms, however produces visually coherent results for larger holes. Using a load-balancing approach and distributing frames across two VINet model instance allows to operate inpainting at 7 FPS. With our adaptions for frame-by-frame real-time inference, LGTSM-based inpainting has an inference latency of only approximately 45 ms. However, while VINet can propagate information of long gone frames to the current frame inpainting through state in the recurrent LSTM units, LGTSM-based inpainting only convolutes on the last three frames for inpainting and fills the remaining information generatively. This results in less coherent results for larger holes. Both frame rates refer to exclusive usage of an NVIDIA Tesla V100 GPU.

4.3.2.5 Construction: 3D Rendering

With the inpainted frame, the detections including the 3D pose information, and the selected theme as an input, we run the 3D scene rendering in the Unity graphics engine to obtain the transformed frame.

4.3.3 Technical Implementation

As all computer-vision models in TransforMR employ convolutional neural networks are therefore computationally intensive, we run them on four NVIDIA Tesla V100 GPUs using CUDA. Each model runs in a separate Docker container on the cloud server.

We implement a central access point to the TransforMR backend, also running in a container that distributes a single request across the different models and integrates their responses in the perception result that is sent back to the client. Client-server and inter-container communication is implemented using gRPC[5]. While the perception is offloaded to the cloud, the transformation and the construction module are running on the terminal client in Unity[6]. This setup allows

[5] gRPC: `https://github.com/grpc/grpc`
[6] Unity: `https://unity.com/`

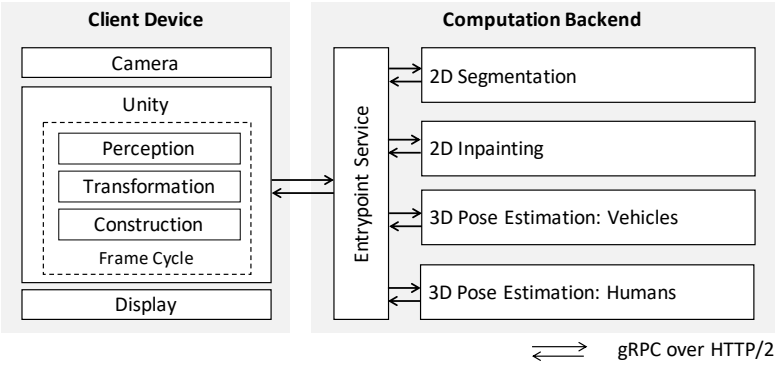


Figure 4.7: Deployment diagram of TransforMR. RGB frames are captured on the client-device camera, shipped over the network to a GPU-accelerated host that performs computation-intensive operations and returns pose estimations for all relevant objects as well as the inpainted frames. Rendering of the virtual 3D objects takes place on the client device.

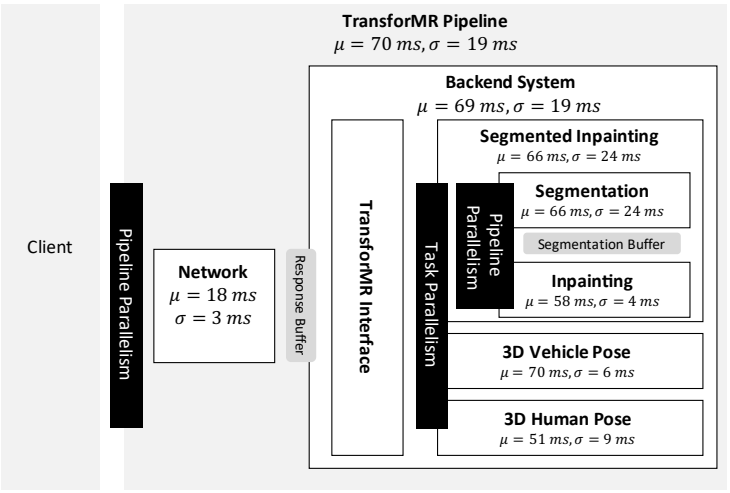


Figure 4.8: Breakdown of the network and inference latencies in TransforMR. Note, that latencies of components in a parallelized pipeline do not sum up, but are given by the maximum of the individual pipeline component latencies. Network RTT was measured from our institutional local area network in Zürich to an AWS EC2 VM in Frankfurt. It was determined in a network test setup by immediately returning the received frame. Benchmarking was performed on the scene that is shown in Figure 4.6.

shifting computationally demanding load from the device into the cloud, thus relaxing the hardware and software requirements on individual users' local devices. Also, the architecture enabled us to quickly test different state-of-the-art algorithms which were not optimized for mobile devices. The downside of this approach is that going over the network adds a lag, depending on the network conditions. Figure 4.7 exhibits a deployment diagram.

Using LGTSM-based inpainting instead of VINet, our computation backend achieves a processing frame rate of approximately 15 FPS and a system lag of 3 frames. We achieve this by a multi-faceted parallelization architecture. *First*, we employ pipeline parallelism between the network and the computation backend, so that we send the next RGB frame while the previous frame is still being processed by the backend. *Second*, we employ threefold task parallelism for a) 3D pose estimation of vehicles, b) 3D pose estimation of humans, and c) the segmentation and inpainting pipeline. As inpainting depends on segmentation, we cannot run segmentation and inpainting in parallel for a single given frame, but we can run instance segmentation, while the previous frame is still being inpainted. That is, *third*, we employ pipeline parallelism between the instance segmentation and the inpainting. In summary, at each frame cycle, the backend runs inference through four neural networks in parallel. We restrict frame buffer size at the central entrypoint service to size 1, so that the backend throttles the client automatically if frames are served faster than they are processed.

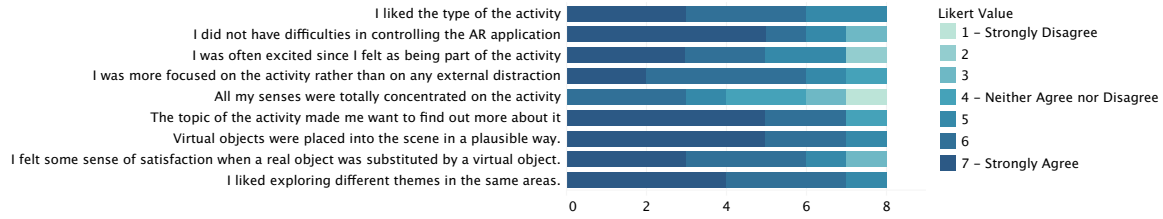


Figure 4.9: Likert ratings of the participants on statements from the ARI questionnaire (upper 6 statements) by Georgiou and Kyza [102] and on TransforMR-specific statements (lower 3 statements)

4.4 Preliminary Evaluation

4.4.1 Participants

We conducted a preliminary user evaluation with 8 participants ($n = 8$, $n_{female} = 2$, ages 21 to 55, $\mu_{age} = 28.8$, $\sigma_{age} = 10.9$), 5 of which from our institution. No one of the participants was involved in the project. Two participants had limited experience with AR applications, another two had extensive experience. The participants received a small gratuity after the evaluation sessions.

4.4.2 Procedure

Participants were first shown a video of a scene that was transformed using TransforMR with the SciFi theme. They were informed that the system is capable of transforming humans and vehicles. We employed quality-optimized inpainting. 5 out of the 8 evaluation sessions were conducted at a busy street near our institution, the other three were conducted at a traffic-calmed street with a bordering park farther away. On-site, participants were instructed to employ an Apple iPad Pro, 12.9 inches, to explore the surroundings at their discretion, switching through all 5 provided themes.

Overall, each evaluation session took 14 to 25 minutes. Afterward, participants were to fill out the Augmented Reality Immersion questionnaire by Georgiou and Kyza [102], extended by TransforMR-specific questions, by stating agreement on a 7-point Likert scale. In addition, we conducted a semi-structured interview to gain insight into their subjective impressions.

4.4.3 Results and Discussion

Figure 4.9 exhibits a subset of the evaluation items. All participants found the experience enjoyable or very enjoyable.

Transformation Themes. Three participants (P1, P6, P7) liked the animal theme most with all of them stating they liked the wing-flapping animation of the bee and the hummingbird. Two participants (P3, P4) expressed that they liked the Halloween theme best, one of them (P3) stating that it suited the atmosphere of the tree-lined road very well. From this, we derive the idea of *context-specific themes* as part of a *consumptive interaction pattern*. E. g., visitors of an amusement park might want to alter their environment by substituting other visitors with suitable cartoon characters. We elaborate on this in the applications subsec. 4.5.2.

Interaction. Most interestingly, three participants using the system asked if were okay for

other participants to join in on the activity as "actors" in order to explore how humans were transformed. From this, we derive the notion of the *director-actor interaction pattern* for *creating narratives*, described in subsec. 4.5.1. Furthermore, P5 described it as "awkward to direct the tablet on random people unknown to him". He noted that this "awkwardness" was reduced when people would approach him or when multiple people were in the scene. Extending the perception range to other objects, e.g., animals, could allow users to focus on other alterations beyond humans. P6 noted that he would have liked to see objects from very up close, but getting too close would make them vanish. Whether the 3D object pose can be inferred from an object depicted in macro-perspective depends on the model. For example, the 3D vehicle pose estimation model we employ requires the vehicles 3D center to be in the frame. On the other hand, the 3D human pose estimation model can also infer the pose, for example, from the head of a human only.

Alteration Experience. Three (P2, P4, P6) out of the 8 participants stated that they would have liked to see augmentation effects known from classic AR apps too, while the other 5 participants (P1, P3, P5, P7, P8) stated they prefer the app to only show virtual objects that have a real correspondence. One participant (P8) who indicated the preference of no augmentations said, that it felt "more real to known that what [she] saw was really there, in a way, and not just imagined". One participant (P2) liked classic augmentations in addition to the real-world object alterations stated that this would be "interesting for situations when there are no cars or humans around". However, with 5 users abstaining from the wish to add Augmented Reality objects to the scene, we conclude that upholding the correspondence awareness between virtual and real objects exhibits a particularly exciting alteration experience. P7 reported in a highly positive mood that he felt as if he was "in an apocalypse movie with mutated bees". However, P6 expressed disconcertment on the fact that the prehistoric theme transforms not only parked cars, but also even moving vehicles to wooden carriages without a draft animal. From this, we see that one has to distinguish between *spatial plausibility* and *semantic plausibility*. The above-mentioned statement on the bee shows that semantic implausibility can be a source of additional excitement or disconcertment. Since users can take agency by selecting certain themes and decide on the camera direction, they can influence the plausibility level achieved.

Alteration Consistency. While all participants considered the alteration consistency positive overall, two participants (P3, P1) noted that sometimes the human objects "were off and jumping around" resp. "switching identities". Incremental improvement of the prediction models could help stabilize prediction. Two participants (P2, P4) said that they noticed significant differences in the object removal quality. P2 recounted that multiple times he "didn't even notice that there was actually a person crossing right in front of them" when at other times it seemed as if "it just blurred the object". While inpainting still remains a difficult problem in computer-vision research, we believe it could be interesting to also add a post-capture AR experience, as known from recent releases in Google ARCore[7]. Here, instead of harder real-time inpainting, this would allow using slower, but better inpainting operating *on all frames* including future ones, thus enriching the optical flow information.

Focus. While all participants stated that they primarily focused on the display, 4 out 8 participants (P1, P5, P6, P8) recalled they would regularly check the real scene and compare it with the transformed scene and the other 4 participants (P2, P3, P4, P7) would compare it to the real scene sometimes. P1 and P2 recounted that they would look up especially to search for new situations. These findings cement our conclusion that users see Correspondence Awareness as

[7] `https://developers.google.com/ar/develop/java/recording-and-playback/introduction`

Figure 4.10: Different scenes transformed towards different themes. TransforMR enables users to roam previously unseen, unbounded, unprepared, and changing environments featuring multiple humans and vehicles, all at the same time, transformed by a user-chosen theme.

an important part of the experience.

4.5 Applications

In the above evaluation, we ascertained that users like to employ TransforMR for real-time in-context character animation with multiple users as well as for the exploration of an alternate mixed reality around them. From this, we derive the usage patterns of *consuming narratives* and *creating narratives*.

4.5.1 Creating Narratives

As seen in Figure 4.11, users can follow a *director-actor interaction pattern*, where one user takes the role of the director and one or more other users take the role of actors. The directing user will then employ their smartphone to capture the acting users in context, thus creating a transformed scene. In so doing, users can collaboratively tell stories about virtual characters, offered in the theme, who walk through parks or school buildings, ride the bus, or do grocery shopping. This kind of role play is only achievable through the concept of pose-equivalent character substitution, allowing interesting new scenes composed of interactions between characters such as anthropomorphic animals, robots, celebrities, or avatars, proxied by real humans.

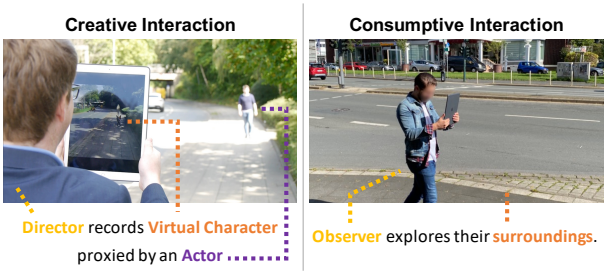


Figure 4.11: **Left**: Following a director-actor interaction pattern, users can perform real-time character animation, reminiscent of motion capturing in professional film making studios, however in real-world context instead, in order to create their own narratives. **Right**: Users can also explore and transform their surroundings single-handedly. In this pattern, interaction is given through theme selection, environment navigation, and camera direction.

4.5.2 Consuming Narratives

Consuming predefined narratives with *context-specific themes*, in particular location as an important context property, could, for example, enable users to experience time-travel through the history of traffic in a certain region e.g., by visualizing evolving eras of urban traffic through-out history. Similarly, visitors at a historic site could transform their environment towards a theme replacing other visitors with models of humans that are in line with the historic culture in question. Or possibly, visitors of an art museum might want to experience the museum alongside the original artists instead of their real fellow visitors. By previously authoring such themes with the corresponding 3D object models, maintainers of a facility could enable visitors to immerse deeper into their stay.

4.6 Limitations

Increasing Prediction Accuracy. Object and pose prediction models as well as inpainting are at the core of the TransforMR system. The plausibility of the transformed scenes is therefore inherently limited by the models' prediction performances. The model for human pose estimation is limited to humans no farther away than approximately 15 meters and unreliable when it comes to correctly detecting keypoints of cyclists. Tracking of objects is also challenging, in particular in if they are located very near to each other. Figure 4.12a) and b) show examples. Improvements could be achieved, e.g., by adding temporal recurrency to the detection models [202], using a learning-based object tracker [138], employing detection models with physical constraints [277], or building on advances in correspondence estimation [324].

Improving Visual Coherence. Modern augmented reality frameworks feature techniques that can improve the visual coherence of the generated imagery, in particular to create realistic illuminations and occlusions [215]. Since shadows are purposefully not removed in the presented pipeline, virtual objects "reuse" the shadow cast by the original objects, thus alleviating part of the illumination problem [46]. However, as seen in Figure 4.12b), occlusions are not detected in TransforMR. Instead, heavily occluded objects are likely to not be detected by the 3D detection models, thus leaving the object untransformed. Adding light source estimation for more realistic illumination and in particular adding a depth map estimation for occlusion shading [331] could improve the geometric plausibility of transformed scenes.

Targeting Head-Mounted Displays. We have designed TransforMR as a mobile AR system so to ensure that users can always keep track of correspondences between virtual and real objects. However, future work might explore the opportunities to employ the pipeline for use in head-mounted displays (HMD). While we hypothesize that such an approach might be beneficial in terms of full immersiveness, it remains an open research question how to minimize latency of such a system generally, and inference latency of all the deep-learning models specifically, so

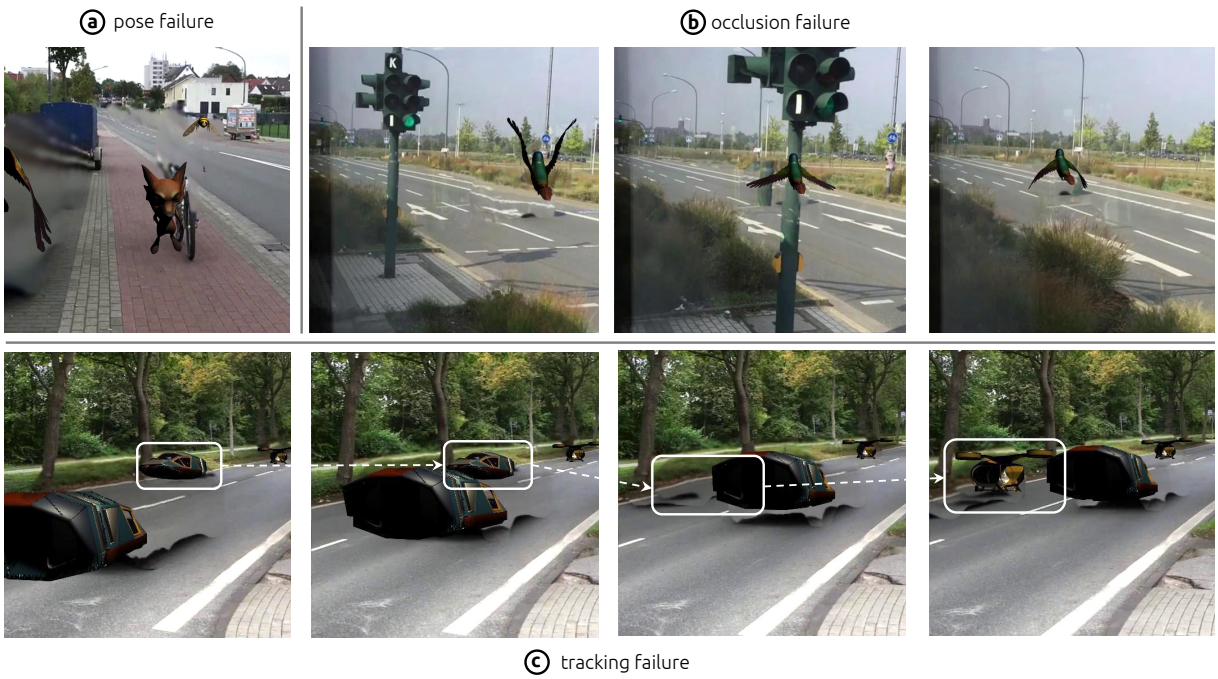


Figure 4.12: Perceptually challenging situations can cause visually incoherent transformed scenes. (a) Intricate poses in humans, e.g., of cyclists, can cause incoherent renderings of the virtual avatars. (b) Heavily occluded objects are likely to not be detected, thus yielding scenes that are mostly coherent with the depth in the scene. However, slightly occluded real objects might be detected anyways, leading to renderings of virtual objects that are negligent of occlusions. (c) Objects that move considerably across a couple of frames, possibly even hidden behind other objects, can cause a discontinuation in the instance tracking chain.

that the higher frame rate demand required for safe and smooth HMD experiences can be met.

Estimating World Coordinates. TransforMR's perception pipeline only estimates the location and pose of objects with respect to the camera, but does not track their location in world coordinates. Therefore, the system does not differentiate between ego motion and object motion. Using SLAM approaches [245], either visually or by means of the devices inertial sensors, could help the system to estimate world-relative object movement. This would allow movement-aware substitutions, e. g. with virtual horses that either stand, walk, trot, or career, depending on the velocity of the real object.

Enriching Themes and Narratives. TransforMR has been implemented with support for detecting and estimating the 3D pose of vehicles and humans. Adding pose estimation for other indoor or outdoor object classes such as animals, trees, chairs, etc. will enable richer experiences. Furthermore, we see a potential for complementing pose-aware object substitution with pose-aware plane substitutions, e.g., to transform streets into water, lava, or lunar soil. Furthermore, we see a high potential for increasing immersiveness through Audio AR features that emphasize narratives for the consumptive usage of TransforMR.

4.7 Conclusion

In this chapter, we have presented TransforMR, a system that performs pose-aware object substitutions to create meaningful alternate mixed reality scenes. Designed under the objectives of environment independence, correspondence traceability, mobile device compatibility, and real-time execution, we proposed a cloud-assisted architecture comprising computer-vision models for 2D instance segmentation, 3D pose estimation for vehicles, 3D pose estimation for humans, speed or quality-optimized alternatives for object removal through inpainting, as well as comprising a semantic mapping procedure and the 3D rendering.

In a preliminary user evaluation, we found that users particularly like to employ TransforMR for real-time character animation, reminiscent of motion capturing in professional filmmaking, however operating without preparation of either the scene or the actor.

While gaming-oriented Mixed Reality applications of today mainly borrow *3D geometry* from the real world and register objects based on planes and visual features, TransforMR heads towards mixed reality experiences that do not only incorporate geometrical but also semantical information from the real-scene context into the composition of the mixed-in virtual scene. In the future, we expect to see more research that explores concepts to realize semantics-driven mixed reality, complementary to or derived from this work on reusing semantical object embeddings.

One major hurdle in realizing semantics-driven mixed reality scenes lies in the computational demand. More specifically, to extract possibly multiple layers of semantics from camera input using demanding neural networks, a single, mobile-device GPU can present an insurmountable bottleneck. In TransforMR, we showcased the integration of four neural networks without being subject to mobile resource limitations. We believe that future research can draw on the feasibility of such a cloud mixed reality approach. To enable the community to reproduce the system and to build on our work, we make all of our source code and assets available on GitHub.

5

Scene Responsiveness for Visuotactile Illusions in Mixed Reality

Manipulating their environment is one of the fundamental actions that humans, and actors more generally, perform. Yet, today's mixed reality systems enable us to situate virtual content in the physical scene but fall short of expanding the visual illusion to believable environment manipulations. In this chapter, we present the concept and system of Scene Responsiveness, the visual illusion that virtual actions affect the physical scene. Using co-aligned digital twins for coherence-preserving just-in-time virtualization of physical objects in the environment, Scene Responsiveness allows actors to seemingly manipulate physical objects as if they were virtual. Based on Scene Responsiveness, we propose two general types of end to-end illusionary experiences that ensure visuotactile consistency through the presented techniques of object elusiveness and object rephysicalization. We demonstrate how our Daydreaming illusion enables virtual characters to enter the scene through a physically closed door and vandalize the physical scene, or users to enchant and summon far-away physical objects. In a user evaluation of our Copperfield illusion, we found that Scene Responsiveness can be rendered so convincingly that it lends itself to magic tricks. We present our system architecture and conclude by discussing the implications of scene-responsive mixed reality for gaming and telepresence.

5.1 Introduction

In most recent research on adaptive Mixed Reality (MR) interfaces, the physical scene dynamically influences the placement of virtual content [59; 271; 213; 303]. In this chapter, we aim to invert the direction of influence between virtuality and physicality, asking "How can virtuality affect the physical world?"

As an answer, we propose *Scene Responsiveness*, the illusion that virtual actions affect the physical world. By altering the visual signal in a video-passthrough MR headset, Scene Responsiveness allows actors to seemingly manipulate their environment, such as a bi-ped character to open the next door for entering or exiting the scene, or a black hole to absorb physical objects

Figure 5.1: We present Scene Responsiveness, the visual illusion that virtual actions affect the physical scene. ⓐ Wearing a video-passthrough Mixed Reality (MR) headset, a user sees an authentic video view of their physical environment, captured with the externally-facing headset cameras. A virtual monkey character is situated and coherently occluded in the otherwise unmodified physical space, grabbing the PHYSICAL cart. ⓑ As the monkey starts dragging the cart, the object toggles its reality state to VIRTUALIZED just-in-time. Just-in-time virtualization is *coherence-preserving*, so the cart still throws its shadow onto the physical scene and the now seemingly empty surface. ⓒ To the user, it appears as if the monkey drags the physical cart and throws it down the staircase. Virtual shadows and collisions render coherently. ⓓ It appears as if the physical cart is gone. ⓔ Of course, what the user sees is just an illusion. The physical object did not move but was just masked from the user's view, blending seamlessly with the video passthrough. Everything seen in the debug capture, incl. the red object guardian outline and the revelation lens in the top-right corner is part of the system debug view. *The figure gives an authentic impression of the visual fidelity in the headset. However, please watch the accompanying video for a full impression of the visual experience.*

from the scene. Also, the user themselves can cast magic enchantments to summon far-away objects. The visual illusion of state responses, shape responses, and pose responses starts by *toggling the reality state* of the manipulated object from PHYSICAL to VIRTUALIZED *just-in-time* as the user or virtual actor begins their manipulation. The response is spatially contained at the involved object, seamlessly blending with the video-passthrough anywhere else and thus maintaining the sensation of being in physical space.

However, the illusionary experience enabled by Scene Responsiveness is not limited to the visually seamless response of physical objects to user input or situated character animations. Instead, we develop and propose *Daydreaming* and *Copperfield illusionary episodes* as self-contained interaction-centered experiences. Daydreaming episodes start with a virtually triggered scene response, and then employ *object elusiveness* and *object rephysicalization* to prevent tactile disillusion: Influenced by the user's behavior, virtualized objects avoid collision with the user by diegetically eluding from the user's body or rephysicalizing in diegetic ways at the pose of their physical counterpart. Copperfield episodes, named after the eponymous vanishing-and-reappearance magician, expand on this by purposefully toggling reality states to make physical objects appear exactly in the pose where a virtual interaction seemingly transported them, thus deepening the illusionary experience of virtual control over physical space.

In addition to the conceptual contribution of Scene Responsiveness, we provide a system *prototype implementation* that draws all required spatial, visual, and semantic information from a co-aligned and unified digital representation of the space and its objects. First, our prototype implementation enables to render virtual content geometrically and physically integrated into the scene respectful of occlusions, collisions, and shadows, across all virtual, physical, and blended areas of the space. Second, the system with its Spielberg component, named after the director and pioneer of character animation, situates virtual characters and their animations as well as gestural user input semantically integrated into space, guided by character and user affordances. Third, apart from the former two system integration efforts, we contribute an integrated spatial computing and shading architecture for *coherence-preserving object virtualization* in world space on video-passthrough MR devices, not yet achievable with the image-space techniques of 2D inpainting in conventional Diminished Reality [237]. Through spatial arrangement of 3D

occluders, and subsequent dynamic assignment of camera layers, manipulation of the rendering order, and purposeful writes to and reads from the depth buffer, we visually remove physical objects by masking them with the environment background, yet render all other virtual content coherently into the scene.

We conducted a user evaluation in two different spaces with 20 participants, 10 participants per space, evaluating the rendering coherence as well as the illusion fidelity in a Copperfield episode. We found out that 18 out of 20 participants were surprised they were able to take a seat on VIRTUALIZED chair that has been dropped by a virtual character in seemingly empty space.

Considering the upcoming releases of video-passthrough MR headsets [225; 7] and the expected industry focus on such devices for the years to follow, we argue that Scene Responsiveness can create captivating MR experiences not only for situated gaming but as a general means for situated MR, such as scene-responsive telepresence or virtual assistants.

Contributions

In summary, we contribute

- the *novel concept of Scene Responsiveness* for high-fidelity illusions in video-passthrough Mixed Reality,
- a *design space of end-to-end illusionary experiences* that prevent disillusion through object elusiveness and rephysicalization to maintain visuotactile consistency, along with the two design samples of *Daydreaming and Copperfield-type epsisodes*,
- the algorithmic *spatial computing and shading architecture RealityToggle* for *coherence-preserving just-in-time virtualization* in 3D world space on video-passthrough Mixed Reality devices,
- an integrated *system architecture and implementation*, comprising our *TwinBuilder Unity plug-in* to obtain, process, and annotate digital representations, as well as our *Spielberg component* for situated character-centric stories and animations,
- a *user evaluation* with 20 participants of the visuotactile Copperfield illusion,
- the applications of *scene-responsive gaming*, *scene-responsive telepresence*, and *scene-responsive television*.

5.2 Background and Related Work

5.2.1 Scene Coherence in Mixed Reality

Under the term of *scene coherence*, foundational work in AR and MR has been dedicated to the graphical techniques needed to render geometrically and physically coherent occlusions [17; 37; 154], collisions, shadows, and reflectance [292; 78; 187; 184], as well as to the question of how to reconstruct the necessary depth [332; 342; 281; 313; 186; 247] and lighting [351; 196; 215; 116] information from the physical scene. Once virtual content can be coherently integrated into the scene, the question arises as to where to render it. FLARE [100] automatically layouts AR content on vertical and horizontal planes, guided by developer-defined rules. SnapToReality [254] automatically aligns virtual content, guided by planes and edges detected in the physical

geometry. Lages et al. [190] investigate how to adapt UI elements in AR to vertical planes. DepthLab [78] not only proposes different depth representations for scene coherence but also considers geometrically coherent path planning for virtual characters.

To ensure illusion fidelity, we integrate occlusion, collision, and shadow coherence through digital space twins, created with our TwinBuilder Unity plug-in. However, our work most importantly differs from previous research in that it must also maintain *manipulation coherence* after the scene has been visually manipulated. To this end, we present a spatial computing and shading architecture with dynamic camera layer assignment, render ordering, and depth buffer access which preserves scene coherence after visually removing physical and inserting virtual objects (Figure 5.2).

5.2.2 Scene Situatedness in Mixed Reality

Scene Responsiveness not only demands scene coherence but also situatedness [353; 38; 359; 92; 292; 115], i.e., a semantic relationship between virtual and physical content.

Digital counterparts co-aligned with physicality through markers or other means have been proposed early on to situate virtual content. A variety of visualization paradigms [325; 208; 153; 154; 391; 208; 234] for situated annotations [145], ghost views [155], virtual paths [214], or magic lenses [211; 210; 21; 22; 28; 338] have since emerged for use in navigation, commerce [144], maintainenance, education [114; 134], tutorial instructions [233; 62], etc. One closely related use of co-aligned representations is found in situated gaming [214; 280]. RoomAlive by Jones et al. [152] allows to create a spatial scan, author experiences therein, and then deploy this experience to a room-scale projection mapping system. We borrow the idea of digital representations that guide content placement, however, use this representation not only for situation, but also for *virtualization* of objects just-in-time by visually removing their physical counterpart first and then inserting a digital object in their place to make them interactable for virtuality.

Adaptive MR in contrast does not rely on a space-specific model but asserts the claim to generalize to different spaces. SemanticAdapt [59] and the approach by Luo et al. [213] investigate adaptive situation of 2D content. Previous work also considers the adaptive situation of virtual 3D actors specifically. Retargetable AR [323] situates virtual characters in a physical scene for situated storytelling. Li et al. [197] build on this concept of situated storytelling, but introduce user interactivity to influence story playback. Liang et al. [199] consider dynamically situating and controlling a virtual pet in MR. SpaceTime [150], Kim et al. [185] and Grønbæk et al. [108] situate avatars based on remote user activities in local space. Schmidt et al. [293] consider physical object manipulation by virtual agents through actuation. Story CreatAR [306], ScalAR [271], and Ng et al. [248] provide frameworks for authoring situated AR experiences. ARAnimator [375] situates animation sequences in space. Shin et al. [303; 302] study the effect of game adaption to different spaces. Scene Responsiveness also situates characters in space guided by objects in the physical scene. However, Scene Responsiveness differs from the above approaches in inverting the order of influence: Not only does the physical scene influence the virtual scene, but also virtual actors and actions can influence the presented physical scene through manipulation of the visual signal in a video-passthrough headset (Figure 5.2).

Blending between physical scene and virtual scenes has been explored by Blending Spaces [67] and RealityCheck [124]. However, physical and virtual scenes are assumed to be structurally and functionally different. In contrast, Scene Responsiveness blends seamlessly between structurally and functionally identical scenes for object virtualization.

Figure 5.2: Levels of scene integration in MR. The concept of Scene Responsiveness aims to add an exciting new level of scene integration beyond coherence and situatedness to MR.

Situated VR co-aligns the virtual scene with the physical scene, either through alignment before the experience as in Reality Skins [299], Oasis [310], Substitutional Reality [305], Tailored Reality [77], or Scenograph [217], through procedural generation as in DreamWalker [373] or VRoamer [58], or through assisted annotation as in ARchitect [201]. While Scene Responsiveness shares the idea of paired physical-virtual object counterparts, it does not pursue showing a functionally different space in VR. Instead, Scene Responsiveness aims to replace a physically captured object with its virtually rendered counterpart as part of the object virtualization step for subsequent interaction. Technically fundamentally different, we target video-passthrough MR, where the user's own hands and body remain normally visible, and they can have face-to-face conversations with other humans in their space, even after manipulating the scene.

5.2.3 Scene Editing in Mixed Reality

Diminished Reality (DR) [239; 236; 235; 238; 60; 113; 183] offers to remove areas from the shown frame. Technically, Scene Responsiveness differs from conventional DR in that it operates through masking in 3D world space, rather than operating in 2D image space, to maintain depth-related scene coherence. Pragmatically, camera frames are not accessible on consumer-grade passthrough headsets to ensure privacy. Conceptually, Scene Responsiveness differs in that it virtualizes objects, i.e., inserts an object replica rather than just removing it. TransforMR [163] is also concerned with inserting virtual counterparts after removing physical objects to produce semantically coherent scenes. However, virtual objects directly follow the pose and articulation of physical objects, prohibitive of independent control over virtual objects in world space. Overlay-based AR such as Annexing Reality [131] also disallows object displacement.

SceneCtrl [379] offers to select, move, delete, and copy objects in an MR scene. It performs editing operations in image space and renders the results in the HoloLens optical see-through display. Remixed Reality [203] shows a rerendered voxel-based representation of the user's environment in VR, captured through Kinect cameras. However, Scene Responsiveness differs from Remixed Reality and SceneCtrl in three fundamental ways.

First, Scene Responsiveness aims for the *imperceptibilty of manipulation* in our Copperfield illusion. Our evaluation indicates that we accomplish this with our spatial computing and shading pipeline for seamlessly blending between co-aligned physical and virtual spaces while maintaining full scene coherence concerning occlusions, collisions, and shadows. In addition, the ability of illusion-quality scene manipulation asks for additional interaction concepts to maintain the illusion. Thus, we contribute the concepts of object elusiveness and rephysicalization as part

of our different illusionary episodes, thereby ensuring visuotactile consistency. These considerations were out of the question in previous research, given the absence of illusion-quality visual manipulation.

Second, Scene Responsiveness aims for *semantic situation and manipulation* through afforded interaction, semantically integrated with the scene and meaningfully related to specific objects. Thus, Scene Responsiveness enables situated character-environment interactions, such as opening an elevator by a button press, or dragging a heavy cart differently than carrying a lighter chair, rather than providing universal yet generic geometric operations on images or voxel grids. A semantic rather than geometric consideration then also allows for operations that target the scene graph such as decomposing an object group to individually apply physics as seen with the coke cans, etc., in Figure 5.1.

Third, Scene Responsiveness aims for a different *interaction paradigm*. Rather than focusing on user input for manipulation only, our concepts of receptive and responsive affordances in space enable story and telepresence modes.

5.2.4 Illusions in Mixed Reality

Actuation to provide haptics in VR has been used for tactile illusions [1; 320]. The dominance of vision over proprioception has also been taken advantage of for perceptual manipulation in VR [250; 328], in particular for haptic retargeting [15; 386]. Scene Responsiveness targets MR rather than VR.

5.3 Scene Responsiveness for Visuotactile Illusions

In the following, we first define the terms central to Scene Responsiveness, and then show how scene-responsive illusions can be maintained and completed through end-to-end *illusionary episodes*.

5.3.1 Scene Responsiveness in Mixed Reality

In situated MR, the physical scene affords the meaningful *placement* of virtual elements. We introduce *Scene Responsiveness* as the illusion that the physical scene *responds* to virtual actions.

Physical and Virtual Actors

Actions are performed by actors. We differentiate between *the user* as a physical actor in local space, *other physical actors* in local space or remote spaces, and *virtual actors* in local space. *Virtual actors* encompass *agents* such as non-player characters (NPCs) in gaming that pursue their own goals, *assistants* that follow the user's goals or instructions, and *avatars* that mimic the behavior of a physical remote actor. In the following, we use the terms "virtual actor" and "character" synonymously.

