Augmented Telepresence based on Multi-Camera Systems

Capture, Transmission, Rendering, and User Experience

Elijs Dima

Department of Information Systems and Technology
Mid Sweden University

Doctoral Thesis No. 345
Sundsvall, Sweden
2021

Mittuniversitetet
Informationssytem och -teknologi
ISBN 978-91-89341-06-7 SE-851 70 Sundsvall
ISNN 1652-893X SWEDEN

Akademisk avhandling som med tillstånd av Mittuniversitetet framlägges till offentlig granskning för avläggande av teknologie doktorsexamen **den 17 Maj 2021** klockan **14:00** i sal **C312**, Mittuniversitetet Holmgatan 10, Sundsvall. Seminariet kommer att hållas på engelska.

Tryck: Tryckeriet Mittuniversitetet

DON'T PANIC!

- Douglas Adams, The Hitchhiker's Guide to the Galaxy

Abstract

Observation and understanding of the world through digital sensors is an ever-increasing part of modern life. Systems of multiple sensors acting together have far-reaching applications in automation, entertainment, surveillance, remote machine control, and robotic self-navigation. Recent developments in digital camera, range sensor and immersive display technologies enable the combination of augmented reality and telepresence into *Augmented Telepresence*, which promises to enable more effective and immersive forms of interaction with remote environments.

The purpose of this work is to gain a more comprehensive understanding of how multi-sensor systems lead to Augmented Telepresence, and how Augmented Telepresence can be utilized for industry-related applications. On the one hand, the conducted research is focused on the technological aspects of multi-camera capture, rendering, and end-to-end systems that enable Augmented Telepresence. On the other hand, the research also considers the user experience aspects of Augmented Telepresence, to obtain a more comprehensive perspective on the application and design of Augmented Telepresence solutions.

This work addresses multi-sensor system design for Augmented Telepresence regarding four specific aspects ranging from sensor setup for effective capture to the rendering of outputs for Augmented Telepresence. More specifically, the following problems are investigated: 1) whether multi-camera calibration methods can reliably estimate the true camera parameters; 2) what the consequences are of synchronization errors in a multi-camera system; 3) how to design a scalable multi-camera system for low-latency, real-time applications; and 4) how to enable Augmented Telepresence from multi-sensor systems for mining, without prior data capture or conditioning.

The first problem was solved by conducting a comparative assessment of widely available multi-camera calibration methods. A special dataset was recorded, enforcing known constraints on camera ground-truth parameters to use as a reference for calibration estimates. The second problem was addressed by introducing a depth uncertainty model that links the pinhole camera model and synchronization error to the geometric error in the 3D projections of recorded data. The third problem was addressed empirically —by constructing a multi-camera system based on off-the-shelf hardware and a modular software framework. The fourth problem was addressed by proposing a processing pipeline of an augmented remote operation system for

augmented and novel view rendering.

The calibration assessment revealed that target-based and certain target-less calibration methods are relatively similar in their estimations of the true camera parameters, with one specific exception. For high-accuracy scenarios, even commonly used target-based calibration approaches are not sufficiently accurate with respect to the ground truth. The proposed depth uncertainty model was used to show that converged multi-camera arrays are less sensitive to synchronization errors. The mean depth uncertainty of a camera system correlates to the rendered result in depth-based reprojection as long as the camera calibration matrices are accurate. The presented multi-camera system demonstrates a flexible, de-centralized framework where data processing is possible in the camera, in the cloud, and on the data consumer's side. The multi-camera system is able to act as a capture testbed and as a component in end-to-end communication systems, because of the general-purpose computing and network connectivity support coupled with a segmented software framework. This system forms the foundation for the augmented remote operation system, which demonstrates the feasibility of real-time view generation by employing on-the-fly lidar de-noising and sparse depth upscaling for novel and augmented view synthesis.

In addition to the aforementioned technical investigations, this work also addresses the user experience impacts of Augmented Telepresence. The following two questions were investigated: 1) What is the impact of camera-based viewing position in Augmented Telepresence? 2) What is the impact of depth-aiding augmentations in Augmented Telepresence? Both are addressed through a quality of experience study with non-expert participants, using a custom Augmented Telepresence test system for a task-based experiment. The experiment design combines in-view augmentation, camera view selection, and stereoscopic augmented scene presentation via a head-mounted display to investigate both the independent factors and their joint interaction. The results indicate that between the two factors, view position has a stronger influence on user experience. Task performance and quality of experience were significantly decreased by viewing positions that force users to rely on stereoscopic depth perception. However, position-assisting view augmentations can mitigate the negative effect of sub-optimal viewing positions; the extent of such mitigation is subject to the augmentation design and appearance.

In aggregate, the works presented in this dissertation cover a broad view of Augmented Telepresence. The individual solutions contribute general insights into Augmented Telepresence system design, complement gaps in the current discourse of specific areas, and provide tools for solving challenges found in enabling the capture, processing, and rendering in real-time-oriented end-to-end systems.

Acknowledgements

First and foremost, I would like to thank my supervisors, Prof. Mårten Sjöström and Dr. Roger Olsson, for their guidance and support, and for both their insights and their example of working through the research process. Next (and of equal importance), a massive "Thank You" to my friends, Yongwei Li and Waqas Ahmad, for their invaluable assistance, support and friendship during these past few years. Thank you for forming the core of a friendly, open, and honest research group that I am glad to have been a part of.

Special thanks to Joakim Edlund, Jan-Erik Jonsson and Martin Kjellqvist here at IST for their help on projects and even more so for the on-topic and off-topic conversations; the workplace environment would not be nearly as good without you. Thanks also to the folks from "one floor above" at IST, past and present: Mehrzad Lavassani, Luca Beltramelli, Leif Sundberg and Simone Grimaldi, thank you all for making the earlier parts of these studies more fun.

Thanks to Prof. Kjell Brunnström for the collaborations and insights into quality assessment. Thanks to the past and present employees at Ericsson Research, both for hosting me in their research environment at Kista near the start of my studies. Thanks to Lars Flodén and Lennart Rasmusson of Observit AB for their insights into the engineering goals and constraints of multi-camera applications, and to Lisa Önnerlöv at Boliden Minerals AB for insight into a particularly hands-on industry. Thanks to Prof. Marek Domański and Prof. Reinhard Koch for hosting me in their respective research groups at Poznan and Kiel; both have been valuable sources of insight into Light Fields and camera systems, and also provided me with exposure to culturally and organizationally diverse research practices and environments. Thanks to Prof. Jenny Read of Newcastle University for the discussions on human vision and perception, and the arcane mechanisms through which we humans create a model of the 3D world.

This work has received funding from: (i) grant 6006-214-290174 from Rådet för Utbildning på Forskarnivå (FUR), Mid Sweden University; (ii) grants nr. 20140200 and nr. 20160194 from the Knowledge Foundation, Sweden; (iii) grant nr. 20201888 from the EU Regional Development Fund; (iv) project nr. 2019-05162 from the Swedish Mining Innovation group.

Contents

List of Papers

This thesis is based on the following papers, herein referred to by their Roman numerals:

PAPER I

E. Dima, M. Sjöström, R. Olsson,
Assessment of Multi-Camera Calibration Algorithms for Two-Dimensional Camera Arrays Relative to Ground Truth Position and Direction,
3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-Con), 2016 . . . . . **??**

PAPER II

E. Dima, M. Sjöström, R. Olsson,
Modeling Depth Uncertainty of Desynchronized Multi-Camera Systems,
International Conference on 3D Immersion (IC3D), 2017 . . . . . **??**

PAPER III

E. Dima, M. Sjöström, R. Olsson, M. Kjellqvist, L. Litwic, Z. Zhang, L. Rasmusson, L. Flodén,
LIFE: A Flexible Testbed for Light Field Evaluation,
3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-Con), 2018 . . . . . **??**

PAPER IV

E. Dima, K. Brunnström, M. Sjöström, M. Andersson, J. Edlund, M. Johanson, T. Qureshi,
View Position Impact on QoE in an Immersive Telepresence System for Remote Operation,
International Conference on Quality of Multimedia Experience (QoMEX),
2019 . . . . . **??**

PAPER V

E. Dima, K. Brunnström, M. Sjöström, M. Andersson, J. Edlund, M. Johanson, T. Qureshi ,
Joint Effects of Depth-aiding Augmentations and Viewing Positions on the

The following papers are not included in the thesis:

List of Figures

List of Tables

Terminology

Abbreviations and Acronyms

2D	Two-Dimensional
3D	Three-Dimensional
4D	Four-Dimensional
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
AT	Augmented Telepresence
DIBR	Depth-Image Based Rendering
ECG	Electro-Cardiography
EEG	Electro-Encephalography
FoV	Field of View
FPS	Frames per Second
GPU	Graphics Processing Unit
HMD	Head-Mounted Display
IBR	Image-Based Rendering
Lidar	Light Detection and Ranging (device)
MBR	Model-Based Rendering
MCS	Multi-Camera System
MOS	Mean Opinion Score
MR	Mixed Reality
MV-HEVC	Multi-View High Efficiency Video Codec
PCM	Pinhole Camera Model
PPA	Psycho-Physiological Assessment
QoE	Quality of Experience
RGB	Color-only (from Red-Green-Blue digital color model
RGB-D	RGB plus Depth
RQ	Research Question
SIFT	Scale-Invariant Feature Transform
SfM	Structure from Motion

SLAM	Simultaneous Localization and Mapping
ToF	Time-of-Flight
UX	User Experience
VR	Virtual Reality

Mathematical Notation

The following terms are mentioned in this work:

λ	Arbitrary scale factor (used in the pinhole camera model)
u, v	Horizontal and vertical coordinate of a 2D point on an image plane
x, y	Coordinates in 2D space
X, Y, Z	Coordinates of a 3D point in any three-dimensional space
f_x, f_y	Focal lengths of a lens in the horizontal and vertical axis scales, respectively
x_0, y_0	The x and y position of a camera's principal point on the camera sensor
s	Skew factor between the x and y axes of a camera sensor
$\mathcal{K}$	Intrinsic camera matrix
C	Camera position in 3D space
$\mathcal{R}$	Camera rotation in 3D space
$\mathbf{H}$	Homography matrix in projective geometry
t	A specific point in time
Δt_n	Synchronization offset (error) between cameras capturing the n-th frame at time t
t_n^{N}	Time when camera 'N' is capturing frame n
Γ	The Plenoptic Function
Υ	Intensity of light
θ, ϕ	Angular directions from a common origin
ξ	Wavelength of light
Δd	Depth uncertainty
$\overrightarrow{r}_{\mathrm{N}}$	Ray cast from camera 'N'
$\vec{E}$	A moving point (object) in 3D space, recorded by a camera or array of cameras
$v_{\vec{E}}$	Movement speed of $\vec{E}$
$\vec{m}$	Shortest vector connecting two rays
$\overline{\Delta d}$	Mean depth uncertainty

Chapter 1

Introduction

This thesis is a comprehensive summary and analysis of the research process behind the works shown in the List of Papers. As such, the following six chapters have a larger emphasis on research questions and methodology than is commonly seen in the listed papers; these chapters are not written to replicate the content of the papers but rather to supplement them.

This chapter defines the overall context and aim of the presented research in light of the importance and timeliness of augmented applications in remote operation that depend on multi-camera systems. The research purpose is defined in two parts, which are supported by a total of six research questions. The scope of this work is described, and a brief summary of the contributions in the form of scientific publications is presented.

1.1 Overall Aim

The overall aim of the research in this thesis is to contribute to a more comprehensive understanding of how multi-camera and multi-sensor systems lead to industrially viable Augmented Telepresence (AT). This aim is investigated by focusing on how cameras and other environment-sensing devices should integrate into capture systems to produce consistent datasets, how those capture systems should be integrated into AT systems within domain-specific constraints, and how such AT systems affect the end-user experience in an industrial context.

1.2 Problem Area

Telepresence and remote working are fast becoming the norm across the world, by choice or necessity. Telepresence for conferences and desk work can be handled sufficiently with no more than a regular Two-Dimensional (2D) camera and display.

However, effective and safe remote working and automation in industrial and outdoor contexts (e.g. logging, mining, construction) requires a more thorough recording, understanding, and representation of the on-site environment. This can be achieved by involving systems of multiple 2D cameras and range sensors such as Light Detection and Ranging (lidar) in the capture process.

Multi-camera and multi-sensor systems already are important tools for a wide range of research and engineering applications, including but not limited to surveillance [OLS+15, DBV16], entertainment [LMJH+11, ZEM+15], autonomous operation [HLP15, LFP13], and telepresence [AKB18]. Recently, immersive Virtual Reality (VR) and Augmented Reality (AR) have gained significant industry traction [KH18] due to advances in Graphics Processing Unit (GPU), Head-Mounted Display (HMD) and network-related (5G) technologies. For industries where human operators directly control industrial machinery on site, there is significant potential in remote, multi-camera based applications that merge immersive telepresence [TRG+17, BDA+19] with augmented view rendering [LYC+18a, VPR+18] in the form of AT.

1.3 Problem Formulation

Augmented Telepresence has the potential to improve user experience and task-based effectiveness, especially when incorporated for industrial applications. In order to achieve immersive AT with seamless augmentation, the geometry and Three-Dimensional (3D) structure of the remote environment needs to be known. Extraction of this geometry is affected by the accuracy of calibration and synchronization of the various cameras and other sensors used for recording the remote locations; a sufficiently large loss of accuracy leads to inconsistencies between the data recorded by different sensors, which propagate throughout the AT rendering chain. Furthermore, multi-sensor systems and the subsequent rendering methods have to be designed for AT within constraints set by the sensors (e.g., inbound data rate, resolution) and the application domains (e.g., no pre-scanned environments in safety-critical areas). Beyond these accuracy and application feasibility problems affecting the system design, the utility of AT depends on how it improves user experience. Guidance via AR has been beneficial in non-telepresence applications, however AT leads to new, open questions about how the separate effects of AR, immersive rendering, and telepresence combine and change the overall user experience.

1.4 Purpose and Research Questions

The purpose driving the research presented in this work is twofold. On the one hand, the focus is on aspects of capture and system design for multi-sensor systems related to AT, and on the other hand the focus is on the resulting user experience formed by applying AT in an industrial context. The purpose of the research is defined by the following points:

P1 *To investigate how multi-camera and multi-sensor systems should be designed for the capture of consistent datasets and use in AT applications.*

P2 *To investigate how user experience is affected by applying multi-sensor based AT in industrial, task-based contexts.*

This twofold research purpose is supported by exploring the following two sets of research questions (RQs):

RQ 1.1 How accurate are the commonly used multi-camera calibration methods, both target-based and targetless, in recovering the true camera parameters represented by the pinhole camera model?

RQ 1.2 What is the relationship between camera synchronization error and estimated scene depth error, and how does camera arrangement in multi-camera systems affect this depth error?

RQ 1.3 What is an appropriate, scalable multi-camera system design for enabling low-latency video processing and real-time streaming?

RQ 1.4 What rendering performance can be achieved by camera-and-lidar-based AT for remote operation in an underground mining context, without data preconditioning?

and

RQ 2.1 What impact does the camera-based viewing position have on user Quality of Experience in an AT system for remote operation?

RQ 2.2 What impact do depth-aiding view augmentations have on user Quality of Experience in an AT system for remote operation?

1.5 Scope

For experimental implementations of multi-camera and AT systems, the implemented systems are built for lab experiments and not for in-field use. The multi-camera video data transfer from capture to presentation devices does not consider state-of-the-art video compression methods, as the focus of the presented research is not data compression. The research includes augmented and multiple-view rendering, but the contributions do not use the 4D Light Field as the transport format or rendering platform for the multi-camera content.

1.6 Contributions

The thesis is based on the results of the contributions listed in the list of papers that are included in full at the end of this summary. As the main author of Papers I, II,

III, IV, V, and VI, I am responsible for the ideas, methods, test setup, implementation, analysis, writing, and presentation of the research work and results. For Paper III, M. Kjellqvist and I worked together on the software implementation, and Z. Zhang and L. Litwic developed the cloud system and contributed to the communication interface definitions for the testbed. The remaining co-authors contributed with research advice and editing in their respective papers.

The general contents of the individual contributions are as follows:

Paper I addresses **RQ 1.1** by comparing calibration accuracy of multiple widely-used calibration methods with respect to ground truth camera parameters.

Paper II addresses **RQ 1.2** by deriving a theoretical model to express the consequences of camera synchronization errors as depth uncertainty, and using the model to show the impact of camera positioning in unsynchronized multi-camera systems.

Paper III addresses **RQ 1.3** by introducing the high-level framework for a flexible end-to-end Light Field testbed and assessing the performance (latency) in the key components used in the framework's implementation.

Paper IV addresses **RQ 2.1** through an experiment design and analysis of the results of using different viewing positions (and therefore camera placement) in an AT remote operation scenario.

Paper V addresses **RQ 2.1** and **RQ 2.2** by analyzing the individual and joint effects of varying viewing positions and augmentation designs on user Quality of Experience in an AT scenario. It also implicitly touches on **P1** by describing the integration of AR elements and the virtual projection approach for AT based on a multi-camera system.

Paper VI addresses **RQ 1.4** by presenting a novel multi-camera and lidar real-time rendering pipeline for multi-sensor based AT for an underground mining context and by analyzing the proposed pipeline's performance under real-time constraints.

1.7 Outline

This thesis is structured as follows. Chapter 2 presents the background of the thesis, covering the major domains of multi-camera capture, view rendering, AT, and Quality of Experience. The specific prior studies that illustrate the state-of-the-art in these domains are presented in Chapter 3. Chapter 4 covers the underlying methodology of the research, and Chapter 5 presents a summary of the results. Chapter 6 presents a discussion of and reflection on the research, including the overall outcomes, impact, and future avenues of the presented work. After the comprehensive summary (Chapters 1 through 6), the bibliography and specific individual contributions (Papers I through VI) are given.

Chapter 2

Background

This chapter covers the four main knowledge domains underpinning the contributions that this thesis is based on. The chapter starts by discussing relevant aspects of multi-camera capture, followed by an overview of view rendering in a multi-view context. After this, the key concepts of AT and Quality of Experience (QoE) are presented.

2.1 Multi-Camera Capture

A Multi-Camera System (MCS) is a set of cameras recording the same scene from different viewpoints. Notable early MCSs were inward-facing systems for 3D model scanning [KRN97] and virtual teleconferencing [FBA+94], as well as planar homogeneous arrays for Light Field dataset capture [WSLH01, YEBM02]. Beyond dataset capture, end-to-end systems such as [YEBM02, MP04, BK10] combined MCS with various 3D presentation devices to show live 3D representations of the observed 3D scene. Since then, MCSs have integrated increasingly diverse sensors and application platforms. Multi-camera systems have been created from surveillance cameras [FBLF08], mobile phones [SSS06], high-end television cameras [FBK10, DDM+15], and drone-mounted lightweight sensors [HLP15] and have included infrared-pattern and Time-of-Flight (ToF) depth sensors [GČH12, BMNK13, MBM16]. Currently, MCS-based processing is common in smartphones [Mö18] and forms the sensory backbone for self-driving vehicles [HHL+17].

Multi-camera capture is a process for recording a 3D environment that simultaneously uses a set of operations with multiple coordinated 2D cameras. Based on the capture process descriptions in [HTWM04, SAB+07, NRL+13, ZMDM+16], these operations can be grouped into three stages of the capture process - pre-recording, recording, and post-recording. The pre-recording stage operations, such as calibration, ensure that the various cameras (and other sensors) are coordinated in a MCS to enable the production of consistent data. The recording stage comprises the actions

of recording image sequences from each camera's sensor to the internal memory, including sensor-to-sensor synchronization between cameras. The post-recording stage contains operations that make the individual image sequences available and convert them to a dataset: the set of consistent information from all cameras that can be jointly used by down-stream applications.

