UNIVERSITÄT
DES
SAARLANDES

Efficient Image-Based Rendering

Dissertation zur Erlangung des Grades des Doktors der Ingenieurwissenschaften der Fakultät für Mathematik und Informatik der Universität des Saarlandes

Vorgelegt von
Mojtaba Bemana

Saarbrücken, 2023

Dean:	Prof. Dr. Jürgen Steimle
Date:	12.07.2023
Chair:	Prof. Dr. Philipp Slusallek
Reviewers:	Dr.-Ing. habil. Karol Myszkowski
	Prof. Dr. Hans-Peter Seidel
	Prof. Dr. Petr Kellnhofer
	Prof. Dr. Tobias Ritschel
Academic Assistant:	Dr. Thomas Leimkühler

Abstract

Recent advancements in real-time ray tracing and deep learning have significantly enhanced the realism of computer-generated images. However, conventional 3D computer graphics (CG) can still be time-consuming and resource-intensive, particularly when creating photo-realistic simulations of complex or animated scenes. Image-based rendering (IBR) has emerged as an alternative approach that utilizes pre-captured images from the real world to generate realistic images in real-time, eliminating the need for extensive modeling. Although IBR has its advantages, it faces challenges in providing the same level of control over scene attributes as traditional CG pipelines and accurately reproducing complex scenes and objects with different materials, such as transparent objects. This thesis endeavors to address these issues by harnessing the power of deep learning and incorporating the fundamental principles of graphics and physical-based rendering. It offers an efficient solution that enables interactive manipulation of real-world dynamic scenes captured from sparse views, lighting positions, and times, as well as a physically-based approach that facilitates accurate reproduction of the view dependency effect resulting from the interaction between transparent objects and their surrounding environment. Additionally, this thesis develops a visibility metric that can identify artifacts in the reconstructed IBR images without observing the reference image, thereby contributing to the design of an effective IBR acquisition pipeline. Lastly, a perception-driven rendering technique is developed to provide high-fidelity visual content in virtual reality displays while retaining computational efficiency.

Zusammenfassung

Jüngste Fortschritte im Bereich Echtzeit-Raytracing und Deep Learning haben den Realismus computergenerierter Bilder erheblich verbessert. Konventionelle 3D-Computergrafik (CG) kann jedoch nach wie vor zeit- und ressourcenintensiv sein, insbesondere bei der Erstellung fotorealistischer Simulationen von komplexen oder animierten Szenen. Das bildbasierte Rendering (IBR) hat sich als alternativer Ansatz herauskristallisiert, bei dem vorab aufgenommene Bilder aus der realen Welt verwendet werden, um realistische Bilder in Echtzeit zu erzeugen, so dass keine umfangreiche Modellierung erforderlich ist. Obwohl IBR seine Vorteile hat, ist es eine Herausforderung, das gleiche Maß an Kontrolle über Szenenattribute zu bieten wie traditionelle CG-Pipelines und komplexe Szenen und Objekte mit unterschiedlichen Materialien, wie z.B. transparente Objekte, akkurat wiederzugeben. In dieser Arbeit wird versucht, diese Probleme zu lösen, indem die Möglichkeiten des Deep Learning genutzt und die grundlegenden Prinzipien der Grafik und des physikalisch basierten Renderings einbezogen werden. Sie bietet eine effiziente Lösung, die eine interaktive Manipulation von dynamischen Szenen aus der realen Welt ermöglicht, die aus spärlichen Ansichten, Beleuchtungspositionen und Zeiten erfasst wurden, sowie einen physikalisch basierten Ansatz, der eine genaue Reproduktion des Effekts der Sichtabhängigkeit ermöglicht, der sich aus der Interaktion zwischen transparenten Objekten und ihrer Umgebung ergibt. Darüber hinaus wird in dieser Arbeit eine Sichtbarkeitsmetrik entwickelt, mit der Artefakte in den rekonstruierten IBR-Bildern identifiziert werden können, ohne das Referenzbild zu betrachten, und die somit zur Entwicklung einer effektiven IBR-Erfassungspipeline beiträgt. Schließlich wird ein wahrnehmungsgesteuertes Rendering-Verfahren entwickelt, um visuelle Inhalte in Virtual-Reality-Displays mit hoher Wiedergabetreue zu liefern und gleichzeitig die Rechenleistung zu erhalten.