Physical and Virtual Actions

We further distinguish between *virtual actions* and *physical actions*. The actions of virtual actors are always virtual. In contrast, actions of the user can be either virtual or physical. When the user takes a seat on a physical chair, their action is fully contained in physical space and thus is physical. However, when the user lights a virtual fire that seemingly burns physical objects, swings a virtual lightsaber to seemingly cut a physical object in half, or uses a physical hand gesture for telekinesis to seemingly summon a remote physical object, we classify these actions as *virtual* because the action's effect is meant to be contained in virtuality and does not affect the physical world *in esse*. However, Scene Responsiveness can create instead an illusion *as if* the virtual action had affected the physical world. This is achieved by manipulating the visual signal that reaches the user's eyes and then maintaining this illusion diegetically by ensuring consistency between the tactile and the visual signal.

Receptive and Responsive Affordances

The meaningful action possibilities that a physical scene offers to an actor can be referred to as affordances [156]. While some actions, such as grabbing and moving an object, modify the physical scene, other actions simply "occupy space" without visibly modifying the physical scene itself, e.g., walking on the floor. To distinguish between these different types of "using" the physical world, we introduce the notion of *receptive affordances* and *responsive affordances* in the MR context.

Receptive affordances represent the meaningful *augmentation possibilities* offered by the physical scene without entailing a modification thereof. Such receptive affordances may invite the simple augmentation of a virtual object on the empty area of a desk. A more involved way of augmentation is situating a virtual avatar during a telepresence session on the spatially most appropriate and available seating accommodation given the local user's current pose and gaze in space.

In contrast, *responsive affordances* refer to the meaningful *manipulation possibilities* offered by the physical scene, entailing modification thereof. For example, a chair might not only offer the receptive affordance of taking a seat but also the responsive affordance of moving it, e.g., pulling it out from under the table.

Affordance Features

Both receptive and responsive affordances exhibit *affordance features* that describe the spatial and operational details of the afforded actor-object interaction, similar to the notion of affordance features in robotics [372]. The receptive affordance of "sitting", provided by a chair, might be described by features such as where to stand immediately before taking the seat, in which direction to look at the time, at which height the seating surface is located, whether there are armrests, and whether the seat is fixed, rotating, or mobile. The responsive affordance of "moving the chair" might be described by features such as the hand and body pose relative to the object when grasping for it and whether the object is held at a close body position or dragged over the floor while walking backward when moving it.

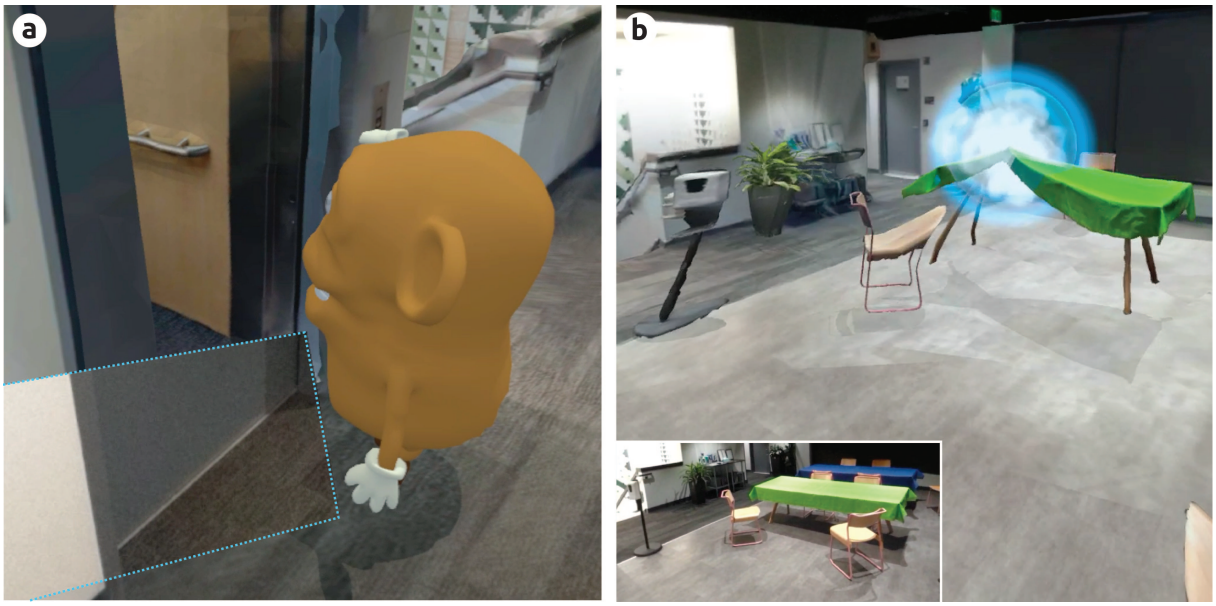


Figure 5.3: State and shape responsiveness. ⓐ Pressing the elevator's call button entails the state-changing response of sliding the elevator door open. ⓑ As the abstract black hole character approaches, objects in its vicinity are deformed, attracted, and even absorbed. Note that the user is still mostly seeing video-passthrough MR and only the visual areas involved along with a small blended neighborhood are visually manipulated. *Please refer to the accompanying video for a more immersive impression.*

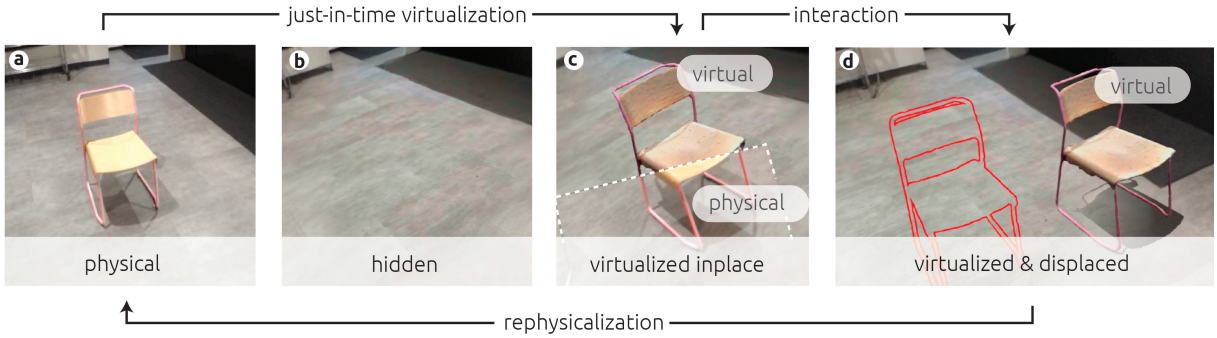


Figure 5.4: Object virtualization. Scene Responsiveness is fundamentally enabled by toggling an object reality state to VIRTUALIZED. ⓐ At first, the object is in PHYSICAL reality state. ⓑ Upon virtualization, we first switch it to HIDDEN reality state, using visual information from a co-aligned digital space twin. We employ alpha blending with a radial gradient for compositing the masking geometry and the passthrough background to obtain a seamless removal effect. ⓒ Then, we insert a digital object twin in the same pose as its physical counterpart, putting it to VIRTUALIZED reality state, but *in-place* manipulation state. Ideally, the visual signals in VIRTUALIZED and PHYSICAL reality state were identical. Therefore, we also preserve scene coherence in VIRTUALIZED reality state, meaning that the virtual counterpart casts a shadow and the masked area receives a shadow. ⓓ A virtual action by an actor leads to a scene response, putting it to *displacing* manipulation state in the process and to *displaced* once the response's target state is reached. ⓐ Rephysicalization in some diegetic way first returns the object to the pose of the physical counterpart and then toggles its reality state back to PHYSICAL.

Situated Animations

These *affordance features* directly parameterize *situated animations* when a character acts upon the affordance. Continuing the example of a responsive affordance for grasping, in our prototype, we use path planning to situate the character at the desired body pose by means of a navigation mesh before grasping Then, we use inverse kinematics to control the hand pose for the actual grasping motion.

Scene Response

Receptive affordances only require the dynamic playback of a situated animation to begin passive object interaction. Responsive affordances are additionally associated with the induced *scene response*. The scene response describes the transformation that the corresponding object undergoes after the user or character acts upon the affordance. This transformation might change spatial properties (Figure 5.1b) or structural properties (Figure 5.3a) or both.

The affordances and therefore the responses in a physical scene are as rich as the objects contained in it and also differ in their specificity. Some affordances such as moving an object, or kicking it over with the foot, are generic and provided by many objects. Other affordances such as opening a fridge by use of its handle or two-handed typing on a keyboard are much more specific to the object. However, all responsive affordances share the general method used to render the scene response as follows.

Toggling the Reality State of an Object By default, we pass through a physical object to the headset displays as captured by the externally facing headset cameras. As soon as a scene response in the scene is triggered, the core illusion of Scene Responsiveness *pretends* that the virtual action stimulates a response in the physical object, however, *actually* we first replace the physical object with a virtual counterpart and then apply the response to the virtual counterpart instead. We refer to this as *virtualization*, i.e., toggling the object's *reality state* from PHYSICAL to VIRTUALIZED. The VIRTUALIZED object can then be spatially detached from the pose of its physical counterpart and transformed in any way virtuality affords. To toggle an object's reality state, we exploit the capacity provided by a video-passthrough MR headset to fully control every light ray that reaches the user's eyes when presenting the scene. Figure 5.4 gives an overview of this process.

Self-contained Illusionary Episodes We propose two types of self-contained *illusionary episodes* that are based on the core illusion of Scene Responsiveness. Both types follow the idea of "What you see is what you feel" to ensure visuotactile consistency, however, toggle reality states according to different rationales. Figure 5.7 provides an overview of the different types.

5.3.2 Scene-Responsive Daydreaming Episodes

A Daydreaming episode aims to unfold more or less surreal happenings while preventing disillusion until completion. It begins with a scene response to a virtual action but then aims to maintain visuotactile consistency to the degree that users are left wondering whether what they saw really happened or if it was just a product of imagination, inspired by *Alice's Adventures*

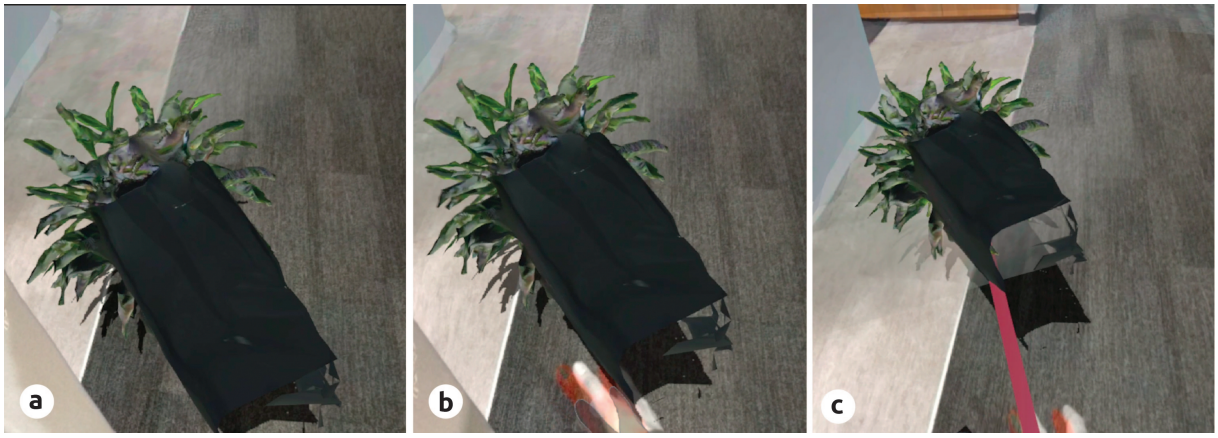


Figure 5.5: Object elusiveness prevents disillusion from visual collisions, where the user perceives the collision visually but not tactilely. In this example, we reuse the same navigation mesh that is used by the character for elusion through object agency.

in Wonderland and *Inception*. In the following section, we present the steps that make up a Daydreaming episode and describe the causes and remedies for disillusion in detail.

5.3.2.1 STEP 1: INITIATING THE ILLUSION

A Daydreaming episode begins as soon as a virtual action triggers a response in a PHYSICAL object. The object's reality state is toggled from PHYSICAL to VIRTUALIZED. Then, the response is applied to the VIRTUALIZED object, as described in Figure 5.4. Figure 5.1 demonstrates how an affordance provided by a single object–the cart–can trigger virtualization of the dependent objects–the coke cans and other items on top–as well.

Within the VIRTUALIZED reality state, we distinguish three *manipulation sub-states* in an object. An object is virtualized *in-place*, i.e., the VIRTUALIZED object's pose, shape, and state initially match those of the PHYSICAL counterpart. A triggered scene response then starts *displacing* the VIRTUALIZED object from its original source pose, shape, or state. Once the direct response ends, the object becomes *displaced* and remains so until changed again. Based on the manipulation state, the object's physics simulation can be intercepted as needed. For example, we override gravity and collision simulation while a character is carrying an object.

5.3.2.2 STEP 2: MAINTAINING THE ILLUSION

Displacing a VIRTUALIZED object elicits a discrepancy between the perceived and the physical reality. This mismatch can cause disillusion when any part of the user's body enters into a *visual collision* with the VIRTUALIZED object: The user perceives the collision visually but not tactilely, providing the user with adamant evidence that they are being tricked. To prevent such disillusion, we must keep the visual and the tactile signal consistent. Tactility is a hard constraint, so visibility must be changed to fit tactility. Therefore, we propose the concept of *virtual object elusiveness* (Figure 5.5): Whenever the user approaches a VIRTUALIZED object too closely, its spatial distance from the user is increased.

We trigger an *elusion event* in an object when the distance between either hand or the headset from the VIRTUALIZED object falls below a specified threshold. To ensure *the lack of tactile feed-*

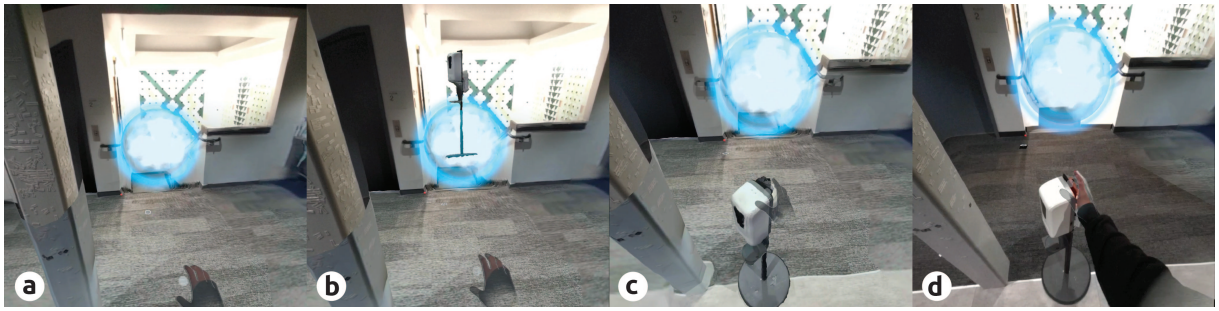


Figure 5.6: Just-in-time rephysicalization in Daydreaming prevents disillusion from tactile collisions. When a tactile collision is imminent, we rephysicalize the object just-in-time to maintain visuotactile consistency. Diagetic in-betweening maintains the illusion.

back, the VIRTUALIZED object can be rendered *anywhere but* in the colliding volume. Therefore, elusion offers two degrees of freedom: *the elusion target* describing where to elude the object to, and the *elusion mechanism* describing how to get it there. The *elusion mechanism* could make the VIRTUALIZED object disappear, disintegrate, or melt, in front of the user's eyes and re-appear, re-integrate, or re-freeze on top of the closest table or shelf. Or the elusion mechanism could give the object some sense of agency, allowing it to innocently slide away a couple of inches as showcased in Figure 5.5c, to jump away, or even, to grow legs and run away. Elusion could also be character-driven. For example, a character might run or jump toward the object and "snatch it away from under the user's nose" in the last moment to then position it somewhere else. Such elusion mechanisms could be chosen depending on the characters and rules of the presented fictional world.

5.3.2.3 Step 3: Completing the Illusion

A second cause of disillusion from visuotactile inconsistency arises from the *tactile collision* between any part of the user's body and the visually hidden, but physically present object: The user perceives the collision tactilely but not visually. Such a collision might even make the user tumble and thus constitutes a safety risk. As a hard constraint, the tactile signal is again determinative. Therefore, we complete the illusion by rephysicalizing the object just-in-time, as either hand or a radius around the headset come close to the hidden object (Figure 5.6). Rephysicaliation brings the VIRTUALIZED object back in-place so that it matches the pose and shape of its physical counterpart by means of a smooth *rephysicalization mechanism* and then toggles the reality state from VIRTUALIZED to PHYSICAL again.

We trigger rephysicalization just-in-time, i.e., when a tactile collision is imminent as estimated by distance. Because the rephysicalization target is defined by the physical object, we can only decide on the mechanism, which–analogously to elusiveness–might again include de- and re-materialization, object agency, character-driven animations. A red object guardian outline (Figure 5.4d) glows up if the distance to the user becomes critically small before the rephysicalization animation is finished.

5.3.3 Scene-Responsive Copperfield Episodes

We design Daydreaming episodes so that users have *no evidence against the illusion*. In a Copperfield episode, we aim to advance the experience by *providing evidence in favor of the*

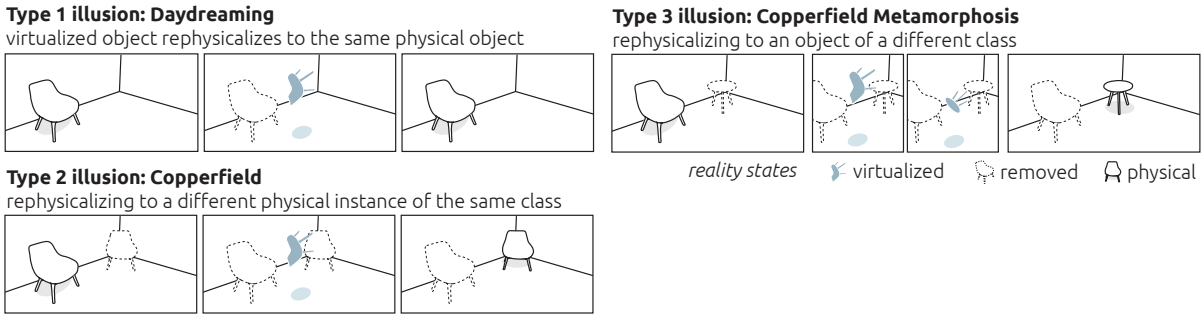


Figure 5.7: Illusion types. In Daydreaming episodes, a one-to-one relationship between the virtual and the physical object is retained. In Copperfield illusions, a single virtual object animates between multiple visually hidden, physically present objects.

illusion through additional touch points with the physical world, similar to the magician who lets the audience stroke the elephant he seemingly made appear "out of thin air".

5.3.3.1 User Experience of Copperfield Episodes

Consider a user wearing an MR headset entering a room that shows multiple chairs. The user can walk up to these chairs and physically interact with them. The user is asked to throw an explosive into the scene by means of a gesture. The explosive destroys some of the chairs while lifting others into the air and tossing them to places distributed across the room. All chairs are still burning a bit. The user walks up to one of the tossed-around but still burning chairs, taking a look at it (*situation* 1). As the user reaches out to touch it, the fire flares up, burning the chair to the ground, thus eluding the user's touch. The user walks up to a different chair tossed somewhere else into the scene by the explosive. The user approaches it and takes a look (*situation* 2). The fire extinguishes. They reach out and find that they can touch the chair despite the fact that it looks slightly virtual. The user takes a seat on the chair. After standing up again, the slightly virtual look is gone. It seems as if the explosive had thrown the physical chair around.

5.3.3.2 Undecidability of Reality States

In the above scenario, the user considers VIRTUALIZED chairs in *situation* 1 and *situation* 2. In *situation* 1, the VIRTUALIZED chair is placed in physically empty space. In *situation* 2, the VIRTUALIZED chair is placed in the same pose and shape as a HIDDEN physical chair. Assuming the lack of visible artifacts, the user has no information to decide from vision only whether the virtual chair will provide tactile feedback or not. The only way of knowing is by trying to touch it. We hypothesize that undecidability may provide an engaging and immersive experience for the following reason.

As a result of undecidability, there is no sense for users in reasoning about its physical existence, because they simply lack the information that would allow them to decide whether an object is physical or not. Instead, they have to trust the system. From this need to develop trust in the system in order to make sense of the tactile world, we hypothesize immersion increases over time, making users forget what they see is a visuotactile illusion rather than actual physics.

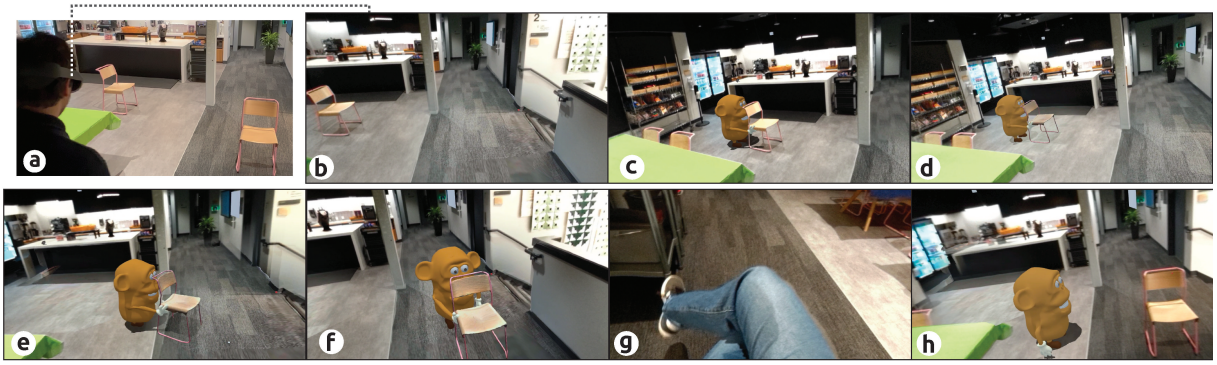


Figure 5.8: Copperfield rephysicalization. ⓐ A physical scene is observed through a video-passthrough MR headset. It features two physical chairs. ⓑ However, the headset only shows the chair *at source* in PHYSICAL reality state while the chair *at target pose* is in a HIDDEN reality state (i.e., diminished from the scene). Even at high zoom in this figure, the removal process is barely noticeable as visual artifacts are drowned out by other artifacts of the video passthrough system such as motion blur. ⓒ Pre-response, a character walks up to the PHYSICAL chair and reaches out to it to grab it. ⓓ As the character reaches the grab pose, the chair toggles from PHYSICAL to VIRTUALIZED reality state and attaches to the character's hand. ⓔ The character starts carrying the chair away. At this point, both the physical source and the physical target chair are hidden from view. ⓕ Controlling the character's hand through inverse kinematics, the character aligns the VIRTUALIZED chair at the target pose, where the still HIDDEN chair is located. ⓖ The user can touch the VIRTUALIZED chair and even sit on it. ⓗ The system toggles the chair's reality state from VIRTUALIZED to PHYSICAL outside the user's view. The chair at the source pose remains HIDDEN.

5.3.3.3 Steps in a Copperfield Episode

Figure 5.8 provides an in-depth description of the steps in a Copperfield episode. Initially, all instances of the same object class are HIDDEN except for one PHYSICAL object. A virtual action triggers virtualization of the PHYSICAL object. Now, however, instead of waiting for just-in-time rephysicalization to its source pose, Copperfield employs story-driven *cross-object rephysicalization*, transporting the VIRTUALIZED object to seemingly empty space, however, in fact aligning it with a visually hidden, but physically present object, the user does not yet know about. Once the user approaches the still VIRTUALIZED, but physically aligned object, they find it will not escape from them.

5.3.3.4 Additional Considerations on Copperfield Rephysicalization

Pseudo-random rephysicalization *Story-driven rephysicalization* in Copperfield is a chance to double down on the illusion and provide faked evidence that virtuality affects physicality. This offers a chance for *pseudo-randomness in the rephysicalization*. Users are conditioned from physics, gaming, and movies that explosives toss around objects at random. Using an animation that looks like a physics simulation but is actually a deterministic process deliberately transporting the object to a defined target position might reinforce the sensation of randomness. The subsequent ability to physically interact with a seemingly randomly placed object might advance the believability of the illusion even further.

Premature and ultimate rephysicalization A Copperfield episode must deviate from an ongoing and longer-running story-narrated rephysicalization, and instead rephysicalize just-in-time if a user is about to run into a HIDDEN object. Given the user has never seen the HIDDEN

targets, this is much more likely to happen than in Daydreaming. Just-in-time blocking the seemingly empty, but physically dangerous space, e.g., through a virtual fire or NPCs might avoid premature rephysicalization. To *ultimately conclude* any Copperfield episode, new instances of the virtual object must enter the scene through different diegetic ways, re-introducing a one-to-one relationship between physical objects and virtual counterparts.

Metamorphosis rephysicalization The above-presented Copperfield episode hinges on the availability of multiple instances of the same object class in the physical scene. By virtually morphing the virtualized object "on its way" (Figure 5.7 bottom), we can broaden the applicability of the Copperfield episode for more diverse scenes. *Morphing mechanisms* could range from a literal mesh-deforming morph operation to computationally less demanding operations like shrinking the object mid-air to a scale of zero and growing it back as a different object, or even abstract ways of changing an object's type, e.g., where the character leaves the room with the object of the source type and comes back with an object of the target type. In particular, the latter approaches alleviate the need for complex, potentially object-specific animation work.

Multi-user rephysicalization In any Copperfield episode, physical objects are hidden before the user sees the scene for the first time. This limits its applicability for single-user home-usage applications because users probably will know their physical space well. However, we imagine the possibility of *multi-user setups* where co-present users manipulate the physical scene, thus making any individual user lose overview of the physicality.

5.4 Architecture and Implementation

In the following, we show how we obtain a rich digital representation of a space and its objects using our *TwinBuilder* component. Then, we present our spatial computing and shading algorithm which makes use of the produced twins to toggle between object reality states and ensures scene coherence, implemented in the *RealityToggle* component. Finally, we present our *Spielberg* component which makes use of RealityToggle in its application flow to control Daydreaming and Copperfield episodes by starting Scene Responsiveness, ensuring elusiveness, and triggering rephysicalizations while controlling the character and providing user interactivity with objects.

5.4.1 *TwinBuilder* Component: Co-aligned and Semantically Rich Space and Object Twins

5.4.1.1 Step 1: Capture Space and Objects in Individual Scans

We obtain a 3D mesh model with textures of the space after moving relevant interactable objects out using Polycam on an iPhone with a LIDAR sensor. We also obtain the models for interactable objects.

5.4.1.2 Step 2: Integrate and Annotate in a Unified Twin Representation

After cropping and converting the models, we import them into Unity and use our custom Unity plugin (Figure 5.9) for further processing. In particular, we integrate objects in space in a single twin by positioning and orienting the objects faithfully. The mesh area navigable by

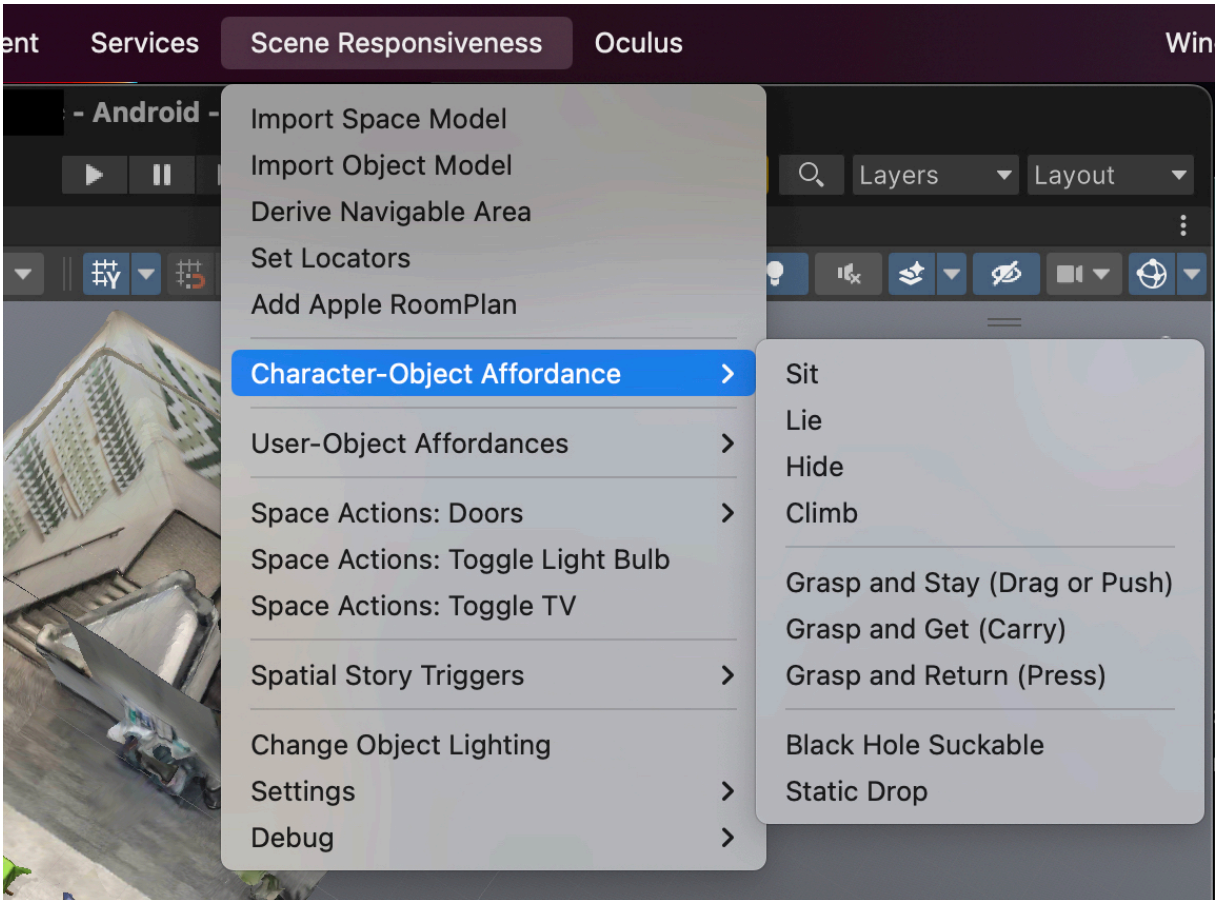


Figure 5.9: Our custom Unity plugin allows semantic annotation of the scanned space, needed to enable Scene Responsiveness.

the character is automatically derived based on the mesh faces' relative orientation to the floor plane using Unity.

Afterward, we annotate character and user affordances. We implemented an exemplary set of receptive character affordances (sit, lie, hide, climb) and responsive affordances (drag/push, carry, press for the bi-ped character, absorbable by the Black Hole). Selecting them in the plug-in adds the needed Unity components and allows parametrization for the associated situated animation. We implemented summoning, disintegration, and repulsion interactions as responsive user affordances. For responsive affordances, the `RealityToggle` Unity component is automatically added. At this point, the resulting digital scene can be executed like a game on the computer or deployed in *interactive mode* to the headset. Alternatively, in addition, affordances and spatial triggers can be connected in a sequence to compose longer-running `Spielberg` stories for *story mode*. This architecture generalizes easily to new scenes, is cleanly extensible, and coherently fits into the 3D application development process.

The resulting digital twin contains all information needed to render coherent, situated, and responsive MR in the passthrough view. Labels in the figure (Figure 5.10) indicate how we use this twin.

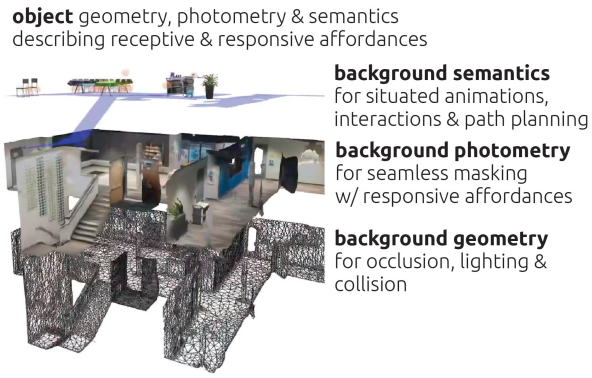


Figure 5.10: Digital twins provide the geometric, visual, and semantic space and object information, needed for Scene Responsiveness.

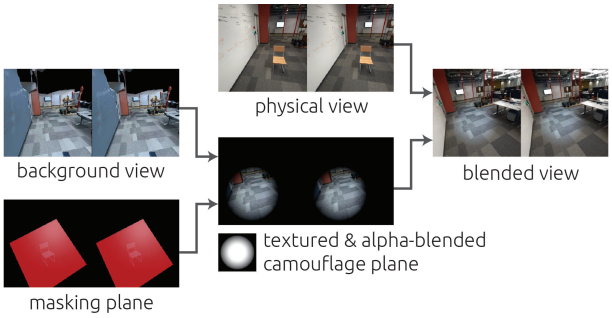


Figure 5.11: Passthrough compositing pipeline. Brightness increased for illustration. We use an architecture of virtual stereo-cameras with render textures to texture the masking occluder, Please refer to Figure 5.13.6 for faithful colors.

5.4.1.3 Step 3: Co-align Digital Twin with Physical Space

We also specify two easily identifiable and stable floor locations in the twin, e.g., wall-wall-floor corners, used for co-alignment in the headset.

5.4.2 *RealityToggle* Component: Spatial Computing and Shading

RealityToggle is our spatial computing and shading component which enables toggling an object's reality state, i.e., virtualizing it for our core illusion of Scene Responsiveness (Figure 5.4), by making use of the previously built twin.

5.4.2.1 Masking Pipeline

First, we *pose a masking occluder* in the 3D scene (Figure 5.11 left). We deliberately do not use a pixel-perfect mask but use a quad as a larger and simpler masking geometry to avoid complicated edges and enable smooth alpha-blending towards the edges. We position the occluder quad on the ray from the headset to the object center, oriented orthogonal to the ray, and tangent with the object's bounding sphere. Heuristically, we set the quad's edge length large enough to fully enclose the object in the conical frustum behind the quad's opaque center area.

Second, we *texture and shade* the masking occluder (Figure 5.11 mid). The quad is textured with a stereoscopic render texture, produced from two additional *masking cameras* with the same calibration as the eye cameras, but rendering the background mesh. We implemented a custom *stereoscopic masking shader* that samples the eye-specific render texture at the corresponding screen coordinates of the respective eye camera. We apply a radial two-step gradient texture map for alpha blending to achieve a smooth fade toward the edges. We render the masking quad with standard depth testing and writing, thus integrating it coherently into the scene.

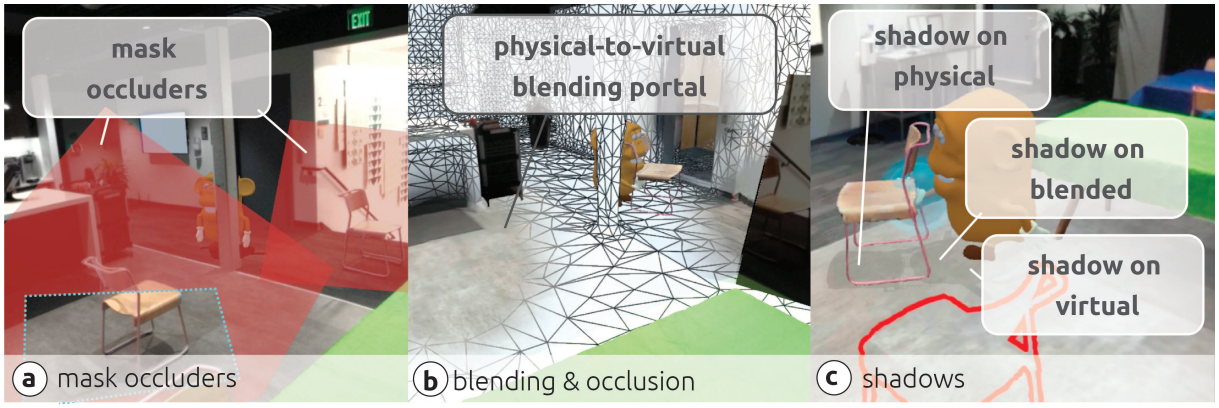


Figure 5.12: Spatial computing and shading architecture. ⓐ We can pose multiple masks in the scene, even in the same line of sight. ⓑ Dynamically assigning rendering layers, rendering queue positions, depth write, and depth test flags, ensure depth coherence for characters, objects, and hands. ⓒ A custom lighting model ensures that shadows are coherently cast onto physical and virtualized objects and surfaces as well as blended areas.