2.1.1 Calibration and Camera Geometry

Camera calibration is a process that estimates camera positions, view directions, and lens and sensor properties [KHB07] through analysis of pixel correspondences and distortions in the recorded image. The results of calibration are camera parameters, typically according to the Pinhole Camera Model (PCM) as defined in the multiple-view projective geometry framework [HZ03], and a lens distortion model such as [Bro66]. The PCM assumes that each point on the camera sensor projects in a straight line through the camera optical center. The mapping between a 3D point at coordinates X, Y, Z and a 2D point on image plane at coordinates u, v is

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = [\mathcal{K}|0_3] \begin{bmatrix} \mathcal{R} & -\mathcal{R}C \\ 0_3^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} . \tag{2.1}$$

The internal camera parameters are focal lengths f_x, f_y, positions of the image central point x_0, y_0, and the skew factor s between the sensor's horizontal and vertical axes. These parameters are enclosed in the *intrinsic matrix* $\mathcal{K}$:

$$\mathcal{K} = \begin{bmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} . \tag{2.2}$$

The camera-to-camera positioning is defined by each camera's position in 3D space C and each camera's rotation $\mathcal{R}$, typically combined as the *extrinsic matrix*:

$$[\mathcal{R}| - \mathcal{R}C] . \tag{2.3}$$

Eq. (2.1) forms the basis for 3D scene reconstruction and view generation from MCS capture. Therefore, parameter estimation errors arising from inaccurate calibration have a direct impact on how accurately the recorded 2D data can be fused [SSO14].

Camera calibration is grouped into two discrete stages, following the PCM: *intrinsic* and *extrinsic* calibration. Intrinsic calibration is a process of estimating the intrinsic matrix $\mathcal{K}$ as well as lens distortion parameters to model the transformation from an actual camera-captured image to a PCM-compatible image. Extrinsic calibration is the estimation of relative camera positions and orientations within a uniform coordinate system, typically with a single camera chosen as the origin. In aggregate, most calibration methods have the following template: 1) corresponding scene points are identified and matched in camera images; 2) point coordinates are used together with projective geometry to construct an equation system where

camera parameters are the unknown variables; and 3) the equation system is solved by combining an analytical solution with a max-likelihood optimization of camera parameter estimates.

The most influential and most cited calibration method is [Zha00]. It relies on a flat 2D target object that holds a grid of easily identifiable points at known intervals (e.g. a non-square checkerboard). The PCM equation is reformulated to establish a homography $\mathbf{H}$ that describes how a 2D calibration surface (nominally at $Z = 0$ plane) is projected onto the camera's 2D image, based on the intrinsic matrix $\mathcal{K}$, camera position C, and the first two columns of the camera rotation matrix ($c_1, c_2 \in \mathcal{R}$):

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathcal{K} \begin{bmatrix} \mathcal{R} & -\mathcal{R}C \end{bmatrix} \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \text{ where } \mathbf{H} = \mathcal{K} \begin{bmatrix} c_1 | c_2 | -\mathcal{R}C \end{bmatrix} \tag{2.4}$$

With at least three observations of the target surface at different positions, the closed-form solution of Eq. (2.4) has a single unique solution up to a scale factor. The scale factor is resolved by the known spacing between points on the target surface. The intrinsic and extrinsic parameter estimates are typically refined together with lens distortion parameters by minimizing the distance between all observed target points and their projections based on the parameter estimates. This calibration method has been incorporated in various computer vision tools and libraries [Bou16, Mat17, Bra00, Gab17] and uses the first few radial and tangential distortion terms according to the Brown-Conrady distortion model [Bro66]. For further details on camera calibration, refer to [KHB07].

Camera calibration is not error-free. One source of error in the calibration process is an incorrect detection and matching of corresponding points between camera views, particularly for calibration methods that rely on ad-hoc scene points and image feature detectors [Low99, BETVG08, RRKB11] instead of a premade calibration target. Another source of error is optical lens system effects such as defocus, chromatic aberration [ESGMRA11], coma, field curvature, astigmatism, flare, glare, and ghosting [TAHL07, RV14], which are not represented by the Brown-Conrady distortion model. Furthermore, the architecture of digital sensor electronics leads to both temporally fluctuating and fixed-pattern noise [HK94, BCFS06, SKKS14], which can affect the recorded image and thus contribute to erroneous estimation of camera parameters.

2.1.2 Synchronization

Synchronization is the measure of simultaneity between the exposure moments of two cameras. Synchronization is parametrized by the synchronization error Δt_n between two cameras (A and B) capturing a frame n at time t:

$$\Delta t_n = \| t_n^{\mathrm{A}} - t_n^{\mathrm{B}} \| \tag{2.5}$$

The multi-view geometry as described in Section 2.1.1 is applicable only if there is no movement within the recorded scene or if all cameras record all scene points at the same moment ($\Delta t_n = 0$). Lack of synchronicity during MCS recording leads to a temporally inconsistent sampling of dynamic scenes, thus breaking the geometry relation. Camera synchronization is therefore a necessary prerequisite for accurate 3D reconstruction and view-to-view projection of dynamic scene content, as well as an important component of multi-view capture [SAB+07].

Synchronous recording can be achieved via external synchronization signaling to the camera hardware or by software instructions through the camera Application Programming Interface (API) [LZT06]. Perfect synchronization can only be guaranteed if an external signal bypasses all on-camera processing and triggers the sensor exposure on all MCS cameras. Such external synchronization is more accurate than software solutions [LHVS14]. A hardware synchronization requirement can affect the camera (and therefore system) cost [PM10] and prevent the use of entire sensor categories like affordable ToF depth cameras [SLK15].

2.1.3 Transmission

The transmission of video recorded by cameras in an MCS is a necessary component for integrating MCS in an end-to-end communication system. In the basic form, transmission consists of video encoding and storage or streaming. Storage, compression, and streaming thus represent the post-recording stage of the capture process, and often define the output interface for an MCS. The choice of using an MCS for recording a 3D scene has traditionally been motivated by the increased flexibility in bandwidth that an MCS offers in comparison to plenoptic cameras [WMJ+17].

A plenoptic camera [NLB+05] uses special optical systems to multiplex different views of the scene onto one sensor, which forces the subsequent signal processing chain to handle the data at the combined bandwidth of all views. Distributing a subset of views from plenoptic capture further requires view isolation, and for video transfer over a network, there is a need for real-time implementations of plenoptic or Light Field video compression. Although efficient Light Field video compression is an active research area (see [AGT+19, LPOS20, HML+19]), the foremost standard for real-time multi-view video compression is the Multi-View High Efficiency Video Codec (MV-HEVC) [HYHL15], which still requires decomposing a single plenoptic image into distinct views.

In contrast, an MCS typically offers one view per camera sensor, with associated image processing; this allows the use of ubiquitous hardware-accelerated single-view video encoders such as HEVC [SBS14] and VP9 [MBG+13], which have been extensively surveyed in [LAV+19, EPTP20]. The multi-camera based capture systems in [MP04, YEBM02, BK10] serve as early examples of bandwidth management that relies on the separated view capture afforded by the MCS design.

2.2 View Rendering

In the broadest sense, *view rendering* is the generation —or synthesis —of new perspectives of a known scene using some form of data describing the scene. View rendering has traditionally been classified into two groups, namely Model Based Rendering (MBR) and Image Based Rendering (IBR) [KSS05]. In this MBR + IBR classification, MBR implies view synthesis from an arrangement of geometric models and associated textures with a scene definition of lights, objects, and virtual cameras. IBR refers to the use of previously recorded 2D images and optional explicit or implicit representations of scene geometry to warp, distort, interpolate or project pixels from the recorded images to the synthesized view.

More recently, this classification has been supplanted by a four-group model that distinguishes between "classical rendering," "light transport," IBR, and "neural rendering" [TFT+20]. Classical rendering essentially refers to MBR from the perspective of computer graphics. Light transport is strongly related to Light Field rendering, which in the MBR + IBR model was classified as a geometry-less type of IBR. Neural rendering is a new approach to view rendering based on either view completion or *de novo* view synthesis through neural network architectures.

Classical a.k.a. Model-Based Rendering is the process of synthesizing an image from a scene defined by virtual cameras, lights, object surface geometries, and associated materials. This rendering is commonly achieved via either rasterization or raytracing [TFT+20]. Rasterization is the process of geometry transformation and pixelization onto the image plane, usually in a back-to-front compositing order known as the painter's algorithm. Rasterization is readily supported by contemporary GPU devices and associated computer graphics pipelines such as DirectX and OpenGL. Raytracing is the process of casting rays from a virtual camera's image pixels into the virtual scene to find ray-object intersections. From these intersections, further rays can be recursively cast to locate light sources, reflections, and so on. Both rasterization and raytracing essentially rely on the same projective geometry as described by Eq. (2.1), albeit with variations in virtual space discretization and camera lens simulation [HZ03, SR11]. The render quality in MBR is dependent on the quality of the scene component models (geometry, textures, surface properties, etc.). These models can be created by artists or estimated from real world data through a process known as inverse rendering [Mar98].

Light Field rendering and Light transport are view rendering approaches that attempt to restore diminished parametrizations of the plenoptic function [AB91]. The plenoptic function Γ is a light-ray based model that describes the intensity Υ of light rays at any 3D position $[X, Y, Z]$, in any direction $[\theta, \phi]$, at any time t, and at any light wavelength ξ:

$$\Upsilon = \Gamma(\theta, \phi, \xi, t, X, Y, Z) \tag{2.6}$$

The Light Field [LH96] is a Four-Dimensional (4D) re-parametrization of the plenoptic function that encodes the set of light rays crossing the space between two planes $[x, y]$ and $[u, v]$. View rendering from the 4D Light Field is the integration of all light rays intersecting a virtual camera's image plane and optical center (assuming a PCM). Light transport refers to a slightly different parametrization of the plenoptic

function, which is based on the rendering equation [Kaj86], that defines light radiance $\Upsilon = \Gamma_0$ from a surface as a function of position, direction, time, and wavelength (same as the plenoptic function), but distinguishes between directly emitted light Γ_e and reflected light Γ_r:

$$\Upsilon = \Gamma_o(\theta, \phi, \xi, t, X, Y, Z) = \Gamma_\mathrm{e}(\theta, \phi, \xi, t, X, Y, Z) + \Gamma_\mathrm{r}(\theta, \phi, \xi, t, X, Y, Z) \tag{2.7}$$

The Light transport rendering often refers to Surface Light Fields [MRP98, WAA+00], which predictably assign an intensity Color-only (RGB) value to every ray that leaves a point on a surface. The 4D Light Field parametrization can be easily adopted to surface light fields by mapping one of the Light Field planes to represent local surface coordinates.

Neural rendering is the collection of rendering techniques that use neural networks to generate a "neural" reconstruction of a scene, and render a novel perspective. The term "neural rendering" was first used in [ERB+18]; however, the fundamental spark for neural rendering was the creation of neural networks such as Generative Adversarial Networks (GANs) [GPAM+14], capable of synthesizing highly realistic, novel images from learned priors. A typical neural rendering process is as follows: 1) Images corresponding to specific scene conditions (lighting, layout, viewpoint) are used as inputs, 2) A neural network uses inputs to "learn" the neural representation of the scene, and 3) Novel perspectives of the scene are synthesized using the learned neural representation and novel scene conditions. As a relatively new field, neural rendering covers a diverse set of rendering methods of varying generality, extent of scene definition, and control of the resulting rendered perspective. The neural synthesis components can also be paired with conventional rendering components to varying extents, spanning the range from rendered image retouching (e.g. [MMM+20]) to complete scene and view synthesis, as seen in [FP18]. For a thorough overview of the state-of-the-art in neural rendering, refer to [TFT+20].

Image-Based Rendering has been used as a catch-all term for any rendering based on some form of scene recording, including Light Field rendering [ZC04]. With an intermediate step of inverse rendering, even MBR could be a subset of IBR; likewise, neural rendering relies on images and thus could be a subset of IBR. To draw a distinction between IBR and "all rendering", in this text IBR specifically refers to rendering through transformation, repeated blending, and resampling of existing images through operations such as blending, warping, and reprojection. As such, IBR relies on implicit or explicit knowledge of the scene geometry and scene recording from multiple perspectives using some form of an MCS. The majority of explicit geometry IBR methods fall under the umbrella of Depth-Image Based Rendering (DIBR) [Feh04]. In DIBR, a 2D image of a scene is combined with a corresponding camera parametrization and a 2D depthmap as an explicit encoding of the scene geometry. As in MBR, projective geometry is the basis for DIBR. DIBR is fundamentally a two-step rendering process: first, the 2D image and 2D depthmap are projected to 3D model using projective geometry and camera parameters; second, the 3D model is projected to a new 2D perspective to render a new view. The second step of the DIBR process is very similar to MBR, especially if the projected 3D model is converted from a collection of points with a 3D position $[X, Y, Z]$ and color $[R, G, B]$ to a 3D mesh with associated vertex colors. There are a number of associated issues

stemming from the point-wise projection used in DIBR, such as ghosting, cracks, disocclusions, and so on. A thorough exploration of DIBR artifacts can be found in [DSF+13, ZZY13, Mud15].

2.3 Augmented Telepresence

Augmented Telepresence is the joint product of conventional telepresence and AR. Specifically, AT denotes immersive video-based communication applications that use view augmentation on the presented output [OKY10]. Augmented Telepresence is a relatively recent term and it therefore lies in a relatively fuzzy area on the immersive environment spectrum. Moreover, AT is defined mainly in reference to two other terms —AR and telepresence —which themselves involve a level of definition uncertainty. To remedy this uncertainty, the concepts of AT, AR, and telepresence are unpacked in the following paragraphs.

Augmented Telepresence is a specific type of virtual environment on the immersive environment spectrum, defined by Milgram *et al.* [MTUK95] as a continuous range spanning from full reality to full virtuality. An additional dimension to this spectrum was added by S. Mann [Man02] to further classify these environments based on the magnitude of alteration ("mediation"), and a more recent attempt to clarify the taxonomy was made in [MFY+18]. In most scenarios, VR is considered as the example of full virtuality, and most of the range between VR and "full reality" is described as Mixed Reality (MR) —the indeterminate blending of real and virtual environments [MFY+18]. Augmented Reality is a subset of MR in which the user generally perceives the real world, with virtual objects superimposed or composited over the real view [Azu97]. The common factor of most MR environments —AR included —is that the user perceives their immediate surroundings, with some degree of apparent modification. In contrast, telepresence primarily implies a displacement of the observed environment. Immersive telepresence systems record and transmit a remote location, generally allowing the user to perceive that location as if they were within it [FBA+94].

Augmented Telepresence is therefore similar to AR in that the perceived real environment is augmented or mediated to some extent. Thus AT fits under the MR umbrella term. Augmented Telepresence differs from AR in that the user's perceived real environment is in a different location and seen from a different viewpoint. In order to preserve the agency of the telepresence user, AT is assumed to only refer to real-time or near real-time representations of the perceived environment, without significant temporal delay between the environment recording and replaying.

2.4 Quality of Experience

QoE is defined as "*the degree of delight or annoyance of the user of an application or service*", and "*results from the fulfillment of the user's expectations . . . of the application or service*" (emphasis added) [MR14, IT17, BBDM+13]. Quality of Experience is an overall measure of any system or application through the lens of user interaction. Although there is a strong overlap between the QoE and User Experience (UX) research traditions [Bev08, HT06, Has08], QoE is typically investigated through controlled experiments and quantitative analysis of collected user opinions, without delving into formative design methods. The results for QoE assessments are reported using Mean Opinion Score (MOS), which is the aggregate parametrization of individual user opinions. These opinions are collected using Likert scales, requiring the user to show their level of agreement (from "Strongly Disagree" to "Strongly Agree") on a linear scale for specific statements [Edm05, JKCP15]. For fields such as video quality assessment, there are standards for conducting such experiments, such as [IT14, IT16].

Evaluation based on MOS is an assessment approach inherently based on subjective participant opinions, despite the rigor of quantitative analysis commonly applied to MOS results. The reliance on subjective metrics (MOS) alone to assess overall QoE has been criticized as an incomplete methodology [KHL+16, HHVM16]. One solution is to use both subjective and objective measurements that together reflect the overall user experience. The objective measurements aimed at QoE assessment can be grouped into two kinds of measurement. One kind of objective measurement is participant-task interaction metrics (such as experimental task completion time, error rates, etc.) as demonstrated in [PPLE12]. The other kind of measurement is participant physiological measurements (such as heart rate, gaze attentiveness, etc.), as demonstrated in [KFM+17, CFM19]. The validity of including physiological assessments as part of the overall QoE is of particular interest for VR-adjacent applications that rely on rendering through HMDs, in no small part due to the phenomenon known as "simulator sickness," as shown in [TNP+17, SRS+18, BSI+18].

It is important to note that, despite inclusion of objective metrics as part of a QoE assessment, there is nonetheless a difference between an objective measurement of an application's performance and a QoE assessment of the same application. More specifically, although the QoE may in part depend on application performance, the overall QoE by definition requires an interaction between the assessed application and a user. There is ongoing research focused on replacing test users with AI agents trained using results from past QoE studies, though such efforts are mainly focused on non-interactive applications such as video viewing, as seen in [LXDW18, ZDG+20].

Chapter 3

Related Works

This chapter presents a discussion on the latest research related to multi-camera calibration and synchronization, augmented view rendering for telepresence applications, and QoE implications of view augmentation in telepresence.

3.1 Calibration and Synchronization in Multi-Camera Systems

Camera calibration and synchronization are necessary for enabling multi-camera capture, as mentioned in Section 2.1. Between the two topics, calibration has received more research attention and is a more mature field. There are notable differences between the state of research on calibration and synchronization; therefore, the following discussion separates the discourses on calibration and synchronization.

3.1.1 Calibration

Calibration between 2D RGB cameras is widely considered a "solved problem," at least concerning parametric camera models (such as the pinhole model) that represent physical properties of cameras, sensors, and lens arrays. This consensus can be readily seen from two aspects of the state of the art in multi-camera calibration publications. First, there are archetype implementations of classic calibration solutions [Zha00, Hei00] in widely used computer vision libraries and toolboxes such as [**?**, Gab17, SMS06, Mat17, SMP05]. Second, a large amount of recent work on camera-to-camera calibration in the computer vision community has been focused on more efficient automation of the calibration process [HFP15, RK18, KCT+19, ZLK18], the use of different target objects in place of the traditional checkerboard [AYL18, GLL13, PMP19, LHKP13, LS12, GMCS12, RK12], or the use of autonomous detectors in identifying corresponding features in scenes without a pre-made target (i.e. targetless

calibration [BEMN09, SSS06, GML+14, DEGH12, SMP05]). A parallel track of calibration research focuses on generic camera models [GN01, RS16], which map individual pixels to associated projection rays in 3D space without parametrizing the cameras themselves. However, as pointed out in [SLPS20], adoption of generic camera models outside the calibration research field is slow.

Extrinsic calibration for multi-sensor systems with RGB cameras and range sensors is a slightly less saturated area compared to camera-to-camera calibration. Mixed sensor calibration methods generally fit into three groups: calibration in the 2D domain, 3D domain, and mixed domain.

Calibration in the 2D domain depends on down-projecting range sensor data (e.g. 3D lidar pointclouds) to 2D depthmaps. The subsequent calibration is equivalent to camera-to-camera calibration, as seen in [BNW+18, N+17]. As shown by Villena-Martínez et al. in [VMFGAL+17], only marginal differences in accuracy exist between 2D domain calibration methods ([Bur11, HKH12, **?**]) when used on RGB and ToF camera data. The 3D to 2D downprojection is also used for neural network architectures to derive camera and depth sensor parameters [SPSF17, IRMK18, CVB+19, SJTC19, SSK+19, PKS19].

Calibration in the 3D domain is commonly used to align two depth-sensing devices, such as a lidar and a stereo camera pair. This problem can be cast as a camera calibration issue using a specific target [GJVDM+17, GBMG17, DCRK17, XJZ+19, ANP+09, NDJRD09] or as finding the rotation and translation transformations between partly overlapping point clouds [SVLK19, WMHB19, Ekl19, YCWY17, XOX18, PMRHC17, NKB19b, NKB19a, ZZS+17, KPKC19, VŠS+19, JYL+19, JLZ+19, PH17, KKL18]. In systems with stereo cameras, conventional 2D camera calibration approaches are used to enable depth estimation from the stereo pair, and in systems with a single RGB camera, a Simultaneous Localization and Mapping (SLAM) process (surveyed in [TUI17, YASZ17, SMT18]) is used to produce a 3D point cloud from the 2D camera.

Finally, calibration in the mixed domain refers to identifying features in each sensors' native domain and finding a valid 2D-to-3D feature mapping. A large number of methods [CXZ19, ZLK18, VBWN19, GLL13, PMP19, VŠMH14, DSRK18, DKG19, SJL+18, TH17, HJT17] solve the registration problem by providing a calibration target with features that are identifiable in both 2D and 3D domains. Other approaches [JXC+18, JCK19, IOI18, DS17, KCC16, FTK19, ZHLS19, RLE+18, CS19] establish 2D-to-3D feature correspondences without a predefined calibration target, relying instead on expected properties of the scene content.

The assessment of camera-to-camera (or camera-to-range-sensor) calibration in the aforementioned literature is typically based on point reprojection error, i.e. the distance between a detected point and its projection from 2D (to 3D) to 2D according to the estimated camera parameters. The reprojection error can also be cast into the 3D domain, verifying point projection in 3D space against a reference measurement of scene geometry, as in [SVHVG+08], or by including a 3D projected position error into the loss function of a neural network for calibration [IRMK18]. In contrast, less focus is placed on verifying the resulting calibration parameters with respect

to the physical camera setup and placement. A notable exception to this trend is the recent analysis by Schöps *et al.* [SLPS20]. In this analysis, both reprojection error and estimated camera positioning were used to argue for the need to adopt generic camera models, relating pixels to their 3D observation lines, as opposed to the commonly chosen parametric models that relate pixels to physical properties of the camera and lens setups. As [SLPS20] observed, although there is potential benefit in adopting generic camera models, the common practice in calibration relies on the standard parametric models and their respective calibration tools. Similarly, the common practice in calibration evaluation relies on the point reprojection error, without considering the *de facto* camera parametrization accuracy.

3.1.2 Synchronization

Camera-to-camera synchronization is not covered as thoroughly as calibration, in part because one can sidestep the synchronization issue by using cameras with externally synchronized sensor shutters, and in part because the temporal offset is not an inherent component of the PCM (described in Section 2.1) or generic camera models applied to MCS, such as [GNN15, LLZC14, SSL13, SFHT16, LSFW14, WWDG13, Ple03]. The existing solutions to desynchronized capture commonly fit in either sequence alignment, wherein a synchronization error is estimated after data capture, or implicit synchronization, where downstream consumers of MCS output expect and accommodate for desynchronized captured data. Additionally, external synchronization is replicated with time-scheduled software triggering as seen in [LZT06, AWGC19], with residual synchronization error dependent on sensor API.

Sequence alignment, also called "soft synchronization" [WX+18], refers to estimating a synchronization error from various cues within the captured data. The estimation is based on best-fit alignment of, for example, global image intensity variation [DPSL11, CI02] or correspondence of local feature point trajectories [ZLJ+19, LY06, TVG04, LM13, EB13, PM10, DZL06, PCSK10]. A handful of methods rely instead on supplementary information such as per-camera audio tracks [SBW07], sensor timestamps [WX+18], or bitrate variation during video encoding [SSE+13, PSG17].

Implicit synchronization is often a side effect of incorporating error tolerance in rendering or 3D mapping processes. In [RKLM12], depthmaps from a desynchronized range sensor are used as a low-resolution guide for image-to-image correspondence matching between two synchronized cameras. The synchronous correspondences are thereafter used for novel view rendering. Two desynchronized moving cameras are used for static scene reconstruction in [KSC15]. Synchronization error is corrected during camera to camera point reprojection, by displacing the origin of one sensor along the estimated camera path through the environment on a least-reprojection-error basis. Similarly, the extrinsic camera calibration methods in [AKF+17, NK07, NS09] handle synchronization error by aligning feature point trajectories over a series of frames rather than matching discrete points per frame.

Throughout all the aforementioned studies, there is the implicit assumption that

synchronization error is undesirable. Unsynchronized data is either used as a rough guide (in implicit synchronization) or aligned to the nearest frame and used as is (in soft synchronization). Yet, neither sequence alignment nor implicit synchronization specifies the consequences of desynchronized capture or demonstrates why synchronization error is undesirable.

3.2 Applications of Augmented Telepresence

Augmented Telepresence applications are fundamentally linked to AR, as defined in Section 2.3. The use of VR, AR and AT in non-entertainment contexts is steadily increasing in education [AA17], healthcare [PM19], manufacturing [MMB20] and construction [NHBH20], and both AR and remote-operation (i.e. telepresence) centers are expected to be key parts of future industry [KH18]. However, AT applications as such are not yet as widespread as VR or telepresence on their own.

The worker safety angle has been a key motivator for AR and particularly VR uptake in industries such as construction and mining. The majority of safety-focused applications have been VR simulations of workspaces designed for worker training, as shown in surveys by Li *et al.* [LYC+18b] and Noghabei *et al.* [NHBH20]. Pilot studies such as [GJ15, PPPF17, Zha17, AGSH20, ID19] have demonstrated the effectiveness of such virtual environments for training purposes. However, VR training does not directly address safety during the actual work tasks; telepresence does.

Applied telepresence is best exemplified by the two systems shown in [TRG+17] and [BBV+20]. Tripicchio *et al.* presented an immersive interface for a remotely controlled crane vehicle in [TRG+17], and Bejczy *et al.* showed a semi-immersive interface and system for remote control of robotic arm manipulators in [BBV+20]. The vehicle control interface is a fully immersive replication of an in-vehicle point of view, with tactile replicas of control joysticks. The robot manipulator interface instead presents multiple disjointed views of the manipulator and the respective environment. The commonality between the two systems is the underlying presentation method: in both examples, directly recorded camera views from a MCS are passed to virtual view panels in a VR environment, presented through a VR headset. Similar interfaces for robot arm control from an ego-centric (a.k.a. "embodied") viewpoint can be seen in [LFS19, BPG+17], while telepresence through robotic embodiment is extensively surveyed in [TKKVE20].

The combination of view augmentation and the aforementioned applied telepresence model forms the archetype for most AT applications. Augmented Telepresence with partial view augmentation is demonstrated in [BLB+18, VPR+18], and AT with complete view replacement can be seen in [ODA+20, LP18]. Bruno *et al.* in [BLB+18] presented a control interface for a robotic arm intended for remotely operated underwater vehicles. View augmentation is introduced by overlaying the direct camera feed with a 3D reconstruction of the observed scene geometry as a false-color depthmap overlay, in addition to showing the non-augmented views and the reconstructed geometry in separate views, similar to the semi-immersive direct views in [BBV+20, YLK20]. Vagvolgyi *et al.* [VPR+18] also showed a depth-overlaid

camera view interface for a robotic arm mounted to a vehicle intended for in-orbit satellite repairs; however, the overlaid 3D depth is taken from a reference 3D model of the target object and registered to the observed object's placement in the scene. Omarali *et al.* [ODA+20] completely replaced the observed camera views with a colored 3D pointcloud composite of the scene recorded from multiple views, and Lee *et al.* [LP18] likewise presented a composite 3D pointcloud with additional virtual tracking markers inserted into the virtual 3D space. Telepresence and AT can manifest through various kinds of view augmentation and rendering, as demonstrated by [BLB+18, VPR+18, BBV+20, YLK20, ODA+20, LP18]. Most activity in telepresence (and, by extension, AT) is related to control interfaces for robotic manipulators; however, as demonstrated by [TRG+17] and [KH18], there is both interest and potential for a broader use of telepresence and AT in industrial applications.

3.3 View Rendering for Augmented Telepresence

View rendering specifically for AT is the process of converting conventional multiple viewpoint capture from an MCS into an immersive presentation of augmented views. Rendering for AT tends to blend image-based and model-based rendering approaches (see Section 2.2) to achieve two separate purposes: an immersive view presentation, and some form of view augmentation.

3.3.1 Immersive View Rendering

Immersive presentation for telepresence is commonly achieved by using an HMD as the output interface and thus has a strong relationship to immersive multimedia presentation, such as 360-degree video rendering. A common presentation method is "surround projection," where camera views are wholly or partly mapped onto a curved surface approximately centered on the virtual position of the HMD, corresponding to the HMD viewport [FLPH19]. To allow for a greater degree of viewer movement freedom, the projection geometry is often modified. In [BTH15], stereo 360-degree panorama views are reprojected onto views corresponding to a narrower baseline, using associated scene depthmaps. In [SKC+19], a spherical captured image is split into three layers (foreground, intermediate background and background) to approximate scene geometry and allow for a wider range of viewpoint translation. In [LKK+16], the projection surface sphere is deformed according to estimated depth from overlap regions of input views to allow for a more accurate parallax for single-surface projection.

Alternative approaches to "surround projection" that appear in the AT context are "direct" and "skeumorphic" projections. "Direct" projection is a straightforward passing of stereo camera views to an HMD's left and right eye images. This projection allows for stereoscopic depth perception, but lacks any degree of freedom for viewer movement, and has mainly been used in see-through AR HMDs [CFF18] or combined with pan-tilt motors on stereo cameras that replicate the VR HMD movement [KF16]. "Skeumorphic" projection is the replication of flat-display in-

terfaces and viewports in a virtual environment with full movement freedom, as seen in [TRG+17, BBV+20], thereby replicating or approximating real-world, non-immersive control interfaces in an immersive environment.

More broadly, any virtual view rendering approach can be adapted as an immersive presentation by rendering novel views of a 3D scene for the left and right eye part of an HMD panel. Starting from free-viewpoint video (surveyed in [LTT15]) rendering, rendering pipelines with live sensor input have recently been presented in [RSA20, MNS19]. In [RSA20], a mesh of a central object was created and iteratively refined from the convex hull obtained from object silhouette projection using multiple recording cameras. Render quality was improved by refinement of mesh normals, temporal consistency, and a surface-normal dependent blending of input view pixels for mesh texturing. In [MNS19], a dense 3D mesh was triangulated from point clouds captured by RGB-D sensors, using a moving least-squares method for joint denoising and rectification of clouds. The subsequent mesh was textured by projecting the mesh vertices onto the camera image plane as texture coordinates.

3.3.2 View Augmentation

View augmentation can be achieved by overlaying an additional, virtual object over the recorded scene view, as seen in [BLB+18, VPR+18, OKY15, RHF+18], or by removing some scene content, as in [WP19, LZS18, OMS17]. In [BLB+18], the augmented overlay (depth colorization) was projected pixel by pixel onto a 2D disparity map coincident with a direct camera view. In [OKY15] and [VPR+18] (with additional details in [PVG+19]), camera views were projected onto a reconstructed ([OKY15]) or prebuilt ([PVG+19]) 3D model of the scene, and virtual fixtures were added to the 3D virtual scene, with optional anchoring to the scene geometry. In [RHF+18], the virtual fixture was projected to 2D and partly overlaid on the camera view, using the camera view's depthmap to block parts of the virtual fixture. When camera views are projected onto a curved surrounding surface for immersive rendering, virtual fixtures are interposed between the curved surface and the virtual HMD render cameras, as seen in [RPAC17]. Content removal from scenes is less common in immersive rendering, but it is typically achieved by replacing the removed scene section from another camera view, which has been displaced either spatially ([LZS18, OMS17]) or temporally ([WP19]).

Both layered surrounding projections and most types of view augmentation depend on access to detailed scene geometry in the form of depth maps. Due to low depth sensor resolution, and errors in image based depth estimation, depth map improvement is an important component of immersive and augmented view rendering. Depth upscaling based on features in corresponding high-resolution RGB images is a prevalent solution. For instance, [FRR+13] and [SSO13] used edge features to limit a global diffusion of sparse projected depth, [ZSK20] added stereoscopic projection consistency to the optimization cost, and [PHHD16] used edges as boundaries for patch-based interpolation between depth points. Neural networks have also been used to refine the upscaling process with high-resolution RGB as a guide. In [NLC+17], a depthmap was upscaled through bicubic upsampling and

refined through a dual stream convolutional neural network sharing weights between the depthmap and edge image refinement streams. In [CG20], the RGB image was downscaled and re-upscaled in one network stream, with image upscaling weights used in a parallel stream for depth upscaling. In [WDG+18], the RGB image was used to directly synthesize a corresponding depthmap with one network, which was then used as an upscaling guide for a depth upscaling network, similar to the upscaling layers in [NLC+17] and [CG20]. For complete surveys of neural- and image-guided depth upscaling, refer to [ECJ17, LJBB20].

3.4 Quality of Experience for Augmented Telepresence

Quality of Experience assessments for AT are closely related to QoE assessment for AR and VR, in large part because of the overlap in chosen display technologies (i.e. VR headsets). A QoE assessment ideally has to involve both subjective and objective metrics, as noted in Section 2.4. The need for subjective participant reporting on experiences has led to the majority of QoE assessments being conducted on test implementations of AT systems, such as [CFM19, BPG+17, PBRA15, PTCR+18, LW15, CFF18], or corresponding VR simulators replicating the live scenarios, as in [BSI+18]. The need for objective metrics has resulted in two intertwined research tracks: the collection of psycho-physiological measurements, and the collection of task completion metrics.

Psycho-Physiological Assessment (PPA) relies on measurements of human physiology through Electro-Encephalography (EEG), Electro-Cardiography (ECG), eye movement registration and gaze tracking, all in an effort to better measure test participants' psychological state during QoE testing. Psycho-Physiological Assessment was proposed as a necessary extension to subjective measurements in [KHL+16, CFM19]. As Kroupi *et al.* [KHL+16] showed, there tends to be a connection between self-reported QoE and physiological measurements. Psycho-Physiological Assessment has been used to directly probe users' level of immersion and sense of realism in immersive video viewing in [BÁAGPB19], and to gauge the effect of transmission delays in remote immersive operation [CFM19]. However, the broad consensus is that PPA is a supplement to —not a replacement for —QoE assessment of immersive multimedia technologies and that PPA should be used to infer the higher cognitive processes of test participants. The PPA methodology and progress towards standardization was extensively surveyed in [EDM+16] and [BÁRTPB18].

Task completion assessment is a QoE measurement specific to interactive applications, the use of which was suggested by Puig *et al.* in [PPLE12] and supported by Keighrey *et al.* in [KFM+17]. The task-related metrics (a.k.a. "implicit metrics") in [PPLE12] were task completion time, error rates and task accuracy; in that pilot study, correlations were found between user reported QoE and the gradual improvement of implicit metrics. In [KFM+17], implicit measurements and PPA measurements were compared in a simultaneous QoE experiment with an interactive system. Correlation between some task completion and PPA measurements was found, with both PPA and implicit metrics hinting at a higher perceived task complexity in VR

compared to the equivalent task in AR. In [RBW+14], implicit metrics together with explicit subjective scores were used to explore the discrepancy between user perception and task performance, finding that test participants subjectively preferred a non-stereoscopic telepresence system, but performed better at depth-reliant tasks in the stereoscopic system equivalent. In summary, [RBW+14, KFM+17, PPLE12] suggest that implicit metrics complete the subjective QoE assessment instead of supplanting it. Implicit, task based metrics have been extensively used to demonstrate the benefits of augmented task guidance as in [ACCM15, BSEN18, VPR+18, WHS19, BPG+17, PBRA15] and to a lesser extent (not via telepresence or immersive environments), to show significance in view position and camera Field of View (FoV) [TSS18, SLZ+18, LTM19].

QoE assessment in immersive environments is affected in particular ways by the use of HMDs as a consequence of HMD technology. The simple choice of using an immersive VR headset instead of conventional displays leads to higher cognitive load for test participants [BSE+17], and a reduction in HMD FoV further increases cognitive load. Latency induced VR sickness is also an important aspect of QoE in setups reliant on VR headsets, as shown in [BDA+19, SRS+18, TNP+17]. A less obvious consequence of VR headsets was shown in [PTCR+18, LW15, CFF18], who found that depth perception can be significantly impaired through HMDs and that people generally tend to underestimate stereoscopic depth in VR and AR environments. As Alnizami *et al.* pointed out in [ASOC17], measuring even passive VR experiences has to go beyond just video quality, and must include consideration of elements such as VR headset ergonomics. The aforementioned studies show that comprehensive QoE assessment becomes even more important for interactive HMD-based experiences, such as task-specific AR and AT applications.

Chapter 4

Methodology

This chapter presents the methodology employed to address the RQs defined in Section 1.4 within the context of the background and related works discussed in Chapters 2 and 3. More specifically, this chapter covers the identification of knowledge gaps, and how the relevant theory, solution synthesis, and assessment approaches were employed to address each research question. Details of the proposed solutions are given in Chapter 5.

4.1 Knowledge Gaps

The RQs in Section 1.4 were formulated as a consequence of the state of the art canvassing and knowledge gap identification, while attempting to address the two-fold research purposes of multi-camera and multi-sensor system design (**P1**) and user experience of multi-sensor based AT (**P2**). This section outlines the identified knowledge gaps associated with the research purpose, and connects to the respective RQs.

4.1.1 Multi-Camera Systems for Augmented Telepresence

Augmented Telepresence based on MCSs invariably requires camera calibration and synchronization to fuse multi-sensor data and render novel or augmented views. The common practice for MCS applications is to pick from a range of standard calibration methods for parametric camera models, as discussed in Section 3.1. However, validation and comparison of such methods relies on pixel reprojection error, and is thus dependent on the quality of the input data and correspondence matching. As [SSO14] highlighted, reprojection (and therefore any image-based rendering) is highly sensitive to errors in camera parametrization. Yet, there is a lack of comparative analysis of calibration methods with respect to camera parameter estimation accuracy, leading to **RQ 1.1**.

There is a distinction between strict (external triggering) and soft (nearest-frame

alignment) synchronization in MCSs (Section 3.1). In the context of designing new MCSs, it is not readily known whether strict synchronization is necessary for a specific application and to what level of accuracy. A parametric model that relates synchronization accuracy to scene depth could be beneficial for determining the thresholds of sufficient synchronization and for constructing MCSs, since many AT- and MCS-based rendering applications rely on scene depth in some form (see Sections 3.3, 3.2); this leads to **RQ 1.2**.

Telepresence applications tend to rely on MCSs that bundle encoding and decoding at central nodes even when view capture is separated [MP04, YEBM02, BK10], leading to potential bandwidth or latency bottlenecks with additional sensors. Hardware accelerated single-view video encoders are readily available and optimized for low latency [LAV+19, SBS15]. Such encoders, together with multimedia transmission frameworks like [tea12], may enable scalable MCS designs for telepresence; this leads to **RQ 1.3**.

Multi-Camera Systems with cameras and lidars are already employed in control, mapping and vehicle automation in various industries [CAB+18, TRG+17, MSV18, ClHClJy19], and there is interest in applying AT for industrial applications [KH18, TRG+17]. Most view augmentation in AT is additive (Section 3.3), but view augmentation through content removal (as seen in [OMS17, LZS18, WP19]) may provide AT users with better awareness of the work environment in safety-critical industries such as mining (Section 3.2). However, such contexts preclude the use of pre-conditioned environment models or plausible-seeming environment synthesis, and the prevalent range sensors (lidars) provide only sparse depth as a basis for view augmentation (Section 3.3). The challenge of depth upscaling, real-time augmentation, and rendering in telepresence without pre-conditioned or hallucinated data leads to **RQ 1.4**.

4.1.2 User Experience of Augmented Telepresence

Augmented Telepresence applications that use immersive rendering can improve user experience by providing augmented guidance, as seen in [ACCM15, BSEN18, VPR+18, WHS19, BPG+18, PBRA15], and by providing stereoscopic views, as seen in [RBW+14]. However, immersive stereoscopic telepresence requires the use of HMDs, which tend to distort users' depth perception [PTCR+18, LW15, CFF18]. At the same time, in non-immersive 2D rendering, view properties such as FoV and viewport arrangement also affect user experience. Camera placement defines viewpoint in immersive and direct projections (Section 3.3); therefore, view properties may effect user experience in AT, leading to **RQ 2.1**. Furthermore, the effects of depth perception in HMDs, view properties in immersive HMD rendering, and augmented guidance may all interact and lead to joint effects; this leads to **RQ 2.2** and to joint exploration of **RQ 2.1** and **RQ 2.2**.