Finally, we composite (Figure 5.11 right) the passthrough layer with the rendered quad (`Blend SrcAlpha OneMinusSrcAlpha, One OneMinusSrcAlpha`). The masking cameras render in a black skybox to maintain the passthrough color faithfully.

5.4.2.2 Scene Coherence in Masked Frustums

Depth coherence beyond the frustum Next, we insert the virtual object twin. Because the masking quad is placed in 3D, a frustum emerges behind it. Virtual content inside the frustum gets culled by occlusion, however should render as if nothing changed. Therefore, we make sure that all virtual content, that shall be rendered in front, is captured by the render texture by dynamically manipulating the camera layering. In particular, the virtual counterpart, the character, and any other virtual object are included in the masking render texture.

Lighting coherence for shape responsiveness In Figure 5.3, we have demonstrated shape responsiveness by means of our abstract Black Hole character and its ability to deform and absorb objects. We implemented the mesh deformations efficiently in a *Black Hole surface shader* with a vertex modifier. The use of a surface shader enables adding and even deforming shadows during the mesh deformation (`fullforwardshadows addshadow` in the surface shader pragma).

5.4.2.3 Scene Coherence in Unmasked Areas

In unmasked areas, virtual content shall be rendered coherently with the physical scene instead. Ensuring both coherence in masked and unmasked areas will elegantly produce the desired coherency in the alpha-blended regions of the view (Figure 5.12).

To enable occlusions, we create a second but invisible instance of the background mesh and all the virtual objects with a custom two-pass "Physical" shader that ensures coherent lighting and occlusion. We ensure coherent *occlusion* in a first *fragment shader pass* by rendering the background mesh as a phantom (i.e., rendering early in the queue filling the depth buffer (`ZWrite On`) but without drawing (`Blend Zero One`)

We ensure coherent *lighting* in a second *surface shader pass*, implementing a custom *ShadowReceiverOnPassthrough* lighting model. This lighting model always renders black but redirects inverted attenuation into the alpha channel, thus allowing to blend shadows with the passthrough view (`Blend SrcAlpha OneMinusSrcAlpha, One OneMinusSrcAlpha` with `keepalpha` surface shader pragma). To ensure shadows cast onto the background mesh are also occluded by the background mesh itself, we exploit that shadows can only exist on a surface (ZTest Equal).

We ensure coherent *collision* with standard physics simulation between any virtual objects and the invisible background mesh.

5.4.2.4 Additional Tools

For illustration in this chapter, we implemented a *revelation lens* that can be pulled up on a controller. Used throughout the figures in this chapter, it renders as an invisible occluder thus revealing the passthrough layer. *Object guardian* as a safety fallback to just-in-time virtualization is implemented as a red outliner where width scales with by distance between user and object. We implemented a variety of *debug shaders*, e.g., to render the mesh's wireframe.

5.4.3 *Spielberg* Component: Character and User Environment Interaction

Our event-driven *Spielberg* component takes care of directing the characters, handling user input, and controlling the behavior, pose, shape, and state of the objects involved.

5.4.3.1 Character-Environment Interaction

Character Model, Rig, and Animations Design We modeled, rigged, and animated our *Bi-Ped character* from scratch [99] in Blender and used Unity's Mecanim animation system. We switch between in-place animations for NavMesh navigation and tree-based blending with root motion for situated character animations.

Control We implemented three ways of controlling the character. In *interactive mode*, the user can raycast onto affordances, visualized through spheres, and character ghost previews upon hovering. In *story mode*, the action target is updated automatically, based on a scripted story and spatial triggers. In *telepresence mode*, the action target is updated by the remote actions of a user. Details on the implementation of the underlying state machines for character control can be found in the supplementary material.

5.4.3.2 User-Environment Interaction

We build on the Quest Hand Tracking and the Oculus Interaction SDK to implement hand tracking, gesture detection, and interaction patterns such as remote object selection and summoning. Hand tracking allowed us to include hand occlusion over virtual content, in particular over masked areas.

5.4.4 Equipment

Design time equipment We use Polycam 3.1 on an iPhone 13 Pro for scanning and Blender 3.3.1 for model conversion. Our annotation plugin runs in Unity 2022.1.18f (Oculus Integration SDK v49.0, OVRPlugin 1.81.0 for OpenXR) on both Windows and Mac.

Run time equipment for single-user experiences From Unity, we deploy the system to a Meta Quest Pro (build v49.0), which offers colored video passthrough. Note that our system works on a standard Quest Pro without the need for jailbreaking or attaching additional custom cameras, because our object masking pipeline does not require raw video frames.

5.5 Preliminary Evaluation

Participants and Apparatus To gain insight into how users perceive the Copperfield illusion, we conducted a user evaluation with 20 participants from our institution. Users wore the Quest Pro and used our app in interactive mode, instructed by the experimenter what to select or do next. We set up an evaluation course, leading through various stations that involved receptive and responsive affordances and ending with the user taking a seat on a previously hidden chair that seemingly has been placed there by the character as part of a Copperfield episode. We adapted the same evaluation course to two spaces: a Corridor space (Figure 5.13) and a Library space (see supplementary material). We performed a Wilcoxon rank-sum test [357] on our between-subjects, post-evaluation Likert questionnaire (Figure 5.14).

Rationale Participants were shortly briefed on the abstract capability of the system to "explore new ways with respect to perception of and interaction between users, virtual characters and their environment".

However, we did not provide specific details on the Scene Responsiveness and Copperfield illusions to prevent tendencies *in favor of the illusion* through social-desirability bias in participant responses, or tendencies *at the expense of the illusion* through attentional bias in participant perception. Participants were fully guided through the evaluation course verbally by the experimenter, gathering qualitative information in a semi-structured fashion and asking for perceptual feedback at specific moments along the course. We debriefed participants at the end of the procedure.

5.5.1 Procedure and Evaluation Course

Because timing, user agency, user knowledge, and user expectation are pivotal to understanding the expressiveness of our illusion perceptibility evaluation, we describe our procedure in detail. Please refer to Figure 5.13 for a visual walk-through.

First phase: On Boarding ① After putting on the headset, the participant learned the basic ray-casting interactions for hovering and selecting affordance visualizers by the example of a receptive hiding affordance. They were trained to control character navigation by pointing and selecting on the ray-interactable nav mesh on the floor. ② We asked them to sit down on

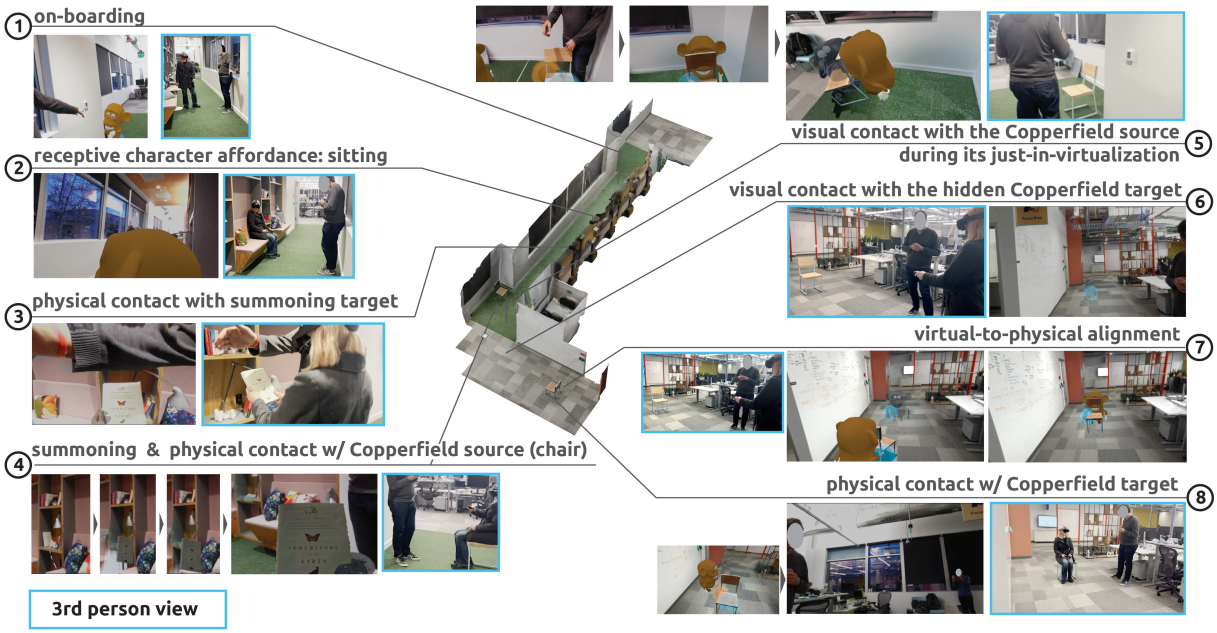


Figure 5.13: Evaluation course, adapted to our *Corridor* space. The core moments are seen in ⑤ and ⑥: ⑤ The participant faces the chair they just sat on as the character grabs it, thus triggering the *just-in-time virtualization*. Notice how the character runs away with it while the physical chair is masked coherently. ⑥-right: As the participant walks around the corner, all they see is empty space and a transparent blue sphere indicating a selectable drop affordance. At that point in the evaluation course (analogous spot for the Library space included), all participants were asked to describe what they saw. All participants described the blue sphere. None of the 20 participants described the existence of anything suspicious at that point in time yet. Notice how the passthrough view still shows the other person in their view but does not show the physical chair visible in the 3rd person view (⑥-left).

a bench. Using a spatial action trigger, the character automatically took a seat next to them, so they learned about its ability to semantically interact with the world on its own.

Second phase: Scene Responsiveness In ③, they had physical contact with a book: They were asked to take it off the shelf and read the title out loud. In ④, they were asked to take a seat, point toward the book, and summon it by selecting it. The book was virtualized just-in-time and levitated towards the participant's selecting hand. They were asked to read out the author.

Third phase: Copperfield illusion Because the participant took a seat, step ④ also represents the first physical contact with the chair as the Copperfield source object. ⑤ The participant was asked to stand up and turn around to face the chair, which featured a visualizer for a responsive affordance at its side. The participant was instructed to select the responsive affordance in order to "command the character to grab the chair". The participant maintained visual contact during the character's grasp and thus observed the just-in-time virtualization from a distance of approximately 1m. Note that the character grasps the chair without moving it yet so that virtualization and displacing are *two fully decoupled steps* with a pause in-between until the user selects the next affordance. This allows us to analyze the perceptibility of virtualization before participants were made aware that the chair is suddenly movable and thus *must* be virtual. Before, they might not even notice the virtualization. Once the character grasped the chair, the participant was asked to describe what happened. Then, they were asked to make the character walk away while grabbing the chair, thus now inevitably exposed to the fact that we substantially manipulate the visual signal.

Next ⑥, the participant was asked to walk around the corner. At this time, they first made visual contact with the hidden Copperfield target chair, more precisely they made visual contact with the background they are presented with. Generally, we expected the participant not to suspect the existence of a second chair.

At the position of the Copperfield target ⑦, participants were shown a blue sphere visualizing an affordance which they were instructed to select to make the character drop the chair. The character walked up to the hidden chair and aligned the virtualized chair it was carrying with the visually hidden, but physically present chair. Participants were asked to follow the character.

Once the character dropped the chair ⑧, participants were asked to take a seat on the still virtualized chair.

5.5.2 Results and Discussion

5.5.2.1 Perception of Copperfield Just-in-Time Virtualization

Initiating a Copperfield episode comprises at least three facets, interesting for evaluation.

Our *object masking coherence* question (Figure 5.14 ⑤ top) shows balanced responses. 50% of participants agreed slightly, mostly, or strongly that they noticed the masking whereas 50% disagreed. Such a balanced result is quite interesting as it certainly indicates masking fidelity can be improved, yet might hint at the fact that personal and situational factors, such as the current focus of attention or immersion might play a role.

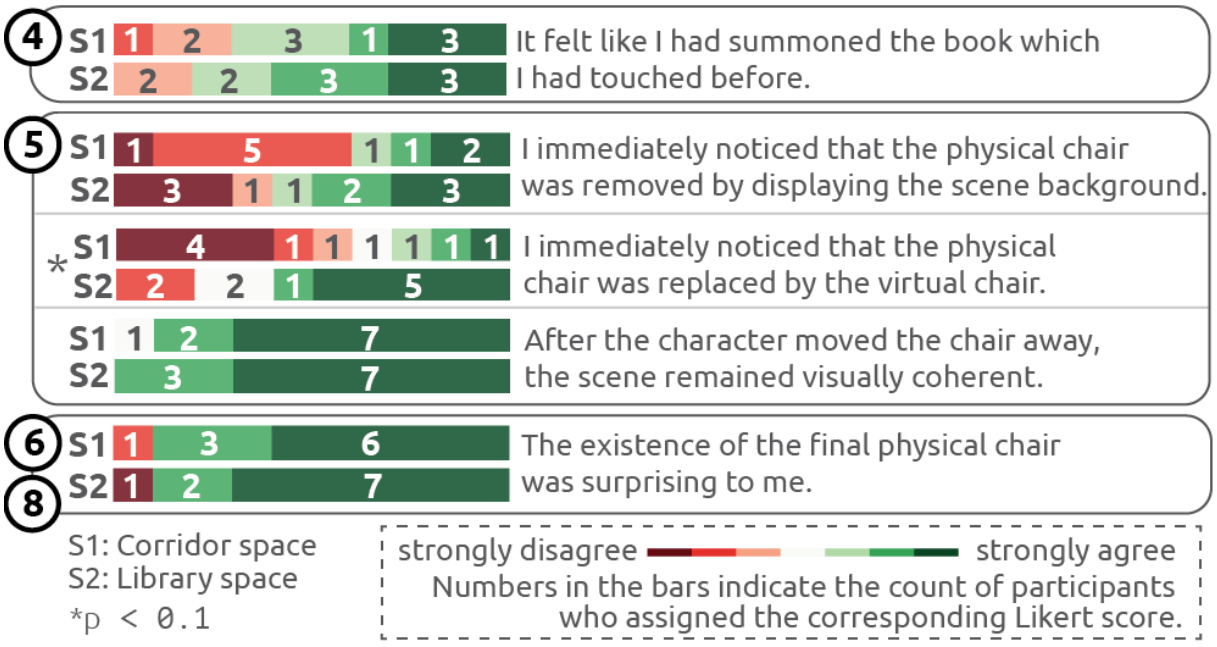


Figure 5.14: Evaluation of a subset of questions from our post-evaluation questionnaire. Numbers correspond to the phases in Figure 5.13.

The *object insertion coherence* question (Figure 5.14 ⑤ mid) is the only one where answers significantly differed between spaces ($p < 0.1$) as revealed by the Wilcoxon rank-sum test. Indeed, considering the scanned object model of the chair in the Library space, its textures are susceptibly brighter than the physical object that appears in the passthrough view. This hints at the potential for either including a way of adapting the texture brightness in the twin-building procedure or a way for real-time adaptivity of the system.

Finally, the *coherence after displacing* question (Figure 5.14 ⑤ bottom) shows that–once the physical object is masked–14 out of 20 participants consider the mixed-reality scene strongly visually coherent, and 5 mostly visually coherent.

5.5.2.2 Perception of Copperfield Target Masking and Rephysicalization

As seen in Figure 5.14 ⑥–⑧, the existence of the physical target chair was surprising to 18 out of 20 participants. A pilot user initially even refused to sit down when asked. We also asked the question "How did you become aware of the physical chair in the final interaction?" offering participants to choose all factors of disillusion. In the following, we only report the answer corresponding to each participant's earliest moment of disillusion. 2 out of 20 participants responded "When I walked up to it, I noticed visual artifacts which gave it away early". 3 out of 20 participants responded "I discovered it through the headset's peripheral gap before the experimenter told me to sit down." 8 out of 20 participants responded "When I walked up to it, but before being asked to sit down, I suspected the existence of the physical chair from the story design." 7 out of 20 participants responded "I was still skeptical after being asked to sit down and only believed the physical chair's existence after touching it or sitting down".

The fact that visual artifacts were selected only twice is promising. In contrast, in the previous paragraph, discussing just-in-time virtualization, 50% of the participants agreed slightly, mostly,

or strongly, that they immediately noticed the scene's background was blended in to visually remove the chair. The perceptual difference between the two is observing the *transition from physical to blended scene* versus only observing the *already blended* scene. It seems natural that the eye can pick up abrupt changes easier than finding artifacts in a static signal (similar to the example of Gestalt emergence in James' optical illusion of a Dalmatian, popularized by Gregory [107]). Such a perceptual phenomenon seems to contribute to the convincing results of the Copperfield illusion.

5.5.2.3 Relevance of Visual Fidelity

The unexpected inflection point in the Copperfield episode hinges on visual fidelity. Interestingly, it seems to be of secondary importance to maintain suspension of disbelief and ensure consistency with the rules of the presented fictional universe in other parts of the experience. In particular, nearly half of the participants expressed unsolicitedly that the character felt "real" (P2, P15), "here" (P2), that they felt a connection to it (P19), that the character seemed aware of the user (P3, P4, P7, P12), that they felt alone after the character left the scene (P18), or similar. These statements indicate presence or connectedness despite a simple character design. In summary, *semantic fidelity*–i.e., awareness of objects and their affordances–drives these experiences, allowing meaningful situated user interactions and character behaviors in the physical scene while *geometric fidelity* ensures virtual content looks as if it is located in the user's space.

5.6 Limitations and Outlook

Increasing visual fidelity The use of 3D scans entails challenges for the representations' visual fidelity.

From a *static* perspective, while the results of LIDAR-based scanning with today's consumer-grade technology are already astonishing, slight geometric distortions across larger spaces, tesselation artifacts at edges, the lack of finer geometrical structures, or mesh holes due to feature sparsity remain perceivable in rendered surfaces and objects. Thus, the use of representations that offer higher visual fidelity such as Neural Radiance Fields, in particular those that aim to infer missing regions [206; 232], can be of interest. This may become more important as the video-passthrough quality increases in the next generations of headsets.

From a *dynamic* perspective, fidelity suffers from changes to the physical scene occurring between scan and use. Moving any object, or even adding clutter to the scene, leads to misalignment. Thus, employing recent methods for 6DoF object tracking is of interest to maintain a real-time understanding. Similarly, adapting light intensity or balancing might help ensure photometric fidelity. Also, except for hands, our system prototype does not take into account dynamic occluders in front of a visually removed object such as humans crossing the line of sight. Adding body detection can help mitigate this.

Generalizing across spaces and objects However, a technical long-term vision at play might consider the more abstract problem of generalizing from static twins to procedural scene understanding and reconstruction for real-time twin generation. Here, we used a static yet integrated representation of geometric, photometric, and semantic layers for space and objects. With this chapter, we hope to provide another reason for advancing efforts in computer perception to decompose reality into logical constituents, thus potentially even allowing to provide

MR experiences that are perceived as passthrough [52; 368], however actually result from generatively re-rendering reality. Apart from this generalization on the *perception* side, it is equally thought-provoking to generalize the *rendering* side through procedurally generating animations of characters.

Applications However, we believe that Scene Responsiveness can already enable a variety of captivating applications today. As insinuated throughout the chapter, situated gaming is a natural fit for Scene Responsiveness. Gameplays that take advantage of Scene Responsiveness could revolve around a wave-like invasion–say, of spiders– which open closed cupboards or press out electrical sockets to enter the scene, and then try to steal physical objects–such as the user's computer to gather intel. Players defend themselves and the object of interest by summoning and throwing objects, blocking the way by enlarging them, closing virtually opened cupboards again, or destroying the cupboards entirely to slow down the invasion. They collect points to make stuffed animals or other objects spring to life and help them in defense. We imagine the possibility of an ecosystem around scanned objects, modeled characters, situated animations, and designed gameplays and see the opportunity for sharing a scanned space and its objects with users in the same space, not only but also for a co-presence multi-player experience.

More broadly, we understand Scene Responsiveness as a general concept with the potential to enable or enhance arbitrary domains of mixed reality. In scene-responsive *telepresence*, activities of remote users can be semantically retargeted onto a virtual avatar interacting with physical objects in local space. This enables avatars to take a seat on a physical chair, even if the chair is pushed under the table and otherwise would not be available for the avatar. A proof-of-concept approach and video, implemented in our system with WebRTC, can be found in the supplementary materials. For *health*, we imagine as the user reaches out to a physical unhealthy chocolate bar, it morphs into something less appealing such as a spider. Or, the chocolate bar grows legs, runs away, morphs into a banana while running, and rephysicalizes at the location of a physical banana. For *learning*, a digital workout coach might use the physically available rowing machine in space to demonstrate correct usage. For *movie entertainment*, 2D content shown in a 2D panel in the headset may semantically affect the physical space, e.g., physical objects in the user's living room start floating when watching a movie situated in outer space. Characters such as a Minion might even step out of the 2D screen and seemingly steal a physical object before stepping back into the frame.

5.7 Conclusion

We presented Scene Responsiveness as a novel concept to increase integration between the virtual and the physical world through high-fidelity illusions. We unfolded the end-to-end illusionary experiences of Daydreaming and Copperfield that maintain visuotactile consistency through elusiveness and rephysicalization. Our evaluation with 20 users suggests that our coherence-preserving spatial computing and shading implementation for just-in-time virtualization enables highly believable visuotactile illusions across different spaces. Considering the increasing industry focus on video-passthrough MR, we believe that Scene Responsiveness can not only become a concept for exciting gaming experiences but for the MR field in general.

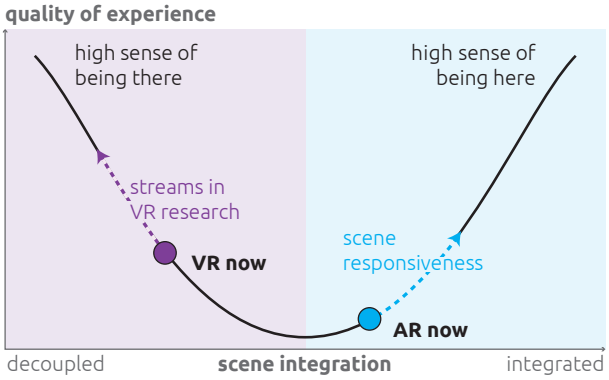


Figure 5.15: Scene Responsiveness as a general concept in Mixed Reality. *Left:* The "sensory Turing test" can be considered one of the grand visions of VR research. It captures the objective to produce artificial signals representing a virtual scene that are indistinguishable from natural signals pertaining to the physical scene. As the technical abilities to decouple virtuality from physicality increases, the user's sensation of being present in the virtual environment also increases, thus presumably improving the quality of the experience. As a result, prominent research streams in VR aim to advance the perceived realism of artificially producable haptic, visual, and auditive signals. *Right:* Similarly, AR research also aims to produce virtual augmentations that are indistinguishable from the natural signal. However, diametrically opposed to VR, the sensation of virtual content being present increases *not as decoupling but as the integration with the physical scene increases.* Scene Responsiveness adds to the ability to integrate virtuality and physicality by enabling virtual content to seemingly affect the physical world.

5.8 Further Details

5.8.1 Application: Scene-Responsive Telepresence

In the chapter, we have described gaming as a promising application of Scene Responsiveness. While Scene Responsiveness aims to be a general concept in Mixed Reality (Figure 5.15), we consider telepresence as a second major application as outlined in the following.

Motivation Previous work [108; 150] has dealt with situating virtual content–in particular the remote user's avatar–in local space while preserving or satisfying certain spatial properties. From this work, we borrow the notion of establishing a semantic binding between local and remote objects. However, Scene Responsiveness lifts the limitations of the local space's physicality when situating avatars.

Spaces In Figure 5.16, consider two roughly similar yet subtly different spaces as they might be found in real-world scenarios between two interlocutors. Both the remote space (left) and the local space (right) feature walkable areas, seating accommodations, tables, doors to enter and exit the scene, etc. However, while the seating accommodations in the remote space are freely accessible, note that they are physically obstructed in local space.

System To demonstrate the concept of scene-responsive telepresence, we monitor the remote space with an iPhone running ARKit and detect a remote user's pose in space using ARKit's body tracking. We stream the detected body pose into the local user's headset to situate the remote user's avatar in local space and render the scene-responsive scene.

Remotely Triggered Avatar Actions and Scene Repsonses Fundamentally, the actions of a remote user are imitated by their localized avatar while semantically retargeting any

remote object usage or manipulations on the objects present in local space, triggering scene repsonses where necessary.

As seen in Figure 5.16, while the remote user is absent (ⓐ-left), we do not show their avatar in local space (ⓐ-right). As the remote user enters and walks in the remote space (ⓑ-left), their avatar is situated in a neutral position in local space (ⓑ-right). Next, the remote user takes a seat (ⓒ-left), however, neither the closest nor any other seating accommodation in local space is directly accessible for the avatar. By building on the concept of Scene Responsiveness, the avatar can still take a seat by manipulating the closest chair in space (ⓒ-right). After manipulating the closest chair, it takes a seat (ⓒ-right). As the remote interlocutor stands up or leaves the space (ⓓ-left), the avatar takes its neutral position again (ⓓ-right).

Limitations and Outlook The procedure described above demonstrates the potential use of scene-responsive telepresence for more integrated avatars in a 1-to-1-object 1-to-1-user setup without additional constraints. As such, our demonstration is limited in a variety of ways. We use an asymmetric setup where only one user wears a headset while the other user is observed through an iPhone, rather than offering bidirectional situated communication. Position in space as detected by body tracking is used as a proxy for object usage, rather than find-grained activity detection, e.g., based on egocentric vision. A cartoonish avatar rather than an avatar with an expressive face is employed, thus prohibiting avatar-driven conversations, and we only consider taking a seat as a responsive affordance in this setup. More importantly however, further research is needed to address the *assignment problem* of mapping objects that are manipulated as part of an activity under space-variant multiplicities, e.g., if objects exist in remote space but not in local space. The inherent use of digital object twins in our implementation offers the potential to "teleport" objects along with the avatar. Despite these limitations, we argue that the presented scenario demonstrates how Scene Responsiveness offers the potential for more integrated avatar interactions. In the future, inspired by work on asynchronous interpersonal communication [85], this might even enable fully asynchronous scene-responsive avatar interaction where human-environment manipulation is retargeted to local space *after* it has been *recorded* in a different space.

5.8.2 Details on the Second Evaluation Space

To evaluate the Copperfield episode, we adapted the same evaluation course to two spaces: a library space and a corridor space. Both featured the same steps in the same order. Figure 5.17 gives an overview of the library space.

5.8.3 Details on the Spielberg Component's Character Control

Our bi-ped character and its interplay with the RealityToggle component is controlled through an integrated set of finite state machines, namely the base state machine and the hand Inverse Kinematics (IK) state machines.

Base State Machine The base state machine controls navigation and situated animation playback. Navigation builds on Unity's navigation mesh (NavMesh) agent, using waypoints and an A* search to navigate on the NavMesh, derived from the scanned mesh. It is triggered either upon selecting a point on the NavMesh directly along with the character's look-at direction (in

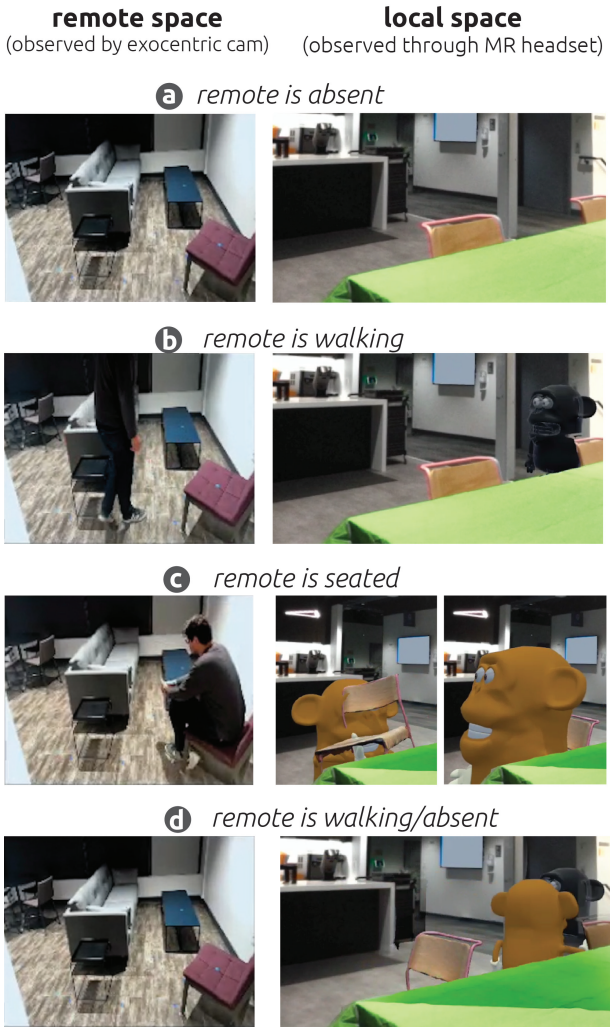


Figure 5.16: Scene Responsiveness for telepresence. The sequence shows how a remote user enters the remote space, takes a seat, and leaves the space again, while their avatar in the local user's space manipulates a chair (©-right) to be able to imitate the remote user's behavior. The phrase in italics is the activity assumed based on remote user body position.

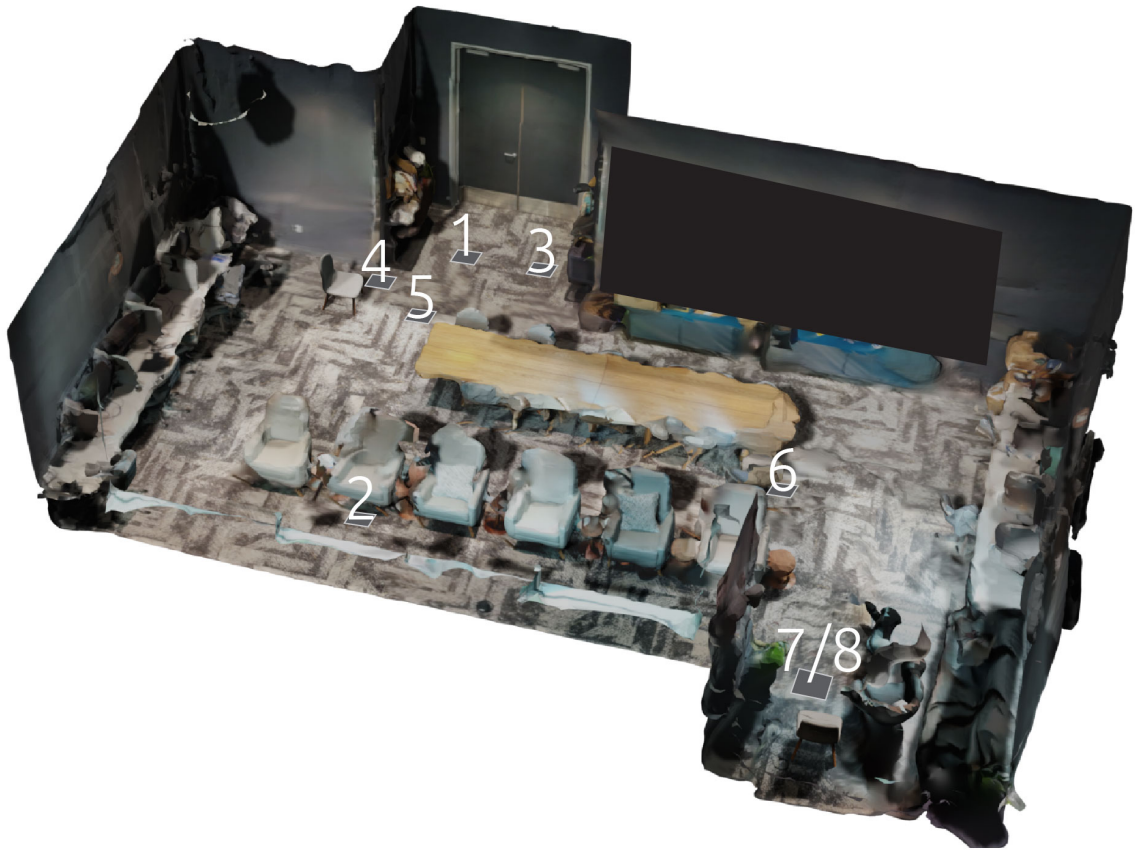


Figure 5.17: Library space as a second condition for our evaluation. The numbering corresponds with the steps in the CORRIDOR, provided in Figure 5.13: 1) on-boarding 2) receptive affordance: sitting 3) physical contact with summoning target 4) summoning and physical contact with Copperfield source 5) visual contact with Copperfield source during just-in-time virtualization 6) visual contact with hidden Copperfield target 7) virtual-to-physical alignment 8) physical contact with Copperfield target Identical chairs were positioned at 4 and 7/8 as Copperfield source resp. target objects.

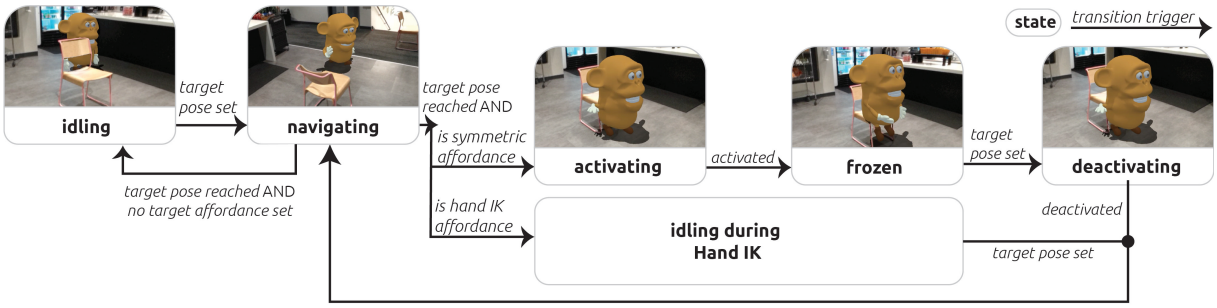


Figure 5.18: Base state machine for situated character-object interaction. It represents the base layer in the character's state machine.

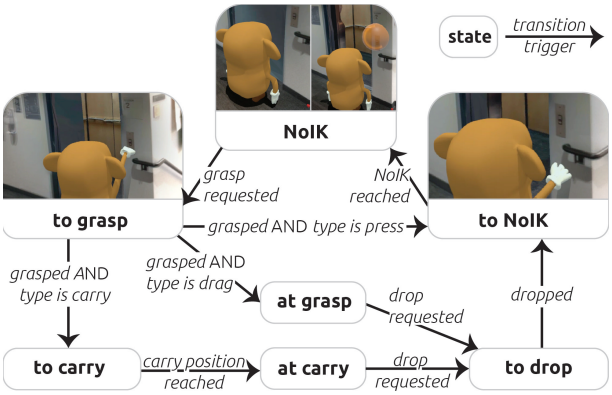


Figure 5.19: The Hand IK state machine overrides the character's base pose from the animatimation state machine's base layer.

our case, through a rotation-aware ray-cast onto the mesh), or by setting a target affordance as part of the interactive, story, or telepresence control mode. To enable the use of NavMesh agents and motion blend trees, navigation animations are built on in-place motion. As shown in Figure 5.18, once, the character has reached its navigation target, it returns to its idle mode or performs the situated animation associated with the reached target affordance. Situated animations that sustainably apply a change to the character pose, such as climbing, turning, or taking a seat, require root motion. We switch between the two motion types based on a predefined animation look-up table. The affordance features directly translate to parameters of the character animation, e.g., the pose from which to start the situated animation. In the state machine, we differentiate between two types of animation flows. For symmetric affordances, the animation is first played-back during the ACTIVATING state, e.g., taking a seat. The character then remains in a FROZEN state until a target affordance and thus the associated target pose is updated again, which triggers the *deactivating* state, and subsequently the navigation to the new target pose. For deactivating a situated character animation, we reverse the animation playback. We implement the character's ability to grasp objects with an additional Hand IK state machine for each hand as a second type of the animation flow.