4.2 Synthesis of Proposed Solutions

This section presents the methodology for addressing the RQs of Section 1.4.

4.2.1 Multi-Camera Systems for Augmented Telepresence

RQ 1.1 required a comparative assessment of existing calibration methods. Both target-based and target-less calibration methods were chosen because target-less calibration has been shown as easier to perform in MCS applications and thus is more appealing under equal calibration accuracy. The ground truth of internal camera parameters of the pinhole-with-distortion model (see Section 2.1.1) cannot be easily obtained for any single camera, but it is necessary for gauging the accuracy of calibration with respect to the ground truth. Hence, a column of three cameras is used to represent a 5-by-3 grid of cameras through horizontal displacement. This adds an identity (equality) constraint for the internal parameters of all views in a grid row belonging to the same real camera, as well as an identity constraint between relative rotation and position of all views in a grid column. Calibration was performed with all test methods as if on a 15-camera dataset; the variations between estimated parameter values under identity constraints were treated as the errors in ground truth estimation.

RQ 1.2 required the combination of synchronization Eq. (2.5) and projective geometry Eq. (2.1) to find the consequences of synchronization error. In an MCS, Eq. (2.1) allows for the triangulation of the 3D position of every scene point at the intersection of rays connecting the scene point with the optical centers of the recording cameras, subject to intrinsic camera parameters and lens distortions. Upon incorrect synchronization, a moving object will be observed at different real positions from different views. In the proposed model of synchronization error consequence, "depth uncertainty" is introduced as a range of plausible depth values along a camera ray. This range is determined by the synchronization error, object movement speed, and —because camera rays are defined through camera intrinsic and extrinsic parameters —by the relative positioning of the unsynchronized cameras. Aggregate depth uncertainty is calculated over all rays of an MCS for a descriptive parametrization of particular MCS arrangements. The proposed model was used to show the resulting depth uncertainty of sample MCS arrangements and to map the effect of varying camera convergence, synchronization error, and object speed on the extent of depth uncertainty. Additionally, a possible reduction in computational complexity was explored because calculating the generic case of the proposed model scales directly with the number of rays (i.e. camera resolution) in the MCS.

RQ 1.3 was addressed by designing and implementing a scalable MCS, and assessing the implementation's performance for video processing and transmission. In contrast to [MP04, YEBM02, BK10], where camera stream processing is centralized, the proposed system assigns a processing device for each camera, thereby approximating an MCS composed entirely of smart cameras without losing features like hardware synchronization. This arrangement allows for parallel video processing

and coding regardless of camera count. Transmission is enabled via [tea12], as it is widely used for multimedia streaming, supports accelerated video encoders on various devices, and enables more flexible data processing by transmission to and video transcoding in cloud-based virtual servers. The implementation was tested for processing latency on the encoding and decoding sides to validate the suitability of [tea12].

RQ 1.4 was addressed by designing and implementing a system for augmented remote operation (non-immersive AT). The capture side of the proposed system was partly based on an extension of the MCS from **RQ 1.3**, with added lidar sensors and optional bypass of video transmission and compression. The sensor data was used to render a plausible remote-operator interface, containing the original camera views as well as augmented and novel views. A non-immersive interface layout was chosen to more closely approximate the existing remote operator interfaces in the given context and to place emphasis on the proposed view generation process. Views are generated at full resolution, and placed in an operator interface via the windowing manager of [SML06]; this allows a separation between the proposed view generation process and the operator interface composition. View generation is based on a combination of image-based rendering (Section 2.2), fast upscaling of sparse depth (Section 3.3), and content replacement through depth-based reprojection. Since lidar depth was used as the basis for projection between views, a fast temporal filtering process was added to reduce lidar measurement oscillation and intermittent measurement drop-out. The proposed view generation process was mostly implemented in CUDA (a parallel processing framework), with parts of the process designed for easier parallelization.

4.2.2 User Experience of Augmented Telepresence

RQ 2.1 and **RQ 2.2** are QoE evaluations. As such, a subjective and objective (task-performance based) assessment was conducted with test participants and a purpose-built prototype AT system. The test system was partly based on the framework of the MCS developed for RQ 1.3 and has added stereoscopic rendering in a VR HMD. A VR HMD was chosen instead of a see-through AR HMD because of a generally wider FoV, which leads to less cognitive load [BSE+17] and less discrepancy between the visible real world and the rendered AR. Two camera pairs were used to provide two different viewing positions to the system users. Both viewing positions are from the third-person perspective since first-person (ego-centric) teleoperation has been covered by [BPG+18, PBRA15]. During rendering, the selected camera pair is projected to curved sections of a projection sphere (see immersive presentation, Section 3.3) to decouple HMD movement from camera movement. Such decoupling is required to support a low-latency response to HMD movement [BSI+18] while having a regular frame rate for cameras recording the remote scene. Each eye's image of the HMD uses its own projection sphere, and the camera baseline within a pair is matched to the average HMD eye baseline to support stereoscopic viewing. The rendered view augmentations are tracked to the content of displayed camera images and rendered separately for each HMD eye at corresponding positions, to enable

stereoscopic augmentation rendering. As [SSR18] found, non-stereoscopic AR over stereoscopic content may damage user QoE. To test the effects of view position and augmentation, the AT system supports depth-aiding view augmentations, and camera pairs are placed at positions that emphasize stereoscopic depth perception of the observed scene to different extents.

4.3 Verification

This section summarizes the verification methods used to address the RQs via the proposed solutions.

RQ 1.1: A dataset of calibration images with 15 view positions was captured with a vertical three-camera stack on a programmable dolly; this forced the calibration methods to estimate the same cameras' parameters five times per calibration attempt. Multiple calibration runs were performed for each calibration method. Target-based calibration ([Zha00] implemented in AMCC [WMU13]) was compared with targetless (Bundler, VisualSFM, BlueCCal [SSS06, Wu13, SMP05]) calibration. These calibration methods were chosen due to their prevalence in related works and the availability of implementations.

The comparison is based on variance of estimated lens distortion coefficients, camera-to-camera distances, and camera-to-camera rotation. For parameters without explicit known ground truth, calibration accuracy was judged by the standard deviation and measurement distribution of repeated parameter estimates for the same physical cameras (or camera pair) at different view positions in the dataset. For camera-to-camera distance, the standard deviation, distribution of repeated parameter estimates, and mean-square-error relative to ground truth was checked. For further details, see Section 4 of Paper I.

RQ 1.2: The derivation of the proposed depth uncertainty model is detailed in Section 3 of Paper II. Three experiments were carried out using a fixed set of camera parameters (sensor size, camera placement, view convergence, synchronization error) to represent a realistic two-camera system for depth uncertainty estimation. In the first experiment, the synchronization error parameter varied from 0 ms (no synchronization error) to 25 ms (half-frame desynchronization at 20 Frames per Second (FPS)), and in-scene movement speed parameter varied from 0.7 to 2.8 m/s, equivalent to half and double average walking speed. In the second experiment, camera convergence angle parameter was varied between 0 and 40 degrees.

Overall depth uncertainty in these experiments was calculated as the mean of the depth uncertainties of all possible intersections of rays from both cameras. The third experiment compared the overall depth uncertainty estimation for all ray intersections and for reduced ray intersections, narrowing to the principal ray of one camera. The resulting overall depth uncertainty and distributions of per-ray uncertainty were used as the basis for comparison. For further details, see Section 4 of Paper II.

RQ 1.3: The system design is detailed in Section 3 of Paper III, and implementa-

tion details for the tested MCS system are given in Section 4 of Paper III. The test system consisted of 11 camera-and-computer pairs, with 10 RGB cameras and one range camera. In each camera pair, image stream was encoded to h.264 video and sent via [tea12] to virtual instances in a private cloud. The video streams were transcoded to a different compression ratio and sent back to a receiving computer, terminating the stream with a video sink element of [tea12]. Communication to and from the cloud took place through the public Internet to represent realistic conditions for the transmission chain components. The cumulative and component-wise latency of frame processing from camera to the Ethernet interface, measured over multiple attempts, was used as basis for validating [tea12] for real-time MCS capture.

RQ 1.4: The technical details and constraints of the test system are detailed in Section 3 of Paper VI. Test recordings were performed in a mine-like lab environment, with cameras and lidars placed at proportional distances and positions as possible on a mining machine. Performance of the view generation process was tested by measuring the execution time for all main components and the mean render time per frame with three sequences of varying amount of in-scene motion. Performance of lidar filtering was measured by the mean and median per-ray variance of lidar depth. The outputs of depth upscaling and view projection were also presented and contrasted with alternate approaches. Section 7 of Paper VI has further details.

RQ 2.1 and **RQ 2.2**: A user test protocol was defined to gather test subject judgement of the AT system, for which the implementation details are given in Section 3.1 in Paper V. The system was tested by 27 non-expert participants, with participants asked to use the AT system to remotely pilot a toy vehicle to reach and accurately touch a number of targets in a random sequence. Each test participant was afforded a training phase to become accustomed to the AT system and given a series of test attempts to complete a navigation task requiring depth judgement. Each attempt presented a different configuration of the test parameters (camera view position, and view augmentation type). The order of parameter permutations for each participant was randomized. The total test duration per participant was kept short to avoid overall fatigue, as suggested in [Cur17]. Participants were also asked to remove the HMD after each test attempt to reduce visual fatigue [GWZ+19].

In line with the QoE methodology discussed in Section 3.4, implicit (task completion) metrics were tracked by the AT system in addition to gathering the explicit, user-reported subjective experience for each test attempt. Simulator sickness questionnaires were also used to assess the changes in participant state caused by the experiment. The implicit system-tracked metrics were the number of targets reached, time to reach target, time spent near target, and accuracy of target touch. The explicit metrics posed questions about task accomplishment, task difficulty, viewpoint helpfulness and augmentation helpfulness on 5-point interval scales. The explicit measurements were aggregated into mean opinion scores for each scale, and implicit measurements were aggregated to mean measurements. The measurement distributions were tested for normality. Paired-sample T-tests were used to determine the significance of differences per each measurement type, and repeated-measures analysis of variance tests were used to investigate the interactions between the different test factors. For further details, see Sec. 3 in Paper IV and Sec. 3.2 to 3.4 in Paper V.

Chapter 5

Results

This chapter covers the main results of addressing the research questions from Section 1.4 via the solutions described in Sections 4.2 and 4.3. One model and three systems were developed over the course of addressing **RQ 1.2**, **RQ 1.3**, **RQ 1.4**, **RQ 2.1** and **RQ 2.2**, and these are summarized in Section 5.1. The main outcomes of the proposed solutions are presented in Section 5.2.

5.1 Proposed Models and Systems

5.1.1 A Model of Depth Uncertainty from Synchronization Error

Depth uncertainty is the range between nearest and farthest possible distances that a moving object can be located in, when observed by an MCS with de-synchronized cameras. Given rays $\overrightarrow{r}_{\mathrm{A}}, \overrightarrow{r}_{\mathrm{B}}$ of cameras 'A' and 'B' with synchronization error Δt, the depth uncertainty Δd of observing an object $\vec{E}$ moving at speed $v_{\vec{E}}$ is

$$\Delta d = \frac{2\sqrt{\left(v_{\vec{E}}\Delta t\right)^2 - \|\vec{m}\|^2}}{\sin(\theta)} \ , \ \theta = \arccos\left(\frac{\vec{r}_{\mathrm{A}} \cdot \vec{r}_{\mathrm{B}}}{\|\vec{r}_{\mathrm{A}}\| \ \|\vec{r}_{\mathrm{B}}\|}\right) \tag{5.1}$$

where $\|\vec{m}\|$ is the nearest distance between $\overrightarrow{r}_{\mathrm{A}}$ and $\overrightarrow{r}_{\mathrm{B}}$, and the vectors $\vec{r}_{\mathrm{A}}, \vec{r}_{\mathrm{B}}$ denote the directions of rays $\overrightarrow{r}_{\mathrm{A}}, \overrightarrow{r}_{\mathrm{B}}$. The general depth uncertainty $\overline{\Delta d}_{\mathrm{A,B}}$ of an MCS with cameras 'A','B' is

$$\overline{\Delta d}_{\mathrm{A,B}} = \frac{1}{n}\sum_{k=1}^{n} \Delta d_k \text{ , where } \Delta d_k \in \{\Delta d \mid \forall \left(\overrightarrow{r}_{\mathrm{A}}, \overrightarrow{r}_{\mathrm{B}} \implies \Delta d \in \mathbb{R}^{+}\right) \} \ . \tag{5.2}$$

A ray $\overrightarrow{r}_{\mathrm{n}}$ can be expressed by intrinsic and extrinsic parameters of camera 'n' (see Section 2.1.1) via

$$\overrightarrow{r}_{\mathrm{n}} = C_{\mathrm{n}} + \lambda \mathcal{R}_{\mathrm{n}}^{-1} \mathcal{K}_{\mathrm{n}}^{-1} \vec{c}_{\mathrm{n}} \ , \tag{5.3}$$

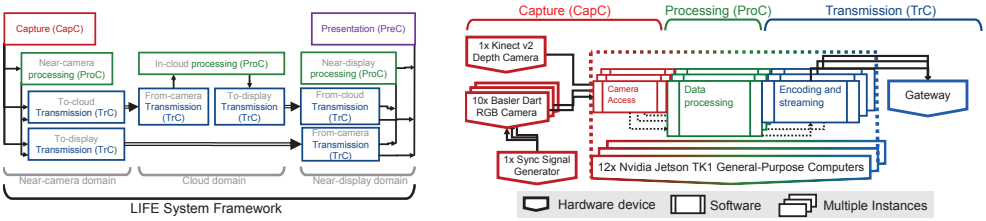

Figure 5.1: Left: High-level view of the scalable end-to-end framework and its components. Right: A multi-camera system implementation of the framework's near-camera domain.

where $\overrightarrow{r}_{\mathrm{n}}$ intersects the camera image plane at pixel coordinate $\vec{c}_{\mathrm{n}} = [u; v; 1]$. Derivation and further details are presented in Section 3 of Paper II.

5.1.2 A Framework for Scalable End-to-End Systems

An end-to-end system framework ("LIFE System framework," Fig. 5.1) was designed by distributing the capture, presentation, processing and transmission across three domains encompassing hardware and software. The capture component encapsulates system cameras and camera control devices. The processing component covers modification of recorded data and generation of supplementary information. The transmission component contains mechanisms such as networking, data stream forming, compression, and decompression. The presentation component encapsulates the rendering process and display hardware and control devices. Dividing the processes among these components enables a degree of independence from technical details such as camera APIs towards the overall end-to-end system, and supports an easier upgrade and extension path for subsequent implementations. Further description and implementation details are in Sections 3 and 4 of Paper III.

5.1.3 A System for Real-Time Augmented Remote Operation

An augmented remote view system was proposed and implemented for rendering augmented and novel views from at least one lidar and two camera inputs. The data transmission is based on the aforementioned framework (Section 5.1.2), and the system as a whole comprises the near-camera and near-display domains. The view generation relies only on inbound sensor data during the live capture, without pre-built models or pre-trained statistical dictionaries. The view generation process, situated in the near-display domain and summarized in Figure 5.2, is split into two simultaneously occurring stages: sensor data accumulation and pre-processing, and the view generation pipeline. As part of pre-processing, lidar data is filtered to reduce static-point oscillation. The proposed filtering is designed to exploit unused time intervals during the lidar frame assembly process, thereby avoiding any filtering-induced delay. The view generation is designed to avoid pre-conditioned data or template dictionaries, and operates entirely on the latest data available from

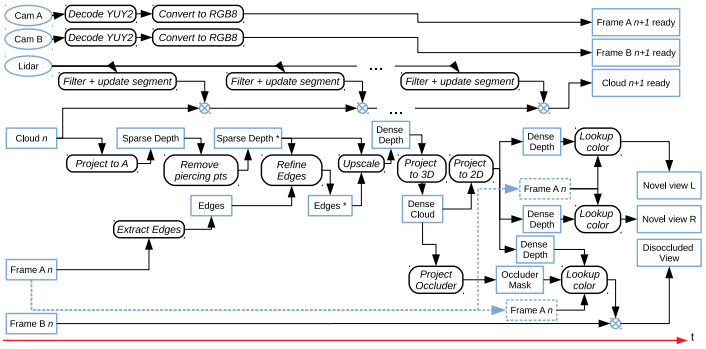

Figure 5.2: High-level overview of view generation process for augmented remote operation.

the sensors (lidar and cameras). One augmented and two novel views of the scene are generated using the lidar frame as the core geometry. View generation relies on projecting sparse lidar points to a camera view, densifying the resulting depthmap, and using that geometry for augmented and novel view creation. Due to the sparseness of the lidar points and their projection to another viewpoint, additional filtering is performed to identify and remove "pierce-through" points that belong to background elements but project inbetween points belonging to continuous foreground elements. The augmented view specifically comprises an in-view occluder removal, as a form of diminished-reality augmentation; the occluder is identified and masked based on lidar point presence in a designated area in the scene's 3D space. The generated views are presented alongside the original views in a flat operator interface without immersive rendering, to be visually consistent with existing user interfaces for remote operation in underground mines. Further details are given in Sections 3 to 5 in Paper VI.

5.1.4 A System for Depth-Aiding Augmented Telepresence

The AT system is based on a combination of the near-camera domain of the MCS system in Section 5.1.2 and an immersive augmented rendering pipeline implemented in OpenVR. Stereoscopic camera images are projected to a virtual sphere, and view content is used to anchor virtual AR elements between the projected views and the HMD position in the virtual render space. For augmentations that track the in-scene objects, the in-view augmentations for each eye are positioned in the virtual space along a line between the optical center of that eye's virtual camera, and the corresponding object pixels of the respective image projections (see Fig. 5.3, right); this ensures a stereoscopically correct AR rendering that allows for HMD movement whilst staying consistent with the real camera stereoscopy. Three kinds of augmentations (A1, A2 and A3 in Fig. 5.3) were used to assist with remote operation, namely a target indicator; a relative target-position grid map; and a visual X, Y, Z distance-to-target indicator. Further details are given in Section 3.1 of Paper V.

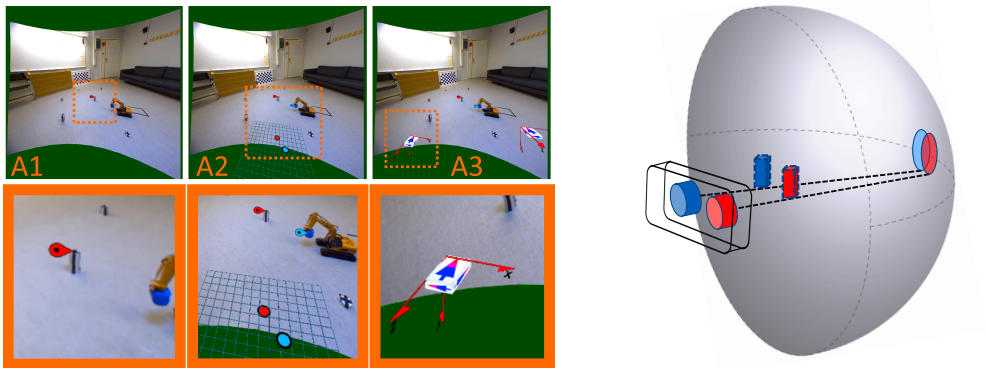

Figure 5.3: Left: Depth-assisting AR designs (A1, A2, A3) used in AT. Right: Principle for stereoscopic rendering of an AR element along view path between left/right HMD eye and anchor object in sphere-projected left/right camera views.

5.2 Verification Results of Proposed Solutions

5.2.1 Accuracy of Camera Calibration

Paper I addresses **RQ 1.1** by evaluating target-based and target-less calibration methods on their accuracy of recovering MCS camera parameters. Analysis of the evaluation results (partly shown in Fig. 5.4 and further detailed in Section 5 of Paper I) indicates that the SIFT [Low99] based target-less calibration methods embedded in Structure from Motion (SfM) tools [SSS06, Wu13] are significantly more accurate than [SMP05], especially for estimation of extrinsic parameters. The assessed target-based calibration method ([Zha00] via [WMU13]) performed no better than [SSS06, Wu13] for all significant camera parameters as identified by Schwartz *et al.* in [SSO14].