HAND IK STATE MACHINE The Hand Inverse Kinematics (IK) state machine is implemented as an overriding animation layer that is applied in an IK pass subsequent to the character's animation base layer. We implement three types of grasping in Hand IK state machine. Figure 5.19 gives an overview and also illustrates the simple case of pressing. Initially, the hands is located at the natural pose as indicated by the base layer in the NOIK state.

To *press* the elevator button, a coffee machine machine button, or similar, the hand transitions into the TO GRASP POSE state. In this state, we linearly interpolate the hand's position and spherically interpolate the hand's rotation from its natural pose to the grasp pose as dictated by the affordance feature. Upon completion, the hand enter's the TO NOIK state which lerps and slerps the hand back to its natural pose until this pose is reached, triggering the transition to the initial THE NOIK state. Instead, to *carry* a smaller object, such as a coke can, the hand "takes a detour" over the TO CARRY state which moves to a predefined pose close to the body and the AT CARRY state. As this state layer is separate from the base layer, the character can carry objects while walking somewhere and even perform other situated animations that do not involve the carrying hand. To *push or drag* heavy objects, the hand simply reaches out, but then stays at the grasp position, e.g., to drag a chair through the room.

In the technical implementation, we extend the above to also enable combined animation flows, e.g., to have the character navigate to a target pose, apply a root motion animation such as

tip-toeing in front of a high table, grasping an object on the table, and then coming back to a normal stand on the ground, with the object in the hand's carrying pose.

6

HandyCast: Smartphone-based Bimanual Input for Virtual Reality in Mobile and Space-Constrained Settings via Pose-and-Touch Transfer

Figure 6.1: HandyCast is a scene-agnostic input technique for bimanual and full-range control in expansive virtual environments under physical space constraints, using touch and motion of a hand-held smartphone. HandyCast enables users to (a) operate rich Virtual Reality environments as in Job Simulator with two virtual hands (b) by sensing embodied interaction from as little physical space as a car's passenger seat as users move, turn, and touch the phone like a controller from the comfort of their lap. (c) HandyCast implements a *pose-and-touch transfer function* that individually fuses and amplifies phone and touch motions into position, rotation, and selection parameters for either virtual hand. Building on visual-inertial odometry to retrieve the 6D phone pose, HandyCast brings the 3D input control common in stationary setups with two hand-held controllers to *mobile settings*, using just a headset and a smartphone anywhere without the need for external tracking.

Despite the potential of Virtual Reality as the next computing platform for general purposes, current systems are tailored to stationary settings to support expansive interaction in mid-air. However, in mobile scenarios, the physical constraints of the space surrounding the user may be prohibitively small for spatial interaction in VR with classical controllers. In this chapter, we present HandyCast, a smartphone-based input technique that enables full-range 3D input with two virtual hands in VR while requiring little physical space, allowing users to operate large virtual environments in mobile settings. HandyCast defines a pose-and-touch transfer function that fuses the phone's position and orientation with touch input to derive two individual 3D hand positions. Holding their phone like a gamepad, users can thus move and turn it to independently control their virtual hands. Touch input using the thumbs fine-tunes

the respective virtual hand position and controls object selection. We evaluated HandyCast in three studies, comparing its performance with that of Go-Go, a classic bimanual controller technique. In our open-space study, participants required significantly less physical motion using HandyCast with no decrease in completion time or body ownership. In our space-constrained study, participants achieved significantly faster completion times, smaller interaction volumes, and shorter path lengths with HandyCast compared to Go-Go. In our technical evaluation, HandyCast's fully standalone inside-out 6D tracking performance again incurred no decrease in completion time compared to an outside-in tracking baseline.

6.1 Introduction

The sinking cost of head-mounted displays is leading to increased adoption of Virtual Reality (VR) with consumers. Experiences range from gaming and fitness to entertainment and virtual sightseeing [255]. Several productivity applications have also emerged for VR, such as 3D modeling and sketching [11].

The uptake of recent VR experiences can also be attributed to the fact that these systems integrate tracking, computation, and interaction all inside just the headset and two hand-held controllers. This makes them portable and suitable outside of controlled home and office environments, providing experiences during parts of the day that offer less entertainment, such as travel and commute.

However, VR scenarios are typically designed for standing interaction in large environments—virtual as well as physical—utilizing the obstacle-free space around the user for input. Thus, while the form factor of VR systems themselves—headsets and controllers—supports mobile operation, not all mobile scenarios support VR use, especially those that constrain the user's space. Some of these may be particularly interesting for VR use, such as in the passenger seat of a car or while traveling on a bus, train, or plane, or simply while waiting in a public space. Such situations offer plenty of time to enjoy virtual experiences, often when seated, yet little space to interact inside them and to perform the required physical motions.

In this chapter, we introduce an interaction technique that retains the immersive spatial 3D interaction around the user while minimizing the demands on physical space. We replace the two VR controllers with a ubiquitous substitute—the personal smartphone—and present *HandyCast*, a smartphone-based input technique to control both virtual hands. HandyCast supports quick and full-range interaction with two-hand control in expansive virtual environments through the inertial and optical sensors inside a phone when held like a gamepad. We disambiguate control over the individual virtual left and right hand from complementary thumb-based touch input.

A single smartphone to control two hands in VR

Figure 6.1a shows a user playing Job Simulator [262], wearing a headset and interacting through his smartphone using HandyCast. While the right virtual hand is interacting far away and the left is operating at medium distance, (b) the user is physically sitting in the passenger seat of a car, controlling both virtual hands in mid-air while resting his physical arms in his lap. HandyCast redirects all motion and touch input on the phone through its (c) *pose-and-touch transfer function* that computes the position, rotation, and manipulation parameters of either virtual hand inside the virtual environment. The user can reach close-by, medium, and distant

objects, as HandyCast amplifies phone motions: We smoothly and instantly translate small phone rotations and movements to larger rotations and movements of both virtual hands.

While controlling both hands with only a single controller removes the independence between the two hands, HandyCast accepts touch input from the left and right thumb as a complement to adjust the positions of the virtual hands. This allows users to embody input through *simultaneous* phone movements and touch motions, resulting in *positional bimanuality* where users can individually control each hand's position. HandyCast thereby builds on users' propensity to unwittingly move physical controllers in video games and, thus, additionally embody their intention through body motions, even though such controller motion yields no effect. HandyCast also leverages users' decade-long experience with touchscreens by evaluating their fine-grained touch motions for accurate 3D cursor control in VR.

Though different from bimanual manipulation with simultaneous rotational and positional control for both hands, HandyCast's positional bimanuality affords users the wide variety of interactions required to operate immersive environments as shown in Figure 6.2.

HandyCast comprises a SteamVR driver that substitutes hand-held controller input with the output of HandyCast's pose-and-touch transfer function. This allows spatially operating any VR application through 3D interaction for selection, placement, and tracing, such as pulling levers, operating sliders, or opening doors—all without the need for external tracking hardware or additional VR controllers. Our video demonstrates these in detail.

We first detail the design rationale of our HandyCast technique and present its three-fold evaluation: 1) Our first user study *in a full-range, seated setup* with 12 participants compared HandyCast, the two-controller baseline Go-Go [269], and the smartphone-based 3D cursor technique Tiltcasting [267] in a unimanual and bimanual task. An HTC VIVE system tracked all spatial input to focus our analysis on the technique-specific differences in task completion. We found that participants completed tasks using HandyCast with no significant difference in completion time compared to the two-controller Go-Go technique, but that HandyCast required significantly less physical travel and control space than the two baselines. Participants reported comparable levels of body ownership. 2) Our second user study *in a space-constrained, seated setup* with 20 participants compared HandyCast with an adapted Space-Constrained Go-Go that uses a level of amplification commensurate to fit the constrained space. We found that HandyCast allowed participants to significantly faster select and place objects than Space-Constrained Go-Go while still requiring significantly less space and mid-air travel. 3) Lastly, our *tracking evaluation* measured the impact of the tracking system on task performance. Four new participants completed this study using HandyCast and repeating the input tasks under outside-in (HTC VIVE) 6D tracking (as used in Studies 1 and 2) and inside-out (phone-only) 6D tracking conditions.

Contributions

With the work in this chapter, we contribute

- HandyCast, a *pose-and-touch transfer function* that we designed for individual control over two virtual hands through input on a single smartphone in space-constrained settings, allowing significantly faster object selection and placement in space-constrained settings than the two-controller Go-Go technique [269],
- a first *interaction user study in a open-space seated setup* to assess completion time, motion

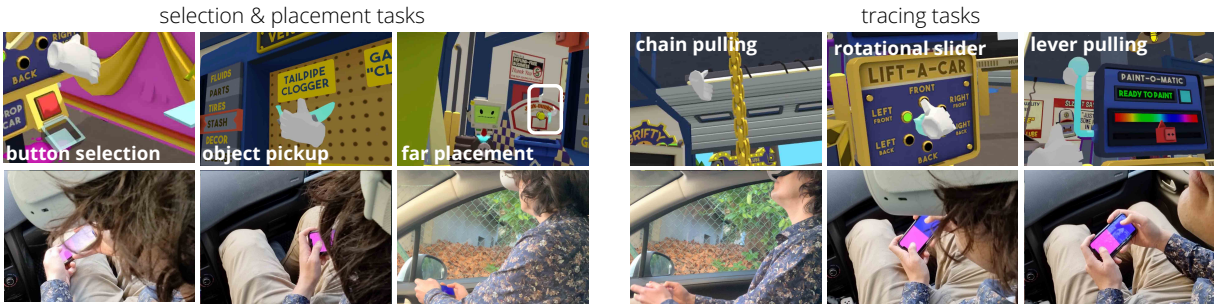


Figure 6.2: Using a single smartphone, HandyCast enables control over bimanual input tasks such as selection, placement, and tracing as found in common VR applications and games. Here we show representative tasks from Job Simulator [262].

paths, and control space of HandyCast compared to a controller-based and a smartphone-based baseline under technique-optimal conditions,

- a second *interaction user study in a space-constrained seated setup* to assess completion time, motion paths, and control space of HandyCast compared to a space-constrained implementation of the two-controller Go-Go baseline,
- a *tracking evaluation* that investigated the drift incurred by HandyCast's inside-out 6D tracking (phone-only) compared to ideal conditions using outside-in tracking (HTC VIVE),
- a SteamVR *controller driver* that brings HandyCast's scene-agnostic quick and reliable bimanual interaction for operating large VR scenarios inside mobile and possibly space-constrained settings.

6.2 Background and Related Work

6.2.1 Interaction Metaphors in VR

Manipulation in 3D user interfaces generally comprises selecting, positioning, rotating, and scaling objects [193; 231; 120; 109; 346]. Several decades of research have brought forward a multitude of input techniques for each of these tasks [269; 35] as well as taxonomies [270; 26; 224; 10] and overviews of design parameters [132; 193]. VR UIs are part of a subset of 3D interaction techniques where the user's body is colocated with the virtual environment.

Two main metaphors exist for manipulating objects in VR [270]. Using the *virtual pointer* metaphor, users can select distant objects by pointing at them [270]. Ray casting faces depth ambiguity, which can be challenging in scenes with densely populated and occluded targets, in turn leading to problem-specific variations [198; 109]. In contrast, the *virtual hand* metaphor [270] provides the user with one, two, or more [291] hand representations that mimic physical hand movement to manipulate objects. In its simplest form, physical and virtual hands are colocated. However, since this rigid coupling limits the reach in VR, researchers have explored techniques to amplify hand positions. For example, Go-Go applies non-linear amplification of hand positions to extend the virtual reach far beyond arm's length [269]. Reach-bounded non-linear (RNL) amplification improves ergonomics while maintaining body ownership [349], exploiting dominance of vision over proprioception [316; 15; 57; 244; 86]. At the core of these amplification techniques are transfer functions that map the position of the physical hands to the virtual

space. HandyCast builds on these amplification techniques for pose transfer to provide initial hand positions before we then apply touch transfer to obtain the final hand positions.

6.2.2 Metaphor Combinations and Extensions

Both metaphors are combined in HOMER [35], which uses a pointer for selection and a hand metaphor for positioning and rotating. Today's commercial VR games typically use hand avatars to power the main gameplay except for targeting tasks, such as shooting. Ray casting is often used for menus (e.g., Resident Evil 4, Warplanes: WW1 Fighters, Half Life: Alyx), except for fully hand-centric games [262] that use virtual hands for menus, too. HandyCast follows the virtual hand metaphor, as pose transfer allows rapid selection of even distant targets, sharing some of the benefits of rays.

Recent research has also studied virtual hand manipulation in VR under specific conditions. Hayatpur et al. investigated how gestural input can be used to specify shape constraints for object manipulation in VR [126]. Yamagami et al. demonstrated how unimanual input can be mapped to bimanual interactions for users who have full use of only one hand [371]. HandyCast builds on this to use a single smartphone for bimanual interaction in VR, especially when augmented with touch input for additional freedom.

Besides controllers, recent projects have leveraged hand pose estimation to define novel transfer functions for object interaction. Force Push detects gestures as input to translate virtual objects [378]. By using a smartphone rather than mid-air hand tracking [119] for input, HandyCast *circumvents classical technical limits* such as noisy estimates, the need for exaggerated gestures irreconcilable with the goal of subtle input, and tracking losses at the edge of the field of view. Additionally, HandyCast *enables novel interaction capabilities* by supporting thumb input, either for refinement or amplification of virtual hand motions. We also incorporate phone-specific affordances such as haptic buttons (e.g., volume buttons) and precise touch gestures.

6.2.3 Smartphone-Based Controllers

Researchers have often utilized smartphones and their sensors for mediating input to remote screens (e.g., [174; 172; 173; 33; 106; 18; 34]) or interaction with spatial AR projections [125; 123]. For VR, researchers have also investigated substituting traditional controllers with phones (e.g., [358], Phonetroller [218], Handymenu [204]) or tablets [317; 74]. Dias et al. proposed a technique to point at objects using headset gaze and using touch on the phone for selection [74], such as for menu interaction in VR. Chen et al. presented two techniques to use smartphone touch in AR with a cursor with 2 degrees of freedom, placed on rigid walls [54]. TMMD maps the pose of an externally tracked phone isomorphically into the virtual space [358], allowing users to select objects either through ray casting or by walking up to the objects and using touch input to attach to them. Touch gestures on the phone then allow translating and rotating the object. HandyCast builds on this notion of interaction, but differs in three regards as it is 1) designed for seated and low-effort interaction, thus using amplification, 2) optimized for simultaneous control over *both* virtual hands, and 3) aimed at maintaining embodiment following the virtual hand metaphor.

In our design, we also built on previous work on remotely controlling cursors on TV screens. Pivot Plane-Casting defines a plane from the phone's orientation, anchored rigidly at the center of the virtual space [175; 176]. Rotating the phone rotates the plane and touching the screen

allows attaching an object intersected by the plane before then translating it. Free Plane-Casting also casts a plane, however anchors the plane in a movable cursor position [175]. Touch input moves the cursor on that plane, thus translating the plane anchor itself and possibly an object along with it if attached. In Free Plane-Casting, phone rotations without concurrent touch have no effect on the cursor position as only the virtual plane orientation changes but not the cursor it is anchored in. INSPECT further extends plane-casting with a rotation mode that switches touch input from translating an attached object on the plane to rotating it [176]. Tiltcasting is related, but accounts for occluded objects [267]. By tilting the phone a plane shown on the screen is tilted correspondingly. Objects in front of the plane vanish and only objects intersected by the plane are selectable by using a cursor controllable by touch. HandyCast reuses the concept of a plane based on phone orientation introduced in Plane-Casting [175], but extends it in several regards: 1) In contrast to previous research, HandyCast is fully agnostic of the application state and does not require a dedicated integration for each VR application. This means, HandyCast works with any existing VR application that can be operated with two virtual hand avatars. 2) HandyCast transfers 10 dimensions (3D phone position, 3D phone orientation, and 2×2D touch), in contrast to 2D touch and 1D rotation as in Tiltcasting or 3D rotation and 2D touch as in Pivot Plane-Casting and Free Plane-Casting. On the one hand, this means users can freely position and orient the plane and thus the virtual hands, following our hypothesis that this increases embodiment in VR. On the other hand, this enables bimanual selection and placement. In our comparison of HandyCast and a version of Tiltcasting adapted for bimanuality as one baseline, free plane placement with motion amplification led to faster task completion.

Pocket6 also uses the phone to select and manipulate 3D objects on a TV screen [18], using the phone to position a scaled 3D cursor and touch input to rotate an attached object (similar to INSPECT [176]). The ASP technique by Bergé et al. follows a similar approach while using external tracking [25]. HandyCast reuses the concept of position tracking as an input into our motion transfer for our hand avatar, but map multitouch input for further spatial hand control, whereas Pocket6 uses single touch for scene rotation and confirmation.

Taken together, HandyCast extends previous concepts, particularly a movable plane anchor as in Free Plane-Casting [175], but is first to use positional input, positional amplification, and up to two touch contacts to drive the 3D positions of two virtual hands in a 3D environment, co-located with the user in VR. This colocation with the user's body further requires spatial registration that depends on the user's seated position (rather than on the scene's or the cursor's center). Additionally, HandyCast targets low-effort input for control without looking at the smartphone itself.

6.3 Pose-and-Touch Transfer for Smartphone-Based Control

6.3.1 Problem and Solution Overview

HandyCast addresses the problem of space-efficient and bimanual object selection and manipulation in VR using a smartphone for input. Fundamentally, HandyCast derives 2×3D virtual hand positions and grab states from fusing the smartphone's 6D pose (i.e., 3D position and 3D orientation) with up to 2×2D touch locations when the phone is held like a gamepad (Figure 6.1b).

The mapping from phone pose+touch to two 3D hand positions in VR and their grab states is defined through our *pose-and-touch transfer function*, combining pose transfer, spatial touch

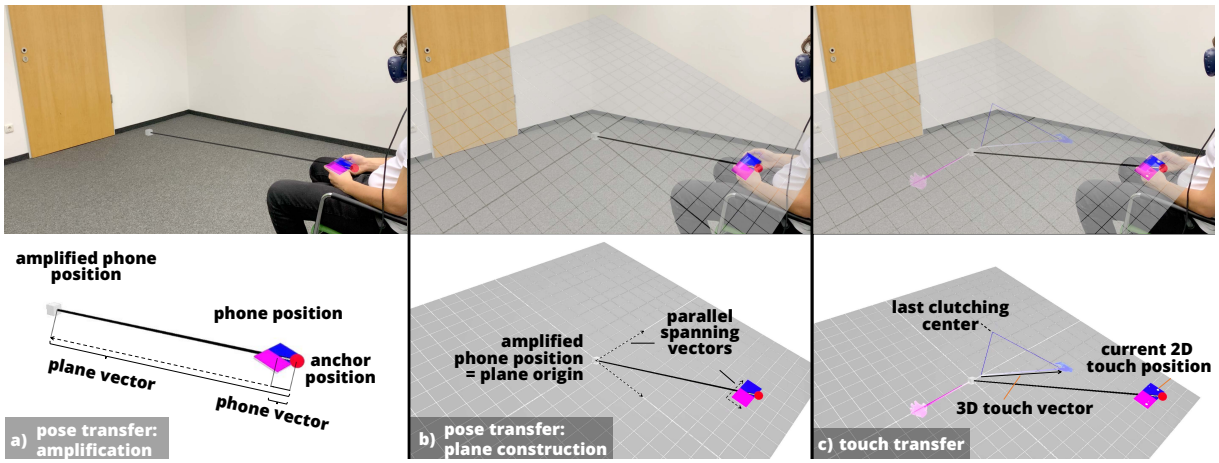


Figure 6.3: Schematic 3D representation of our *HandyCast* transfer function. For pose transfer, HandyCast spans a plane in 3D space, originating in the distance-amplified smartphone position. For touch transfer, the hands are moved from the plane origin (adjusted by a constant shoulder-width offset) along the gain-accelerated touch vector. Plane and vectors shown for illustration only.

transfer, and touch presence transfer. In pose transfer, we first compute a 3D plane, rotated according to the phone's rotation, and anchored in the amplified phone position. The amplified phone position is computed by non-linearly scaling the vector that leads from a calibrated neutral position to the current phone position. In spatial touch transfer, we then position two cursors on that plane based on the relative touch cursor input given to each touch zone on the phone's screen. Finally, we transfer touch presence for each touch zone, i.e., the binary state whether touch is applied or not, to the corresponding hand avatar's grab state. These three components can be computed and integrated at a single point in time, thus allowing simultaneous usage of phone and touch motion, individually for both hands.

6.3.2 Transfer Functions to Map Phone Input to Two Hand Avatars in VR

In the following, we give further details on pose transfer, spatial touch transfer, and touch presence transfer.

6.3.2.1 Pose Transfer

As shown in Figure 6.3, we constantly derive a plane in 3D space from the 6D smartphone pose, linking the real world to the virtual world in three steps.

1) Neutral pose anchoring. The user defines a neutral 3D phone position by pressing the volume-down button of the phone, e.g., when hands are recumbent in their lap. This calibrates a fixed spatial anchor $\mathbf{p}_{\text{anchor}}$ in world space (red sphere, 6.3b).

2) Phone tracking and phone vector computation. Moving the phone in the constrained space above the lap results in the phone vector $\mathbf{v}_{\text{phone}}$, leading from the calibrated anchor $\mathbf{p}_{\text{anchor}}$ to the phone's position. The physical movement is small (very short, white vector 6.3b). In HandyCast, phone motion can be tracked either through an outside-in (i.e., external) system, tracking a phone-mounted tracker with base stations, or through inside-out (i.e. standalone phone-only) tracking using the phone's sensors. To unify the phone's inside-out and the headset's coordinate system, we specify a registration procedure, as detailed in the supplementary material. By

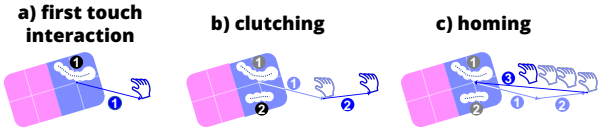


Figure 6.4: In 2D input space, we implement two gain-accelerated cursors that enable clutching and automatically perform homing.

simply holding the phone in front of the headset and then pressing the volume-up button, a registration transform is created, used to convert from phone to headset coordinates.

3) Vector amplification allows the small phone vector to contribute to larger motions. We obtain the amplification vector $\mathbf{v}_{\text{plane}}$ by non-linearly scaling the phone vector $\mathbf{v}_{\text{phone}}$ as

$$\mathbf{v}_{\text{plane}} = \lambda \|\mathbf{v}_{\text{phone}}\| \, \mathbf{v}_{\text{phone}}.$$

Starting in the calibrated anchor $\mathbf{p}_{\text{anchor}}$ and following the amplification vector $\mathbf{v}_{\text{plane}}$ leads to the amplified phone position $\mathbf{p}_{\text{amplified}}$ (black vector in Figure 6.3b), serving as the plane's origin. λ is a constant, that can be set dependent on the desired strength of the amplification effect. In our studies, we choose $\lambda = 2.5$. From the phone orientation, we can easily obtain the phone's horizontal and vertical directional vectors (Figure 6.3c). Even without any touch applied, we offset the right hand position along the horizontal direction of the plane, and the left hand along the inverted horizontal direction, so to position both hands at shoulder width.

6.3.2.2 Spatial Touch Transfer

For spatial touch transfer, we divide the touchscreen into two separate control zones, allowing the thumb of either hand to independently control the respective hand avatar through touch transfer in five steps.

1) Initial touch transfer is zero. When no touch is present, virtual hand avatar position is determined through pose transfer only.

2) Gain-accelerated touch offset. We implemented a classical relative 2D cursor function computing the next 2D cursor offset from the current cursor offset and the current touch motion with gain acceleration following prior work [219; 48; 246]. Gain acceleration affords fine-tuning as well as longer-distance adjustments of hand avatars. Encroaching touch paths that cross the boundary between both touch zones, are maintained disruption-free for the thumb of the zone in which they originated. We implemented it as a finite state machine, instantiated once for each thumb, to compute each offset.

3) Transferring 2D touch offsets to 3D hand adjustments, we map the current 2D cursor offset to 3D touch vectors by rotating the offset vector according to the phone orientation.

Figure 6.3d shows an example of the resulting 3D touch vector.

4) Optional clutching supports continued touch transfer, which extends the virtual position, reachable by touch. When users need to move hand avatars beyond the constrained touchscreen input surface, they can lift off the thumb, and reposition it while the virtual hand stays in place. In our implementation of touch transfer, we allow for clutch timeout (set to 0.2s), i.e., a brief period of time within which a user may touch down again somewhere else on the screen and continue from the previous hand position. This updates the reference center for the next touch motion without affecting the already accumulated offset.

5) Automatic homing finally compensates for potential clutching. To spare the user the manual effort of resetting hand avatars back to the initial position, the accumulated offset returns to 0

following a homing timeout (set to 2.5s). We avoid a sudden jump of hand avatars by lerping avatars ($t = 0.01, \Delta = 15$ms; ≈51% per second) back to their neutral position, defined by pose transfer only.

6.3.2.3 Touch Presence Transfer for Commands

While VR controllers feature a trigger button to mimic grasping, smartphones offer no equivalent counterpart. Thus, we integrated grasp and release as part of touch contact itself—following the metaphor of direct touch interaction. If a touch-up event is not followed by another touch-down event in the respective control zone (i.e., either left or right) within the clutching delay, HandyCast triggers a release event.

6.3.3 Design Decisions

6.3.3.1 Plane Origin

HandyCast defines a plane in 3D space that is anchored in a position derived by amplifying the phone vector. This allows users to anchor the plane wherever optimal for the current bimanual task. In contrast, Tiltcasting [267] and Pivot Plane-Casting [175] anchor the plane origin in a fixed point of the 3D scene, so that the users can only touch and turn to move the cursor. Free Plane-Casting [175] makes no use of the phone position either and anchors the plane origin at the touch-controlled cursor unlike HandyCast's use of a point derived from the phone position.

TMMD [358] is the only technique where the user's body and the virtual scene are colocated, anchoring the plane origin with the phone. However, this still limits touch to object selection or requires physically walking up to the object, which is why TMMD supports raycasting. In contrast, HandyCast anchors the plane at a dynamic location computed through positional amplification, which can be controlled with little effort.

6.3.3.2 Pose Transfer Amplification

Go-Go uses isomorphic mapping when close to the body and non-linearly amplification beyond a threshold [269]. To maximize space efficiency, we design for immediate non-linear amplification, using touch transfer to allow for linear corrections at all times—even at distant locations.

6.3.3.3 Amplification Anchoring

HandyCast bases pose transfer on real-world anchor positions. Using the head or a derived point (e.g., the chest) as a moving anchor similar to previous techniques leads to unintended hand avatar motions, particularly with large amplification. Thus, we initially create an anchor to world coordinates, defined as the user comfortably holds their phone in their lap.

6.3.3.4 Relative Touch Cursor Mapping

HandyCast maintains a relative cursor for each touch zone where the touch-down location defines the origin of the touch interaction, rather than absolutely mapping the touch-down position to a

coordinate. We chose this design as users have no visual control over their touch-down location on the screen when operating in VR.

6.3.3.5 Touch Transfer Acceleration

Touch transfer integrates gain acceleration for two reasons. 1) The constrained size of the screen requires touch motions to be small. Translating touch to medium and far distances would require high gains, but this contradicts our design goals to allow fine-tuning pose-transferred locations, which requires low gains. Gain acceleration offers both low and high gains, based on touch velocity. 2) Previous research has found gain acceleration advantageous over constant gain [48].

6.3.3.6 Clutching

Without clutching, lifting the thumb would reset the touch offset, thus instantly moving back the hand avatar to the pose-transferred location. Such liftoffs may even be unintentional, such as when sliding the thumb towards the screen edges.

6.3.3.7 Target-Agnostic Operation

HandyCast receives application-independent events as input, allowing our technique to operate on a standalone controller (i.e., phone), oblivious to the state of the virtual scene. That is, it operates without knowledge of target presence or proximity. While such information could improve the technique (e.g., whether the hand is hovering over an object), it would limit general applicability for real-world apps.

6.4 Implementation

For the study setup, we implemented our transfer function in a VR evaluation environment, featuring objects, placement zones, and task protocols in Unity with the SteamVR framework and the Unity XR Interaction Toolkit. For input, we implemented an iOS 15 app. We smooth hand avatar positions using a 1€ filter [47]. The virtual environment supports user input from controllers and our phone app. The phone connects to the Unity server via TCP. In coordinates of the virtual environment, the phone pose can be determined from either the 6-DoF outside-in tracking by the VIVE system, or by the 6-DoF inside-out tracking provided by the phone only and also sent to the Unity server via TCP. The Unity app logs all transform updates of the HMD, the VIVE Tracker (Figure 6.5), the VIVE Pro Controllers, both hand avatars and all incoming pose and touch events from the smartphone. We also log all trial-related events with a variety of event parameters (e.g., positions, states, timestamps). The Unity app can also generate debug views, illustrating pose-and-touch transfer with planes, vectors, and ellipses in real-time, and shown throughout this chaper. Separately and independently of our study environment, we developed a low-level controller driver for SteamVR, thus allowing backward-compatible drop-in use with standard VR apps on a Quest 2 with AirLink. Please refer to the supplementary material for a more detailed description of the controller implementation.

Figure 6.5: Users hold the HandyCast input phone like a gamepad, with each thumb operating in a single touch zone. For usage in common VR games, HandyCast can use inside-out tracking with no additional tracking hardware needed. For our two evaluations, a VIVE tracker is be mounted to the phone allowing us to separately investigate interaction performance and tracking effects.

Inside-Out Tracking for Mobile Use

We integrated spatial tracking into HandyCast, such that our technique affords mobile use and needs not rely on an external tracking system. We build on ARKit [5] to track the phone inside-out, which performs visual-inertial odometry using the phone's rear-facing camera and the IMUs.

6.5 User Study 1: All Techniques under Technique-Optimal Space Setup

The first study compared participants' performance in a series of target acquisition and placement tasks during unimanual and bimanual use in a within-subjects design. Our goal was to analyze the effect of technique on task completion time, physical space requirements, and trajectories. Participants also rated techniques on perceived exertion, perceived workload, and body ownership.

6.5.1 Input Techniques

Participants completed all conditions using our *HandyCast* technique as well as two baseline techniques.[1] As a controller baseline, we included the state-of-the-art *Controller-based Go-Go* technique [269], where participants used *two* controllers and Go-Go's non-linear amplification for the tasks.

To maintain spatial comparability between subjects for the statistical analysis, we hold amplification constant across participants. Go-Go's amplification parameters D and k [269] specify the trade-off between 1) control volume, 2) accuracy, and 3) embodiment at distant locations. For this first user study, we set $D = 40\text{cm}$, and the amplification factor to $k = 0.5$, 1) ensuring that users can very comfortably reach all target selection and placement locations at short arm length, 2) with practical accuracy in the distance, 3) while maintaining the clear association between physical and virtual hand in an initially linear mapping, so to use Go-Go as a reference

[1] In this study, we had included a second custom technique as a further condition, which turned out inferior to *HandyCast* and thus is not further reported on in this section. Note however that it was included in the statistical analysis to follow. Please refer to the supplementary materials for a technique description as well as the statistical analysis, also comparing it to HandyCast and the two baselines.

for embodiment across distant interaction. For *HandyCast* we specify $\lambda = 2.5$ (e.g., a forward motion of 15 cm is amplified to 5.56 m). Please refer to the accompanying video for a visual impression on the control volume.

Choosing a smartphone baseline requires more subtle consideration. The problem of two-hand input with a smartphone under space constraints in VR is unexplored. Thus, no readily applicable baseline for comparison with our novel technique exists. Instead, we have to adapt candidates from literature for bimanual usage, either based on a fixed-anchor technique such as Tiltcasting [267] and Pivot Plane-Casting [175] or a moveable-anchor technique such as Free Plane-Casting [175]. We chose a fixed-anchor technique over Free Plane-Casting for two reasons: 1) Pivot Plane-Casting as fixed-anchor technique is reported in the original paper to be faster than Free Plane-Casting. 2) Free Plane-Casting is ambiguous to adapt for two-hand input, given the need for two instead of one cursors, inducing interesting design questions. Should such a bimanual version of Free Plane-Casting anchor its freely moveable plane in a midpoint between both hands, making it most similar to our HandyCast technique? Or should this adaption maintain two fully independent planes? While any such design would be interesting to explore, it creates a novel method diffcult to consider an objective baseline.

In the space of fixed-anchor techniques, we choose Tiltcasting as a baseline, as we seek a baseline to provide fast and embodied input, both of which promised in the absolute mapping of Tiltcasting: 1) It promises fast input as touch-down positions define the target position directly reducing time needed for touch distances. 2) It promises embodied input as the neutral position is immediately assumed upon touch-up, not requiring thoughtful user input to go back to neutral.

Having chosen Tiltcasting, we extended it to *Bimanual Tiltcasting*, which anchors the plane fixed in space, and, therefore, uses a high control-display gain ratio for touch transfer. This allows reaching all objects in the scene. As described, we keep an absolute cursor for touch input. To enable bimanuality, we position both hand avatars on the same Tiltcasting plane with a small distance in between.

6.5.2 Apparatus: Seated, Input Tracked Outside-In

To remove the impact of tracking performance from this study, we used the HTC VIVE external tracking system for all techniques. (We separately evaluate the inside-out tracking of HandyCast and its effect on task completion in a tracking study.)

As shown in Figure 6.6, participants wore a VIVE Pro Eye for all tasks, sitting on a fixed chair. For *Controller-based Go-Go*, participants used both VIVE Pro controllers for input. For the smartphone-based *HandyCast* and *Bimanual Tiltcasting*, we mounted a VIVE tracker to an iPhone 11 Pro, which forwarded all touch events registered on the phone to a PC. Participants do not see the debug view (showing planes or vectors) before or during the study but only see the hand avatars as such when training and using our technique.

6.5.3 Task

Participants completed two tasks in the study. For both, they were instructed to complete them as fast as possible.

Figure 6.6: Apparatus for our full-range study. a) Our Unity app rendered the study environment, displayed through a VIVE Pro Eye. b) For *Go-Go* input, participants used the VIVE controllers. c) For all phone techniques, a VIVE tracker provided the pose of an iPhone 11 Pro, which relayed touches to Unity.

Figure 6.7: Participants completed unimanual and bimanual tasks, a) selecting from a $H_3 \times W_5 \times D_3$ grid and placing at a different grid location (every second target and placement shown for clarity.) Targets and placement zones were equally distributed across trials. b) Acquisition during Task 1 (unimanual), c) first placement of Task 2 (bimanual). The left hand is colored in magenta, the right in blue. The target and placement zones are color-coded correspondingly when highlighted.

Task 1: Unimanual Object Selection and Placement During each trial, participants were instructed to grab a highlighted object in the virtual scene with the specified hand. After grabbing it, they moved it to the indicated placement zone. Releasing the object within the zone completed the trial and advanced to the next. When participants erroneously dropped the object, they could grab and move it again until they succeeded.

To acquire a target, participants either pressed the trigger on the controller or touched and held down on the smartphone. Releasing the trigger or the touch dropped the virtual object.

An object counted as correctly placed if its center was within the placement zone, moved with the instructed hand. Participants received visual cues as soon as a release would be correct (Figure 6.7c). An acoustic cue then confirmed a successful release.

Task 2: Bimanual Selection and Placement Participants grabbed two targets during each trial, one after the other, and then placed them in the corresponding placement zones. Consistent coloring indicated which target to grab with which hand (left: pink, right: blue) and where to place it. Our video figure shows examples of bimanual trials, each of which followed one of the following combinations: $\{\text{grab}_L, \text{grab}_R, \text{drop}_L, \text{drop}_R\}$, or $\{\text{g}_R, \text{g}_L, \text{d}_R, \text{d}_L\}$.

In both tasks, for each trial, participants had 15 seconds to acquire a target and 15 seconds to place it. The remaining time was indicated by a countdown timer. If participants ran out of time, the trial counted as an error.

Target Arrangement Figure 6.7 shows the grid arrangement of locations for targets, which were shown one and two at a time for Task 1 and 2, respectively. Placement zones were at the same grid locations, one or two highlighted depending on the task.

The location of targets built on previous studies (e.g., Erg-O [244] and other VR input techniques [349; 86]), but added placement zones, distant object interaction, and bimanual interaction. In this study, the target size is set to 25 cm. The final grid (height = 3 × width = 5 × depth = 3) contained cells of 1.6 m ×1.6 m ×3 m with a 0.4 m spacing in 2D and 1 m spaces in depth.

Technique Rating After completing all trials for a technique, participants filled out a short questionnaire in VR: 2 Borg CR-10 ratings (exertion in hands/lower arms, exertion in upper arms) from 0 (nothing at all) to 10 (extremely strong), 4 TLX subscales (mental demand, physical demand, effort, frustration) from 1 (very low) to 20 (very high), and 4 avatar embodiment questions (ownership, double hand, control, interference) [105] on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree).

6.5.4 Procedure and Design

Before the evaluation phase, participants received training with all techniques in both tasks for approx. 10 minutes. After finishing a technique in a task, they answered the questionnaire. Before starting the measured study trials for the next technique, they were allowed at least two more training trials to remember it. The evaluation part took approximately 40 min per participant.

Independent Variables

The study followed a within-subjects design with two independent variables: Operation and Technique. Operation had two levels: *Acquisition* and *Placement*. Technique is considered with three levels: *HandyCast*, *Controller-based Go-Go*, *Bimanual Tiltcasting*. Technique order was counterbalanced across participants and both tasks with a Latin square.

For each task, we precomputed randomly selected target and placement locations, which were counterbalanced across participants. Even though we did not explicitly analyze them as independent variables, we ensured equal distribution of Hand (Task 1, *left* or *right*) or Hand Combination (Task 2, start with grab_L or grab_R, see above), and Depth delta with three levels (0: same layer, +1: pushing back, −1: bringing forward).

Trial Repetitions

For Task 1, participants repeated acquisition and placement 7 × for 2 Hand × 3 Depth delta × 4 Techniques × 2 Operations = 336 trials per participant. For Task 2, they repeated the task 4 × for 2 Hand × 3 Depth delta × 4 Techniques × 2 bimanual Operations = 192 trials per participant.

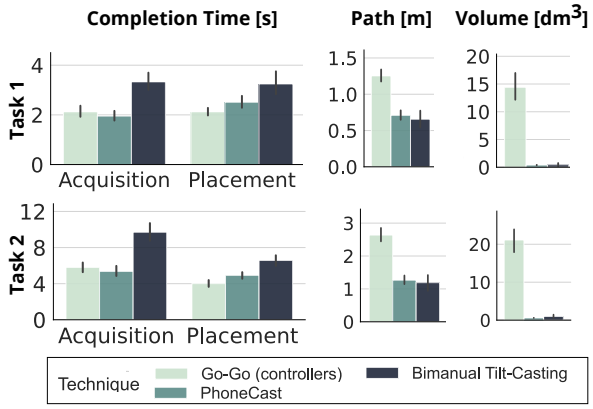


Figure 6.8: Aggregated results of our study by task and operation, and the dependent variables completion time, length of input paths, and physical control space volume. Error bars indicate 95% confidence intervals.

Dependent Variables

To analyze differences between TECHNIQUES and OPERATIONS, we logged task completion times (i.e., selection and placement time), as well as physical and virtual motion paths. For *Controller-based Go-Go* in Task 2, we measured the sum of both controllers' motions; for the smartphone techniques, we doubled the length of paths traveled to account for movement of both real hands holding the smartphone. From the recorded motion paths, we derived the required volume for operation (i.e., control space) from the enclosing world-oriented cuboid, based on the 5th to 95th percentile of points per axis.

6.5.5 Participants

We recruited 12 participants (2 female, 10 male, ages=23–35, M=28.4, SD=4.1). 4 participants had never worn a VR headset before, 5 used one less than 5 times, 2 occasionally, and 1 on a weekly basis.

6.5.6 Results

6.5.6.1 Completion Time

We performed a two-way repeated-measures ANOVA on task completion time for TECHNIQUE × OPERATION with participant as the random variable.

In Task 1, participants completed acquisition on average in $2.49\,s$ ($\sigma = 0.72$) and placement in $2.74\,s$ ($\sigma = 0.73$). We found a significant main effect of TECHNIQUE on time ($F_{3,33} = 27.331, p < .001, \eta^2 = .713$) as well as of OPERATION on time ($F_{1,11} = 30.592, p < .001, \eta^2 = .736$). We also found an interaction between both variables ($F_{3,33} = 11.472, p < .001, \eta^2 = .510$). Post-hoc t-tests using Bonferroni-adjusted confidence intervals on TECHNIQUE showed significant differences between *Controller-based Go-Go* and *Bimanual Tiltcasting* as well as between *HandyCast* and *Bimanual Tiltcasting*, but not between *Controller-based Go-Go* and *HandyCast*. As shown in Figure 6.8, *Controller-based Go-Go* was 1.2 s faster than *Bimanual Tiltcasting* on average (154.8%, $p < .001$), whereas *HandyCast* was 1.1 s faster than *Bimanual Tiltcasting* (147.4%, $p < .001$). We found no significant difference between *Controller-based Go-Go* and *HandyCast*.

In Task 2, participants completed acquisition (accumulated for both objects) on average in 6.91 s ($\sigma = 2.16$) and placement in 5.43 s ($\sigma = 1.35$). We found a significant main effect of TECHNIQUE on

time ($F_{3,33} = 60.028, p < .001, \eta^2 = .845$) and of Operation on time ($F_{1,11} = 111.130, p < .001, \eta^2 = .910$). We also found an interaction effect between both ($F_{3,33} = 38.793, p < .001, \eta^2 = .779$). Post-hoc *t*-tests using Bonferroni-adjusted confidence intervals showed a significant difference between all comparisons involving *Bimanual Tiltcasting*. We found no significant difference between *Controller-based Go-Go* and *HandyCast*.

6.5.6.2 Travel Length of Physical Motions

For Task 1, a one-way repeated-measures ANOVA on traveled motion path for Technique showed a significant main effect ($F_{3,33} = 79.053, p < .001, \eta^2 = .878$). Post-hoc *t*-tests using Bonferroni-adjusted confidence intervals do reveal significant differences between *Controller-based Go-Go* and both of *HandyCast* and *Bimanual Tiltcasting*, but not between *HandyCast* and *Bimanual Tiltcasting*. As shown in Figure 6.8b, the average distance traveled using *HandyCast* was 0.544 m shorter than *Controller-based Go-Go* (56.6%, $p < .001$). Using *Bimanual Tiltcasting*, the average distance was 0.597 m shorter than *Controller-based Go-Go* (52.4%, $p < .001$).

For Task 2, we also found a significant main effect ($F_{3,33} = 104.622, p < .001, \eta^2 = .905$). Post-hoc *t*-tests using Bonferroni-adjusted confidence intervals showed significant differences in the same comparisons as in Task 1.

6.5.6.3 Required Physical Volume (Control Space)

For Task 1, a one-way repeated-measures ANOVA for Technique on the required volume found a significant main effect ($F_{3,33} = 97.763, p < .001, \eta^2 = .895$). As shown in Figure 6.8c, post-hoc *t*-tests using Bonferroni-adjusted confidence intervals revealed significant differences between all techniques except between *HandyCast* and *Bimanual Tiltcasting*. *HandyCast* required 14014.1 cm^3 less control volume than *Controller-based Go-Go* (2.72%, $p < .001$).

For Task 2, we also found a significant main effect ($F_{3,33} = 123.414, p < .001, \eta^2 = .918$). Post-hoc *t*-tests using Bonferroni-adjusted confidence intervals showed significant differences between the same combination of techniques as in Task 1. *HandyCast* required 20547.9 cm^3 less control volume than *Controller-based Go-Go* (2.79%, $p < .001$).

6.5.6.4 Questionnaires

Next, we analyze the participants' questionnaire responses. Please refer to the supplementary material for a detailed break-down of answers for all questions.

For perceived exertion (Borg CR-10), pairwise Bonferroni-adjusted Wilcoxon signed-rank tests showed no significant differences between participants' ratings ($p < .1$).

In terms of perceived workload (NASA TLX), pairwise Bonferroni-adjusted Wilcoxon signed-rank tests reveal a significant difference in *effort* between *Controller-based Go-Go* and *Bimanual Tiltcasting* (7.25 ± 4.0). On average, effort in *Controller-based Go-Go* was perceived lower by 3.75 than in *Bimanual Tiltcasting* ($p = .033$).

For avatar embodiment, pairwise Bonferroni-adjusted Wilcoxon signed-rank tests showed significant differences between *Controller-based Go-Go* and *Bimanual Tiltcasting* for both *Ownership* and *Control*. Reported *Ownership* was on average 2.5 points higher in *Controller-based Go-Go*

($p = .029$), and reported *Control* was on average 2.3 points higher in *Controller-based Go-Go* ($p = .031$).

6.5.6.5 Error Rates

Participants rarely ran out of time: 3 times of all 4032 trials conducted in Task 1 (once with *Controller-based Go-Go*, twice with *Bimanual Tiltcasting*), and 6 times in Task 2 (among these: once with *Controller-based Go-Go*, and 4 times with *Bimanual Tiltcasting*).

6.5.7 Discussion

Our first evaluation revealed several interesting insights about participants' performance during interaction with objects in VR using the different techniques.

Insight 1a: Controller-based Go-Go and HandyCast achieved comparable completion times, but HandyCast requires less motion Performance of HandyCast was closest to the well-established Go-Go with no significant difference between them. However, despite the comparable completion times, HandyCast required significantly less travel on average, both for unimanual (0.54 m, 56.6% of Go-Go) and bimanual tasks (1.37 m, 48% of Go-Go). The volume enclosed by HandyCast was equivalent to a cube with edge length 8.4 cm. Go-Go's volume is equivalent to a cube with edge length 28 cm ($> 3\times$). These results are particularly interesting, since the indicated path lengths already represent the phone trajectory multiplied by two to account for the movements of both hands, holding the phone. These results support our design intentions, in particular for HandyCast's transfer function.

Insight 1b: participants do not report significant differences between Controller-based Go-Go and HandyCast Participants' questionnaire responses are similarly promising, as HandyCast showed no significant loss in *ownership* or *control* in the embodiment questionnaire ratings, compared to the individual controllers. At the same time, despite the reduction in control space, participants' ratings also showed no difference in perceived fatigue or physical demand between Go-Go and HandyCast. It was one of our hypotheses that HandyCast would reduce fatigue. Taken together, this indicates that ownership, control, and demand in our HandyCast technique are similar to input with two controllers in current VR systems, yet with the benefit of ubiquitous use, in mobile settings, and at a fraction of the required space for reliable operation.

Insight 1c: Touch presence transfer for commands makes placement slower than selection As described, earlier, we found an interaction effect between Technique and Operation on *completion time*. For Task 1, i.e., a unimanual operation of first selecting a single target and the placing it, *Controller-based Go-Go* showed no difference between acquisition and placement completion time (both $\mathcal{M} = 2.1$ s), suggesting that there is no structural difference between the two operations. However, in *HandyCast*, we observe a difference of 0.5 s between selection ($\mathcal{M} = 2.0$s) and placement ($\mathcal{M} = 2.5$s). We rationalize this with two effects. First, HandyCast is more susceptible to "loosing" objects "on the way", entailing a reacquisition. While a trigger button in the *Controller-based Go-Go* controllers can be hold tight when moving the virtual hand with the object attached from the selection location to the target location,

HandyCast employs touch for both attaching to the object and then also to move it, together with pose transfer. When users need to clutch on the touchscreen, the objects remains attached for 0.2 s before being released. If clutching in motor space takes longer and the user exceeds this timeout, the object is dropped, necessitating a reacquisition, and thus slowing down overall completion time. Second, when participants had not sufficiently internalized the 0.2 s delay, they pulled the phone back too quickly (with the object still attached), thus requiring a re-acquisition and another placement attempt for the object.

Insight 1d: Independent Go-Go controllers allow concurrent movement and thus slightly faster placement Comparing the results of Task 1—in particular the interaction effects—with Task 2 also reveals interesting aspects. Bimanual acquisition is slower than double the unimanual acquisition time, likely because 1) users need to connect the color coded outline to the color-coded hand avatar, inducing cognitive processing time, and 2) users need to coordinate which hand to move to the targets, and how to position the other hand in the mean time. *HandyCast* was faster than *Controller-based Go-Go* for acquisition (−0.4 s), but slower for placement (+0.9 s).

Beyond the reason above, the motion logs showed another reason for Go-Go's advantage. Participants often acquired targets sequentially, but often *concurrently* moved both hands to the target zones. While *HandyCast* provides the same effect during pose transfer, touch transfer is slowed down when the pose transfer moves in an opposite direction.

Insight 1e: Bimanual Tiltcasting requires the least volume and the least motion paths of all techniques, but is slower having only touch and orientation input available. Given that all differences between smartphone techniques were significant with respect to completion time, we can establish a ranking: *HandyCast* was fastest, followed by *Bimanual Tiltcasting*, in both Tasks 1 and 2. The mean acquisition time of approx. 3.4 s, which we measured in Task 1 under the *Bimanual Tiltcasting* technique is similar to the acquisition measured at 3.6 s in the original Tiltcasting paper under the target-agnostic, standard-display, small-target condition ([267], Table 1). While the differences in the concrete task and technique (we evaluated our *bimanual extension* of the original technique, see Subsection 6.5.1) do not allow for further comparison, the equivalent magnitude in completion time is an indication of external validity. The fact that completion under the bimanual task is more than twice the unimanual time might follow the same considerations described in *Insight 1d*.

In terms of control volume and motion paths, *Bimanual Tiltcasting* was the best-performing technique, making it a suitable technique for interaction in space-constrained settings. Yet, without a significant increase in control volume, *HandyCast* was significantly faster (1.1 s in Task 1 and 2.99 s in Task 2). This indicates that *HandyCast* manages to leverage our assumption that users *inadvertently* use body motions, even when the controller is not motion-sensitive.

Bimanual Tiltcasting differs from *HandyCast* in two aspects, which might contribute to *Bimanual Tiltcasting*'s slower task completion time. First, users can only use one channel–touch–to move hands forward in the former, while they can use two channels–touch and pose–in HandyCast. This, effectively reduces the bandwidth of information the user can input at a given point in time. Second, the absolute cursor, designed in Tiltcasting is less effective than our relative and gain-accelerated spatial touch transfer: touch-down events carry uncertainty where the hand will actually jump to, and accidental touch-up events reset the hand to the world-anchored neutral position, requiring to reacquire the object starting from the neutral position again.

In summary, either changing touch control or adding pose control or changing both made the difference. Thus, the question might arise if changing touch control only would have provided sufficient improvement of *Bimanual Tiltcasting*, rendering the addition of pose control superfluous. To understand the relative important between the input modes, we consider Figure 6.8, Task 2, which reveals that the physical motion path span approx. 60 cm with *HandyCast*. Given that a physical forward motion of 15 cm amplies to approx. 5.5m of forward motion in virtual space in studyplanecast, we can conclude that pose input was the driving input mode to our technique in order to cover the long-ranging virtual distances.

Taken together, using touch and orientation only without our full 6DoF pose transfer to move the hands was detrimental to the performance. We interpret this ranking as a promising indicator of our pose-and-touch transfer function.

Summary

Taken together, *HandyCast* yields completion times comparable with the best baseline *Controller-based Go-Go*, but significantly reduces motion and control space. Yet, we do not observe a loss in reported ownership, control, demand, or effort, between the two techniques. Compared to using touch and orientation only, our concept of 10D pose-and-touch transfer enables significantly faster completion times with only an insignificant increase in control volume.

6.6 User Study 2: Space-Constrained Setup

In our first study, we evaluated user performance for the controller-based Go-Go baseline, a smartphone baseline, and *HandyCast* proposed in this chapter. In accordance with the original idea of Go-Go, the Go-Go amplification parameters were configured such that the farthest target in the scene could be reached by an arm length. However, in our envisioned scenario of mobile VR usage, space might be significantly smaller. Therefore, we conduct a second study in a space-constrained setup, comparing a more sensitive configuration of Go-Go ($D = 35$cm, $k = 4$) with earlier and stronger amplification against *HandyCast* (same parameters as in study 1).

6.6.1 Procedure, and Apparatus: Seated, Space-Constrained Input Tracked Outside-In

For this second study, we physically constrain the space with cardboard to the sides and a wall to the front of the seated user, mimicking the available space on a bus or airplane seat as shown in Figure 6.9. Again, we use the outside-in VIVE tracking system for both techniques, making sure that the base stations can see into the boxed setup by mounting them on the ceiling.

Figure 6.9: Space-constrained setup for the second user study.

6.6.2 Task, Procedure, and Participants

In this study, 20 participants (3 female, 17 male, ages=21-41, M=29.0, SD=6.3), complete Task 2 from our first study (bimanual acquisition and placement). 6 of the participants never used

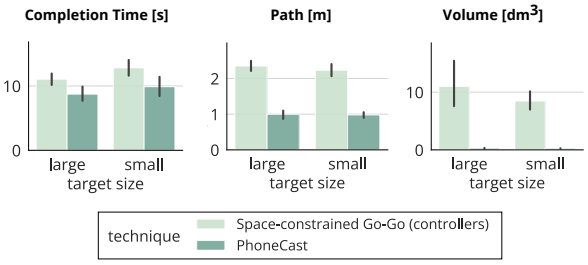


Figure 6.10: Aggregated results of our space-constrained, seated study by target size. Error bars indicate 95% confidence intervals.

VR, 10 use it a few times a quarter or less, and 4 weekly. We follow the same procedure as in the first user study.

6.6.3 Design

This study follows a mixed design, with Target Size as a between-subject variable, and Technique as a within-subjects variable. Target Size has two subject groups: *large* (25 cm) and *small* (20 cm) with 10 subjects each. technique has two levels: *Space-Constrained Go-Go* and *HandyCast*. We're interested in the same dependent variables as in user study 1, i. e., *completion time*, *interaction volume*, and *motion path length*. We counterbalance the order of technique by alternation.

6.6.4 Results

We performed a mixed-design ANOVA on task completion time with technique as independent within-subjects variable and target size as independent between-subject variable. Figure 6.10 gives an overview.

We found a significant main effect of technique on time ($F_{1,18} = 36.9$, $p < .00001$, $\eta^2 = .67$) as well as of target size on time ($F_{1,18} = 3.3$, $p < .1$, $\eta^2 = .15$). Post-hoc t-tests using Bonferroni-adjusted confidence intervals showed a significant difference between large and small targets when using Space-Constrained Go-Go ($p < .1$) but not when using HandyCast.

With respect to motion, we found significant main effects of technique on path length ($F_{1,18} = 453.4$, $p < .00001$, $\eta^2 = .962$) as well as on volume ($F_{1,18} = 71.9$, $p < .0001$, $\eta^2 = .8$)

6.6.5 Discussion

This second study reveals several insights, in particular against the backdrop of the findings of study 1.

Insight 2a: In a space-constrained setup, HandyCast enables faster object selection and placement than Space-Constrained Go-Go. While there was no significant difference between HandyCast and unconstrained Go-Go in the previous study, in the space-constrained setup of this second study, HandyCast ($M = 9.2s$) is significantly faster than Space-Constrained Go-Go ($M = 11.8s$) by 22.0%. This drop in performance of the Go-Go technique results from the loss of control due to increased amplification compared to constrained Go-Go. While this increase in amplification is required to stay within the space constraints, it entails at least two effects detrimental to Go-Go's performance: 1) While Go-Go's amplification function

is smooth and continuous by its mathematical formulation, the more aggressive amplification accelerates the virtual hand much quicker, making it harder for users to supervise ballistic movements, and thus increasing the risk of overshooting. In HandyCast, the same or even a smaller volume can be operated at lower motion amplification because input is complemented by touch. 2) At distance, the Go-Go technique operates at high amplification which then also impedes the corrective movement after the initial ballistic movement. In HandyCast, corrective movements with touch operate at constant speed, independent of distance, therefore always guaranteeing the ability of fine-tuning.

Insight 2b: Even with its aggressive amplification, Space-Constrained Go-Go uses significantly more space than HandyCast By design, Space-Constrained Go-Go requires less interaction volume and shorter path lengths than standard Go-Go in the previous study. Interestingly, despite not changing the parameterization of HandyCast, users also travel less with the smartphone and occupy less interaction volume than in the previous study, likely due to their awareness of being constrained and thus intuitively relying on touch input more to avoid touching the obstacles. However, HandyCast still requires significantly less interaction space ($\mathcal{M} = 0.3dm^3$) and path lengths ($\mathcal{M} = 0.99m$) than Space-Constrained Go-Go ($\mathcal{M} = 9.95dm^3$ and $\mathcal{M} = 2.3m$ resp.). From this, we conclude that HandyCast exhibits a fundamental space-efficiency advantage, independent of the specific Go-Go parameterization. This advantage results partly from HandyCast's ability to accept touch input, but also from the specific designs concerning amplification reference points: While Go-Go computes *two amplification vectors* relative to *a single point* at the chest, HandyCast computes *a single amplification vector* relative to *a single point*, namely the custom-defined home position in a neutral posture.

Because both Go-Go controllers share the same reference point, Go-Go introduces a radius of linear amplification in front of the chest which smoothly transitions to a non-linear amplification. If non-linear amplification was to kick-in immediately, only one controller could be set to a neutral position at the reference point, blocking the other controller from the neutral position, forcing it to operate at non-linear amplification. Using a single smartphone allows for a natural neutral position from which non-linear amplification can kick-in immediately while still maintaining bimanuality through touch.

Insight 2c: Decreasing the interaction volume in Go-Go induces more frequent controller collsions In contrast to the previous study, in this space-constrained setup, multiple participants (P3, P4, P8, P19) reported that it was annoying or irritating that they often hit one controller with the other when selecting or placing objects. This effectively hints at the reduced bimanuality when using controllers in a shared small volume, esp. when oriented around a shared reference point. By design, a single smartphone as in HandyCast cannot suffer from this problem.

Insight 2d: Space-Constrained Go-Go benefits from larger target sizes due to larger amplification inaccuracies at distance. As indicated by the interactions reported above, Space-Constrained Go-Go benefits from larger targets with respect to completion times whereas HandyCast does not. We interpret these results as follows: A trial is composed of 1) pointing towards the target 2) attaching to it, 3) pointing towards the drop zone with the attached target and 4) detaching from the target for each corresponding hand. Pointing in turn is composed of initial and final movement. The target size only impacts the final movement in

the pointing steps. Thus, the fact that Space-Constrained Go-Go's performance is significantly impaired by reducing target size reveals that the final movement step is impaired. The fact that HandyCast is not equally impaired indicates that the final movement is less impacted here, either due to the ability to fine-tune virtual hand positions quickly by touch input, or due to the lower amplification, afforded by the two input modalities that can be combined to high amplifications if and only if desired.

Summary

While there was no significant difference between HandyCast and standard Go-Go in the previous study with respect to completion time, this second study has revealed that HandyCast outperforms Space-Constrained Go-Go in a spatially constrained setup by unfolding its advantage of decomposing non-linear and linear amplification into two different, separately controllable modalities of hand motion and thumb touch input. It's fundamental property of simulating bimanuality through touch-augmented pose transfer becomes an advantage over Space-Constrained Go-Go which suffers from controller collisions reducing independent input in a constrained space.

6.7 Tracking Study: Phone vs. VIVE

In our previous studies, we evaluated participants' performance using the input techniques under ideal tracking conditions. The purpose of this evaluation was to assess the effect of tracking technology on performance. HandyCast has the potential to run entirely on the user's phone, where inside-out tracking may cause less accuracy in task completion and thus reduced completion speed.

Task, Procedure, and Apparatus

In this study, participants completed Task 2 from our first study (bimanual target acquisition and placement). Using *HandyCast*, we compared two tracking methods: outside-in tracking and inside-out tracking. Outside-in tracking used the VIVE's surrounding base stations and a tracker attached to the phone as in our first study (Figure 6.5). Inside-out tracking used our ARKit-based implementation.

The evaluation implemented a within-subjects design with Tracking Method as the independent variable. Participants repeated bimanual acquisitions and placements 24 × (2 Tracking Method × 2 blocks) = 96 trials. The order of Tracking Method was counterbalanced across both tasks and participants. Participants received the same instructions and training as in our first study, with the addition not to accidentally occlude the phone's back camera.

The apparatus was the same as in our first study, only that inside-out tracking as well as outside-in tracking were active throughout all trials for later comparison. In the *Outside-in* condition, the VIVE tracker drove the transfer function as in our first study. In the *Inside-out* condition, the tracker merely served to record ground-truth positions and orientations, but the transfer function was exclusively driven by the phone-reported 6D poses without input from the VIVE system. Participants performed this spatial registration before each of the two *Inside-out* blocks.

Our analysis is two-fold. 1) We investigate the total drift across a full block of 24 trials with

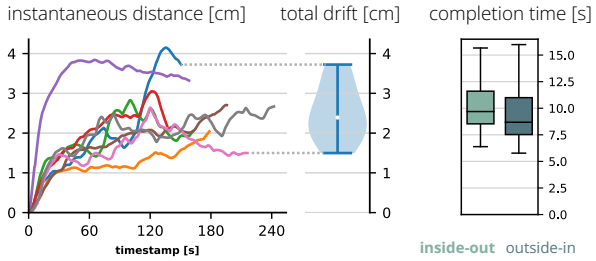


Figure 6.11: Tracking evaluation. Left: Drift that emerged from phone-based *Inside-out* tracking during a block. Right: Impact of tracking method on task completion ($p = .1$).

inside-out tracking. 2) We analyze the effect of Tracking on Completion Time as a dependent variable.

Participants

We recruited another 4 participants (1 female, 3 male, ages=27–32, M=28.8, SD=2.2). 1 participant had used VR occasionally, 2 less than 5 times, and 1 never.

Results—Tracking effect on performance

We performed a one-way repeated-measures ANOVA on *completion time* for Tracking Method. Figure 6.11 shows the distribution of completion times, showing a lower mean for *Outside-in* ($\mathcal{M}$=9.85 vs. *inside-out* $\mathcal{M}$=10.49), though differences were not significant.

Results—Spatial drift

As shown in Figure 6.11a, using *Inside-out tracking* incurred an average total drift of 25 mm ($SD = 7.3$) during a full block. With an amplification factor $\lambda = 2.5$, a difference of 25mm creates an error offset of approx. 12 cm in virtual space, when reaching out exactly 10 cm. For each *Inside-out* block, we computed the Euclidean distance between the VIVE tracker and the phone-reported position, offset to zero upon the start of the first trial. Figure 6.11 left shows a plot of all 8 blocks. For visual clarity only, in the plot, we also apply exponential smoothing with a factor of .001. The distribution of accumulated drift at the end of each block is shown in Figure 6.11 middle. As measured by the tracker, the inside-out-tracked pose drifted 0.25 cm/m on average ($SD = 0.09$), computed as accumulated error divided by total distance traveled.

Discussion

Our analysis showed the impact of drift as part of *Insight-out* tracking on the performance of using *HandyCast*. Over the course of just one block, phone-reported positions drifted up to 4 cm. The only small difference in completion time indicates that participants were able to compensate for this amount of drift, possibly because of the dominance of vision over proprioception during interaction in VR.

Depending on a participant's speed, a block lasted 2.5–4 minutes. A gameplay session in VR may last much longer and more drift may accumulate as a result. HandyCast could compensate for such drift by adjusting the respective anchor position to calculate pose transfer. This is currently static and future iterations of HandyCast will need to account for dynamically updated anchors

(e.g., by monitoring maximum proximity to the torso or detecting resting arms in the lap), which would also allow resetting drift.

Some participants' drift curves show sudden moments (e.g., violet or blue trace). During the trial, these occasional moments manifested as visual jumps and thus interfered with the participant's current input motion. This demanded participants' manual counteraction, which slowed down the trial in these specific moments.

Overall, our evaluation has established the feasibility of phone-based inside-out tracking for the use of HandyCast. This also makes the use of our phone-only controller driver practical, which allows control over existing VR apps and games. We also believe that with increasingly powerful tracking methods on today's consumer phones, drift will decline in the future.

6.8 Limitations and Future Work

Improving Tracking In this chapter, we have proposed full inside-out tracking based on the phone's sensors for pose-and-touch transfer and compared it to outside-in tracking for reference. To improve the inside-out tracking accuracy, future research lies in using a headset-based tracking, e.g. by camera-based phone pose estimation, leveraging hand tracking as a proxy to then estimate the phone pose from wrist poses, by displaying an optical marker to the blindly operated touchscreen, or even using acoustic tracking [343; 149].

Studying Transfer Functions under Different Parameters Since HandyCast is scene-agnostic, all parameters can be specified on the controller side, similar to setting the gain acceleration of a mouse in the operating system. For specific VR apps, users might choose a different amplification factor in HandyCast, depending on the virtual and physical environment. In this chapter, we have specified values through pilots so that users can comfortably reach all targets in the task with all pose-aware techniques. Future research lies in understanding the effects of different parameters within our design, e.g., the relative importance between pose and touch by studying it under varying gain factors, amplification factors or even other amplification function families (e.g., Hermite curves, [349]) in the case of pose transfer. Furthermore, exploring the coupling of positional or positionally amplified input with other plane-anchoring techniques such as Free Plane-Casting [175], e.g. maintaing one plane per cursor, offers promising research directions.

Designing for More Complex Interactions Using the 10 DoF from pose and touch in a single smartphone constrains HandyCast to positional bimanuality, not sufficient for full 12-DoF bimanuality, e.g., required to hold a bottle in one hand and opening it with the other. Other more complex interactions, however, could be enabled by processing further input signals. In particular, we use atomic touch-down and touch-up events to attach to and detach from objects, e.g., leaving double taps unused for further mappings. In the future, double taps might be used to enter the teleportation mode, thus enabling rested locomotion.

6.9 Conclusion

We have presented HandyCast, a smartphone-based input technique designed to control bimanual input in large VR environments through small motions in space-constrained environments.

HandyCast's core concept is its pose-and-touch transfer function that fuses the smartphone's 3D position, 3D orientation, and 2D touch motions as input to jointly output two independent 3D positions to place virtual hand avatars. The touch-transfer component of our technique allows users to clutch during input, thereby repeatedly readjusting avatar locations, and thus affording navigation inside infinite spaces.

In our *space-constrained user study*, we found that HandyCast requires significantly less completion time, path length and interaction volume compared to the space-constrained Go-Go implementation. In a *tracking evaluation*, we compared the performance of HandyCast during externally tracked and phone-only tracked operation. We found that despite small inaccuracies and drift, HandyCast affords operation on today's smartphones and therefore enable fully mobile use together with untethered VR systems.

Because HandyCast is completely scene-agnostic, our low-level SteamVR driver is fully compatible with existing VR applications and games, which we demonstrated at the example of Job Simulator. We conclude that HandyCast brings comfortable, full-range, and bimanual 3D input to mobile VR by retrofitting the user's smartphone as an ubiquitous controller. In comparison to controller-based state-of-the art baselines, HandyCast does not only reduce interaction volume and improve completion times in space-constraint settings, but 1) does not require dedicated hardware, 2) can be highly amplified for usage in very small volumes without suffering from inter-controller collisions, 3) allows unlimited reach without parameter re-adjustment through touch, and 4) allows for robust refinement of virtual hand position through touch, independent of the virtual distance between user and object.

6.10 Further Details

6.10.1 Overview

So far, we presented HandyCast, a smartphone-based input technique designed for small motion that enables full-range 3D input for two virtual hands in VR. HandyCast builds on a pose-and-touch transfer function that fuses the phone's position and orientation with touch input to derive individual hand poses. We have described our HandyCast technique in detail, and reported and discussed its performance in the three different studies (full-range user study, space-constrained user study, tracking study).

In this supplement, we first provide further insights into the performance of HandyCast in user study 1, in particular by analyzing the interplay of touch and motion. In user study 1, we also studied Elliptic Arm Extension a second custom technique we developed but which turned out inferior to HandyCast, as a fourth condition. In Section 6.10.3 of these supplementary materials, we give a detailed description and report on its performance, as measured in user study 1. Third, we explicate additional implementation details concerning the overall system, the spatial registration procedure designed for inside-out tracking, and our stand-alone controller driver for SteamVR.

6.10.2 Details on Full-Range, Seated User Study 1

In our first user study with a full-range, seated setup, we compared our present HandyCast, our Elliptic Arm Extension technique, and the two baselines against each other to understand their performance in terms of completion time and space. Here, we want to understand participants'

usage patterns with our best technique, HandyCast, and how they leveraged simultaneous motion and touch input. Therefore, we first analyze our recordings of touch and motion input. Then, we show the breakdown of the questionnaire in that study.

6.10.2.1 Detailed Touch-and-Motion Interplay Analysis

The effective touch screen size measures 150mm of width × 77mm of height. From our touch input logs, we see that all bimanual trials with *HandyCast* (n=288), half of them show that the users applied a long touch stroke ($M = 59\,\text{mm}$, $SD = 80\,\text{mm}$). In the other half of trials, user used touch primarily for acquiring and releasing objects, not for moving the virtual hand avatars.

Figure 6.12 shows a single trial with *HandyCast* (P4, Task 2, Trial 8). First, the user was asked to select the highlighted object with the right (blue) hand. As seen in Figure 6.12a, in the first phases of the velocity profile, they quickly moved the phone (ⓐ, red peak at 0.6 s), then slowly approached the target (ⓑ, soft red peak at 1.2 s). Then, they used the right thumb to slowly approach the target (ⓒ, blue peak at 1.5 s). Figure 6.12b also shows this touch sequence in the blue input area (ⓒ) and Figure 6.12c shows the resulting motions of the virtual hand avatar. Movements like these suggest that participants performed acquisitions in two phases; first, they approached the target coarsely by moving the phone and then fine-tuned through touch.

Considering phase ⓔ, we see that velocity for both phone motion and left thumb peak at the same time at 3.1 s, before the participant grabs the object at 3.7 s. A similar pattern can be observed in phases ⓕ, ⓖ and ⓗ.

Taken together, the data suggests two usage patterns concerning the interplay of motion and touch: 1) *touch-based refinement* and 2) *simultaneous touch-and-pose control*. Both usage patterns are unique to pose-and-touch transfer functions, enabling accuracy as well as space efficiency.

6.10.2.2 Detailed Questionnaire Responses

Figure 6.13 left shows participants' ratings for mental and physical demand, effort and frustration. Figure 6.13 right shows the results of participants' ratings for ownership ("I felt as if the virtual hands were my hands."), Double Hand ("It seemed as if I might have more than two hands."), Control ("It felt like I could control the virtual hands as if it were my own hands.") and Interference ("The movements of the virtual hands were caused by my movements") [105]. Figure 6.14 shows participants' ratings concerning their lower arms and upper arms, respectively.

6.10.3 Details on Elliptic Arm Extension

In our first user study, comparing our two custom techniques HandyCast and Elliptic Arm Extension, as well as the baselines Full-Range Go-Go and Tiltcasting, we found that HandyCast is our stronger technique. Therefore, we provide the details on Elliptic Arm Extension only in this supplement. In the following, we first describe its design, then report the ANOVA with respect to completion time and space usage, and finally discuss its performance properties.