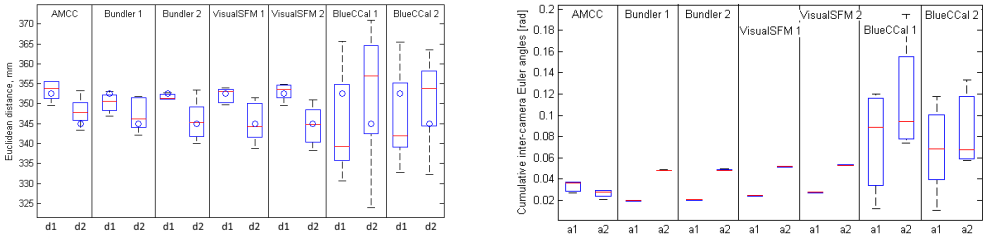

Figure 5.4: Comparison of target-based (AMCC [Zha00]) and targetless (Bundler, VisualSFM, BlueCCal [SSS06, Wu13, SMP05]) camera calibration methods, measured on a rigid 3-camera rig. Left: estimated distances between camera centers. Circle shows ground truth. Right: estimated rotation difference a_n between rigidly mounted cameras n and $n+1$. Box plots show median, 25th and 75th percentile, whiskers show minimum and maximum.

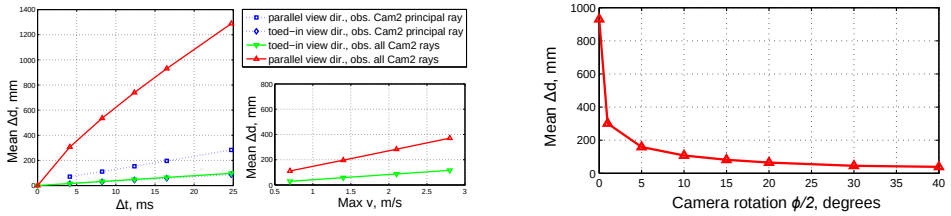

Figure 5.5: Left: Depth uncertainty Δd, given varying camera desynchronization and varying maximum speed of scene elements for parallel and $\phi = 20°$ -convergent view directions. Right: Mean Δd along all rays of camera 1, for varying convergence ϕ of both cameras (indicated rotation $\phi/2$ for camera 1, with simultaneous negative rotation $-\phi/2$ on camera 2).

5.2.2 Consequences of Synchronization Error

Paper II addresses **RQ 1.2** by applying the model of Section 5.1.1 to a range of synchronization delays and camera arrangements to quantify a loss of accuracy in depth estimation as an increase in depth uncertainty. Simulation results (in Fig. 5.5 and Section 5 of Paper II) show that the overall depth uncertainty of a system is directly proportional to synchronization error. Depth uncertainty is significantly affected by the angle of convergence between cameras; more specifically, cameras in parallel arrangement have significantly larger depth uncertainty compared to toed-in cameras.

5.2.3 Latency in the Scalable End-to-End System

Paper III addresses **RQ 1.3** via the proposed framework described in Section 5.1.2 and a latency analysis of the video processing components (see Fig. 5.6 and Section 5 in Paper III). Results indicate that a scalable implementation based on transmission via [tea12] can support operation within the real-time requirement of 40 ms, set by the 25 FPS frame rate of the cameras. Overheads for video stream formatting components are negligible, and the majority of time to process each frame depends on the latencies of the selected video encoder and decoder.

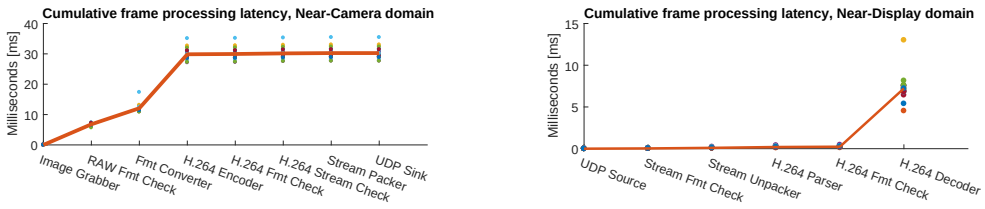

Figure 5.6: Cumulative latency for video frame processing in the scalable end-to-end system. The line shows average frame latency; dots show individual latency measurements.

5.2.4 Performance of the Augmented Remote Operation System

Paper VI addresses **RQ 1.4** via a performance assessment of the system described in Section 5.1.3. Results show that the proposed lidar filtering halves the amplitude of static point oscillation in both frame-to-frame measurement and overall per-point oscillation, and reduces the amount of intermittently missing lidar points (see Table 5.1 and Tables 3, 4 and 5 in Paper VI). The view generation process itself takes an approximate average of 50 ms per frame to create one augmented and two novel views (see Table 5.2 and Section 7.A in Paper VI). The majority of that time (37 ms) is used on sparse depth point filtering and upscaling, which scales with inbound image and lidar resolution, but does not scale with the number of synthesized output views. While this per-frame rendering time is not as low as [RSA20], it fits within the constraints set by the inbound lidar data rate (10 Hz) as well as within the feasible remote operation constraint (frame rate > 15 FPS) outlined in [YLK20]. Further results and details are found in Section 7 of Paper VI.

Table 5.1: Lidar point oscillation amplitude (meters) in the augmented remote operation system for a motionless scene

	Excluding missing points			**Including missing points**		
	avg	min	median	avg	min	median
Unfiltered	0.076	0.001	0.066	0.077	0.020	0.080
Filtered	0.039	0.0002	0.034	0.039	0.006	0.043

Table 5.2: Frame render time (ms) in the augmented remote operation system with varying apparent sizes (amount of pixels) of the disoccluded scene object

Amount of disoccluded pixels	**2.0%**	**2.9%**	**8.7%**
Avg. total time per frame (ms)	49.6	50.1	51.7

5.2.5 Effects of View Positions and Depth-Aiding Augmentations

Papers IV and V address **RQ 2.1** and **RQ 2.2** by a QoE study using the test system described in Section 5.1.4. During the test, only one participant had a strong simulator sickness response, and there were no significant correlations between test sequence order and participant responses. The explicit results, shown in Fig. 5.7 and in Papers IV and V, indicate that AR design and viewing position had noticeable effects on the experiment task. Participant QoE dropped by 1 to 2 units when using the ground viewing position, which requires stereoscopic depth perception for task completion. Likewise, implicit measurements of task performance showed a negative effect from the ground viewing position.

Depth-aiding AR reduced the difference in user performance between the viewing positions for the explicit task accomplishment and task difficulty scores, implying that AR can reduce the negative effect of a compromised viewing position. The

variation of depth-aiding AR presentation only affected the explicit task difficulty to a significant degree, but participants generally rated the helpfulness of two active-assistance AR designs as "Fair" to "Good", implying some perceived benefit towards the overall QoE. The results described in Section 4 of Papers IV and V indicate that a significant loss in QoE can be seen when users have to rely on stereoscopic depth perception in HMD-based telepresence. Depth-aiding AR can be used to mitigate this loss; however, the choice of camera placement (and therefore viewing position) is more impactful for the overall QoE in AT.

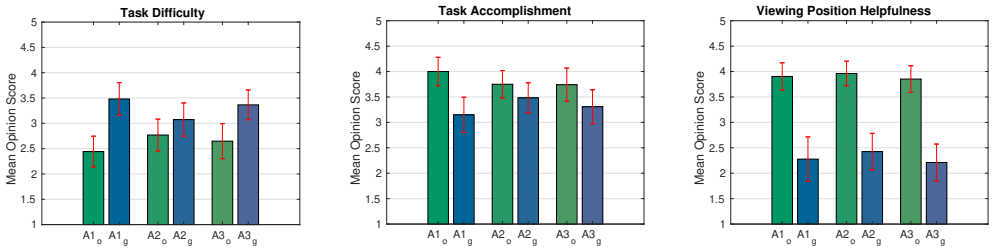

Figure 5.7: The MOS and 95% confidence intervals, for three depth-aiding AR designs (A1, A2, A3) and two viewpoint positions ([o]verhead, [g]round).

Chapter 6

Discussion

This chapter presents a retrospective on the results and outcome of addressing the research questions (RQs), an overall reflection on the methodology used to conduct the research, and a discussion of the context and further challenges related to the research outcomes.

6.1 Reflections on Results

The work presented in this thesis fulfils the research purpose from Section 1.4 by addressing six RQs. The specific results of each question are already described in Section 5.2; this section offers a discussion of the overall outcome of addressing the RQs. The RQs are restated here for reading convenience.

6.1.1 Accuracy of Camera Calibration

***RQ 1.1**: How accurate are the commonly used multi-camera calibration methods, both target-based and targetless, in recovering the true camera parameters represented by the pinhole camera model?*

Calibration was found to be a relatively mature field with widely used methods readily integrated into image processing tool collections, as described in Section 3.1. A gap was identified regarding strict comparisons of target-based and target-less calibration methods on the basis of the ground truth accuracy of camera parameters. A ground-truth based comparison was performed and described in Paper I. The results revealed a parity between the accuracy of the tested target-based and target-less methods. Most of the tested methods had a low degree of error, but one of the tested methods performed significantly less well.

These results supplement the existing literature directly by the performed comparison, and indirectly by proposing ground-truth based assessment of camera pa-

rameter estimation as an alternative approach to evaluating camera calibration. These findings were also used to inform calibration choices for the systems described in Sections 5.1.2, 5.1.3 and 5.1.4. Further development of calibration methods was not pursued, given the maturity of the field and the abundance of existing solutions targeting parametric camera models.

6.1.2 Consequences of Synchronization Error

***RQ 1.2**: What is the relationship between camera synchronization error and estimated scene depth error, and how does camera arrangement in multi-camera systems affect this depth error?*

In contrast to calibration, camera synchronization is a less explored area, with a notable gap in relating synchronization accuracy to geometric multi-camera models (see Section 3.1). A new model for mapping synchronization error to depth estimation error was proposed in Paper II to address this gap and to define the concept of depth uncertainty. The model was subsequently used to show the impact of synchronization error and of ancillary parameters such as camera convergence. The findings from this investigation were used to motivate the hardware choices for the system described in Section 5.1.4. The non-converged layout of cameras for stereoscopic pass-through viewing had the least tolerance for synchronization error and therefore justifies a hardware synchronization solution. The proposed model can be applied in the design process of an MCS, such as to set a desired depth estimation accuracy and determine the necessary level of synchronization accuracy. The description in Paper II uses the pinhole camera model, but any ray-based generic multi-camera model can be substituted for depth uncertainty estimation.

6.1.3 A Framework for Scalable End-to-End Systems

***RQ 1.3**: What is an appropriate, scalable multi-camera system design for enabling low-latency video processing and real-time streaming?*

RQ 1.3 led to a new proposed framework for scalable end-to-end systems, described in Section 5.1.2. The framework places emphasis on scalability and flexibility by means of compartmentalization of processing, and the use of modular computing platforms in the MCS implementation. This sets the framework (and implementation) apart from MCSs described in the literature [MP04, YEBM02, BK10] and places greater emphasis on component-agnostic MCS design. The flexibility of the proposed framework can be seen in the following properties. (1) Devices and processes of the proposed system are separated into framework domains and components based on their purpose and role in the end-to-end processing chain; this allows the changing of system capabilities at the hardware and software level on a component by component basis. (2) The use of per-camera, fully connected computers allows for any distribution of processing operations on the available platforms ("domains"). (3) The implemented MCS uses off-the-shelf cameras and computers, and manages transmission via an open-source media streaming framework; this increases com-

patibility between the MCS and third-party processing or rendering applications.

RQ 1.3 is, admittedly, an open-ended research question that does not permit an all-encompassing, single answer. Rather, the proposed framework and corresponding implementation serve as one specific, viable solution for real-time capture. The suitability of the proposed system design and the selected transmission platform was verified via streaming latency tests, detailed in Paper III. The MCS implemented for the latency tests was subsequently used as the basis for capture, processing (specifically image rectification and image stream compression), and transmission in the systems built to investigate **RQ 1.4**, **RQ 2.1**, and **RQ 2.2**. Those systems, as described in Sections 5.1.3 and 5.1.4, further reinforce the suitability of the proposed framework.

6.1.4 Augmented Remote Operation

***RQ 1.4**: What rendering performance can be achieved by camera-and-lidar-based AT for remote operation in an underground mining context, without data preconditioning?*

The results of **RQ 1.4** demonstrated that camera-and-lidar-based AT for remote operation is feasible within the specified context without relying on pre-conditioned data. The feasibility condition was set by the inbound data rate of the slowest sensor and the minimum frame rate that allows remote operation, as identified in [YLK20]. The proposed rendering pipeline integrated concepts from related literature for e.g. fast depth upscaling [PHHD16] and introduced new solutions to resolve issues (such as lidar point oscillation, irregularity and sparseness of projected lidar points) related to the specific application.

The view composition shown in Paper VI corresponds to "skeumorphic" rather than "immersive" view presentation (for clarification of view presentation types, see Section 3.3.1). This choice was made to better relate the results to current remote operation solutions in the mining industry, which set the context and constraints for **RQ 1.4**. The proposed solution in Paper VI describes the generation of independent views, which can be composed at will independent of any specific display technologies. The proposed solution can therefore be readily generalized to other types of view presentations for AT.

6.1.5 Quality of Experience in Augmented Telepresence

***RQ 2.1**: What impact does the camera-based viewing position have on user Quality of Experience in an AT system for remote operation?*

***RQ 2.2**: What impact do depth-aiding view augmentations have on user Quality of Experience in an AT system for remote operation?*

The results of **RQ 2.1** and **RQ 2.2** show that the choice of viewing position significantly affects QoE in immersive AT to a greater extent than the tested in-view augmentations. The importance of viewing positions for non-immersive, non-telepresence user interfaces was discussed in [TSS18, SLZ+18, LTM19]; the results in

Papers IV and V show the impact in AT. The augmentation-free control case in the tests also demonstrated that viewing position has a significant impact on QoE for HMD-based telepresence in general.

Depth-aiding augmentations were found to have a significant, but less pronounced effect on QoE. This outcome, together with the findings in [BKRB14, DWSS17], indicates that in-view augmentation does not take precedence over monoscopic and stereoscopic depth cues. Notably, the tested augmentations did not remove other depth cues, nor entirely replace the target objects. Pervasive view augmentation across the majority of the presented view may still have a dominant impact on user QoE in AT in proportion with the diminished presentation of other depth cues. The results in Papers IV and V are relevant to AT applications where depth perception plays a role, such as navigation, positioning, interaction with a 3D environment and so forth, and are likely less applicable to passive immersive experiences.

6.2 Reflections on Methodology

6.2.1 Connection between Research Questions and Purpose

This work and its contributions are aimed at a broad section of the video-based communication process, starting from aspects of capture systems and ending at the user experience of communication applications. The approach was limited from the outset to systems with multiple cameras and telepresence applications to make the work more focused and manageable. However, that still covers the entire range from capture technology to user experience of whole systems. The research purpose was stated in two parts as a way of separating the investigations of technology from the investigations of user experience.

The first part, ***P1****: To investigate how multi-camera and multi-sensor systems should be designed for the capture of consistent datasets and use in AT applications*, encompassed the goal of investigating the technical aspects of AT systems and the components thereof, from capture to rendering. The second part, ***P2****: To investigate how user experience is affected by applying multi-sensor based AT in industrial, task-based contexts*, completes the remainder of the purpose and corresponds to the goal of investigating how such systems (as covered through **P1**) can benefit an end-user. The two-fold research purpose was supported by the two sets of RQs, defined in Section 1.4, with the first set of RQs corresponding to **P1** and the second set to **P2**.

RQ 1.1 and **RQ 1.2** were formulated to isolate a specific aspect of MCS as an entry point into the broader problem of MCS design and use. Calibration and synchronization were specifically selected as entry points because both are necessary to have a functioning MCS. **RQ 1.3** was formulated to investigate the transmission component of end-to-end systems and to determine a suitable transmission approach for real time MCS applications. At the same time, **RQ 1.3** aimed to address **P1** in a wider sense, via a focus on the design of MCS for low-latency processing and streaming —both important prerequisites for enabling AT. Finally, **RQ 1.4** completed the scope

of **P1** by focusing on the entire rendering chain of an end-to-end telepresence system. These four RQs cover the technology-focused part of the research purpose by investigating the end-to-end process of MCS-based telepresence through its key components, namely —capture, transmission, and rendering.

RQ 2.1 and **RQ 2.2** complement the technology-focused investigations by focusing on user interaction with AT applications. **RQ 2.1** also supports the purpose of investigating multi-camera and multi-sensor system design by focusing on the user experience impact of an MCS design aspect (camera positioning).

6.2.2 Adequacy of Methodology

Research Question 1.1

RQ 1.1 was addressed by a comparative assessment of a select few calibration methods. The methods were selected based on both prevalence in literature, and availability of functioning reference implementations, to reduce the chance of errors caused by faulty re-implementations. The selected methods are representative of commonly used calibration solutions, but they do not comprise the full set of existing calibration methods. In retrospect, having more calibration solutions would provide better support for the generalization of the conclusions, especially regarding the target-based calibration group which was represented by a single (though widely used) method. A new test dataset was captured for the assessments, because existing calibration datasets do not normally provide constraints on the parameter ground truth. The conducted assessment was based on parameter identity constraints in order to exclude any dependence on tertiary parameter measurement, which would be a source of unknown error in the ground truth.

Research Question 1.2

RQ 1.2 was addressed by deriving a theoretical model, and using that model to demonstrate the effects of synchronization error through simulations. The simulation parameters were chosen to represent conventional stereo-camera setups. The proposed model relies on two assumptions: 1) movement of scene elements can be sufficiently approximated by constant speed in a straight line at the small timescales between successive frames; and 2) scene element depth is determined from two cameras, without adding constraints from additional cameras. These assumptions do affect the generalizability of the model as presented in Paper II. The model was not verified through experimental setup of de-synchronized cameras and predictably moving scene objects. Such experimental verification would lend support to the solution of **RQ 1.2**, but sources of error in such an experimental setup would have to be addressed. Furthermore, the derived depth uncertainty model is based on exactly those multi-view geometry equations that would have been used to calculate the scene element depth; therefore the main contribution from an experimental verification setup would be the sources of parameter and measurement error.

Research Question 1.3

RQ 1.3 was addressed by proposing a framework for and end-of-end system, and evaluating the processing latency of camera data. This verification method was used to validate the implementation choices (i.e. the solutions for transmission and processing) in the specific context of low-latency video processing and streaming. The latency measurements were obtained using the debugging tools of the transmission solution [tea12]. The scalability of the framework was not experimentally verified, since the framework was defined at a high level of abstraction and the scalability property is directly evident (as explained in Section 6.1.3). Such verification would have been necessary if a reference implementation of the proposed framework had been published, which was not deemed necessary at the time.

Research Question 1.4

RQ 1.4 was addressed by proposing and implementing an augmented remote operation system, as described in Paper VI. The rendering performance was primarily defined as the time necessary to process one frame of input data and create all output content. This time was measured across multiple repetitions of test recordings to account for the natural variance of software execution timing in a non-real-time operating system. The alternatives for end-to-end augmented remote operation in the related literature were either unavailable for re-use, or did not correspond to the constraints of the problem setting that Paper VI addressed, thereby preventing off-the-shelf whole-system comparisons. In an effort to compensate for this, detailed process descriptions, step by step measurements and comparisons for key stages of the rendering process were used in Paper VI. In general, time-based performance assessments of computational tasks such as rendering depend not only on the implementation and algorithm design choices, but also on the underlying tools and technologies. This dependence inevitably causes complications for direct comparisons between solutions, especially for complete end-to-end systems. From a methodology standpoint, such complications can serve as an argument for a more compartmentalized approach involving independent investigations and solutions to specific subsets of the overall problem.