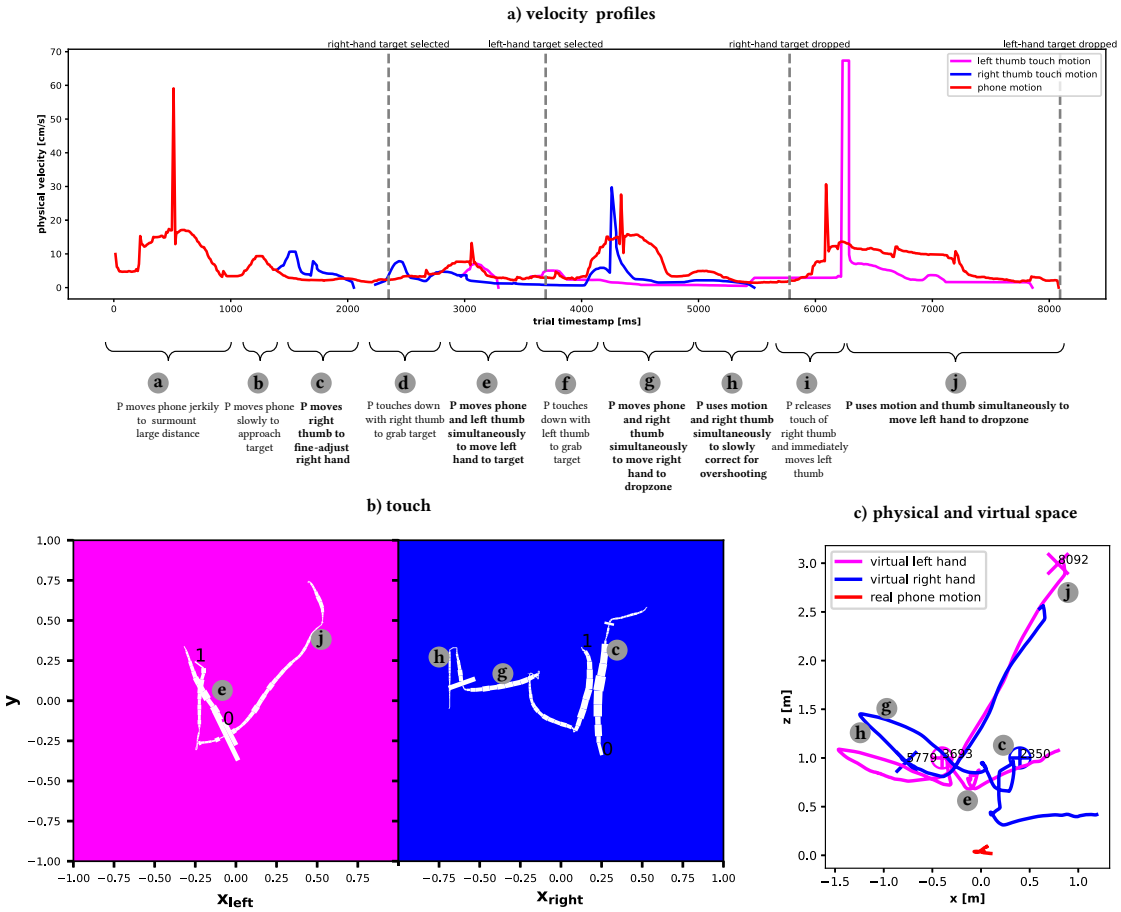


Figure 6.12: Touch-and-Motion Interplay Analysis. All three subfigures describe the same trial. The central text row gives our interpretion of distinct phases in the trial. The textcircled letters above each phase are reused in all the other subfigures, and in the text. Subfigure a) shows the measured velocity [cm/s] of the physical phone motion (red), computed from the VIVE tracker trajectory, and the touch stroke for the left (magenta) and right (blue) thumb over time [s]. Touch stroke velocity can only be computed and plotted if the thumb is touching the screen, resulting in a single red but multiple of blue and magenta line segments. Vertical dotted lines indicate the timestamp of a grab or drop object interaction. Subfigure b) shows the corresponding touch strokes on the respective left (magenta) and right (blue) touch zones. Black digits indicate the touch stroke ordering on that touch zone. Subfigure c) shows a top-down view of the scene, showing the physical phone motion (red) spanning only a fraction of space and movement of the virtual left (magenta) and right (blue) hands with pose-and-touch transfer.

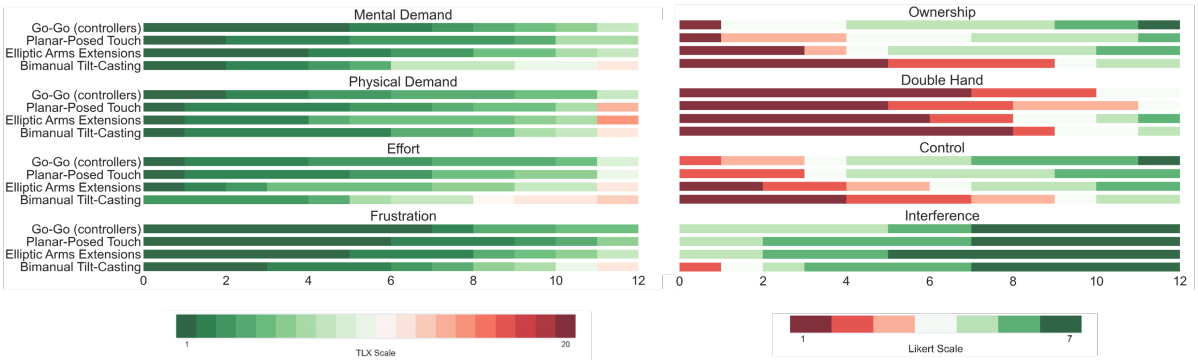


Figure 6.13: Embodiment results. TLX questionnaire and Avatar Embodiment questionnaire responses by technique and response level.

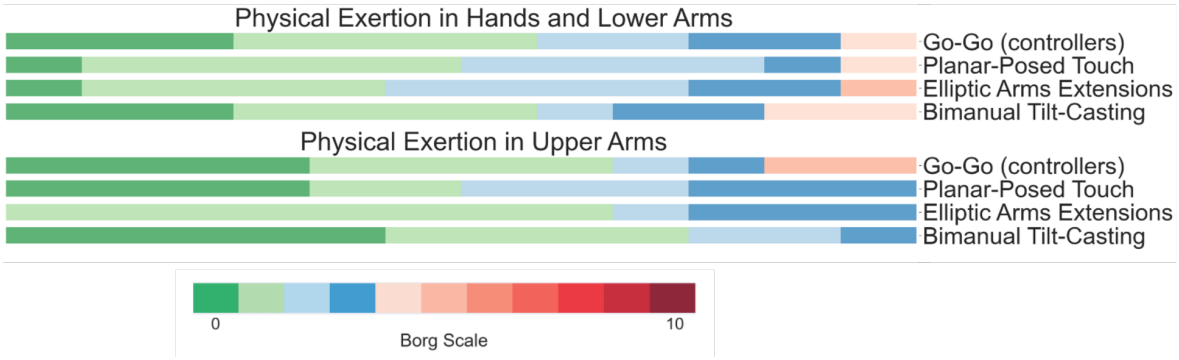


Figure 6.14: Participants' exertion ratings (Borg CR10 scale).

6.10.3.1 Design

HandyCast and Elliptic Arm Extension implement separate pose transfer functions, but share the same touch transfer mechanism. In contrast to the definition of a plane in HandyCast, the transfer function of Elliptic Arm Extension is inspired by the motion of physically reaching out to grab an object. To that end, we amplify a small yawing gesture in the phone so to position the hands on a large semi-ellipse, as shown in Figure 6.15. The length of the semi-ellipse is specified by the constant e_{forward}, the ellipse width by the constant e_{lateral}. When rotating the phone clockwise about the upwards pointing vector $\bar{\mathbf{v}}_{\text{up}}$ (indicated in green in Figure 6.15), the left hand extends out, following an elliptic trajectory. Turning the phone counterclockwise extends the right hand instead. The yawing angle α indicates how much the user yaws the phone around the up-vector $\bar{\mathbf{v}}_{\text{up}}$.

In summary, using the equations in Figure 6.15, we can compute the left hand's position p_{L} and the right hand's position p_{R} in Elliptic Arm Extension as

$$p_{\text{L}} = \mathbf{p} + \cos(\alpha)\,\bar{\mathbf{v}}_{\text{lateral}}\,e_{\text{lateral}} + \sin(\alpha)\,\bar{\mathbf{v}}_{\text{phone}}\,e_{\text{forward}}$$
$$p_{\text{R}} = \mathbf{p} - \cos(\alpha)\,\bar{\mathbf{v}}_{\text{lateral}}\,e_{\text{lateral}} - \sin(\alpha)\,\bar{\mathbf{v}}_{\text{phone}}\,e_{\text{forward}}$$

6.10.3.2 Results of User Study 1 with Respect to Elliptic Arm Extension

Completion time *For Task 1*, post-hoc *t*-tests using Bonferroni-adjusted confidence intervals on Technique showed significant differences between all comparisons with *Bimanual Tiltcasting* or *Elliptic Arm Extension*, but not among both. *Controller-based Go-Go* was 0.7 s faster than *Elliptic Arm Extension* on average (133.4%, $p < .001$). *HandyCast* was 0.6 s faster than *Elliptic Arm Extension* (127.0%, $p < .01$). *For Task 2*, post-hoc *t*-tests using Bonferroni-adjusted confidence intervals showed a significant difference between all comparisons involving either *Bimanual Tiltcasting* or *Elliptic Arm Extension* or both.

Travel length of physical motions *For Task 1*, post-hoc *t*-tests using Bonferroni-adjusted confidence intervals do reveal significant differences between any of {*Controller-based Go-Go*, *Elliptic Arm Extension*} and any of {*HandyCast*, *Bimanual Tiltcasting*}, but not within the two sets. The average distance traveled using *HandyCast* was 0.634 m shorter than *Elliptic Arm*

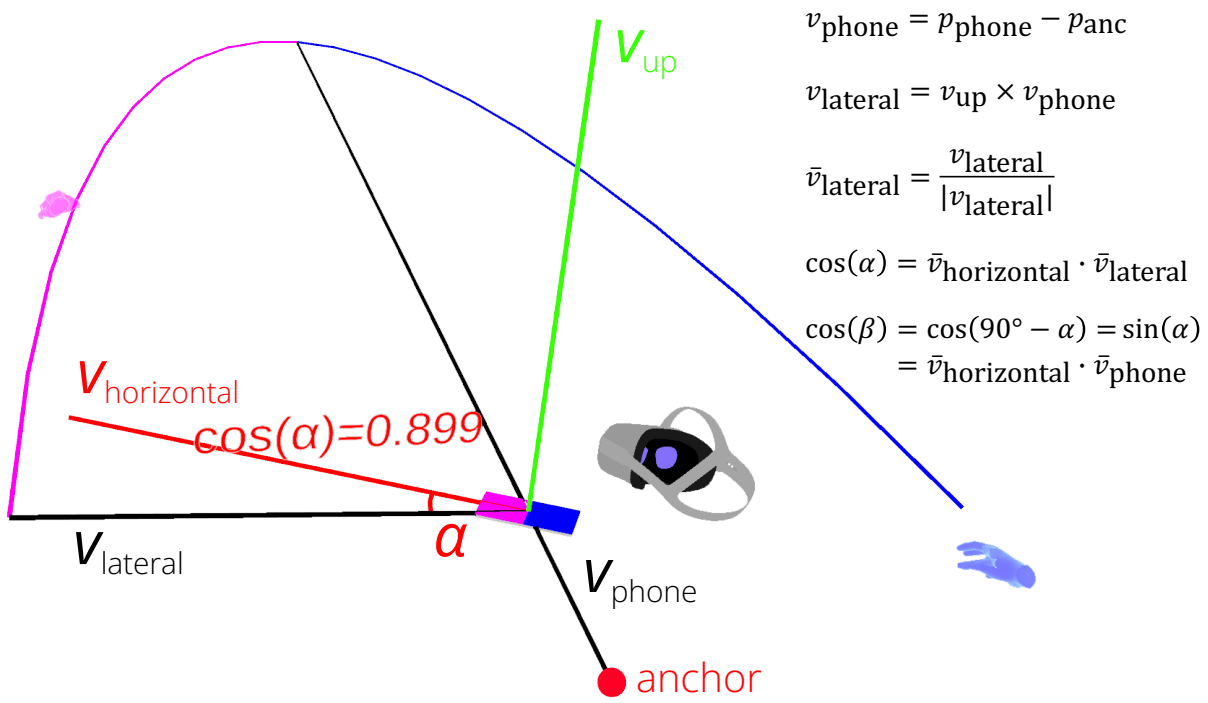


Figure 6.15: Schematic 3D representation of our *Elliptic Arm Extension* transfer function. Pose transfer in Elliptic Arm Extension is modeled after the motion of reaching out an arm. By the subtle motion of yawing the phone around its vertical axis, the corresponding hand shoots forward quickly on an elliptic trajectory. When the phone is fully rotated by 90 resp. -90 degrees, the right resp. left hand is fully extended and touch transfer can be added to reach objects which are even farther away.

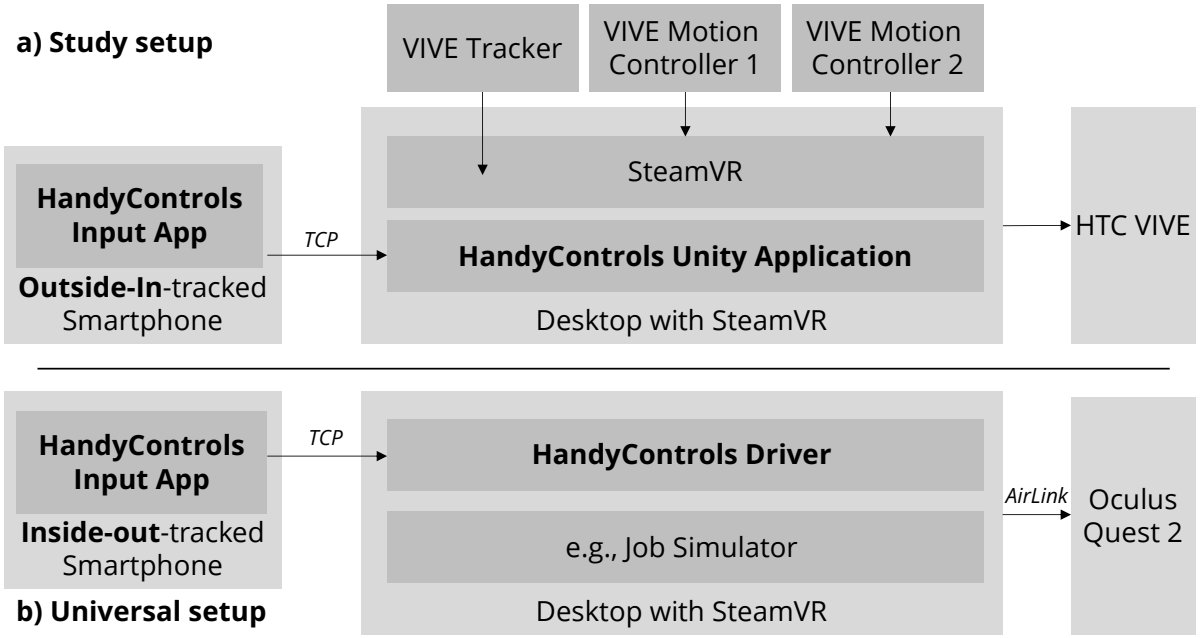


Figure 6.16: We implemented two instances of HandyCast: a) inside Unity to study performance with different tracking and input conditions and b) as a low-level controller driver for universal use in SteamVR, using inside-out tracking and transferring pose-and-touch events directly to the app.

Extension (52.8%, $p < .001$). Using *Bimanual Tiltcasting*, the average distance was 0.687 m shorter than *Elliptic Arm Extension* (48.8%, $p < .001$).

Required physical volume (control space) *For Task 1*, *HandyCast* required 2964.9 cm³ less control volume than *Elliptic Arm Extension* (11.5%, $p < .001$). *Elliptic Arm Extension* required 11049 cm³ less control volume than *Controller-based Go-Go* (23.3%, $p < .001$). *For Task 2*, *HandyCast* required 4180.6 cm³ less control volume than *Elliptic Arm Extension* (12.3%, $p < .001$). *Elliptic Arm Extension* required 16367.3 cm³ less control volume than *Controller-based Go-Go* (22.6%, $p < .001$).

Questionnaires *In terms of perceived workload* (NASA TLX), using pairwise Bonferroni-adjusted Wilcoxon signed-rank tests, we found a significant difference in *effort* between *Controller-based Go-Go* (3.5 ± 2.1), and *Elliptic Arm Extension* (5.5 ± 2.9). On average, participants rated effort for *Controller-based Go-Go* 2.0 points lower (= better) on the TLX scale than *Elliptic Arm Extension* ($p = .03$).

Error rates Participants only ran out of time in Task 2: once with *Elliptic Arm Extension*.

6.10.3.3 Comparison Details

Summary 1: the intricate pose-and-touch transfer combination in Elliptic Arm Extension is not as intuitive and less efficient than HandyCast *Elliptic Arm Extension* was not as efficient, and resulted in significantly slower completion times and a smaller

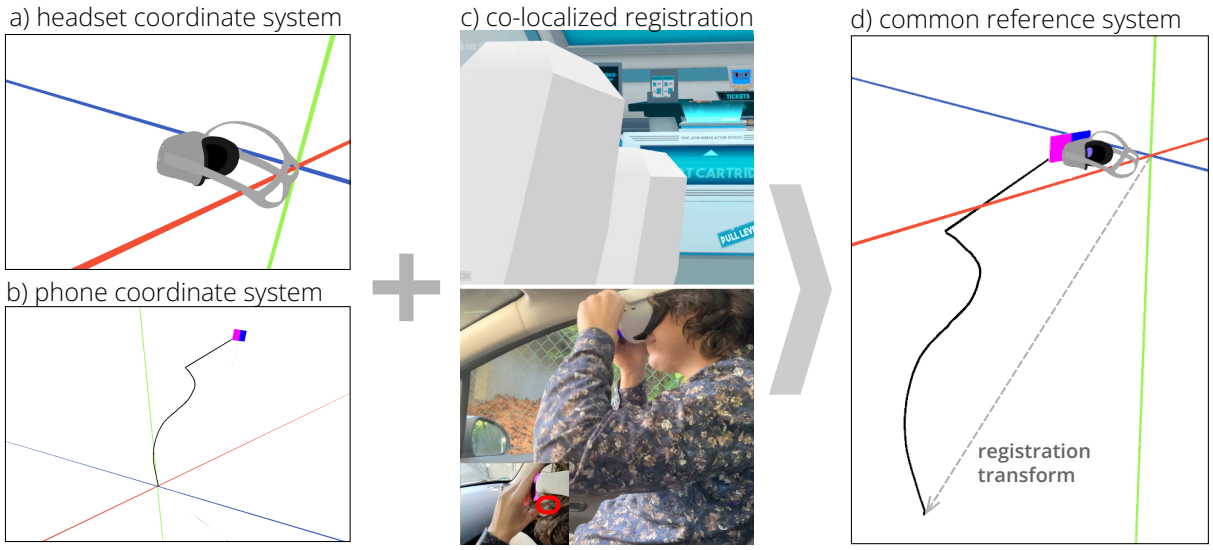


Figure 6.17: Spatial registration for co-located inside-out tracking. HandyCast can use full inside-out tracking through visual-inertial odometry, and thus does not require mounting a tracking ring of LEDs or similar to the phone for on-the-go usage.

reduction of path length (88.6% of Go-Go) and control space (22.6%). Also surprisingly, participants rated Elliptic Arm Extension more tedious than Go-Go. We rationalize this finding with qualitative feedback of two participants (P1, P9), who indicated that it is not fully intuitive when to use touch transfer and when to use pose in Elliptic Arm Extension, because pose and touch were ”doing totally different things” (P9). Indeed, while touch translates the hand avatars based on the direction of the touch path, for pose transfer a rotation of the smartphone is required, similar to rotating a shoulder to reaching out with one hand. So, fusing two very different concepts results in a less intuitive design than, for example, HandyCast, where forward movement in both pose and touch lead add up to a fast forward movement of the hand avatar. P1 expressed that they ”sometimes forgot to rotate the phone to reach out more and instead tried to “move the phone farther forward,” when they “thought too much about it.” This might hint at the idea of extending the design of Elliptic Arm Extension to compute the forward reaching gesture not only from the phone yaw, but also from the phone’s distance to the anchor (as in HandyCast).

Summary 2: Bimanual Tiltcasting requires the least volume and the least motion paths of all techniques, but is slower having only touch and orientation input available. Given that all differences between smartphone techniques were significant with respect to completion time, we can establish a ranking: *HandyCast* was fastest, followed by *Elliptic Arm Extension* and *Bimanual Tiltcasting*, in both Tasks 1 and 2.

6.10.4 Details on the Technical Implementation

6.10.4.1 Details on the System Architecture

As stated before, we implemented both a study environment in Unity with C# supporting all four techniques as well as a stand-alone controller driver, written in C++, that supports the usage of HandyCastin any SteamVR app. Figure 6.16 gives an overview. All smartphone techniques

in the study environment can be driven inside-out or outside-in tracking, also allowing to record both at the same time and thereby enabling our tracking evaluation. The study was powered by a Unity app that we developed to present tasks, log timestamped input events, motions, and trajectories, and prompt participants for ratings. Our app ran at 90 fps on an Intel Core i7-9700K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3080 GPU. The study app processed online input for task completion, advanced participants through the study procedure, and we analyzed all logs offline to compare input techniques.

6.10.4.2 Details on the Stand-alone Controller Driver Implementation

Separately and independently of our study environment, we developed a low-level controller driver for SteamVR, thus allowing backward-compatible drop-in use with standard VR apps on a Quest 2 with AirLink. For this, we implemented HandyCast to run natively on the iPhone. Our implementation registers touch and hard-button events as well as the inside-out tracking from the phone, relays it to the PC via TCP, where they are processed and fused for transfer by a C++ controller driver that interfaces with SteamVR. Our controller implementation also handles spatial registration and anchoring. No external tracking is required in this implementation, allowing it to be fully mobile.

6.10.4.3 Details on Spatial Registration for Inside-out Tracking

To enable inside-out tracking, our HandyCast smartphone app builds on ARKit's visual-inertial odometry using one of the iPhone's back-facing cameras. While ARKit provides 6DoF translation and rotation of the phone with respect to an arbitrary physical scene origin, this physical scene origin is in itself unrelated to the virtual scene. We align the physical coordinate system, tracked by ARKit, with the virtual coordiante system, by asking the user to co-locate the headset and the smartphone and pressing the volume-up button. This procedure bridges the physical and virtual coordinate systems by means of the headset position. Once registered, the physical phone coordinates can be translated to virtual coordinates. In case of drift or tracking errors, the user can re-position . We implemented this procedure for both the SteamVR controller (running on Quest2 via AirLink and SteamVR) and the Unity environment (running on the HTC VIVE).

The larger the island of knowledge, the longer the shoreline of wonder.

Ralph W. Sockman, *as cited in* 274

7

Conclusion: Approaching MR that Seamlessly Blends with the User's Space and Mind

7.1 Summary

In this dissertation, new concepts and technologies for deepening the integration between virtual scenes with the physical scene were conceived, developed, and studied. Leveraging user and environment sensing, these concepts and technologies established a digital awareness of the physical scene in order to situate virtual signals usefully and interpretably therein in alignment with the user's perception thereof.

With SoundsRide, we have contributed a concept and a technology the recursive-predictive use of semantics in the form of sound affordances to produce and integrate auditive signals, meaningful to the user, with the physical scenery in real-time as it changes under user and environment-variant conditions.

With TransforMR, we have contributed a concept and a technology for the use of semantics with object and subject poses enabling the construction of alternative worlds in real-time, interpretable by the user, in open-ended, unprepared, and previously unseen environments, based on a comprehensive visual perception, scene composition, and graphics rendering pipeline.

With the Scene Responsiveness work and its underlying RealityToggle, Spielberg, and Twin Builder subsystems, we have contributed concepts and technologies for the use of semantics in the form of character-object and user-object affordances, complemented with expansive geometric and photometric scene properties, to enable the interactive real-time in-situ composition of visuotactilly consistent possible worlds, elevating the achievable levels of scene integration between physicality and virtuality.

With HandyCast, we have contributed a concept and a technology that enables situated input respectful of the limitations in space and hardware for on-the-go MR interaction.

The presented ensemble of concepts and technologies corroborates that situatedness and semantic awareness in MR provide a means of enabling exciting and meaningful experiences, appli-

cations, and interactions for a user's specific situation, characterized by their seamless blend with the user's space and mind, spanning different modalities, form factors, and mechanisms of sensing, transduction, and rendering, far beyond visually displaying a 3D object in space.

To conclude, we take the opportunity to adopt an integrative perspective that overarchs the individual contributions in this dissertation to derive broader implications as an outlook on the future of foundational MR platforms as well as situated and semantics-aware MR experiences, applications, and interactions built upon them.

7.2 Implications and Outlook

7.2.1 Phase 1: Enhancing Partial Awareness

While all of our projects have already exhibited more or less comprehensive primitives of semantic awareness, the pool of semantic concepts to be considered in any single application or framework offers to be increased ad libitum. As observable today, current research and product efforts are concerned with the conception and realization of methods that increase the degree of awareness for individual situational parameters.

On the environment-sensing side, today, this is predominantly reflected in the proliferation of visual perception systems, both in hardware and software. The cameras required might be pragmatically positioned for egocentric vision on the user's head moving in lockstep with the user's eyes following the always-on AR vision, mounted exocentrically in the environment for home usage, or even carried by moving platforms up to swarms of drones for highly specialized applications, coordinated by a user-worn device. The high density of information comprised in optical signals allows for a multitude of 3D computer vision methods such as object detection, 6D and 9D object pose estimation, 3D keypoint point estimation, shape reconstruction, surface reconstruction, semantic segmentation, instance segmentation, human pose estimation, hand pose estimation, head and eye gaze estimation, facial reconstruction, or activity recognition. Each of these methods are receiving attention in research to increase their traditional quality measures while also increasingly turning toward problem framings that accommodate larger or even open vocabularies, spatiotemporal consistency, a low computational footprint, diverse environmental robustness constraints, and long-lived session-spanning dynamics. At the same time, semantic awareness is often required to be complemented with high-fidelity geometric and photometric reconstructions, which are the subject of research in the ever-evolving field of SLAM and SfM and, impelled by the progress in implicit scene representations, newer areas such as neural radiance fields and gaussian splatting. Research challenges remain in the integration across these layers of interest, i.e., in the unification of representations that allow for localization while incrementally modeling textural, structural, and semantic aspects over time in dynamic scenes.

Similarly, on the user-sensing side, cameras also gain attention. For example, wrist-worn cameras can be used to sense the users's hand poses and mid-air gestures [180; 367; 376], on-surface gestures, activities [258; 73], and full-body poses [136; 200]. Chest-worn and shoulder-worn cameras as found recently in the product landscape [142] and in classical research [121] offer to understand user gestures and to contextualize voice commands. Head-worn cameras, oriented toward the user's face enable reconstruction of high-resolution facial activations and motions as well as eye gaze directions for a multitude of applications, ranging from head avatar rigging for telecommunication to natural input techniques. At the same time, highly specialized sen-

sors, fitted to body-worn and body-integrated devices offer solutions to more specific problems, ranging from health applications to input techniques.

The above-discussed problems for gaining awareness of select aspects of physical reality encompass some of the most active and innovative research areas in the field. Extrapolating these research trends promises to gain insights into momentary aspects of the physical environment and the user's body with high expressiveness and detail. However, they share that they aim to map received signals to a logical unit that directly corresponds with the signal's source. In contrast, despite the many open questions regarding human perceptual and cognitive processes, there seems little doubt that humans not only reconstruct a scene as an unordered set of isolated units but instead maintain a more holistic awareness of the surroundings and themselves therein.

7.2.2 Phase 2: Evolving from Partial to Holistic Awareness

7.2.2.1 Overview

Based on the insights gathered from our works in TransforMR, Scene Responsiveness, and SoundsRide, we argue that holistic awareness arises not only from recognizing individual logical units in the physical scene for which signals can be processed locally and in isolation but instead is established through additional systemic qualitative properties of the system's perceptual process. More specifically, we argue that the desire for holistic awareness requires, first, designing for uncertainty, second, designing under a constructivist paradigm, third, designing with strong priors, and, fourth, designing from an integrative perspective.

7.2.2.2 Designing for Uncertainty

As demonstrated by the use of visual scene completion techniques employed in the alternative worlds and possible worlds projects in this dissertation, establishing a more holistic awareness is fundamentally bound to uncertainty. The reason lies in the fact that scene completion problems, such as the ones faced in TransforMR and Scene Responsiveness, are inherently ill-posed, as they allow for an infinite number of valid yet more or less likely and more or less complex solutions. Imagine a painting on a gray wall. Removing this painting could reveal the wall of the exact same color as seen on the rest of the wall. Or it could reveal a rectangular shape on the wall that is slightly brighter than the rest of the wall because this area was protected from dirt and sunlight, depending on the duration the painting had been hanging there. It might also reveal a hole in the wall that was intentionally covered by the painting or even a hidden safe box.

Humans have an innate ability to select and reason about relevant aspects of reality, while abstracting away less relevant aspects behind varying degrees of certainty. Subconsciously, they will have an implicit, vague expectation of what they might find when taking away the painting, even without ever perceiving a direct signal thereof. At the same time, they can easily integrate new information that meets their expectations, while–being aware of the uncertain nature of their expectations–they can also quickly integrate mildly surprising information, or trigger a longer-running thread of thought for new information that does not confirm their expectations. Mixed Reality systems that aim to model a scene holistically in a similar way, therefore need to accommodate aspects of uncertainty in purposeful ways. In TransforMR, the uncertainty of how the scene looks behind a removed object is resolved by a neural generative inpainting procedure that reflects the underlying dataset, regularizing the output toward the least complex fill. In Scene Responsiveness, the uncertainty of when the user's movement will cause a collision

with either the physical object or its virtually displaced twin is resolved via the mechanisms of object rephysicalization and object elusiveness. In SoundsRide, the uncertainty of when the user will reach a sound affordance location is accounted for by the mechanism of corrective resynchronization. It can be expected that future research will focus on this uncertainty property to approach holistic perception. Research questions include, first, how to allow for uncertainty, but then also, second, how to resolve uncertainty as new information is obtained.

7.2.2.3 Designing under a Constructivist Paradigm

As evidenced by the above-reviewed scene completion problem, any estimation of the physical scene has the nature of a construction, informed by more or less sparse samples of physical reality. This means establishing a more holistic awareness is not merely descriptive but also a fundamentally *creative act*. This creative act is characterized by an interpretative process that subjectively interpolates and extrapolates perceived signals to a more complete, yet "more constructed", model of reality, certainly often assumed to offer the highest degrees of correspondence with elements of physical reality given the perceived signals. This constructivist nature is at the core of any model of physical reality that is produced in real-time as new sensor data is processed, however, it is of particular importance for a holistic understanding of the world as sensor data rarely reflects the entirety of relevant aspects of the world.

The necessity of a constructivist view is, among other factors such as information reduction and the human tendency to make sense of new information, a result of the above-described uncertainty on the nature of reality. However, in contrast to the human mind, in which uncertainties often remain in an abstract or vague form, and where this vagueness might even suppress surfacing competing concretizations, MR systems as implemented in TransforMR and Scene Responsiveness often require some assignment of visual representations to these fundamental uncertainty-induced "vaguenesses". As known from the problem of "medium specificity" in the arts, where cinematic media requires more concrete depictions of sceneries, while literature remains abstract in words to exploit the reader's fantasy, the need for an assignment of a visual representation in MR immediate elicits the need to resolve uncertainty by constructing a plausible visualization for the sake of a most useful concretization.

As a result of this consideration, in the future, we expect to see such MR systems striving for a more holistic representation of reality by taking a constructivist or imaginative component that substantiates areas of uncertainty. This expectation is also nurtured by the trends toward the wider use of generative AI.

7.2.2.4 Designing with Priors

The notion of holistic semantic awareness implies the need for an interpretative context that corresponds with the user's perception of reality. All situated and semantics-aware systems presented in this thesis have corroborated this need for strong assumptions of how the user and the world "work". While this has always been a foundational principle in any kind of user interface design, the long-term ambition to build systems that blend with the user's space and mind further cements this principle.

Rather than building an understanding of the world and the user on the fly, an information-efficient procedure of interpreting incoming signals is bound to prior assumptions of how the world works–a notion that has a long-standing tradition in statistics and machine learning and

is also found in neurally based computer vision [49]. This is in line with the many perceptual mechanisms employed by humans to perceive a situation such as filling-in, predictive coding, and contextual inference. At the same time, to construct models that not only correspond in an arbitrary way with physical reality but also correspond with the current user's specific ways of reconstructing reality, these priors themselves must correspond in some ways with the priors applied by humans more generally and the current user specifically.

LLMs might offer a convenient way of "initializing" an understanding of the user and their behavior, the physical environment, and the relationship between the two. However, one of their own main deficiencies currently lies in a lack of a consolidated model of the physical and metaphysical principles in existence. Therefore, while LLMs already penetrated many areas of user interface design, their increasing adoption will only reinforce the need for well-designed user interfaces that have been built with strong assumptions regarding the in-situ use of semantics-aware MR systems.

7.2.2.5 Designing from an Integrative Perspective

Finally, striving for MR systems that feature a more holistic awareness of the physical world and the user therein is built upon the emergentistic acknowledgment that the overall semantics of a scene is more than the sum of its elements.

While MR platforms will inevitably increase their capabilities of gaining awareness of individual logical units with implementations of well-defined computer vision problems as described above, from objects in indoor sceneries up to a fine-grained understanding of animals and human bodies, along with the ability to potentially even assign specific states and shapes of the unit of focus, the question of how to model object-to-object relationships and user-to-object relationships purposefully remains an open research question. While scene graphs that model the relationships between objects offer a useful formal structure, the nature of what actually constitutes individual relationships between objects is highly coupled to the specific purpose of an application. Even more, the relationship between a user and the objects surrounding them is highly dependent on situational parameters. As discussed in the Scene Responsiveness project, understanding affordance in objects and parameterizing these affordances is bound to an understanding of an object's potential functions which directly translates to the question of how a specific user might intend to make use of an object. Again, progress in AI promises to enable to statistically learn from data how users make of use objects, what triggers their behavior, and how sets of objects are used together.

7.2.2.6 Conclusion

In conclusion, to approach the envisioned future of a more holistic approach in semantics-aware scene understanding and reconstruction for MR experiences, applications, and interactions, we expect a shift in MR systems that follow a constructivist paradigm interpreting signals against the backdrop of expectations, integrating insights into a broader aggregate of scene constituents, and taking into account uncertainties involved, even in aspects that are momentarily or always invisible to sensors. This desire to reconstruct reality, disregarding this lack of directly related sensor information, adds a foundational new quality in perspective. The system is required to transcend its momentary state, elevating itself from a mere instantaneous signal processor to a long-lived modeler of physical reality, similar in some aspects to the user's perceptual process and different in other aspects.

7.2.3 Phase 3: Complementing Semantic Awareness with Semantically Informed Physical Control

As corroborated by this dissertation's progress, the ability to situationally and semantically control the information flow from physical reality into the human perceptual system is an intimate and powerful technical tool to shape a user's perceived reality meaningfully. At the same time, semantic awareness can be leveraged not only to change informational input but also physical configuration directly by means of physical manipulation through robots.

From our lens of semantics-aware and situated MR, such robots not only act as a disparate entity next to the MR system but are instead part of a distributed MR system, partially worn and integrated on and with the user's body, and partially placed in the environment. As a result of this integrative perspective, the real-time digital representation of the environment, the user, and the relationship between both is unified and shared between the user device on the one hand and the robot on the other hand. In consequence, robotic action can be directly aligned with human need, intention, and understanding as far as the presented notion of a situated computing device enables it.

As soon as the user-borne device estimates the intention of grasping a far-away object, the robot can spring to action and hand it over. Similarly, many of the myriad applications that exist at the interface of MR and robotics [321] may benefit from a semantic awareness that is aligned with the user's understanding of reality. In particular, however, we also see a potential to directly transfer concrete concepts presented in this dissertation to the field of semantics-informed robotic mixed reality.

Borrowing SoundsRide's concept for estimating user motion through space while being semantically aware of the physical scenery opens the field for a plethora of robotic applications, e.g., employing a quadcopter to automatically capture cinematic camera shots with high dynamics involving both the subject and the environmental points of interest, or to scout the route ahead in curvy environments. Integrating concepts of visual scene manipulation as contributed in TransforMR and Scene Responsiveness with the robotic ability for physical scene manipulation can go beyond the classical incorporation of robots in MR. Instead of either fully eliminating the physical scene and thus the robot by immersing the user in a disparate VR environment, or fully showing the robot as in classical robotic AR visualization, the concept of semantic transformation as presented in TransforMR or reality state toggling as presented in Scene Responsiveness enable the robot to become a transformed or even hidden actor. We envision a variety of applications for such an approach ranging from "subliminal robots" that help realize the ambition of responsive architecture to intriguing user experiences where physical items seem to appear out of thin air wherever and whenever they are needed.

Resulting from the perspective of making the robot part of a larger distributed MR system, the design of such systems can also benefit from the above-discussed insights for holistically semantics-aware MR experiences, by accounting for uncertainty, constructivism, prior knowledge, and integration. In particular, as LLMs and AI more generally hope to offer solutions to some of the long-standing problems in the field of robotics [55], the need for situated and semantics-aware MR systems development specifically will increase in importance.

7.3 Macro-Conclusion

In 1965, Sutherland formulated what we call VR today as the quest for the ultimate display that "can control the existence of matter" [318] to visually, haptically, and acoustically render any desired scenery.

As if that weren't enough, situated and semantics-aware MR as presented in this dissertation complements this quest with the need for the "ultimate digital replica" that perfectly represents the world and the user therein. To this end, at any given point in time, the MR system shall model the user's relevant environment and what governs it; it shall model what the user thinks, wants, and does; it shall model the relationship between the user and the environment; it shall situate informational and physical elements inside that environment or at least at the interface between the user and the environment (i.e., at the user's sensory system); for many applications, it shall project the aforementioned aspects for the next seconds or minutes or even longer using information from the past; and then, it shall produce exactly the perceivable reality as it is needed for the user's concrete problem to be solved. All of this shall be executed on hardware that is comfortable, fast, non-obtrusive, cheap, available when needed, and undisturbing when not.

The pursuit of these thought experiments[1] brings us to the limits of physics, biology, engineering, philosophy, psychology, neuroscience, computer science, and AI. Virtually every major hardware component in the most modern MR platforms of today, astonishing as they might be already and nearly unbelievable artifacts of human creativity they already are, are dissatisfactory from batteries to chips to display, from sensors to algorithms to interactions, when compared to the above vision, motivating many more research breakthroughs and product innovations in the future.

However, even without an ultimate display and without an ultimate replica, we can explore, design, implement, study, and utilize MR systems that manifest parts of the potential of the outlined vision by specifically designing for select aspects of reality, the user, and the relationship between the two to control what the user senses and perceives as reality. In this dissertation, we presented ideas and technologies that we believe to be useful on the path toward situated computing systems that perceive, process, and produce reality, seamlessly integrating with the user's space and mind.

[1] Both the description of the ultimate display and the ultimate replica take the role of thought experiments. The ultimate display is not in sight, neither as a "force" that manipulates objective physicality nor as an illusion in subjectively perceived reality, and the ultimate replica cannot exist independent of subjective construction.

Bibliography

[1] Abtahi, P., Landry, B., Yang, J., Pavone, M., Follmer, S., and Landay, J. A. Beyond the Force: Using quadcopters to appropriate objects and the environment for haptics in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[2] Albert, H. *Treatise on critical reason.* Princeton University Press, 1985.

[3] Albrecht, R., Väänänen, R., and Lokki, T. Guided by music: pedestrian and cyclist navigation with route and beacon guidance. *Personal and Ubiquitous Computing 20*, 1 (2016), 121–145.

[4] Anson, E. The semantics of graphical input. *ACM SIGGRAPH Computer Graphics 13*, 2 (1979), 113–120.

[5] Apple. ARKit. https://developer.apple.com/augmented-reality/arkit/, 2022.

[6] Apple. Apple Watch Ultra 2 - Technical Specifications. https://support.apple.com/en-us/111832, 2023.

[7] Apple. Introducing Apple Vision Pro: Apple's first spatial computer. https://www.apple.com/newsroom/2023/06/introducing-apple-vision-pro/, 2023.

[8] Apple. iPhone 15 Pro and iPhone 15 Pro Max - Technical Specification. https://www.apple.com/iphone-15/specs/, 2023.

[9] Apple. Apple Vision Pro - Technical Specifications. https://www.apple.com/apple-vision-pro/specs/, 2024.

[10] Argelaguet, F., and Andujar, C. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics 37* (2013), 121 – 136.

[11] Arora, R., Kazi, R. H., Anderson, F., Grossman, T., Singh, K., and Fitzmaurice, G. W. Experimental evaluation of sketching on surfaces in VR. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 5643–5654.

[12] Auda, J., Gruenefeld, U., Faltaous, S., Mayer, S., and Schneegass, S. A scoping survey on cross-reality systems. *ACM Computing Surveys 56*, 4 (2023), 1–38.

[13] Avetisyan, A., Khanova, T., Choy, C., Dash, D., Dai, A., and Niessner, M. SceneCAD: Predicting object alignments and layouts in RGB-D scans. In *Proceedings of the European Conference on Computer Vision* (2020).

[14] Ayer, A. J. Editor's instructions. In *Logical positivism*, A. J. Ayer, Ed., vol. 2. Simon and Schuster, 1959.

[15] Azmandian, M., Hancock, M., Benko, H., Ofek, E., and Wilson, A. D. Haptic Retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 1968–1979.

[16] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., and MacIntyre, B. Recent advances in augmented reality. *IEEE Computer Graphics and Applications 21*, 6 (2001), 34–47.

[17] Azuma, R. T. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments 6*, 4 (1997), 355–385.

[18] Babic, T., Reiterer, H., and Haller, M. Pocket6: A 6DoF controller based on a simple smartphone application. In *Proceedings of the 2018 ACM Symposium on Spatial User Interaction* (2018), pp. 2–10.

[19] Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Aydin, A., Lüke, K. H., and Schwaiger, R. InCarMusic: Context-aware music recommendations in a car. *Lecture Notes in Business Information Processing 85 LNBIP* (2011), 89–100.

[20] Bar-Hillel, Y. Logical syntax and semantics. *Language 30*, 2 (1954), 230–237.

[21] Baričević, D., Lee, C., Turk, M., Höllerer, T., and Bowman, D. A. A handheld AR magic lens with user-perspective rendering. In *Proceedings of the 2012 IEEE International Symposium on Mixed and Augmented Reality* (2012), pp. 197–206.

[22] Barnum, P., Sheikh, Y., Datta, A., and Kanade, T. Dynamic Seethroughs: Synthesizing hidden views of moving objects. In *Proceedings of the 2009 IEEE International Symposium on Mixed and Augmented Reality* (2009), pp. 111–114.

[23] Barwise, J., and Perry, J. Situations and attitudes. *The Journal of Philosophy 78*, 11 (1981), 668–691.

[24] Bederson, B. B. Audio augmented reality. In *Proceedings of the 1994 CHI Conference on Human Factors in Computing Systems* (1995), pp. 210–211.

[25] Bergé, L.-P., Dubois, E., and Raynal, M. Design and evaluation of an "around the smartphone" technique for 3D manipulations on distant display. In *Proceedings of the 3rd ACM Symposium on Spatial User Interaction* (2015), pp. 69–78.

[26] Bergström, J., Dalsgaard, T.-S., Alexander, J., and Hornbæk, K. How to evaluate object selection and manipulation in VR? Guidelines from 20 years of studies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–20.

[27] Biederman, I. On the semantics of a glance at a scene. In *Perceptual organization*. Routledge, 2017, pp. 213–253.

[28] Bier, E. A., Stone, M. C., Pier, K., Buxton, W., and DeRose, T. D. ToolGlass and magic lenses: the see-through interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (1993), pp. 73–80.

[29] Billinghurst, M., Clark, A., Lee, G., et al. A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction 8*, 2-3 (2015), 73–272.

[30] Billinghurst, M., Poupyrev, I., Kato, H., and May, R. Mixing realities in shared space: An augmented reality interface for collaborative computing. In *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo.* (2000), vol. 3, pp. 1641–1644.

[31] Bimber, O., and Raskar, R. *Spatial augmented reality: merging real and virtual worlds.* CRC Press, 2005.

[32] Bolt, R. A. "Put-that-there" voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and interactive Techniques* (1980), pp. 262–270.

[33] Boring, S., Baur, D., Butz, A., Gustafson, S., and Baudisch, P. Touch Projector: Mobile interaction through video. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems* (2010), pp. 2287–2296.

[34] Boring, S., Jurmu, M., and Butz, A. Scroll, Tilt or Move it: Using mobile phones to continuously control pointers on large public displays. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7* (2009), pp. 161–168.

[35] Bowman, D. A., and Hodges, L. F. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Proceedings of the 1997 ACM Symposium on Interactive 3D Graphics* (1997).

[36] Braunhofer, M., Kaminskas, M., and Ricci, F. Location-aware music recommendation. *International Journal of Multimedia Information Retrieval 2*, 1 (2013), 31–44.

[37] Breen, D. E., Whitaker, R. T., Rose, E., and Tuceryan, M. Interactive occlusion and automatic object placement for augmented reality. *Computer Graphics Forum 15*, 3 (1996), 11–22.

[38] Bressa, N., Korsgaard, H., Tabard, A., Houben, S., and Vermeulen, J. What's the situation with situated visualization? A survey and perspectives on situatedness. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (Jan. 2022), 107–117.

[39] Brodsky, W. The effects of music tempo on simulated driving performance and vehicular control. *Transportation Research Part F: Traffic Psychology and Behaviour 4*, 4 (2001), 219–241.

[40] Brooks, F. P. Grasping reality through illusion—interactive graphics serving science. In *Proceedings of the 1998 CHI Conference on Human Factors in Computing Systems* (1988), pp. 1–11.

[41] Brooks, F. P., Ouh-Young, M., Batter, J. J., and Jerome Kilpatrick, P. Project GROPE: Haptic displays for scientific visualization. *ACM SIGGRAPH Computer Graphics 24*, 4 (1990), 177–185.

[42] Brown, E., and Cairns, P. A grounded investigation of immersion in games. In *Extended Abstracts of the 2004 CHI Conference on Human Factors in Computing Systems* (2004), pp. 31–32.

[43] Burnett, G., Hazzard, A., Crundall, E., and Crundall, D. Altering speed perception through the subliminal adaptation of music within a vehicle. In *Proceedings of the 9th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2017), pp. 164–172.

[44] Carnap, R. *Meaning and necessity: A study in semantics and modal logic*, vol. 30. University of Chicago Press, 1947.

[45] Carnap, R. Empiricism, semantics, and ontology. *Revue internationale de philosophie* (1950), 20–40.

[46] Casas, L., Fauconneau, M., Kosek, M., Mclister, K., and Mitchell, K. Image based proximate shadow retargeting. In *Proceedings of the Conference on Computer Graphics and Visual Computing* (2018), pp. 43–50.

[47] Casiez, G., Roussel, N., and Vogel, D. 1 Euro Filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems* (2012), pp. 2527–2530.

[48] Casiez, G., Vogel, D., Balakrishnan, R., and Cockburn, A. The impact of control-display gain on user performance in pointing tasks. *Human-Computer Interaction 23* (2008), 215–250.

[49] Cavalli, L. *On Priors for Robust Pose Estimation*. PhD thesis, ETH Zurich, 2023.

[50] Chalmers, D. J. *Reality+: Virtual worlds and the problems of philosophy*. Penguin UK, 2022.

[51] Chang, Y.-L., Liu, Z. Y., Lee, K.-Y., and Hsu, W. Learnable gated temporal shift module for deep video inpainting. In *Proceedings of the 30th British Machine Vision Conference* (2019).

[52] Chaurasia, G., Nieuwoudt, A., Ichim, A.-E., Szeliski, R., and Sorkine-Hornung, A. Passthrough+: Real-time stereoscopic view synthesis for mobile mixed reality. *Proceedings of the ACM on Compututer Graphics and Interactive Technologies 3*, 1 (2020).

[53] Chen, D., Liao, J., Yuan, L., Yu, N., and Hua, G. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1105–1114.

[54] Chen, Y., Katsuragawa, K., and Lank, E. Understanding viewport-and world-based pointing with everyday smart devices in immersive augmented reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13.

[55] Cheng, G., Zhang, C., Cai, W., Zhao, L., Sun, C., and Bian, J. Empowering large language models on robotic manipulation with affordance prompting. *arXiv preprint arXiv:2404.11027* (2024).

[56] Cheng, L., Ofek, E., Holz, C., and Wilson, A. D. VRoamer: Generating on-the-fly VR experiences while walking inside large, unknown real-world building environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces* (2019), pp. 359–366.

[57] Cheng, L.-P., Ofek, E., Holz, C., Benko, H., and Wilson, A. D. Sparse Haptic Proxy: Touch feedback in virtual environments using a general passive prop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 3718–3728.

[58] Cheng, L.-P., Ofek, E., Holz, C., and Wilson, A. D. VRoamer: Generating on-the-fly VR experiences while walking inside large, unknown real-world building environments. In *Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces* (2019), pp. 359–366.

[59] Cheng, Y., Yan, Y., Yi, X., Shi, Y., and Lindlbauer, D. SemanticAdapt: Optimization-based adaptation of mixed reality layouts leveraging virtual-physical semantic connections. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology* (2021), pp. 282–297.

[60] Cheng, Y. F., Yin, H., Yan, Y., Gugenheimer, J., and Lindlbauer, D. Towards understanding diminished reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–16.

[61] Cheng, Z., and Shen, J. On effective location-aware music recommendation. *ACM Transactions on Information Systems 34*, 2 (2016).

[62] Chidambaram, S., Reddy, S. S., Rumple, M., Ipsita, A., Villanueva, A., Redick, T., Stuerzlinger, W., and Ramani, K. EditAR: A digital twin authoring environment for creation of AR/VR and video instructions from a single demonstration. In *Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality* (2022), pp. 326–335.

[63] Chomsky, N. Logical syntax and semantics: Their linguistic relevance. *Language 31*, 1 (1955), 36–45.

[64] Chomsky, N. Linguistic contributions to the study of mind: Future. *Language and thinking* (1968), 323–364.

[65] Chomsky, N. *Language and mind*. Cambridge University Press, 2006.

[66] Collins, K. An introduction to procedural music in video games. *Contemporary Music Review 28*, 1 (2009), 5–15.

[67] Cools, R., Esteves, A., and Simeone, A. L. Blending Spaces: Cross-reality interaction techniques for object transitions between distinct virtual and augmented realities. In *Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality* (2022), pp. 528–537.

[68] Čopič Pucihar, K., Coulton, P., and Alexander, J. The use of surrounding visual context in handheld ar: device vs. user perspective rendering. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems* (2014), pp. 197–206.

[69] Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., and Hart, J. C. The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM 35*, 6 (1992), 64–73.

[70] Ctrl-labs at Reality Labs, Sussillo, D., Kaifosh, P., and Reardon, T. A generic noninvasive neuromotor interface for human-computer interaction. *bioRxiv* (2024), 2024–02.

[71] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 764–773.

[72] Davidson, D. *Subjective, intersubjective, objective*, vol. 3. OUP Oxford, 2001.

[73] Devrio, N., and Harrison, C. DiscoBand: Multiview depth-sensing smartwatch strap for hand, body and environment tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (2022), pp. 1–13.

[74] Dias, P., Afonso, L., Eliseu, S., and Santos, B. S. Mobile devices for interaction in immersive virtual environments. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces* (2018).

[75] DigiLens Inc. Principles of DigiLens waveguides. `https://www.youtube.com/watch?v=aZum_COPUek`, 2019.

[76] Ding, Y., Yang, Q., Li, Y., Yang, Z., Wang, Z., Liang, H., and Wu, S.-T. Waveguide-based augmented reality displays: perspectives and challenges. *eLight 3*, 1 (2023), 24.

[77] Dong, Z.-C., Wu, W., Xu, Z., Sun, Q., Yuan, G., Liu, L., and Fu, X.-M. Tailored Reality: Perception-aware scene restructuring for adaptive VR navigation. *ACM Transactions on Graphics 40*, 5 (2021), 1–15.

[78] Du, R., Turner, E., Dzitsiuk, M., Prasso, L., Duarte, I., Dourgarian, J., Afonso, J., Pascoal, J., Gladstone, J., Cruces, N., et al. DepthLab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (2020), pp. 829–843.

[79] Eckert Jr, J. P., Weiner, J. R., Welsh, H. F., and Mitchell, H. F. The UNIVAC system. In *Joint AIEE-IRE Computer Conference: Review of Electronic Digital Computers* (1951), pp. 6–16.

[80] Ehrlich, S. K., Agres, K. R., Guan, C., and Cheng, G. A closed-loop, music-based brain-computer interface for emotion mediation. *PLoS ONE 14*, 3 (2019), 1–24.

[81] Elliott, G. T., and Tomlinson, B. PersonalSoundtrack: Context-aware playlists that adapt to user pace. *Extended Abstracts of the 2006 CHI Conference on Human Factors in Computing Systems* (2006), 736–741.

[82] Fakhrhosseini, S. M., Landry, S., Tan, Y. Y., Bhattarai, S., and Jeon, M. If you're angry, turn the music on: Music can mitigate anger effects on driving performance. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2014).

[83] Feiner, S., MacIntyre, B., and Seligmann, D. Knowledge-based augmented reality. *Communications of the ACM 36*, 7 (1993), 53–62.

[84] Feiner, S. K. Augmented reality: A new way of seeing. *Scientific American 286*, 4 (2002), 48–55.

[85] Fender, A. R., and Holz, C. Causality-preserving asynchronous reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–15.

[86] Feuchtner, T., and Müller, J. Ownershift: Facilitating overhead interaction in virtual reality with an ownership-preserving hand space shift. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018), pp. 31–43.

[87] Feyerabend, P. Wittgenstein's philosophical investigations. *The Philosophical Review 64*, 3 (1955), 449–483.

[88] Feyerabend, P. Materialism and the mind-body problem. *The Review of Metaphysics* (1963), 49–66.

[89] Feyerabend, P. *Against method: Outline of an anarchistic theory of knowledge*. Verso Books, 197.

[90] Fisher, S. S., McGreevy, M., Humphries, J., and Robinett, W. Virtual environment display system. In *Proceedings of the 1986 workshop on Interactive 3D graphics* (1987), pp. 77–87.

[91] Fitzmaurice, G. W. Situated information spaces and spatially aware palmtop computers. *Communications of the ACM 36*, 7 (1993), 39–49.

[92] Fleck, P., Sousa Calepso, A., Hubenschmid, S., Sedlmair, M., and Schmalstieg, D. RagRug : A toolkit for situated analytics. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[93] Forster, C., Carlone, L., Dellaert, F., and Scaramuzza, D. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics 33*, 1 (2016), 1–21.

[94] Forster, C., Zhang, Z., Gassner, M., Werlberger, M., and Scaramuzza, D. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics 33*, 2 (2016), 249–265.

[95] Frege, G. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik 100* (1892), 25–50.

[96] Frege, G. *Logische Untersuchungen*, 4 ed. Vandenhoeck und Ruprecht, Göttingen, 1992.

[97] Freudenthal, H. *LINCOS: Design of a language for cosmic intercourse. Part I.* North-Holland Publishing Company, 1960.

[98] Frid, E., Gomes, C., and Jin, Z. Music Creation by Example. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–13.

[99] Gabriel, R. 3D character creation in Blender. `https://www.youtube.com/playlist?list=PL5TzNMWlAlkqmoaVC7teQsI9B8LG6Upf5`, 2020.

[100] Gal, R., Shapira, L., Ofek, E., and Kohli, P. FLARE: Fast layout for augmented reality applications. In *Proceedings of the 2014 IEEE international Symposium on Mixed and Augmented Reality* (2014), pp. 207–212.

[101] Gao, C., Saraf, A., Huang, J.-B., and Kopf, J. Flow-edge guided video completion. In *Proceedings of the European Conference on Computer Vision* (2020).

[102] Georgiou, Y., and Kyza, E. A. The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings. *International Journal of Human-Computer Studies 98* (2017), 24–37.

[103] Gibson, J. J. The theory of affordances. In *The Ecological Approach to Visual Perception*. 1977, pp. 67–82.

[104] Goldman, A. I. *Epistemology and cognition*. Harvard University Press, 1986.

[105] Gonzalez-Franco, M., and Peck, T. C. Avatar embodiment. Towards a standardized questionnaire. *Frontiers in Robotics and AI 5* (2018), 74.

[106] Graf, H., and Jung, K. The smartphone as a 3D input device: Using accelerometer and gyroscope for 3D navigation with a smartphone. In *Proceedings of the IEEE International Conference on Consumer Electronics* (2012), pp. 254–257.

[107] Gregory, R. L. *The Intelligent Eye*. McGraw-Hill, 1970.

[108] Grønbæk, J. E., Pfeuffer, K., Velloso, E., and Gellersen, H. Partially blended realities: Aligning dissimilar spaces for distributed mixed reality meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).

[109] Grossman, T., and Balakrishnan, R. The design and evaluation of selection techniques for 3D volumetric displays. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (2006), pp. 3–12.

[110] Grubert, J., Langlotz, T., Zollmann, S., and Regenbrecht, H. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE Transactions on Visualization and Computer Graphics 23*, 6 (2016), 1706–1724.

[111] Gruen, R., Ofek, E., Steed, A., Gal, R., Sinclair, M., and Gonzalez-Franco, M. Measuring system visual latency through cognitive latency on video see-through AR devices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces* (2020), pp. 791–799.

[112] Guan, P., Penner, E., Hegland, J., Letham, B., and Lanman, D. Perceptual requirements for world-locked rendering in AR and VR. In *Proceedings of the SIGGRAPH Asia 2023 Conference* (2023), pp. 1–10.

[113] Guida, J., and Sra, M. Augmented reality world editor. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology* (2020), pp. 1–2.

[114] Guven, S., and Feiner, S. Interaction techniques for exploring historic sites through situated media. In *2006 IEEE Symposium on 3D User Interfaces* (2006), pp. 111–118.

[115] Guven, S., Feiner, S., and Oda, O. Mobile augmented reality interaction techniques for authoring situated media on-site. In *2006 IEEE International Symposium on Mixed and Augmented Reality* (2006), pp. 235–236.

[116] Haefner, B., Green, S., Oursland, A., Andersen, D., Goesele, M., Cremers, D., Newcombe, R., and Whelan, T. Recovering real-world reflectance properties and shading from HDR imagery. In *2021 International Conference on 3D Vision* (2021), pp. 1075–1084.

[117] Hagerer, G. J., Lux, M., Ehrlich, S., and Cheng, G. Augmenting affect from speech with generative music. In *Extended Abstracts of the 2015 CHI Conference on Human Factors in Computing Systems* (2015), vol. 18, pp. 977–982.

[118] Han, K., Rezende, R. S., Ham, B., Wong, K.-Y. K., Cho, M., Schmid, C., and Ponce, J. SCNet: Learning semantic correspondence. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1831–1840.

[119] Han, S., Liu, B., Cabezas, R., Twigg, C. D., Zhang, P., Petkau, J., Yu, T.-H., Tai, C.-J., Akbay, M., Wang, Z., et al. MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics 39*, 4 (2020), 87–1.

[120] Hancock, M., Carpendale, S., and Cockburn, A. Shallow-Depth 3D Interaction: Design and evaluation of one-, two-and three-touch techniques. In *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems* (2007), pp. 1147–1156.

[121] Harrison, C., Benko, H., and Wilson, A. D. OmniTouch: Wearable multitouch interaction everywhere. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (2011), pp. 441–450.

[122] Harrison, S., Tatar, D., and Sengers, P. The three paradigms of HCI. *Virgina Tech* (2007).

[123] Hartmann, J., Gupta, A., and Vogel, D. Extend, Push, Pull: Smartphone mediated interaction in spatial augmented reality via intuitive mode switching. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction* (2020), pp. 1–10.

[124] Hartmann, J., Holz, C., Ofek, E., and Wilson, A. D. RealityCheck: Blending virtual environments with situated physical reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.

[125] Hartmann, J., and Vogel, D. An examination of mobile phone pointing in surface mapped spatial augmented reality. *International Journal of Human-Computer Studies 153* (2021).

[126] Hayatpur, D., Heo, S., Xia, H., Stuerzlinger, W., and Wigdor, D. Plane, Ray, and Point: Enabling precise spatial manipulations with shape constraints. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), pp. 1185–1195.

[127] He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2961–2969.

[128] Helmholz, P., Vetter, S., and Robra-Bissantz, S. AmbiTune: Bringing context-awareness to music playlists while driving. *Lecture Notes in Computer Science 8463* (2014), 393–397.

[129] Helmholz, P., Ziesmann, E., and Robra-Bissantz, S. Context-awareness in the car: Prediction, evaluation and usage of route trajectories. *Lecture Notes in Computer Science 7939* (2013), 413–419.

[130] Herling, J., and Broll, W. Advanced self-contained object removal for realizing real-time diminished reality in unconstrained environments. In *Proceedings of the 2010 IEEE International Symposium on Mixed and Augmented Reality* (2010), pp. 207–212.

[131] Hettiarachchi, A., and Wigdor, D. Annexing Reality: Enabling opportunistic use of everyday objects as tangible proxies in augmented reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 1957–1967.

[132] Hinckley, K., Pausch, R., Goble, J. C., and Kassell, N. F. A survey of design issues in spatial input. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology* (1994), pp. 213–222.

[133] Hirzle, T., Müller, F., Draxler, F., Schmitz, M., Knierim, P., and Hornbæk, K. When XR and AI meet: A scoping review on extended reality and artificial intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–45.

[134] Hollerer, T., Feiner, S., and Pavlik, J. Situated documentaries: Embedding multimedia presentations in the real world. In *Digest of Papers. Third International Symposium on Wearable Computers* (1999), pp. 79–86.

[135] Hollerer, T. H. *User interfaces for mobile augmented reality systems.* PhD thesis, Columbia University, 2004.

[136] Hori, R., Hachiuma, R., Saito, H., Isogawa, M., and Mikami, D. Silhouette-based synthetic data generation for 3D human pose estimation with a single wrist-mounted 360° camera. In *Proceedings of the 2021 IEEE International Conference on Image Processing* (2021), pp. 1304–1308.

[137] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[138] Hu, H.-N., Cai, Q.-Z., Wang, D., Lin, J., Sun, M., Krahenbuhl, P., Darrell, T., and Yu, F. Joint monocular 3D vehicle detection and tracking. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 5390–5399.

[139] Hu, X., Deng, J., Zhao, J., Hu, W., Ngai, E. C., Wang, R., Shen, J., Liang, M., Li, X., Leung, V. C., and Kwok, Y. K. SAfeDJ: A crowd-cloud codesign approach to situation-aware music delivery for drivers. *ACM Transactions on Multimedia Computing, Communications and Applications 12*, 1 (2015).

[140] Huang, H., Solah, M., Li, D., and Yu, L. F. Audible Panorama: Automatic spatial audio generation for panorama imagery. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Chi (2019), 1–11.

[141] Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., Li, Z., and Liu, W. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 783–791.

[142] Humane, Inc. Humane AI Pin. https://humane.com/aipin, 2024.

[143] Hume, D. *An enquiry concerning human understanding.* Open Court Publishing Company, 1748. Reprint 1900.

[144] Huynh, B., Ibrahim, A., Chang, Y. S., Höllerer, T., and O'Donovan, J. User perception of situated product recommendations in augmented reality. *International Journal of Semantic Computing 13*, 03 (2019), 289–310.

[145] Huynh, B., Orlosky, J., and Höllerer, T. Designing a multitasking interface for object-aware ar applications. In *Adjunct Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (2020), pp. 39–40.

[146] iFixit. Vision Pro Teardown Part 2 - Is the Apple Vision Pro Really 4K? https://youtu.be/wt22M5nWJ4Q, 2024.

[147] Izadinia, H., Shan, Q., and Seitz, S. M. IM2CAD. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5134–5143.

[148] Jadid, A., Rudolph, L., Pankratz, F., and Klinker, G. Utilizing multiple calibrated imus for enhanced mixed reality tracking. In *Adjunct Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (2019), pp. 384–386.

[149] Jin, H., Holz, C., and Hornbaek, K. Tracko: Ad-hoc Mobile 3D Tracking Using Bluetooth Low Energy and Inaudible Signals for Cross-Device Interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (2015), pp. 147–156.

[150] Jo, D., Kim, K.-H., and Kim, G. J. SpaceTime: Adaptive control of the teleported avatar for improved AR tele-conference experience. *Computer Animation and Virtual Worlds 26*, 3-4 (2015), 259–269.

[151] Johanson, M. The Turing test for telepresence. *arXiv preprint arXiv:1511.02590* (2015).

[152] Jones, B., Sodhi, R., Murdock, M., Mehra, R., Benko, H., Wilson, A., Ofek, E., MacIntyre, B., Raghuvanshi, N., and Shapira, L. RoomAlive: Magical experiences enabled by scalable, adaptive projector-camera units. In *Proceedings of the 27th Annual ACM Aymposium on User Interface Software and Technology* (2014), pp. 637–644.

[153] Kalkofen, D., Mendez, E., and Schmalstieg, D. Interactive focus and context visualization for augmented reality. In *Proceedings of the 2007 IEEE International Symposium on Mixed and Augmented Reality* (2007), pp. 191–201.

[154] Kalkofen, D., Sandor, C., White, S., and Schmalstieg, D. Visualization techniques for augmented reality. In *Handbook of Augmented Reality*, B. Furht, Ed. Springer, 2011, pp. 65–98.

[155] Kalkofen, D., Veas, E., Zollmann, S., Steinberger, M., and Schmalstieg, D. Adaptive ghosted views for augmented reality. In *Proceedings of the 2013 IEEE International Symposium on Mixed and Augmented Reality* (2013), pp. 1–9.

[156] Kaptelinin, V., and Nardi, B. Affordances in hci: Toward a mediated action perspective. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems* (2012), pp. 967–976.

[157] Kari, M. Method for supporting a user of a motor vehicle, 2022. DPMA, DE10 2022 126 876.4.

[158] Kari, M., and Bethge, D. Devices and methods for joint route guidance, 2021. DPMA, DE10 2021 130 939.5.

[159] Kari, M., and Bethge, D. Procedure for determining spare parts requirements, 2021. DPMA, DE10 2021 129 094.5.

[160] Kari, M., and Bethge, D. Method and device for monitoring objects, 2022. DPMA, DE10 2022 110 349.8.

[161] Kari, M., and Bethge, D. Method for operating an automated maneuvering system of a vehicle, 2022. DPMA, DE10 2022 105 245.1, filed and under review.

[162] Kari, M., and Dreharov, N. Method for providing an evaluation system, 2022. DPMA, DE10 2022 120 936.9.

[163] Kari, M., Grosse-Puppendahl, T., Coelho, L. F., Fender, A. R., Bethge, D., Schütte, R., and Holz, C. TransforMR: Pose-Aware Object Substitution for Composing Alternate Mixed Realities. In *Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality* (2021), pp. 69–79.

[164] Kari, M., Grosse-Puppendahl, T., Jagaciak, A., and Bethge, D. Method and system for scene-synchronous selection and playback of audio sequences for a motor vehicle, 2021. DPMA, DE10 2021 110 268.5.

[165] Kari, M., Grosse-Puppendahl, T., Jagaciak, A., Bethge, D., Schütte, R., and Holz, C. SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology* (2021), Best Paper Award, pp. 118–133.

[166] Kari, M., and Holz, C. HandyCast: Phone-based Bimanual Input for Virtual Reality in Mobile and Space-Constrained Settings via Pose-and-Touch Transfer. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–15.

[167] Kari, M., and Isele, S. T. Vehicle and control device and procedures for operating the vehicle, 2022. DPMA, DE10 2022 105 155.2.

[168] Kari, M., and Kerl, P. Computer-implemented method for transmitting information about a headlight of a motor vehicle with multiple light sources, 2022. DPMA, DE10 2022 113 682.5.

[169] Kari, M., Schütte, R., and Sodhi, R. Scene Responsiveness for Visuotactile Illusions in Mixed Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023), Best Paper Honorable Mention Award, pp. 1–15.

[170] Kari, M., and Sodhi, R. Techniques and graphics-processing aspects for enabling scene responsiveness in mixed-reality environments, including by using situated digital twins, and systems and methods of use thereof, 2023. USPTO, US 63/494,449, filed and under review.

[171] Kari, M., and Thum, M. Method, system and computer program product for analyzing time series datasets of an entity, 2021. DPMA, DE10 2021 130 938.7.

[172] Katzakis, N., and Hori, M. Mobile phones as 3-DOF controllers: A comparative study. In *Proceedings of the 8th IEEE International Symposium on Dependable, Autonomic and Secure Computing* (2009), pp. 345–349.

[173] Katzakis, N., and Hori, M. Mobile devices as multi-DOF controllers. In *2010 IEEE Symposium on 3D User Interfaces* (2010), pp. 139–140.

[174] Katzakis, N., Hori, M., Kiyokawa, K., and Takemura, H. Smartphone game controller. *ITE Technical Report 35*, 25 (2011), 55–60.

[175] Katzakis, N., Hori, M., Kiyokawa, K., and Takemura, H. Plane-Casting: 3D cursor control with a smartphone. In *Proceedings of The 3rd Dimension of CHI: Touching and Designing 3D User Interfaces* (2012), pp. 13–21.

[176] Katzakis, N., Teather, R. J., Kiyokawa, K., and Takemura, H. INSPECT: Extending plane-casting for 6-DOF control. *Human-centric Computing and Information Sciences 5* (2015), 22.

[177] Kawai, N., Sato, T., and Yokoya, N. Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics 22*, 3 (2016), 1236–1247.

[178] Kerl, C., Sturm, J., and Cremers, D. Dense visual SLAM for RGB-D cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2013), pp. 2100–2106.

[179] Kido, D., Fukuda, T., and Yabuki, N. Diminished reality system with real-time object detection using deep learning for onsite landscape simulation during redevelopment. *Environmental Modelling and Software 131*, June (2020), 104759.

[180] Kim, D., Hilliges, O., Izadi, S., Butler, A. D., Chen, J., Oikonomidis, I., and Olivier, P. Digits: Freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (2012), pp. 167–176.

[181] Kim, D., Woo, S., Lee, J.-Y., and Kweon, I. S. Recurrent temporal aggregation framework for deep video inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence 42*, 5 (2019), 1038–1052.

[182] Kim, D., Woo, S., Lee, J.-Y., and So Kweon, I. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5792–5801.

[183] Kim, H., Kim, T., Lee, M., Kim, G. J., and Hwang, J.-I. Don't bother me: How to handle content-irrelevant objects in handheld augmented reality. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology* (2020), pp. 1–5.

[184] Kim, K., Billinghurst, M., Bruder, G., Duh, H. B.-L., and Welch, G. F. Revisiting trends in augmented reality research: A review of the 2nd decade of ISMAR (2008–2017). *IEEE Transactions on Visualization and Computer Graphics 24*, 11 (2018), 2947–2962.

[185] Kim, K., Maloney, D., Bruder, G., Bailenson, J. N., and Welch, G. F. The effects of virtual human's spatial and behavioral coherence with physical objects on social presence in ar. *Computer Animation and Virtual Worlds 28*, 3-4 (2017).

[186] Klein, G., and Murray, D. Parallel tracking and mapping for small AR workspaces. In *Proceedings of the 2007 IEEE International Symposium on Mixed and Augmented Reality* (2007), pp. 225–234.

[187] Kruijff, E., Swan, J. E., and Feiner, S. Perceptual issues in augmented reality revisited. In *Proceedings of the 2010 IEEE International Symposium on Mixed and Augmented Reality* (2010), pp. 3–12.

[188] Krzyzaniak, M., Frohlich, D., and Jackson, P. J. Six types of audio that DEFY reality!: A taxonomy of audio augmented reality with examples. *Proceedings of the 14th International Audio Mostly Conference* (2019), 160–167.