Research Questions 2.1 and 2.2

RQ 2.1 and **RQ 2.2** were addressed through a single experiment using a custom AT solution developed for the research purpose. The investigation of two factors (viewing positions and augmentations) was combined to more effectively use a limited number of test participants, and to explore the joint interaction of the two factors. The experiment design was based on the QoE and general user-based testing methodology from the related literature, but no PPA was conducted because access to suitable equipment, lab space and willing test participants was limited. Instead, task-completion related metrics were used to supplement the participant opinion scores. The data analysis was performed in accordance with the methodology of

related works. In retrospect, the visual design of AR information probably had a notable effect on user experience; in the test system, the augmentations were designed for functionality rather than aesthetic appearance. In similar future investigations, involving iterative UX design methods for augmentation design would be beneficial.

6.3 Impact and Significance

Multi-camera systems are prevalent in modern day-to-day life, with applications in surveillance, entertainment production, vehicle autonomy, and much more. The recent advances in consumer-grade VR and AR headsets, the ubiquity of multi-camera and multi-sensor platforms, and the increasing need for remote-work solutions place telepresence and AT at the forefront of relevant topics for numerous industries. As such, there is a need for corresponding investigations in how to enable AT, and how to effectively apply AT in the aforementioned industries.

The research described in this thesis contributes to the knowledge base on both the technical feasibility and user experience of AT and scales to the broader context of MCS applications. Paper I provides an additional perspective for the calibration research community regarding the choice of evaluation metrics for assessing calibration accuracy. Paper II introduces a new model for consequences of camera synchronization that can serve as an additional method for assessing synchronization, a way of categorizing MCS solutions, and a tool for MCS design. Paper III presents a framework (and an implementation example) for an end-to-end multi-camera based system that can be applied for AT and general multi-view video communication solutions. Paper VI demonstrates the feasibility of AT in an industrial application within the constraints imposed by the application setting. Papers IV and V show the user experience effect of AT and highlight the interaction between stereoscopic perception, AR, viewing position, and immersive rendering through an HMD. In aggregate, these papers contribute to the future of better AT by introducing new models, frameworks, and assessments to the research community, and by providing the basis of new MCS design tools and AT systems for industries interested in AT for practical applications.

6.4 Risks and Ethical aspects

The work presented in this thesis is primarily a study of technological artifacts, namely MCSs and AT systems or components thereof. The outcomes of this work will, at best, contribute to better MCSs and to a larger adoption of AT for non-entertainment applications. There is a distant risk that this work could indirectly contribute to potentially problematic or harmful applications of multi-camera based sensing technology, but the presented research does not directly enable such applications nor defines any clear paths to the misuse of the research results.

Augmented Telepresence was investigated in the context of operator safety, as

part of the background and motivation of this work. To manage the risks derived from AT applications based on the work in this thesis, on-site safety and integrity testing should precede actual deployment. In the course of user QoE assessment, human participants were recruited to assist in testing an AT system. Participant involvement was voluntary, and all participants were informed about the test procedure and the use of the results, as well as given a choice to interrupt the test at any moment for any reason, without needing to provide any justification. The test duration per participant was kept short (within 30 minutes) to avoid fatigue. Participant responses and AT system usage metrics were anonymized, and informed consent was obtained from all participants. All user tests related to the research presented took place well before the outbreak of COVID-19; any subsequent or future tests would likely have to follow a strict system of precautions, as suggested in [BSDH20].

6.5 Future Work

Given the broad scope of the research purpose driving this study, and the broad range of the investigated problems regarding capture, transmission, rendering, system design, and user experience of AT, the scope for future work is vast. One path is to expand and build upon the proposed synchronization model, such as by adding parametrization of rolling sensor shutter, shutter speed, and motion blur, or by using said model in a cost function for multi-camera layout optimization. Another path is to further develop the telepresence systems described in Sections 5.1.3 and 5.1.4. Rendering methods can be improved by including virtual surface illumination, scattering, and environment lighting techniques used by the computer graphics community. Designs for depth-aiding augmentations can be explored more thoroughly through UX-design methodology, with greater focus on user needs analysis and formative evaluation as inputs to the design process. Similarly, QoE assessments of remote augmented operation for mining are still needed; as [SPG+19] indicates, there are open questions about whether augmentation design should prioritize visual appearance (thus improving user aesthetic experience) or task performance (improving user control). The proposed systems, and the scalable end-to-end framework can act as a technical base for such studies, and similarly support investigations of other applications of MCS and AT.

From a computer vision research perspective, the single most notable gap in the work presented in this thesis is the absence of neural rendering in the proposed systems. The direct path would be to apply neural rendering to increase the render fidelity (both resolution and frame rate) and to improve the camera-to-camera (and lidar-to-camera) correspondences. Use of neural rendering would also allow to decouple the capture and render resolutions in end-to-end real-time systems, for AT or otherwise. Furthermore, predictive models can be applied to selectively improve the presentation quality at the point of the user's attention; combining such models with adaptive in-view augmentations for improved AT is an open research area.

Bibliography

[AA17] Murat Akçayır and Gökçe Akçayır. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, 20:1–11, 2017.

[AB91] Edward H Adelson and James R Bergen. *The Plenoptic Function and the Elements of Early Vision*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

[ACCM15] Andrea Albarelli, Augusto Celentano, Luca Cosmo, and Renato Marchi. On the interplay between data overlay and real-world context using see-through displays. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*, pages 58–65. ACM, 2015.

[AGSH20] Kurt Andersen, Simone José Gaab, Javad Sattarvand, and Frederick C Harris. Mets vr: Mining evacuation training simulator in virtual reality for underground mines. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*, pages 325–332. Springer, 2020.

[AGT+19] Waqas Ahmad, Mubeen Ghafoor, Syed Ali Tariq, Ali Hassan, Mårten Sjöström, and Roger Olsson. Computationally efficient light field image compression using a multiview hevc framework. *IEEE access*, 7:143002–143014, 2019.

[AKB18] David Anton, Gregorij Kurillo, and Ruzena Bajcsy. User experience and interaction performance in 2d/3d telecollaboration. *Future Generation Computer Systems*, 82:77–88, 2018.

[AKF+17] Cenek Albl, Zuzana Kukelova, Andrew Fitzgibbon, Jan Heller, Matej Smid, and Tomas Pajdla. On the two-view geometry of unsynchronized cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[ANP+09] Hadi Aliakbarpour, Pedro Nunez, Jose Prado, Kamrad Khoshhal, and Jorge Dias. An efficient algorithm for extrinsic calibration between a 3d laser range finder and a stereo camera for surveillance. In *2009 International Conference on Advanced Robotics*, pages 1–6. IEEE, 2009.

[ASOC17] Hanan Alnizami, James Scovell, Jacqueline Ong, and Philip Corriveau. *Measuring Virtual Reality Experiences is more than just Video Quality*, volume 3 Issue 1, pages 9–17. Video Quality Experts Group (VQEG), www.vqeg.org, 2017.

[AWGC19] Sameer Ansari, Neal Wadhwa, Rahul Garg, and Jiawen Chen. Wireless software synchronization of multiple distributed cameras. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2019.

[AYL18] Firas Abedi, You Yang, and Qiong Liu. Group geometric calibration and rectification for circular multi-camera imaging system. *Optics express*, 26(23):30596–30613, 2018.

[Azu97] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

[BÁAGPB19] Miguel Barreda-Ángeles, Sara Aleix-Guillaume, and Alexandre Pereda-Baños. Users' psychophysiological, vocal, and self-reported responses to the apparent attitude of a virtual audience in stereoscopic 360-video. *Virtual Reality*, pages 1–14, 2019.

[BÁRTPB18] Miguel Barreda-Ángeles, Rafael Redondo-Tejedor, and Alexandre Pereda-Baños. Psychophysiological methods for quality of experience research in virtual reality systems and applications. *IEEE COMSOC MMTC Communications - Frontiers*, 4(1):14–20, 2018.

[BBDM+13] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, Patrick Le Callet, et al. *Qualinet white paper on definitions of quality of experience*. Qualinet, Lausanne, Switzerland, 2013.

[BBV+20] Bence Bejczy, Rohat Bozyil, Evaldas Vaičekauskas, Sune Baagø Krogh Petersen, Simon Bøgh, Sebastian Schleisner Hjorth, and Emil Blixt Hansen. Mixed reality interface for improving mobile manipulator teleoperation in contamination critical applications. *Procedia Manufacturing*, 51:620–626, 2020.

[BCFS06] M Bigas, Enric Cabruja, Josep Forest, and Joaquim Salvi. Review of CMOS image sensors. *Microelectronics Journal*, 37(5):433–451, 2006.

[BDA+19] K. Brunnström, E. Dima, M. Andersson, M. Sjöström, T. Qureshi, and M. Johanson. Quality of experience of hand controller latency in a virtual reality simulator. In D.M. Chandler, M. McCourt, and J.B. Mulligan, editors, *Human Vision and Electronic Imaging 2019*, pages HVEI–218. Society for Imaging Science and Technology, 2019.

[BEMN09] Rune H Bakken, Bjørn G Eilertsen, Gustavo U Matus, and Jan H Nilsen. Semi-automatic camera calibration using coplanar control points. In *Proceedings of NIK Conference*, pages 37–48, 2009.

[BETVG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[Bev08] Nigel Bevan. Classifying and selecting ux and usability measures. In *COST294-MAUSE Workshop: Meaningful Measures: Valid Useful User Experience Measurement.*, 2008.

[BK10] Tibor Balogh and Péter Tamás Kovács. Real-time 3D light field transmission. In *Real-Time Image and Video Processing*, volume 7724, page 772406. International Society for Optics and Photonics, 2010.

[BKRB14] Matthias Berning, Daniel Kleinert, Till Riedel, and Michael Beigl. A study of depth perception in hand-held augmented reality using autostereoscopic displays. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 93–98. IEEE, 2014.

[BLB+18] Fabio Bruno, Antonio Lagudi, Loris Barbieri, Domenico Rizzo, Maurizio Muzzupappa, and Luigi De Napoli. Augmented reality visualization of scene depth for aiding rov pilots in underwater manipulation. *Ocean Engineering*, 168:140–154, 2018.

[BMNK13] Kai Berger, Stephan Meister, Rahul Nair, and Daniel Kondermann. A state of the art report on Kinect sensor setups in computer vision. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 257–272. Springer, 2013.

[BNW+18] Koyel Banerjee, Dominik Notz, Johannes Windelen, Sumanth Gavarraju, and Mingkang He. Online camera lidar fusion and object detection on hybrid data for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1632–1638. IEEE, 2018.

[Bou16] Jean-Yves Bouguet. Camera calibration toolbox for matlab. *URL http://www. vision. caltech. edu/bouguetj/calib_doc*, 2016.

[BPG+17] Filippo Brizzi, Lorenzo Peppoloni, Alessandro Graziano, Erika Di Stefano, Carlo Alberto Avizzano, and Emanuele Ruffaldi. Effects of augmented reality on the performance of teleoperated industrial assembly tasks in a robotic embodiment. *IEEE Transactions on Human-Machine Systems*, 48(2):197–206, 2017.

[BPG+18] F. Brizzi, L. Peppoloni, A. Graziano, E. D. Stefano, C. A. Avizzano, and E. Ruffaldi. Effects of augmented reality on the performance of teleoperated industrial assembly tasks in a robotic embodiment. *IEEE Transactions on Human-Machine Systems*, 48(2):197–206, 2018.

[Bra00] Gary Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[Bro66] Duane C Brown. Decentering distortion of lenses. *Photogrammetric Engineering and Remote Sensing*, 1966.

[BSDH20] Kjell Brunnström, Bo Schenkman, Anders Djupsjöbacka, and Omar Hamsis. Covid-19 precautions for lab experiments involving test persons, 2020.

[BSE+17] James Baumeister, Seung Youb Ssin, Neven AM ElSayed, Jillian Dorrian, David P Webb, James A Walsh, Timothy M Simon, Andrew Irlitti, Ross T Smith, Mark Kohler, et al. Cognitive cost of using augmented reality displays. *IEEE Transactions on Visualization and Computer Graphics*, 23(11):2378–2388, 2017.

[BSEN18] Felix Bork, Christian Schnelzer, Ulrich Eck, and Nassir Navab. Towards efficient visual guidance in limited field-of-view head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2983–2992, 2018.

[BSI+18] Kjell Brunnström, Mårten Sjöström, Muhammad Imran, Magnus Pettersson, and Mathias Johanson. Quality of experience for a virtual reality simulator. *Electronic Imaging*, 2018(14):1–9, 2018.

[BTH15] Dash Bodington, Jayant Thatte, and Matthew Hu. Rendering of stereoscopic 360◦ views from spherical image pairs. Technical report, Tech. rep, 2015.

[Bur11] Nicolas Burrus. Kinect rgb demo. *Manctl Labs. Available online: http://rgbdemo. org/(accessed on 21 January 2017)*, 2011.

[CAB+18] Guglielmo Carra, Alfredo Argiolas, Alessandro Bellissima, Marta Niccolini, and Matteo Ragaglia. Robotics in the construction industry: state of the art and future opportunities. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, volume 35, pages 1–8. IAARC Publications, 2018.

[CFF18] Fabrizio Cutolo, Umberto Fontana, and Vincenzo Ferrari. Perspective preserving solution for quasi-orthoscopic video see-through hmds. *Technologies*, 6(1):9, 2018.

[CFM19] David Concannon, Ronan Flynn, and Niall Murray. A quality of experience evaluation system and research challenges for networked virtual reality-based teleoperation applications. In *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*, pages 10–12. ACM, 2019.

[CG20] Ruijin Chen and Wei Gao. Color-guided depth map super-resolution using a dual-branch multi-scale residual network with channel interaction. *Sensors*, 20(6):1560, 2020.

[CI02] Yaron Caspi and Michal Irani. Spatio-temporal alignment of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(11):1409–1424, 2002.

[ClHClJy19] Luo Chun-lei, Sha Hao, Ling Chun-lai, and Li Jin-yang. Intelligent detection for tunnel shotcrete spray using deep learning and lidar. *IEEE Access*, 2019.

[CS19] Hongyu Chen and Sören Schwertfeger. Heterogeneous multi-sensor calibration based on graph optimization. *arXiv preprint arXiv:1905.11167*, 2019.

[Cur17] Igor D.D. Curcio. *On Streaming Services for Omnidirectional Video and its Subjective Assessment*, volume 3 Issue 1, pages 26–32. Video Quality Experts Group (VQEG), www.vqeg.org, 2017.

[CVB+19] Daniele Cattaneo, Matteo Vaghi, Augusto Luis Ballardini, Simone Fontana, Domenico Giorgio Sorrenti, and Wolfram Burgard. Cmrnet: Camera to lidar-map registration. *arXiv preprint arXiv:1906.10109*, 2019.

[CXZ19] Chen Chen, Guangming Xiong, and Sen Zhu. Outdoor 3d environment reconstruction based on multi-sensor fusion for remote control. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 1753–1757. IEEE, 2019.

[DBV16] Ruofei Du, Sujal Bista, and Amitabh Varshney. Video fields: Fusing multiple surveillance videos into a dynamic virtual environment. In *Proceedings of the 21st International Conference on Web3D Technology*, pages 165–172. ACM, 2016.

[DCRK17] Ankit Dhall, Kunal Chelani, Vishnu Radhakrishnan, and K Madhava Krishna. Lidar-camera calibration using 3d-3d point correspondences. *arXiv preprint arXiv:1705.09785*, 2017.

[DDM+15] Marek Domański, Adrian Dziembowski, Dawid Mieloch, Adam Łuczak, Olgierd Stankiewicz, and Krzysztof Wegner. A practical approach to acquisition and processing of free viewpoint video. In *Picture Coding Symposium (PCS), 2015*, pages 10–14. IEEE, 2015.

[DEGH12] Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. Faster and more accurate feature-based calibration for widely spaced camera pairs. In *Second International Conference on Digital Information and Communication Technology and it's Applications (DICTAP)*, pages 87–92. IEEE, 2012.

[DKG19] Joris Domhof, Julian F. P. Kooij, and Dariu M. Gavrila. An extrinsic calibration tool for radar, camera and lidar. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8107–8113. IEEE, 2019.

[DPSL11] Ferran Diego, Daniel Ponsa, Joan Serrat, and Antonio M López. Video alignment for change detection. *IEEE Transactions on Image Processing*, 20(7):1858–1869, 2011.

[DS17] Li Ding and Gaurav Sharma. Fusing structure from motion and lidar for dense accurate depth map estimation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1283–1287. IEEE, 2017.

[DSF+13] Ismael Daribo, Hideo Saito, Ryo Furukawa, Shinsaku Hiura, and Naoki Asada. Hole filling for view synthesis. In *3D-TV System with Depth-Image-Based Rendering*, pages 169–189. Springer, 2013.

[DSRK18] Varuna De Silva, Jamie Roche, and Ahmet Kondoz. Robust fusion of lidar and wide-angle camera data for autonomous mobile robots. *Sensors*, 18(8):2730, 2018.

[DWSS17] C. Diaz, M. Walker, D. A. Szafir, and D. Szafir. Designing for depth perceptions in augmented reality. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 111–122, 2017.

[DZL06] Congxia Dai, Yunfei Zheng, and Xin Li. Subframe video synchronization via 3d phase correlation. In *IEEE International Conference on Image Processing (ICIP)*, pages 501–504. IEEE, 2006.

[EB13] Georgios D Evangelidis and Christian Bauckhage. Efficient subframe video alignment using short descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2371–2386, 2013.

[ECJ17] Ivan Eichhardt, Dmitry Chetverikov, and Zsolt Janko. Image-guided tof depth upsampling: a survey. *Machine Vision and Applications*, 28(3-4):267–282, 2017.

[Edm05] Diane R Edmondson. Likert scales: A history. In *Proceedings of the 12th conference on historical analysis and research in marketing (CHARM)*, pages 127–133, 2005.

[EDM+16] Ulrich Engelke, Daniel P Darcy, Grant H Mulliken, Sebastian Bosse, Maria G Martini, Sebastian Arndt, Jan-Niklas Antons, Kit Yan Chan, Naeem Ramzan, and Kjell Brunnström. Psychophysiology-based qoe assessment: A survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21, 2016.

[Ekl19] Vebjørn Fossheim Eklo. Slam-driven localization and registration. Master's thesis, NTNU, 2019.

[EPTP20] Gangadharan Esakki, Andreas Panayides, Sravani Teeparthi, and Marios Pattichis. A comparative performance evaluation of vp9, x265, svt-av1, vvc codecs leveraging the vmaf perceptual quality metric. In *Applications of Digital Image Processing XLIII*, volume 11510, page 1151010. International Society for Optics and Photonics, 2020.

[ERB+18] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[ESGMRA11] FC Estrada-Silva, J Garduño-Mejía, and M Rosete-Aguilar. Third-order dispersion effects generated by non-ideal achromatic doublets on sub-20 femtosecond pulses. *Journal of Modern Optics*, 58(10):825–834, 2011.

[FBA+94] Henry Fuchs, Gary Bishop, Kevin Arthur, Leonard McMillan, Ruzena Bajcsy, Sang Lee, Hany Farid, and Takeo Kanade. Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, volume 26, 1994.

[FBK10] Anatol Frick, Bogumil Bartczack, and Reinhard Koch. 3D-TV LDV content generation with a hybrid ToF-multicamera rig. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-Con)*, pages 1–4, 2010.

[FBLF08] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2008.

[Feh04] Christoph Fehn. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3d-tv. In *Electronic Imaging 2004*, pages 93–104. International Society for Optics and Photonics, 2004.

[FLPH19] Ching-Ling Fan, Wen-Chih Lo, Yu-Tung Pai, and Cheng-Hsin Hsu. A survey on 360 video streaming: Acquisition, transmission, and display. *ACM Computing Surveys (CSUR)*, 52(4):71, 2019.

[FP18] Timothy Forbes and Charalambos Poullis. Deep autoencoders with aggregated residual transformations for urban reconstruction from remote sensing data. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 23–30. IEEE, 2018.

[FRR+13] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013.

[FTK19] Robert Frohlich, Levente Tamas, and Zoltan Kato. Absolute pose estimation of central cameras using planar regions. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[Gab17] Bernat Gabor. Camera calibration with opencv. `https://docs.opencv.org/2.4/doc/tutorials/calib3d/camera_calibration/camera_calibration.html`, 2017.