[189] Ku, J., Pon, A. D., and Waslander, S. L. Monocular 3D object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 11867–11876.

[190] Lages, W. S., and Bowman, D. A. Walking with adaptive augmented reality workspaces: Design and usage patterns. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019), pp. 356–366.

[191] Land, M. Z., and McConnell, P. N. Method and apparatus for dynamically composing music and sound effects using a computer entertainment system, 1994. USPTO, US 5,315,057.

[192] Laput, G. *Context-Driven Implicit Interactions*. PhD thesis, Carnegie Mellon University, 2019.

[193] LaViola, J. J., Kruijff, E., McMahan, R. P., Bowman, D. A., and Poupyrev, I. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman, 2017.

[194] Lee, S., Oh, S. W., Won, D., and Kim, S. J. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 4413–4421.

[195] Lee, W. P., Chen, C. T., Huang, J. Y., and Liang, J. Y. A smartphone-based activity-aware system for music streaming recommendation. *Knowledge-Based Systems 131* (2017), 70–82.

[196] LeGendre, C., Ma, W.-C., Fyffe, G., Flynn, J., Charbonnel, L., Busch, J., and Debevec, P. DeepLight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5918–5928.

[197] Li, C., Li, W., Huang, H., and Yu, L.-F. Interactive augmented reality storytelling guided by scene semantics. *ACM Transactions on Graphics 41*, 4 (2022), 1–15.

[198] Liang, J., and Green, M. JDCAD: A highly interactive 3D modeling system. *Computers and Graphics 18*, 4 (1994), 499–506.

[199] Liang, W., Yu, X., Alghofaili, R., Lang, Y., and Yu, L.-F. Scene-aware behavior synthesis for virtual pets in mixed reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–12.

[200] Lim, H., Li, Y., Dressa, M., Hu, F., Kim, J. H., Zhang, R., and Zhang, C. BodyTrak: Inferring full-body poses from body silhouettes using a miniature camera on a wristband. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6*, 3 (2022), 1–21.

[201] Lin, C.-e., Cheng, T. Y., and Ma, X. ARchitect: Building interactive virtual experiences from physical affordances by bringing human-in-the-loop. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13.

[202] Lin, M., Lin, L., Liang, X., Wang, K., and Cheng, H. Recurrent 3D pose sequence machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 810–819.

[203] Lindlbauer, D., and Wilson, A. D. Remixed Reality: Manipulating space and time in augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–13.

[204] Lipari, N. G., and Borst, C. W. Handymenu: Integrating menu selection into a multifunction smartphone-based VR controller. In *Proceedings of the 2015 IEEE Symposium on 3D User Interfaces* (2015), pp. 129–132.

[205] Litany, O., Remez, T., Freedman, D., Shapira, L., Bronstein, A. M., and Gal, R. ASIST: Automatic semantically invariant scene transformation. *Computer Vision and Image Understanding 157* (2017), 284–299.

[206] Liu, H.-K., Shen, I., Chen, B.-Y., et al. NeRF-In: Free-form NeRF inpainting with RGB-D priors. *IEEE Computer Graphics and Applications 44*, 2 (2022).

[207] Liu, Z., Wu, Z., and Tóth, R. SMOKE: Single-stage monocular 3D object detection via keypoint estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 996–997.

[208] Livingston, M. A., Swan, J. E., Gabbard, J. L., Hollerer, T. H., Hix, D., Julier, S. J., Baillot, Y., and Brown, D. Resolving multiple occluded layers in augmented reality. In *Proceedings of the 2003 IEEE International Symposium on Mixed and Augmented Reality* (2003), pp. 56–65.

[209] Locke, J. *An essay concerning human understanding.* The Pennsylvania State University, 1689. reprint 1999.

[210] Looser, J., Billinghurst, M., and Cockburn, A. Through the looking glass: The use of lenses as an interface tool for augmented reality interfaces. In *Proceedings of the 2nd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia* (2004), pp. 204–211.

[211] Looser, J., Grasset, R., and Billinghurst, M. A 3D flexible and tangible magic lens in augmented reality. In *Proceedings of the 2007 IEEE International Symposium on Mixed and Augmented Reality* (2007), pp. 51–54.

[212] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. MediaPipe: A framework for building perception pipelines.

[213] Luo, W., Lehmann, A., Widengren, H., and Dachselt, R. Where should we put it? Layout and placement strategies of documents in augmented reality for collaborative sensemaking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–16.

[214] Magerkurth, C., Cheok, A. D., Mandryk, R. L., and Nilsen, T. Pervasive games: Bringing computer entertainment back to the real world. *Computers in Entertainment (CIE) 3*, 3 (2005), 4–4.

[215] Mandl, D., Yi, K. M., Mohr, P., Roth, P. M., Fua, P., Lepetit, V., Schmalstieg, D., and Kalkofen, D. Learning lightprobes for mixed reality illumination. In *Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality* (2017), pp. 82–89.

[216] Mann, S. Mediated reality with implementations for everyday life. *Presence Connect 1* (2002).

[217] Marwecki, S., and Baudisch, P. Scenograph: Fitting real-walking VR experiences into various tracking volumes. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018), pp. 511–520.

[218] Matulic, F., Ganeshan, A., Fujiwara, H., and Vogel, D. Phonetroller: Visual representations of fingers for precise touch input with mobile phones in VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[219] McCallum, D. C., and Irani, P. ARC-Pad: Absolute+ relative cursor positioning for large displays with a mobile touchscreen. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology* (2009), pp. 153–156.

[220] McGill, M., Brewster, S., McGookin, D., and Wilson, G. Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–16.

[221] Meerits, S., and Saito, H. Real-time diminished reality for dynamic scenes. In *2015 IEEE International Symposium on Mixed and Augmented Reality Workshops* (2015), pp. 53–59.

[222] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.-P., Rhodin, H., Pons-Moll, G., and Theobalt, C. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Transactions on Graphics 39*, 4 (2020), 82–1.

[223] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. Single-shot multi-person 3D pose estimation from monocular RGB. In *Proceedings of the 2018 International Conference on 3D Vision* (2018), pp. 120–130.

[224] Mendes, D., Caputo, F. M., Giachetti, A., Ferreira, A., and Jorge, J. A survey on 3D virtual object manipulation: From the desktop to immersive virtual environments. *Computer Graphics Forum 38*, 1 (2019), 21–45.

[225] Meta. Introducing Meta Reality: A look at the technologies necessary to convincingly blend the virtual and physical worlds. https://www.meta.com/blog/quest/mixed-reality-definition-passthrough-scene-understanding-spatial-anchors/, 2022.

[226] Miksik, O., Vineet, V., Lidegaard, M., Prasaath, R., Niessner, M., Golodetz, S., Hicks, S. L., Pérez, P., Izadi, S., and Torr, P. H. The Semantic Paintbrush: Interactive 3D mapping and recognition in large outdoor spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), pp. 3317–3326.

[227] Milgram, P., Colquhoun, H., et al. A taxonomy of real and virtual world display integration. *Mixed reality: Merging real and virtual worlds 1* (1999), 1–26.

[228] Milgram, P., and Colquhoun Jr, H. W. A framework for relating head-mounted displays to mixed reality displays. 1177–1181.

[229] Milgram, P., and Kishino, F. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems 77*, 12 (1994), 1321–1329.

[230] Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. Augmented Reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and telepresence technologies* (1995), vol. 2351, Spie, pp. 282–292.

[231] Mine, M. R. Virtual environment interaction techniques. Tech. rep., University of North Carolina at Chapel Hill, 1995.

[232] Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K. G., Kelly, J., Brubaker, M. A., Gilitschenski, I., and Levinshtein, A. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2023).

[233] Mohr, P., Mandl, D., Tatzgern, M., Veas, E., Schmalstieg, D., and Kalkofen, D. Retargeting video tutorials showing tools with surface contact to augmented reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 6547–6558.

[234] Mohr, P., Tatzgern, M., Grubert, J., Schmalstieg, D., and Kalkofen, D. Adaptive user perspective rendering for handheld augmented reality. In *2017 IEEE Symposium on 3D User Interfaces* (2017), pp. 176–181.

[235] Mori, S., Erat, O., Broll, W., Saito, H., Schmalstieg, D., and Kalkofen, D. InpaintFusion: Incremental RGB-D inpainting for 3D scenes. *IEEE Transactions on Visualization and Computer Graphics 26*, 10 (2020), 2994–3007.

[236] Mori, S., Herling, J., Broll, W., Kawai, N., Saito, H., Schmalstieg, D., and Kalkofen, D. 3D PixMix: Image inpainting in 3D environments. In *Adjunct Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (2018), pp. 1–2.

[237] Mori, S., Ikeda, S., and Saito, H. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications 9*, 1 (2017), 1–14.

[238] Mori, S., Schmalstieg, D., and Kalkofen, D. Good keyframes to inpaint. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[239] Mori, S., Shibata, F., Kimura, A., and Tamura, H. Efficient use of textured 3D model for pre-observation-based diminished reality. In *Adjunct Proceedings of the 2015 IEEE International Symposium on Mixed and Augmented Reality Workshops* (2015), pp. 32–39.

[240] Mousavian, A., Anguelov, D., Flynn, J., and Kosecka, J. 3D bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 7074–7082.

[241] Mueller, F. F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., Weigel, M., Nanayakkara, S., Obrist, M., Li, Z., Delfa, J., et al. Next steps for human-computer integration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–15.

[242] Muender, T., Reinschluessel, A. V., Drewes, S., Wenig, D., Döring, T., and Malaka, R. Does it feel real? Using tangibles with different fidelities to build and explore scenes in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.

[243] Murase, R., Zhang, Y., and Okatani, T. Video-rate video inpainting. In *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision* (2019), pp. 1553–1561.

[244] Murillo, R. A. M., Subramanian, S., and Plasencia, D. M. Erg-O: Ergonomic optimization of immersive virtual environments. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017), pp. 759–771.

[245] Nakajima, Y., Mori, S., and Saito, H. Semantic object selection and detection for diminished reality based on SLAM with viewpoint class. In *Adjunct Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (2017), pp. 338–343.

[246] Nancel, M., Chapuis, O., Pietriga, E., Yang, X.-D., Irani, P. P., and Beaudouin-Lafon, M. High-precision pointing on large wall displays using small handheld devices. In *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems* (2013), pp. 831–840.

[247] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 IEEE International Symposium on Mixed and Augmented Reality* (2011), pp. 127–136.

[248] Ng, G., Shin, J. G., Plopski, A., Sandor, C., and Saakes, D. Situated game level editing in augmented reality. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (2018), pp. 409–418.

[249] Nieto, O., and Bello, J. P. Systematic exploration of computational music structure research. In *Proceedings of the 17th International Society for Music Information Retrieval Conference* (2016), pp. 547–553.

[250] Nilsson, N. C., Zenner, A., Simeone, A. L., Degraen, D., and Daiber, F. Haptic proxies for virtual reality: Success criteria and taxonomy. In *Proceedings of the 1st Workshop on Everyday Proxy Objects for Virtual Reality* (2021).

[251] Norman, D. A. Some observations on mental models. In *Mental Models*, D. Genter and A. L. Stevens, Eds. Lawrence Erlbaum, 1984, pp. 15–34.

[252] North, A. C., and Hargreaves, D. J. Music and driving game performance. *Scandinavian Journal of Psychology 40*, 4 (1999), 285–292.

[253] Ntavelis, E., Romero, A., Kastanis, I., Van Gool, L., and Timofte, R. SESAME: Semantic editing of scenes by adding, manipulating or erasing objects. In *Proceedings of the European Conference on Computer Vision* (2020), pp. 394–411.

[254] Nuernberger, B., Ofek, E., Benko, H., and Wilson, A. D. SnapToReality: Aligning augmented reality to the real world. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 1233–1244.

[255] Oculus. National Geographic: Explore VR. https://www.oculus.com/experiences/quest/2046607608728563/, 2019.

[256] Ogden, C. K., and Richards, I. A. *The meaning of meaning: A study of the influence of thought and of the science of symbolism.* Harcourt, Brace & World, Inc., 1923.

[257] Oh, S. W., Lee, S., Lee, J.-Y., and Kim, S. J. Onion-peel networks for deep video completion. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 4403–4412.

[258] Ohnishi, K., Kanehira, A., Kanezaki, A., and Harada, T. Recognizing activities of daily living with a wrist-mounted camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 3103–3111.

[259] Olsen Jr, D. R., and Dempsey, E. P. Syngraph: A graphical user interface generator. In *Proceedings of the 10th annual conference on Computer Graphics and Interactive Techniques* (1983), pp. 43–50.

[260] Osokin, D. Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods* (2018).

[261] Osokin, D., and Ageeva, M. Real-time 3D multi-person pose estimation demo. https://github.com/Daniil-Osokin/lightweight-human-pose-estimation-3d-demo.pytorch, 2019.

[262] Owlchemy Labs. Job Simulator. https://jobsimulatorgame.com/, 2016.

[263] Pausch, R., Proffitt, D., and Williams, G. Quantifying immersion in virtual reality. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques* (1997), pp. 13–18.

[264] Pausch, R., Shackelford, M. A., and Proffitt, D. A user study comparing head-mounted and stationary displays. In *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium* (1993), pp. 41–45.

[265] PC Magazine. Best Products of 1994. https://archive.org/details/pc-magazine-best-products-of-1994-jan-1995/page/n131/mode/2up, 1995.

[266] Peerdeman, P. Sound and music in games. Tech. rep., Vrije Universiteit Amsterdam, 2006.

[267] Pietroszek, K., Wallace, J. R., and Lank, E. Tiltcasting: 3D interaction on large displays using a mobile device. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (2015), pp. 57–62.

[268] Popper, K. R. Epistemology without a knowing subject. In *Studies in Logic and the Foundations of Mathematics*, vol. 52. Elsevier, 1968, pp. 333–373.

[269] Poupyrev, I., Billinghurst, M., Weghorst, S., and Ichikawa, T. The Go-Go Interaction Technique: Non-linear mapping for direct manipulation in VR. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology* (1996), pp. 79–80.

[270] Poupyrev, I., and Ichikawa, T. Manipulating objects in virtual worlds: Categorization and empirical evaluation of interaction techniques. *Journal of Visual Languages & Computing 10*, 1 (1999), 19–35.

[271] Qian, X., He, F., Hu, X., Wang, T., Ipsita, A., and Ramani, K. ScalAR: Authoring semantically adaptive augmented reality experiences in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–18.

[272] Qin, H., Patrick Rau, P. L., and Salvendy, G. Measuring player immersion in the computer game narrative. *International Journal of Human-Computer Interaction 25*, 2 (2009), 107–133.

[273] Quine, W. V. On what there is. *The review of metaphysics* (1948), 21–38.

[274] Ratcliffe, S. *Oxford Essential Quotations*. Oxford University Press, 2018.

[275] Rekimoto, J., Ayatsuka, Y., and Hayashi, K. Augment-able reality: Situated communication through physical and digital spaces. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)* (1998), pp. 68–75.

[276] Rekimoto, J., and Nagao, K. The world through the computer: Computer augmented interaction with real world environments. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology* (1995), pp. 29–36.

[277] Rempe, D., Guibas, L. J., Hertzmann, A., Russell, B., Villegas, R., and Yang, J. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision* (2020).

[278] Rensink, R. A. Scene perception. In *Encyclopedia of Psychology*, A. Kazdin, Ed. Oxford University, 2000, pp. 151–155.

[279] Robinett, W. Synthetic experience: A proposed taxonomy. *Presence: Teleoperators and Virtual Environments 1*, 2 (1992), 229–247.

[280] Rompapas, D., Sandor, C., Plopski, A., Saakes, D., Yun, D. H., Taketomi, T., and Kato, H. HoloRoyale: A large scale high fidelity augmented reality game. In *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018), pp. 163–165.

[281] Roxas, M., Hori, T., Fukiage, T., Okamoto, Y., and Oishi, T. Occlusion handling using semantic segmentation and visibility-based rendering for mixed reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (2018), pp. 1–8.

[282] Rubin, S., and Agrawala, M. Generating emotionally relevant musical scores for audio stories. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (2014), pp. 439–448.

[283] Rubin, S., Berthouzoz, F., Mysore, G. J., Li, W., and Agrawala, M. UnderScore: Musical underlays for audio stories. In *Proceedings of the 25th Annual ACM Symposium on User interface Software and Technology* (2012), pp. 359–366.

[284] Ruder, M., Dosovitskiy, A., and Brox, T. Artistic style transfer for videos. In *German Conference on Pattern Recognition* (2016), Springer, pp. 26–36.

[285] Russell, B. On denoting. *Mind 14*, 56 (1905), 479–493.

[286] Russell, B. *The problems of philosophy*. Online Edition Project Gutenberg, 1912.

[287] Russell, B. Mr. Strawson on referring. *Mind 66*, 263 (1957), 385–389.

[288] Salas-Moreno, R. F., Newcombe, R. A., Strasdat, H., Kelly, P. H., and Davison, A. J. SLAM++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1352–1359.

[289] Sato, H., Hirai, T., Nakano, T., Goto, M., and Morishima, S. A music video authoring system synchronizing climax of video clips and music via rearrangement of musical bars. In *ACM SIGGRAPH 2015 Posters* (2015), p. 2010.

[290] Savage, N. S., Baranski, M., Chavez, N. E., and Höllerer, T. I'm feeling LoCo: A location based context aware recommendation system. *Lecture Notes in Geoinformation and Cartography*, 199599 (2012), 37–54.

[291] Schjerlund, J., Hornbæk, K., and Bergström, J. Ninja Hands: Using many hands to improve target selection in VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–14.

[292] Schmalstieg, D., and Höllerer, T. *Augmented Reality: Principles and Practice*. Addison-Wesley, 2016.

[293] Schmidt, S., Nunez, O. J. A., and Steinicke, F. Blended agents: Manipulation of physical objects within mixed reality environments and beyond. In *Proceedings of the 2019 ACM Symposium on Spatial User Interaction* (2019), pp. 1–10.

[294] Schoop, E., Smith, J., and Hartmann, B. HindSight: Enhancing spatial awareness by sonifying detected objects in real-time 360-degree video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), vol. 2018-April, pp. 1–12.

[295] Schulz, P., Alexandrovsky, D., Putze, F., Malaka, R., and Schöning, J. The role of physical props in VR climbing environments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[296] Schütte, R. Basispositionen in der Wirtschaftsinformatik—ein gemäßigt-konstruktivistisches Programm. In *Wirtschaftsinformatik und Wissenschaftstheorie: Bestandsaufnahme und Perspektiven*, J. Becker, W. König, R. Schütte, O. Wendt, and S. Zelewski, Eds. Springer, 1999, pp. 211–241.

[297] Schütte, R., and Zelewski, S. Wissenschafts-und erkenntnistheoretische Probleme beim Umgang mit Ontologien. *Tagung Wirtschaftsinformatik und Wissenschaftstheorie 99* (1999).

[298] Schütte, R., and Zelewski, S. Epistemological problems in working with ontologies. In *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics* (2002).

[299] Shapira, L., and Freedman, D. Reality Skins: Creating immersive and tactile virtual environments. In *Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality* (2016), pp. 115–124.

[300] Shen, V., Rae-Grant, T., Mullenbach, J., Harrison, C., and Shultz, C. Fluid Reality: High-resolution, untethered haptic gloves using electroosmotic pump arrays. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023), pp. 1–20.

[301] Sheridan, T. B., et al. Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments 1*, 1 (1992), 120–125.

[302] Shin, J.-e., Kim, H., Parker, C., Kim, H.-i., Oh, S., and Woo, W. Is any room really ok? The effect of room size and furniture on presence, narrative engagement, and usability during a space-adaptive augmented reality game. In *Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality* (2019), pp. 135–144.

[303] Shin, J.-e., Yoon, B., Kim, D., and Woo, W. A user-oriented approach to space-adaptive augmentation: The effects of spatial affordance on narrative experience in an augmented reality detective game. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[304] Shotton, J., Winn, J., Rother, C., and Criminisi, A. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European Conference on Computer Vision* (2006), pp. 1–15.

[305] Simeone, A. L., Velloso, E., and Gellersen, H. Substitutional Reality: Using the physical environment to design virtual reality experiences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), pp. 3307–3316.

[306] Singh, A., Kaur, R., Haltner, P., Peachey, M., Gonzalez-Franco, M., Malloch, J., and Reilly, D. Story CreatAR: A toolkit for spatially-adaptive augmented reality storytelling. In *2021 IEEE Virtual Reality and 3D User Interfaces* (2021), pp. 713–722.

[307] Skarbez, R., Smith, M., and Whitton, M. C. Revisiting Milgram and Kishino's reality-virtuality continuum. *Frontiers in Virtual Reality 2* (2021), 647997.

[308] Slater, M. Immersion and the illusion of presence in virtual reality. *British Journal of Psychology 109*, 3 (2018), 431–433.

[309] Speicher, M., Hall, B. D., and Nebeling, M. What is Mixed Reality? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–15.

[310] Sra, M., Garrido-Jurado, S., and Maes, P. Oasis: Procedurally generated social virtual spaces from 3D scanned real spaces. *IEEE Transactions on Visualization and Computer Graphics 24*, 12 (2017), 3174–3187.

[311] Staggers, N., and Norcio, A. F. Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies 38*, 4 (1993), 587–605.

[312] Stalnaker, R. C. Possible worlds. *Noûs* (1976), 65–75.

[313] State, A., Hirota, G., Chen, D. T., Garrett, W. F., and Livingston, M. A. Superior augmented reality registration by integrating landmark tracking and magnetic tracking. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (1996), pp. 429–438.

[314] Storey, V. C. Understanding semantic relationships. *The Very Large Databases Journal 2* (1993), 455–488.

[315] Strawson, P. F. On referring. *Mind 59*, 235 (1950), 320–344.

[316] Suhail, M., Sargunam, S. P., Han, D. T., and Ragan, E. D. Redirected reach in virtual reality: Enabling natural hand interaction at multiple virtual locations with passive haptics. pp. 245–246.

[317] Surale, H. B., Gupta, A., Hancock, M., and Vogel, D. TabletInVR: Exploring the design space for using a multi-touch tablet in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[318] Sutherland, I. E. The ultimate display. In *Proceedings of the Congress of the International Federation of Information Processing* (1965), vol. 2, New York, pp. 506–508.

[319] Sutherland, I. E. A head-mounted three dimensional display. In *Proceedings of the AFIPS Joint Computer Conference* (1968), pp. 757–764.

[320] Suzuki, R., Hedayati, H., Zheng, C., Bohn, J. L., Szafir, D., Do, E. Y.-L., Gross, M. D., and Leithinger, D. RoomShift: Room-scale dynamic haptics for VR with furniture-moving swarm robots. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–11.

[321] Suzuki, R., Karim, A., Xia, T., Hedayati, H., and Marquardt, N. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–33.

[322] Szeto, R., El-Khamy, M., Lee, J., and Corso, J. J. HyperCon: Image-to-video model transfer for video-to-video translation tasks. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision* (2021).

[323] Tahara, T., Seno, T., Narita, G., and Ishikawa, T. Retargetable AR: Context-aware augmented reality in indoor scenes based on 3D scene graph. In *Adjunct Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (2020), pp. 249–255.

[324] Tan, F., Tang, D., Dou, M., Guo, K., Pandey, R., Keskin, C., Du, R., Sun, D., Bouaziz, S., Fanello, S., Tan, P., and Zhang, Y. HumanGPS: Geodesic preserving feature for dense human correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2021), pp. 1820–1830.

[325] Tatzgern, M., Grasset, R., Veas, E. E., Kalkofen, D., Seichter, H., and Schmalstieg, D. Exploring distant objects with augmented reality. In *Proceedings of the 2013 Joint Virtual Reality Conference between the European Association of Virtual Reality and Augmented Reality and the Eurographics Symposium on Virtual Environments* (2013), pp. 49–56.

[326] Techmonitor. Apple launches Audiovisual Mac, Caspar PlainTalk. `https://techmonitor.ai/technology/apple_launches_audiovisual_macs_caspar_plaintalk`, 1993.

[327] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Niessner, M. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (June 2016).

[328] Tseng, W.-J., Bonnail, E., Mcgill, M., Khamis, M., Lecolinet, E., Huron, S., and Gugenheimer, J. The dark side of perceptual manipulations in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–15.

[329] Turing, A. M. *Computing machinery and intelligence*. Springer, 1950. Reprint 2009.

[330] Unity. Unity Docs - About AR Foundation. `https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@4.0/manual/index.html`, 2019.

[331] Valentin, J., Kowdle, A., Barron, J. T., Wadhwa, N., Dzitsiuk, M., Schoenberg, M., Verma, V., Csaszar, A., Turner, E., Dryanovski, I., Afonso, J., Pascoal, J., Tsotsos, K., Leung, M., Schmidt, M., Guleryuz, O., Khamis, S., Tankovitch, V., Fanello, S., Izadi, S., and Rhemann, C. Depth from motion for smartphone AR. *ACM Transactions on Graphics 37*, 6 (Dec. 2018).

[332] Valentin, J., Kowdle, A., Barron, J. T., Wadhwa, N., Dzitsiuk, M., Schoenberg, M., Verma, V., Csaszar, A., Turner, E., Dryanovski, I., et al. Depth from motion for smartphone AR. *ACM Transactions on Graphics 37*, 6 (2018), 1–19.

[333] van der Zwaag, M. D., Dijksterhuis, C., de Waard, D., Mulder, B. L., Westerink, J. H., and Brookhuis, K. A. The influence of music on mood and performance while driving. *Ergonomics 55*, 1 (2012), 12–22.

[334] van Orman Quine, W. Two dogmas of empiricism. *The Philosophical Review 60*, 1 (1951), 20–43.

[335] Vande Veire, L., and De Bie, T. From raw audio to a seamless mix: creating an automated DJ system for Drum and Bass. *Eurasip Journal on Audio, Speech, and Music Processing 2018*, 1 (2018).

[336] Veluri, B., Itani, M., Chan, J., Yoshioka, T., and Gollakota, S. Semantic Hearing: Programming acoustic scenes with binaural hearables. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023), pp. 1–15.

[337] Vidal, J. J. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering 2*, 1 (1973), 157–180.

[338] Viega, J., Conway, M. J., Williams, G., and Pausch, R. 3D Magic Lenses. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology* (1996), pp. 51–58.

[339] Viirre, E., Pryor, H., Nagata, S., and Furness III, T. A. The virtual retinal display: a new technology for virtual reality and augmented vision in medicine. In *Medicine Meets Virtual Reality* (1998), IOS Press, pp. 252–257.

[340] Wakim, S., and Grewal, M. Human organs and organ systems. In *Human Biology*. 2021.

[341] Wald, J., Tateno, K., Sturm, J., Navab, N., and Tombari, F. Real-time fully incremental scene understanding on mobile platforms. *IEEE Robotics and Automation Letters 3*, 4 (2018), 3402–3409.

[342] Walton, D. R., and Steed, A. Accurate real-time occlusion for mixed reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (2017), pp. 1–10.

[343] Wang, A., and Gollakota, S. MilliSonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–11.

[344] Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., and Catanzaro, B. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems* (2019).

[345] Wang, Y., Liang, W., Li, W., Li, D., and Yu, L.-F. Scene-Aware Background Music Synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 1162–1170.

[346] Ware, C., and Jessome, D. R. Using the Bat: A six-dimensional mouse for object placement. *IEEE Computer Graphics and Applications 8* (1988), 65–70.

[347] Wellner, P., Mackay, W., and Gold, R. Back to the real world. *Communications of the ACM 36*, 7 (1993), 24–26.

[348] Wen, H., Sze, N. N., Zeng, Q., and Hu, S. Effect of music listening on physiological condition, mental workload, and driving performance with consideration of driver temperament. *International Journal of Environmental Research and Public Health 16*, 15 (2019).

[349] Wentzel, J., d'Eon, G., and Vogel, D. Improving virtual reality ergonomics through reach-bounded non-linear input amplification. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–12.

[350] Whalen, Z. Play along - An approach to videogame music. *Game Studies 4*, 1 (2004), 1–28.

[351] Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., and Davison, A. ElasticFusion: Dense slam without a pose graph. In *Proceedings of the 2015 Conference on Robotics: Science and Systems* (2015).

[352] Whelan, T., Salas-Moreno, R. F., Glocker, B., Davison, A. J., and Leutenegger, S. ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research 35*, 14 (2016), 1697–1716.

[353] White, S., and Feiner, S. SiteLens: Situated visualization techniques for urban site visits. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems* (2009), pp. 1117–1120.

[354] White, S. M. *Interaction and presentation techniques for situated visualization.* PhD thesis, Columbia University, 2009.

[355] Whitmire, E. M. *Input Devices for the Next Generation of Computing Platforms.* PhD thesis, University of Washington, 2019.

[356] Wiemer, M. Mojo Vision: Designing anytime, anywhere AR contact lenses with Mojo lens. In *AVR21 Industry Talks II of the SPIE International Society for Optics and Photonics* (2021).

[357] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (1992).

[358] Wilkes, C. B., Tilden, D., and Bowman, D. A. 3D user interfaces using tracked multi-touch mobile devices. In *Proceedings of the 2012 Joint Virtual Reality Conference between the European Association of Virtual Reality and Augmented Reality and the Eurographics Symposium on Virtual Environments* (2012).

[359] Willett, W., Jansen, Y., and Dragicevic, P. Embedded data representations. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 461–470.

[360] Winograd, T. Towards a procedural understanding of semantics. *Revue internationale de philosophie* (1976), 260–303.

[361] Winograd, T. Moving the semantic fulcrum. *Linguistics and Philosophy* (1985), 91–104.

[362] Winograd, T., and Flores, F. *Understanding computers and cognition: A new foundation for design.* Ablex Publishing, 1986.

[363] Wittgenstein, L. *Tractatus logico-philosophicus.* 1922.

[364] Wittgenstein, L. *Philosophical investigations*. 1953.

[365] Wohlscheid, J. P. A Look at Connectix QuickCam — One of the First Webcams in the World. https://hackernoon.com/a-look-at-connectix-quickcam-one-of-the-first-webcams-in-the-world, 2023.

[366] Wright, E. Perception, pretence and reality. *Poetics Today 4*, 3 (1983), 513–542.

[367] Wu, E., Yuan, Y., Yeo, H.-S., Quigley, A., Koike, H., and Kitani, K. M. Back-Hand-Pose: 3D hand pose estimation for a wrist-worn camera via dorsum deformation network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (2020), pp. 1147–1160.

[368] Xiao, L., Nouri, S., Hegland, J., Garcia, A. G., and Lanman, D. Neural-Passthrough: Learned real-time view synthesis for VR. In *Proceedings of the ACM SIGGRAPH 2022 Conference* (2022).

[369] Xu, R., Li, X., Zhou, B., and Loy, C. C. Deep flow-guided video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3723–3732.

[370] Xu, Y., Zhu, S., and Tung, T. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *Proceedings of the International Conference on Computer Vision (ICCV)* (10 2019), pp. 7759–7769.

[371] Yamagami, M., Junuzovic, S., Gonzalez-Franco, M., Ofek, E., Cutrell, E., Porter, J., Wilson, A., and Mott, M. Two-In-One: A design space for mapping unimanual input into bimanual interactions in vr for users with limited movement. *ACM Transactions on Accessible Computing* (2021).

[372] Yamanobe, N., Wan, W., Ramirez-Alpizar, I. G., Petit, D., Tsuji, T., Akizuki, S., Hashimoto, M., Nagata, K., and Harada, K. A brief review of affordance in robotic manipulation research. *Advanced Robotics 31*, 19-20 (2017), 1086–1101.

[373] Yang, J., Holz, C., Ofek, E., and Wilson, A. D. DreamWalker: Substituting real-world walking experiences with a virtual reality. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), pp. 1093–1107.

[374] Yang, X., Zhou, L., Jiang, H., Tang, Z., Wang, Y., Bao, H., and Zhang, G. Mobile3DRecon: Real-time monocular 3D reconstruction on a mobile phone. *IEEE Transactions on Visualization and Computer Graphics 26*, 12 (2020), 3446–3456.

[375] Ye, H., Kwan, K. C., Su, W., and Fu, H. ARAnimator: In-situ character animation in mobile AR with user-defined motion gestures. *ACM Transactions on Graphics 39*, 4 (2020), 83–1.

[376] Yeo, H.-S., Wu, E., Lee, J., Quigley, A., and Koike, H. Opisthenar: Hand poses and finger tapping recognition by observing back of hand using embedded wrist camera. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), pp. 963–971.

[377] Yu, F., Wang, D., Shelhamer, E., and Darrell, T. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2403–2412.

[378] Yu, R., and Bowman, D. A. Force Push: Exploring expressive gesture-to-force mappings for remote object manipulation in virtual reality. *Frontiers in ICT 5* (2018), 25.

[379] Yue, Y.-T., Yang, Y.-L., Ren, G., and Wang, W. SceneCtrl: Mixed reality enhancement via efficient scene editing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017), pp. 427–436.

[380] Zelewski, S. *Das Leistungspotential der Künstlichen Intelligenz: eine informationstechnisch-betriebswirtschaftliche Analyse*. Wehle, 1986.

[381] Zelewski, S. Grundlagen. In *Betriebswirtschaftslehre Band 1*, H. Corsten, Ed., 4 ed. Oldenbourg, 2008, pp. 1–97.

[382] Zepf, S., Hernandez, J., Dittrich, M., and Schmitt, A. Towards empathetic car interfaces: Emotional triggers while driving. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–6.

[383] Zhang, H., Mai, L., Xu, N., Wang, Z., Collomosse, J., and Jin, H. An internal learning approach to video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 2720–2729.

[384] Zhang, R., Li, W., Wang, P., Guan, C., Fang, J., Song, Y., Yu, J., Chen, B., Xu, W., and Yang, R. AutoRemover: Automatic object removal for autonomous driving videos. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34.

[385] Zhao, Y. *A Quest for Visual Commonsense: Scene Understanding by Functional and Physical Reasoning*. PhD thesis, University of California, Los Angeles, 2015.

[386] Zhao, Y., and Follmer, S. A functional optimization based approach for continuous 3D retargeted touch of arbitrary, complex boundaries in haptic virtual reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–12.

[387] Zhao, Y., Kupferstein, E., Castro, B. V., Feiner, S., and Azenkot, S. Designing AR visualizations to facilitate stair navigation for people with low vision. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), pp. 387–402.

[388] Zhao, Y., Lindberg, D., Cleary, B., Mercier, O., Mcclelland, R., Penner, E., Lin, Y.-J., Majors, J., and Lanman, D. Retinal-Resolution Varifocal VR. In *ACM SIGGRAPH 2023 Emerging Technologies*. 2023, pp. 1–3.

[389] Zhou, X., Wang, D., and Krähenbühl, P. Objects as points. *arXiv preprint arXiv 1904.07850* (2019).

[390] Zhu, Y., Wang, Y., Li, G., and Guo, X. Recognizing and releasing drivers' negative emotions by using music: Evidence from driver anger. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2016), pp. 173–178.

[391] Zollmann, S., Langlotz, T., Grasset, R., Lo, W. H., Mori, S., and Regenbrecht, H. Visualization techniques in augmented reality: A taxonomy, methods and patterns. *IEEE Transactions on Visualization and Computer Graphics 27*, 9 (2020), 3808–3825.

[392] Žurauskas, M., El-Haddad, M., Silverstein, B., Lanman, D., and Cavin, R. Achieving the visual Turing test: Integrated display and eye tracking technologies. https://research.facebook.com/publications/achieving-the-visual-turing-test-integrated-display-and-eye-tracking-technologies/, 2023. Facebook Research.

Eidesstattliche Erklärung

Ich gebe folgende eidesstattliche Erklärung ab:

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig ohne unzulässige Hilfe Dritter verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen unter der Angabe der Quelle als solche gekennzeichnet habe.

Die Grundsätze für die Sicherung guter wissenschaftlicher Praxis an der Universität Duisburg-Essen sind beachtet worden.

Ich habe die Arbeit keiner anderen Stelle zu Prüfungszwecken vorgelegt.

_____, Essen, Mai 2024

Mohamed Kari