[GBMG17] Carlos Guindel, Jorge Beltrán, David Martín, and Fernando García. Automatic extrinsic calibration for lidar-stereo vehicle sensor setups. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2017.

[GČH12] Vineet Gandhi, Jan Čech, and Radu Horaud. High-resolution depth maps based on ToF-stereo fusion. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4742–4749, 2012.

[GJ15] Andrzej Grabowski and Jarosław Jankowski. Virtual reality-based pilot training for underground coal miners. *Safety science*, 72:310–314, 2015.

[GJVDM+17] Trevor Gee, Jason James, Wannes Van Der Mark, Alfonso Gastelum Strozzi, Patrice Delmas, and Georgy Gimel'farb. Estimating extrinsic parameters between a stereo rig and a multi-layer lidar using plane matching and circle feature extraction. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 21–24. IEEE, 2017.

[GLL13] Xiaojin Gong, Ying Lin, and Jilin Liu. 3d lidar-camera extrinsic calibration using an arbitrary trihedron. *Sensors*, 13(2):1902–1918, 2013.

[GMCS12] Andreas Geiger, Frank Moosmann, Ömer Car, and Bernhard Schuster. Automatic camera and range sensor calibration using a single shot. In *International Conference on Robotics and Automation (ICRA)*, St. Paul, USA, May 2012.

[GML+14] Patrik Goorts, Steven Maesen, Yunjun Liu, Maarten Dumont, Philippe Bekaert, and Gauthier Lafruit. Self-calibration of large scale camera networks. In *IEEE International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, pages 107–116. IEEE, 2014.

[GN01] Michael D Grossberg and Shree K Nayar. A general imaging model and a method for finding its parameters. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 108–115. IEEE, 2001.

[GNN15] Ginni Grover, Ram Narayanswamy, and Ram Nalla. Simulating multi-camera imaging systems for depth estimation, enhanced photography and video effects. In *Imaging Systems and Applications*, pages IT3A–2. Optical Society of America, 2015.

[GPAM+14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua

Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[GWZ+19] Jie Guo, Dongdong Weng, Zhenliang Zhang, Yue Liu, Henry B.-L. Duh, and Yongtian Wang. Subjective and objective evaluation of visual fatigue caused by continuous and discontinuous use of hmds. *Journal of the Society for Information Display*, 27(2):108–119, 2019.

[Has08] Marc Hassenzahl. *User experience (UX): Towards an experiential perspective on product quality*, volume 339. Association Francophone d'Interaction Homme-Machine, 2008.

[Hei00] Janne Heikkilä. Geometric camera calibration using circular control points. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1066–1077, 2000.

[HFP15] Lionel Heng, Paul Furgale, and Marc Pollefeys. Leveraging image-based localization for infrastructure-based calibration of a multi-camera rig. *Journal of Field Robotics*, 32(5):775–802, 2015.

[HHL+17] Christian Häne, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, Paul Furgale, Torsten Sattler, and Marc Pollefeys. 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68:14–27, 2017.

[HHVM16] Tobias Hoßfeld, Poul E Heegaard, Martín Varela, and Sebastian Möller. Qoe beyond the mos: an in-depth look at qoe via better metrics and their relation to mos. *Quality and User Experience*, 1(1):2, 2016.

[HJT17] Chih-Hung Huang, Shang-Jhih Jhang, and Chi-Yi Tsai. An efficient rgb-d camera based point cloud registration algorithm. In *2017 International Conference on Applied System Innovation (ICASI)*, pages 558–561. IEEE, 2017.

[HK94] Glenn E Healey and Raghava Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994.

[HKH12] Daniel Herrera, Juho Kannala, and Janne Heikkilä. Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2058–2064, 2012.

[HLP15] Lionel Heng, Gim Hee Lee, and Marc Pollefeys. Self-calibration and visual SLAM with a multi-camera system on a micro aerial vehicle. *Autonomous Robots*, 39(3):259–277, 2015.

[HML+19] Saghi Hajisharif, Ehsan Miandji, Per Larsson, Kiet Tran, and Jonas Unger. Light field video compression and real time rendering. In *Computer Graphics Forum*, volume 38, pages 265–276. Wiley Online Library, 2019.

[HT06] Marc Hassenzahl and Noam Tractinsky. User experience – a research agenda. *Behaviour and Information Technology*, 25(2):91–97, 2006.

[HTWM04] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, 2004.

[HYHL15] Miska M Hannuksela, Ye Yan, Xuehui Huang, and Houqiang Li. Overview of the multiview high efficiency video coding (mv-hevc) standard. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2154–2158. IEEE, 2015.

[HZ03] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[ID19] Ergin Isleyen and H Sebnem Düzgün. Use of virtual reality in underground roof fall hazard assessment and risk mitigation. *International Journal of Mining Science and Technology*, 29(4):603–607, 2019.

[IOI18] Ryoichi Ishikawa, Takeshi Oishi, and Katsushi Ikeuchi. Lidar and camera calibration using motions estimated by sensor fusion odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7342–7349. IEEE, 2018.

[IRMK18] Ganesh Iyer, R Karnik Ram, J Krishna Murthy, and K Madhava Krishna. Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1110–1117. IEEE, 2018.

[IT14] ITU-T. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. Report ITU-T Rec. P.913, International Telecommunication Union, Telecommunication standardization sector, 1/2014 2014.

[IT16] ITU-T. Subjective assessment methods for 3d video quality. Report ITU-T Rec. P.915, International Telecommunication Union, 2016/03 2016.

[IT17] ITU-T. Vocabulary for performance, quality of service and quality of experience. Report ITU-T Rec. P.10/G.100, International Telecommunication Union (ITU), ITU Telecommunication Standardization Sector, 2017.

[JCK19] Jinyong Jeong, Younghun Cho, and Ayoung Kim. The road is enough! extrinsic calibration of non-overlapping stereo camera and lidar using road information. *IEEE Robotics and Automation Letters*, 4(3):2831–2838, 2019.

[JKCP15] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *Current Journal of Applied Science and Technology*, pages 396–403, 2015.

[JLZ+19] Jianhao Jiao, Qinghai Liao, Yilong Zhu, Tianyu Liu, Yang Yu, Rui Fan, Lujia Wang, and Ming Liu. A novel dual-lidar calibration algorithm using planar surfaces. *arXiv preprint arXiv:1904.12116*, 2019.

[JXC+18] Jingjing Jiang, Peixin Xue, Shitao Chen, Ziyi Liu, Xuetao Zhang, and Nanning Zheng. Line feature based extrinsic calibration of lidar and camera. In *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pages 1–6. IEEE, 2018.

[JYL+19] Jianhao Jiao, Yang Yu, Qinghai Liao, Haoyang Ye, and Ming Liu. Automatic calibration of multiple 3d lidars in urban environments. *arXiv preprint arXiv:1905.04912*, 2019.

[Kaj86] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986.

[KCC16] Pileun Kim, Yong Kwon Cho, and Jingdao Chen. Target-free automatic registration of point clouds. In *33rd International on Symposium on Automation and Robotics in Construction (ISARC 2016)*, pages 686–693, 2016.

[KCT+19] Ehsan Khoramshahi, Mariana Batista Campos, Antonio Maria Garcia Tommaselli, Niko Vilijanen, Teemu Mielonen, Harri Kaartinen, Antero Kukko, and Eija Honkavaara. Accurate calibration scheme for a multi-camera mobile mapping system. *Remote Sensing*, 11(23):2778, 2019.

[KF16] Sven Kratz and Fred Rabelo Ferriera. Immersed remotely: Evaluating the use of head mounted devices for remote collaboration in robotic telepresence. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 638–645. IEEE, 2016.

[KFM+17] Conor Keighrey, Ronan Flynn, Siobhan Murray, Sean Brennan, and Niall Murray. Comparing user qoe via physiological and interaction measurements of immersive ar and vr speech and language therapy applications. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 485–492. ACM, 2017.

[KH18] Vanessa Kohn and David Hardborth. Augmented reality - a game changing technology for manufacturing processes? *Twenty-Sixth European Conference on Information Systems (ECIS2018)*, 2018.

[KHB07] Juho Kannala, Janne Heikkilä, and Sami S Brandt. Geometric camera calibration. *Wiley Encyclopedia of Computer Science and Engineering*, pages 1–11, 2007.

[KHL+16] Eleni Kroupi, Philippe Hanhart, Jong-Seok Lee, Martin Rerabek, and Touradj Ebrahimi. Modeling immersive media experiences by sensing impact on subjects. *Multimedia Tools and Applications*, 75(20):12409–12429, 2016.

[KKL18] Julius Kümmerle, Tilman Kühner, and Martin Lauer. Automatic calibration of multiple cameras and depth sensors with a spherical target. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[KPKC19] G Ajay Kumar, Ashok Kumar Patil, Tae Wook Kang, and Young Ho Chai. Sensor fusion based pipeline inspection for the augmented reality system. *Symmetry*, 11(10):1325, 2019.

[KRN97] Takeo Kanade, Peter Rander, and PJ Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997.

[KSC15] Christian Kerl, Jorg Stuckler, and Daniel Cremers. Dense continuous-time tracking and mapping with rolling shutter rgb-d cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2264–2272, 2015.

[KSS05] Peter Kauff, Oliver Schreer, and Thomas Sikora. *3D Videocommunication: Algorithms, Concepts, and Real-time Systems in Human Centred Communication*. Wiley, 2005.

[LAV+19] Thorsten Laude, Yeremia Gunawan Adhisantoso, Jan Voges, Marco Munderloh, and Jörn Ostermann. A comprehensive video codec comparison. *APSIPA Transactions on Signal and Information Processing*, 8, 2019.

[LFP13] Gim Hee Lee, Friedrich Faundorfer, and Marc Pollefeys. Motion estimation for self-driving cars with a generalized camera. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2746–2753. IEEE, 2013.

[LFS19] Chunxu Li, Ashraf Fahmy, and Johann Sienz. An augmented reality based human-robot interaction interface using kalman filter sensor fusion. *Sensors*, 19(20):4586, 2019.

[LH96] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 31–42. ACM, 1996.

[LHKP13] Bo Li, Lionel Heng, Kevin Koser, and Marc Pollefeys. A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1301–1307. IEEE, 2013.

[LHVS14] Richard Latimer, Jason Holloway, Ashok Veeraraghavan, and Ashutosh Sabharwal. Socialsync: Sub-frame synchronization in a smartphone camera network. In *European Conference on Computer Vision*, pages 561–575. Springer, 2014.

[LJBB20] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[LKK+16] Jungjin Lee, Bumki Kim, Kyehyun Kim, Younghui Kim, and Junyong Noh. Rich360: optimized spherical representation from structured panoramic camera arrays. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.

[LLZC14] Xinzhao Li, Yuehu Liu, Shaozhuo Zhai, and Zhichao Cui. A structural constraint based dual camera model. In *Chinese Conference on Pattern Recognition*, pages 293–304. Springer, 2014.

[LM13] Cheng Lu and Mrinal Mandal. A robust technique for motion-based video sequences temporal alignment. *IEEE Transactions on Multimedia*, 15(1):70–82, 2013.

[LMJH+11] Jorge Lopez-Moreno, Jorge Jimenez, Sunil Hadap, Ken Anjyo, Erik Reinhard, and Diego Gutierrez. Non-photorealistic, depth-based image editing. *Computers & Graphics*, 35(1):99–111, 2011.

[Low99] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, 1999.

[LP18] Donghyeon Lee and Young Soo Park. Implementation of augmented teleoperation system based on robot operating system (ros). In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5497–5502. IEEE, 2018.

[LPOS20] Mikael Le Pendu, Cagri Ozcinar, and Aljosa Smolic. Hierarchical fourier disparity layer transmission for light field streaming. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2606–2610. IEEE, 2020.

[LS12] Yuankun Liu and Xianyu Su. Camera calibration with planar crossed fringe patterns. *Optik-International Journal for Light and Electron Optics*, 123(2):171–175, 2012.

[LSFW14] Junbin Liu, Sridha Sridharan, Clinton Fookes, and Tim Wark. Optimal camera planning under versatile user constraints in multi-camera image processing systems. *IEEE Transactions on Image Processing*, 23(1):171–184, 2014.

[LTM19] Mårten Lager, Elin A Topp, and Jacek Malec. Remote supervision of an unmanned surface vessel - a comparison of interfaces. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 546–547, 2019.

[LTT15] Chuen-Chien Lee, Ali Tabatabai, and Kenji Tashiro. Free viewpoint video (fvv) survey and future research direction. *APSIPA Transactions on Signal and Information Processing*, 4, 2015.

[LW15] Chiuhsiang Joe Lin and Bereket Haile Woldegiorgis. Interaction and visual performance in stereoscopic displays: A review. *Journal of the Society for Information Display*, 23(7):319–332, 2015.

[LXDW18] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 932–940, 2018.

[LY06] Cheng Lei and Yee-Hong Yang. Tri-focal tensor-based multiple video synchronization with subframe optimization. *IEEE Transactions on Image Processing*, 15(9):2473–2480, 2006.

[LYC+18a] Xiao Li, Wen Yi, Hung-Lin Chi, Xiangyu Wang, and Albert P.C. Chan. A critical review of virtual and augmented reality (vr/ar) applications in construction safety. *Automation in Construction*, 86:150 – 162, 2018.

[LYC+18b] Xiao Li, Wen Yi, Hung-Lin Chi, Xiangyu Wang, and Albert PC Chan. A critical review of virtual and augmented reality (vr/ar) applications in construction safety. *Automation in Construction*, 86:150–162, 2018.

[LZS18] Shuai Li, Ce Zhu, and Ming-Ting Sun. Hole filling with multiple reference views in dibr view synthesis. *IEEE Transactions on Multimedia*, 20(8):1948–1959, 2018.

[LZT06] Georgios Litos, Xenophon Zabulis, and Georgios Triantafyllidis. Synchronous image acquisition based on network synchronization. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 167–167. IEEE, 2006.

[Man02] Steve Mann. Mediated reality with implementations for everyday life. *Presence Connect*, 1, 2002.

[Mar98] Stephen Robert Marschner. *Inverse rendering for computer graphics*. Citeseer, 1998.

[Mat17] MathWorks. Matlab | camera calibration. `https://se.mathworks.com/help/vision/camera-calibration.html`, 2017.

[MBG+13] Debargha Mukherjee, Jim Bankoski, Adrian Grange, Jingning Han, John Koleszar, Paul Wilkins, Yaowu Xu, and Ronald Bultje. The latest open-source video codec vp9-an overview and preliminary results. In *2013 Picture Coding Symposium (PCS)*, pages 390–393. IEEE, 2013.

[MBM16] Matteo Munaro, Filippo Basso, and Emanuele Menegatti. Openptrack: Open source multi-camera calibration and people tracking for rgb-d camera networks. *Robotics and Autonomous Systems (RAS)*, 75:525–538, 2016.

[MFY+18] Steve Mann, Tom Furness, Yu Yuan, Jay Iorio, and Zixin Wang. All reality: Virtual, augmented, mixed (x), mediated (x, y), and multimediated reality. *arXiv preprint arXiv:1804.08386*, 2018.

[MMB20] Ali Ahmad Malik, Tariq Masood, and Arne Bilberg. Virtual reality in manufacturing: immersive and collaborative artificial-reality in design of human-robot workspace. *International Journal of Computer Integrated Manufacturing*, 33(1):22–37, 2020.

[MMM+20] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12826–12835, 2020.

[MNS19] Siim Meerits, Vincent Nozick, and Hideo Saito. Real-time scene reconstruction and triangle mesh generation using multiple rgb-d cameras. *Journal of Real-Time Image Processing*, 16(6):2247–2259, 2019.

[MP04] Wojciech Matusik and Hanspeter Pfister. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 814–824. ACM, 2004.

[MR14] Sebastian Möller and Alexander Raake. *Quality of Experience - Advanced Concepts, Applications and Methods*. T-Labs Series in Telecommunication Services. Springer International Publishing, Switzerland, 2014.

[MRP98] Gavin Miller, Steven Rubin, and Dulce Ponceleon. Lazy decompression of surface light fields for precomputed global illumination. In *Eurographics Workshop on Rendering Techniques*, pages 281–292. Springer, 1998.

[MSV18] Giovanni Mastrorocco, Riccardo Salvini, and Claudio Vanneschi. Fracture mapping in challenging environment: a 3d virtual reality approach combining terrestrial lidar and high definition images. *Bulletin of Engineering Geology and the Environment*, 77(2):691–707, 2018.

[MTUK95] Paul Milgram, Haruo Takemura, Akira Utsumi, and Fumio Kishino. Augmented reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and telepresence technologies*, volume 2351, pages 282–292. International Society for Optics and Photonics, 1995.

[Mud15] Suryanarayana M Muddala. Free view rendering for 3d video. *Edge-Aided Rendering and Depth-Based", Doctoral Thesis*, (226), 2015.

[Mö18] Stefan Möllenhoff. Beautiful blur for smartphone portraits: How bokeh effect works. `https://www.androidpit.com/how-bokeh-effect-works-with-smartphones`, 2018.

[N+17] Sergiu Nedevschi et al. Online cross-calibration of camera and lidar. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 295–301. IEEE, 2017.

[NDJRD09] Pedro Núñez, Paulo Drews Jr, Rui P Rocha, and Jorge Dias. Data fusion calibration for a 3d laser range finder and a camera using inertial data. In *ECMR*, pages 31–36, 2009.

[NHBH20] Mojtaba Noghabaei, Arsalan Heydarian, Vahid Balali, and Kevin Han. Trend analysis on adoption of virtual and augmented reality in the architecture, engineering, and construction industry. *Data*, 5(1):26, 2020.

[NK07] Mami Noguchi and Takekazu Kato. Geometric and timing calibration for unsynchronized cameras using trajectories of a moving marker. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 20–20. IEEE, 2007.

[NKB19a] Balázs Nagy, Levente Kovács, and Csaba Benedek. Online targetless end-to-end camera-lidar self-calibration. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[NKB19b] Balázs Nagy, Levente Kovács, and Csaba Benedek. Sfm and semantic information based online targetless camera-lidar self-calibration. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1317–1321. IEEE, 2019.

[NLB+05] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005.

[NLC+17] Min Ni, Jianjun Lei, Runmin Cong, Kaifu Zheng, Bo Peng, and Xiaoting Fan. Color-guided depth map super resolution using convolutional neural network. *IEEE Access*, 5:26666–26672, 2017.

[NRL+13] Rahul Nair, Kai Ruhl, Frank Lenzen, Stephan Meister, Henrik Schäfer, Christoph S Garbe, Martin Eisemann, Marcus Magnor, and Daniel Kondermann. A survey on time-of-flight stereo fusion. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 105–127. Springer, 2013.

[NS09] Michael Nischt and Rahul Swaminathan. Self-calibration of asynchronized camera networks. In *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2164–2171. IEEE, 2009.

[ODA+20] Bukeikhan Omarali, Brice Denoun, Kaspar Althoefer, Lorenzo Jamone, Maurizio Valle, and Ildar Farkhatdinov. Virtual reality based telerobotics framework with depth cameras. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1217–1222. IEEE, 2020.

[OKY10] Fumio Okura, Masayuki Kanbara, and Naokazu Yokoya. Augmented telepresence using autopilot airship and omni-directional camera. In *IEEE International Symposium on Mixed and Augmented Reality 2010*, volume IEEE International Symposium on Mixed and Augmented Reality 2010 Science and Technolgy Proceedings, pages 259–260. IEEE Xplore, 2010.

[OKY15] Fumio Okura, Masayuki Kanbara, and Naokazu Yokoya. Mixed-reality world exploration using image-based rendering. *Journal on Computing and Cultural Heritage (JOCCH)*, 8(2):1–26, 2015.

[OLS+15] Shun-Hsing Ou, Chia-Han Lee, V Srinivasa Somayazulu, Yen-Kuang Chen, and Shao-Yi Chien. On-line multi-view video summarization for wireless video sensor network. *IEEE Journal of Selected Topics in Signal Processing*, 9(1):165–179, 2015.

[OMS17] Kei Oishi, Shohei Mori, and Hideo Saito. An instant see-through vision system using a wide field-of-view camera and a 3d-lidar. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 344–347. IEEE, 2017.

[PBRA15] Lorenzo Peppoloni, Filippo Brizzi, Emanuele Ruffaldi, and Carlo Alberto Avizzano. Augmented reality-aided tele-presence

system for robot manipulation in industrial manufacturing. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, VRST '15, pages 237–240, New York, NY, USA, 2015. ACM.

[PCSK10] Flavio Padua, Rodrigo Carceroni, Geraldo Santos, and Kiriakos Kutulakos. Linear sequence-to-sequence alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):304–320, 2010.

[PH17] Zoltan Pusztai and Levente Hajder. Accurate calibration of lidar-camera systems using ordinary boxes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 394–402, 2017.

[PHHD16] Hannes Plank, Gerald Holweg, Thomas Herndl, and Norbert Druml. High performance time-of-flight and color sensor fusion with image-guided depth super resolution. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1213–1218. IEEE, 2016.

[PKS19] Kihong Park, Seungryong Kim, and Kwanghoon Sohn. High-precision depth estimation using uncalibrated lidar and stereo fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[Ple03] Robert Pless. Using many cameras as one. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–587. IEEE, 2003.

[PM10] Dmitry Pundik and Yael Moses. Video synchronization using temporal signals from epipolar lines. In *European Conference on Computer Vision*, pages 15–28. Springer, 2010.

[PM19] Anitha S Pillai and Prabha Susy Mathew. Impact of virtual reality in healthcare: a review. *Virtual and Augmented Reality in Mental Health Treatment*, pages 17–31, 2019.

[PMP19] Juraj Peršić, Ivan Marković, and Ivan Petrović. Extrinsic 6dof calibration of a radar–lidar–camera system enhanced by radar cross section estimates evaluation. *Robotics and Autonomous Systems*, 114:217–230, 2019.

[PMRHC17] Alba Pujol-Miro, Javier Ruiz-Hidalgo, and Josep R Casas. Registration of images to unorganized 3d point clouds using contour cues. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 81–85. IEEE, 2017.

[PPLE12] J. Puig, A. Perkis, F. Lindseth, and T. Ebrahimi. Towards an efficient methodology for evaluation of quality of experience in augmented reality. In *Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, volume Proc Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012), pages 188–193. IEEE Xplore, 2012.

[PPPF17] Shiva Pedram, Pascal Perez, Stephen Palmisano, and Matthew Farrelly. Evaluating 360-virtual reality for mining industry's safety training. In *International Conference on Human-Computer Interaction*, pages 555–561. Springer, 2017.

[PSG17] Igor Pereira, Luiz F Silveira, and Luiz Gonçalves. Video synchronization with bit-rate signals and correntropy function. *Sensors*, 17(9):2021, 2017.

[PTCR+18] Grant Pointon, Chelsey Thompson, Sarah Creem-Regehr, Jeanine Stefanucci, and Bobby Bodenheimer. Affordances as a measure of perceptual fidelity in augmented reality. *2018 IEEE VR 2018 Workshop on Perceptual and Cognitive Issues in AR (PERCAR)*, pages 1–6, 2018.

[PVG+19] Will Pryor, Balazs P Vagvolgyi, William J Gallagher, Anton Deguet, Simon Leonard, Louis L Whitcomb, and Peter Kazanzides. Experimental evaluation of teleoperation interfaces for cutting of satellite insulation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4775–4781. IEEE, 2019.

[RBW+14] Hadi Rizek, Kjell Brunnström, Kun Wang, Börje Andrén, and Mathias Johanson. Subjective evaluation of a 3d videoconferencing system. *Proceedings Volume 9011, Stereoscopic Displays and Applications XXV*, 2014.

[RHF+18] Menandro Roxas, Tomoki Hori, Taiki Fukiage, Yasuhide Okamoto, and Takeshi Oishi. 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*, pages 1–8, 2018.

[RK12] Taufiqur Rahman and Nicholas Krouglicof. An efficient camera calibration technique offering robustness and accuracy over a wide range of lens distortion. *IEEE Transactions on Image Processing*, 21(2):626–637, 2012.

[RK18] Pavel Rojtberg and Arjan Kuijper. Efficient pose selection for interactive camera calibration. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 31–36. IEEE, 2018.

[RKLM12] Kai Ruhl, Felix Klose, Christian Lipski, and Marcus Magnor. Integrating approximate depth data into dense image correspondence estimation. In *Proceedings of the 9th European Conference on Visual Media Production*, pages 26–31. ACM, 2012.

[RLE+18] Radhika Ravi, Yun-Jou Lin, Magdy Elbahnasawy, Tamer Shamseldin, and Ayman Habib. Simultaneous system calibration of a multi-lidar multicamera mobile mapping platform. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 11(5):1694–1714, 2018.

[RPAC17] Taehyun Rhee, Lohit Petikam, Benjamin Allen, and Andrew Chalmers. Mr360: Mixed reality rendering for 360 panoramic videos. *IEEE transactions on visualization and computer graphics*, 23(4):1379–1388, 2017.

[RRKB11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[RS16] Srikumar Ramalingam and Peter Sturm. A unifying model for camera calibration. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1309–1319, 2016.

[RSA20] Sverker Rasmuson, Erik Sintorn, and Ulf Assarsson. A low-cost, practical acquisition and rendering pipeline for real-time free-viewpoint video communication. *The Visual Computer*, pages 1–13, 2020.

[RV14] Dikpal Reddy and Ashok Veeraraghavan. Lens flare and lens glare. In *Computer Vision*, pages 445–447. Springer, 2014.

[SAB+07] Elena Stoykova, A Ayd, Philip Benzie, Nikos Grammalidis, Sotiris Malassiotis, Joern Ostermann, Sergej Piekh, Ventseslav Sainov, Christian Theobalt, Thangavel Thevar, et al. 3-D time-varying scene capture technologies — a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1568–1586, 2007.

[SBS14] Vivienne Sze, Madhukar Budagavi, and Gary J Sullivan. High efficiency video coding (hevc). In *Integrated circuit and systems, algorithms and architectures*, volume 39, page 40. Springer, 2014.

[SBS15] G Sundari, T Bernatin, and Pratik Somani. H. 264 encoder using gstreamer. In *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, pages 1–4. IEEE, 2015.

[SBW07] Prarthana Shrstha, Mauro Barbieri, and Hans Weda. Synchronization of multi-camera video recordings based on audio. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 545–548. ACM, 2007.

[SFHT16] Chris Sweeney, Victor Fragoso, Tobias Hollerer, and Matthew Turk. Large scale SfM with the distributed camera model. *arXiv preprint arXiv:1607.03949*, 2016.

[SJL+18] Rihui Song, Zhihua Jiang, Yanghao Li, Yunxiao Shan, and Kai Huang. Calibration of event-based camera and 3d lidar. In *2018 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, pages 289–295. IEEE, 2018.

[SJTC19] Babak Shahian Jahromi, Theja Tulabandhula, and Sabri Cetin. Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles. *Sensors*, 19(20):4357, 2019.

[SKC+19] Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia. Motion parallax for 360 rgbd video. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1817–1827, 2019.

[SKKS14] Tamara Seybold, Marion Knopp, Christian Keimel, and Walter Stechele. Beyond standard noise models: Evaluating denoising algorithms with respect to realistic camera noise. *International Journal of Semantic Computing*, 8(02):145–167, 2014.

[SLK15] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight Kinect. *Computer Vision and Image Understanding (CVIU)*, 139:1–20, 2015.

[SLPS20] Thomas Schops, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. Why having 10,000 parameters in your camera model is better than twelve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2544, 2020.

[SLZ+18] Hongling Sun, Yue Liu, Zhenliang Zhang, Xiaoxu Liu, and Yongtian Wang. Employing different viewpoints for remote guidance in a collaborative augmented environment. In *Proceedings of the Sixth International Symposium of Chinese CHI*, ChineseCHI '18, pages 64–70, New York, NY, USA, 2018. ACM.

[SML06] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit (4th ed.)*. Kitware, 2006.

[SMP05] Tomáš Svoboda, Daniel Martinec, and Tomáš Pajdla. A convenient multicamera self-calibration for virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 14(4):407–422, 2005.

[SMS06] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5695–5701. IEEE, 2006.

[SMT18] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. Visual slam and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[SPG+19] Alistair G Sutcliffe, Charalambos Poullis, Andreas Gregoriades, Irene Katsouri, Aimilia Tzanavari, and Kyriakos Herakleous. Reflecting on the design process for virtual reality applications. *International Journal of Human–Computer Interaction*, 35(2):168–179, 2019.

[SPSF17] Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. Regnet: Multimodal sensor registration using deep neural networks. In *2017 IEEE intelligent vehicles symposium (IV)*, pages 1803–1810. IEEE, 2017.

[SR11] Peter Sturm and Srikumar Ramalingam. *Camera models and fundamental concepts used in geometric computer vision*. Now Publishers Inc, 2011.

[SRS+18] Raimund Schatz, Georg Regal, Stephanie Schwarz, Stefan Suettc, and Marina Kempf. Assessing the qoe impact of 3d rendering style in the context of vr-based training. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2018.

[SSE+13] Florian Schweiger, Georg Schroth, Michael Eichhorn, Anas Al-Nuaimi, Burak Cizmeci, Michael Fahrmair, and Eckehard Steinbach. Fully automatic and frame-accurate video synchronization using bitrate sequences. *IEEE Transactions on Multimedia (TMM)*, 15(1):1–14, 2013.

[SSK+19] Christoph Schöller, Maximilian Schnettler, Annkathrin Krämmer, Gereon Hinz, Maida Bakovic, Müge Güzet, and Alois Knoll. Targetless rotational auto-calibration of radar and camera for intelligent transportation systems. *arXiv preprint arXiv:1904.08743*, 2019.

[SSL13] Hooman Shidanshidi, Farzad Safaei, and Wanqing Li. A method for calculating the minimum number of cameras in a light field based free viewpoint video system. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.

[SSO13] Sebastian Schwarz, Mårten Sjöström, and Roger Olsson. A weighted optimization approach to time-of-flight sensor fusion. *IEEE Transactions on Image Processing*, 23(1):214–225, 2013.

[SSO14] Sebastian Schwarz, Mårten Sjöström, and Roger Olsson. Multivariate sensitivity analysis of time-of-flight sensor fusion. *3D Research*, 5(3):18, 2014.

[SSR18] Junwei Sun, Wolfgang Stuerzlinger, and Bernhard E. Riecke. Comparing input methods and cursors for 3d positioning with head-mounted displays. In *Proceedings of the 15th ACM Symposium on Applied Perception*, SAP '18, pages 8:1–8:8, New York, NY, USA, 2018. ACM.

[SSS06] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 835–846. ACM, 2006.

[SVHVG+08] Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration

and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. Ieee, 2008.

[SVLK19] Weizhao Shao, Srinivasan Vijayarangan, Cong Li, and George Kantor. Stereo visual inertial lidar simultaneous localization and mapping. *arXiv preprint arXiv:1902.10741*, 2019.

[TAHL07] Eino-Ville Talvala, Andrew Adams, Mark Horowitz, and Marc Levoy. Veiling glare in high dynamic range imaging. In *ACM Transactions on Graphics (TOG)*, volume 26, page 37. ACM, 2007.

[tea12] The Gstreamer team. Gstreamer: Open Source Multimedia Framework. `https://gstreamer.freedesktop.org/`, 2012.

[TFT+20] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020.

[TH17] Chi-Yi Tsai and Chih-Hung Huang. Indoor scene point cloud registration algorithm based on rgb-d camera calibration. *Sensors*, 17(8):1874, 2017.

[TKKVE20] Alexander Toet, Irene A Kuling, Bouke N Krom, and Jan BF Van Erp. Toward enhanced teleoperation through embodiment. *Front. Robot. AI 7: 14. doi: 10.3389/frobt*, 2020.

[TNP+17] Huyen TT Tran, Nam Pham Ngoc, Cuong T Pham, Yong Ju Jung, and Truong Cong Thang. A subjective study on qoe of 360 video for vr communication. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017.

[TRG+17] Paolo Tripicchio, Emanuele Ruffaldi, Paolo Gasparello, Shingo Eguchi, Junya Kusuno, Keita Kitano, Masaki Yamada, Alfredo Argiolas, Marta Niccolini, Matteo Ragaglia, and Carlo Alberto Avizzano. A stereo-panoramic telepresence system for construction machines. *Procedia Manufacturing*, 11:1552–1559, 2017. 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy.

[TSS18] Gabriel Mamoru Nakamura Taira, Antonio Carlos Sementille, and Silvio Ricardo Rodrigues Sanches. Influence of the camera viewpoint on augmented reality interaction. *IEEE Latin America Transactions*, 16(1):260–264, 2018.

[TUI17] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: a survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9(1):16, 2017.

[TVG04] Tinne Tuytelaars and Luc Van Gool. Synchronizing video sequences. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE, 2004.

[VBWN19] Surabhi Verma, Julie Stephany Berrio, Stewart Worrall, and Eduardo Nebot. Automatic extrinsic calibration between a camera and a 3d lidar using 3d point and plane correspondences. *arXiv preprint arXiv:1904.12433*, 2019.

[VMFGAL+17] Víctor Villena-Martínez, Andrés Fuster-Guilló, Jorge Azorín-López, Marcelo Saval-Calvo, Jeronimo Mora-Pascual, Jose Garcia-Rodriguez, and Alberto Garcia-Garcia. A quantitative comparison of calibration methods for rgb-d sensors using different technologies. *Sensors*, 17(2):243, 2017.

[VPR+18] Balazs P Vagvolgyi, Will Pryor, Ryan Reedy, Wenlong Niu, Anton Deguet, Louis L Whitcomb, Simon Leonard, and Peter Kazanzides. Scene modeling and augmented virtuality interface for telerobotic satellite servicing. *IEEE Robotics and Automation Letters*, 3(4):4241–4248, 2018.

[VŠMH14] Martin Velas, Michal Španěl, Zdeněk Materna, and Adam Herout. Calibration of rgb camera with velodyne lidar. In *2014 International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, pages 135–144. Václav Skala-UNION Agency, 2014.

[VŠS+19] Martin Velas, Michal Španěl, Tomas Sleziak, Jiri Habrovec, and Adam Herout. Indoor and outdoor backpack mapping with calibrated pair of velodyne lidars. *Sensors*, 19(18):3944, 2019.

[WAA+00] Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 287–296, 2000.

[WDG+18] Chamara Saroj Weerasekera, Thanuja Dharmasiri, Ravi Garg, Tom Drummond, and Ian Reid. Just-in-time reconstruction: Inpainting sparse maps using single view depth predictors as priors. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

[WHS19] Michael E Walker, Hooman Hedayati, and Daniel Szafir. Robot teleoperation with augmented reality virtual surrogates. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 202–210, 2019.

[WMHB19] Celyn Walters, Oscar Mendez, Simon Hadfield, and Richard Bowden. A robust extrinsic calibration framework for vehicles with unscaled sensors. *Towards a Robotic Society*, 2019.

[WMJ+17] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017.

[WMU13] Michael Warren, David McKinnon, and Ben Upcroft. Online calibration of stereo rigs for long-term autonomy. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3692–3698. IEEE, 2013.

[WP19] Meng-Lin Wu and Voicu Popescu. Rgbd temporal resampling for real-time occlusion removal. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–9, 2019.

[WSLH01] Bennett S Wilburn, Michal Smulski, Hsiao-Heng Kelin Lee, and Mark A Horowitz. Light field video camera. In *Media Processors 2002*, volume 4674, pages 29–37. International Society for Optics and Photonics, 2001.

[Wu13] Changchang Wu. Towards linear-time incremental structure from motion. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-Con)*, pages 127–134. IEEE, 2013.

[WWDG13] Huogen Wang, Jiachen Wang, Zhiyong Ding, and Fei Guo. Self-converging camera arrays: Models and realization. In *2013 Ninth International Conference on Natural Computation (ICNC)*, pages 338–342. IEEE, 2013.

[WX+18] Jianhua Wu, Zhenhua Xiong, et al. A soft time synchronization framework for multi-sensors in autonomous localization and navigation. In *2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 694–699. IEEE, 2018.

[XJZ+19] Bohuan Xue, Jianhao Jiao, Yilong Zhu, Linwei Zheng, Dong Han, Ming Liu, and Rui Fan. Automatic calibration of dual-lidars using two poles stickered with retro-reflective tape. *arXiv preprint arXiv:1911.00635*, 2019.

[XOX18] Yinglei Xu, Yongsheng Ou, and Tiantian Xu. Slam of robot based on the fusion of vision and lidar. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 121–126. IEEE, 2018.

[YASZ17] Georges Younes, Daniel Asmar, Elie Shammas, and John Zelek. Keyframe-based monocular slam: design, survey, and future directions. *Robotics and Autonomous Systems*, 98:67–88, 2017.

[YCWY17] Fang Yin, Wusheng Chou, Dongyang Wang, and Guang Yang. A novel fusion method of 3d point cloud and 2d images for 3d environment reconstruction. In *Ninth International Conference on Digital Image Processing (ICDIP 2017)*, volume 10420, page 1042020. International Society for Optics and Photonics, 2017.

[YEBM02] Jason C Yang, Matthew Everett, Chris Buehler, and Leonard McMillan. A real-time distributed light field camera. *Rendering Techniques*, 2002:77–86, 2002.

[YLK20] Yeohun Yun, Seung Joon Lee, and Suk-Ju Kang. Motion recognition-based robot arm control system using head mounted display. *IEEE Access*, 8:15017–15026, 2020.

[ZC04] Cha Zhang and Tsuhan Chen. A survey on image-based rendering—representation, sampling and compression. *Signal Processing: Image Communication*, 19(1):1–28, 2004.

[ZDG+20] Huaizheng Zhang, Linsen Dong, Guanyu Gao, Han Hu, Yonggang Wen, and Kyle Guan. Deepqoe: A multimodal learning framework for video quality of experience (qoe) prediction. *IEEE Transactions on Multimedia*, 22(12):3210–3223, 2020.

[ZEM+15] Matthias Ziegler, Andreas Engelhardt, Stefan Müller, Joachim Keinert, Frederik Zilly, Siegfried Foessel, and Katja Schmid. Multi-camera system for depth based visual effects and compositing. In *Proceedings of the 12th European Conference on Visual Media Production*, page 3. ACM, 2015.

[Zha00] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[Zha17] Hui Zhang. Head-mounted display-based intuitive virtual reality training system for the mining industry. *International Journal of Mining Science and Technology*, 27(4):717–722, 2017.

[ZHLS19] Weikun Zhen, Yaoyu Hu, Jingfeng Liu, and Sebastian Scherer. A joint optimization approach of lidar-camera fusion for accurate dense 3-d reconstructions. *IEEE Robotics and Automation Letters*, 4(4):3585–3592, 2019.

[ZLJ+19] Shuya Zhou, Hanxi Li, Haiqiang Jin, Jianyi Wan, and Jihua Ye. Accurate camera syncronization using deep-shallow mixed models. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 123–128. IEEE, 2019.

[ZLK18] Lipu Zhou, Zimo Li, and Michael Kaess. Automatic extrinsic calibration of a camera and a 3d lidar using line and plane correspondences. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5562–5569. IEEE, 2018.

[ZMDM+16] Pietro Zanuttigh, Giulio Marin, Carlo Dal Mutto, Fabio Dominio, Ludovico Minto, and Guido Maria Cortelazzo. Data fusion from depth and standard cameras. In *Time-of-Flight and Structured Light Depth Cameras*, pages 161–196. Springer, 2016.

[ZSK20] Faezeh Sadat Zakeri, Mårten Sjöström, and Joachim Keinert. Guided optimization framework for the fusion of time-of-flight with stereo depth. *Journal of Electronic Imaging*, 29(5):053016, 2020.

[ZZS+17] Liang Zhang, Xiao Zhang, Juan Song, Peiyi Shen, Guangming Zhu, and Shaokai Dong. Viewpoint calibration method based on point features for point cloud fusion. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2224–2228. IEEE, 2017.

[ZZY13] Yin Zhao, Ce Zhu, and Lu Yu. Virtual view synthesis and artifact reduction techniques. In *3D-TV System with Depth-Image-Based Rendering*, pages 145–167. Springer, 2013.