Acknowledgments

I greatly appreciate all the individuals who have contributed to the successful completion of this work. My deepest gratitude goes to my advisor Karol Myszkowski for his consistent guidance and support throughout the entire process. I am also grateful to Tobias Ritschel for his valuable insights and guidance during our collaborative projects. His vision, advice, and creativity have shaped me into a researcher. I was fortunate to work with Hyeonseung Yu and Piotr Didyk on early projects. They were among the most talented and inspiring people I have ever met. I would like to thank Jeppe Revall Frisvad for sharing his expertise and knowledge that led to the success of the Eikonal field project. Special thanks go to my oldest peer at MPI, Krzysztof Wolski, for the enjoyable time we shared. I also enjoyed the opportunity to work closely with other amazing researchers, especially Uğur Çoğalan, Lingyan Ruan, and Bin Chen. I would like to thank Thomas Leimkühler for proofreading the introduction of this thesis. I am deeply indebted to all members of AG4 for enriching my academic and personal life, and to Hans-Peter Seidel, for providing the best possible opportunities. Lastly, I would like to express my sincere gratitude to my wife, Zahra, as well as my family, for providing continuous support during my Ph.D. journey.

Contents

Chapter 1

Introduction

This thesis proposes several approaches which leverage the power of deep learning to broaden the scope of image-based rendering algorithms. This chapter describes the motivation (Section 1.1), introduces the main contributions (Section 1.2), and gives an overview of the whole thesis (Section 1.3).

1.1 Motivation

The role of computer graphics has become increasingly tangible in our everyday lives. Today, computer graphics play a pivotal role in the film industry[1], video games, digital photography, smartphones, and finally, in building the metaverse, a new world for interacting with technology [Zhao et al., 2022]. The powerful user interfaces empowered by computer graphics allow artists and individual users to transform their creative ideas into spectacular digital content quickly [Rombach et al., 2022]. Recent advances in artificial intelligence (AI) and deep learning techniques have undoubtedly contributed to the democratization of computer graphics[2] by reducing costs and enhancing the quality of computer-generated images. Nevertheless, real-time photorealistic rendering remains a significant challenge and is typically only possible with specialized and expensive graphics hardware[3]. Moreover, creating realistic 3D content and animating it in virtual reality is a laborious task that often requires extensive time and effort by a skilled artist.

While the current graphics pipeline allows for the manipulation of the physical attributes of a scene, such as geometry, lighting, and material appearance, there may be instances where the goal is to reproduce real-world objects rather than digitally crafting them. Examples of this include digitizing historical landmarks and artifacts for preservation, which may motivate the cultural heritage industry to explore virtual tourism opportunities. Another potential application is the creation of digital humans, which could be more efficiently achieved by scanning actual bodies rather than constructing them from scratch.

In fact, sensor data of real-world objects and environments obtained with a smartphone or DSLR camera can be processed and modeled to create realistic virtual environments in an easy and efficient way, and that's what *Image-Based Rendering* (IBR) does. IBR is an alternative approach that holds promise for producing photorealistic images bypassing costly physical simulations. IBR techniques start with a set of observations of real-world scenes and enable the generation of new images from the captured images alone (Figure 1.1). Unlike the 3D model-based rendering in traditional graphics, the rendering complexity for IBR is typically less dependent on

[1] www.collider.com: 8 most realistic CGI characters in movies

[2] www.ibc.org: How AI is reinventing visual effects

[3] www.blogs.nvidia.com: Leading Lights: NVIDIA researchers showcase groundbreaking advancements for real-Time graphics

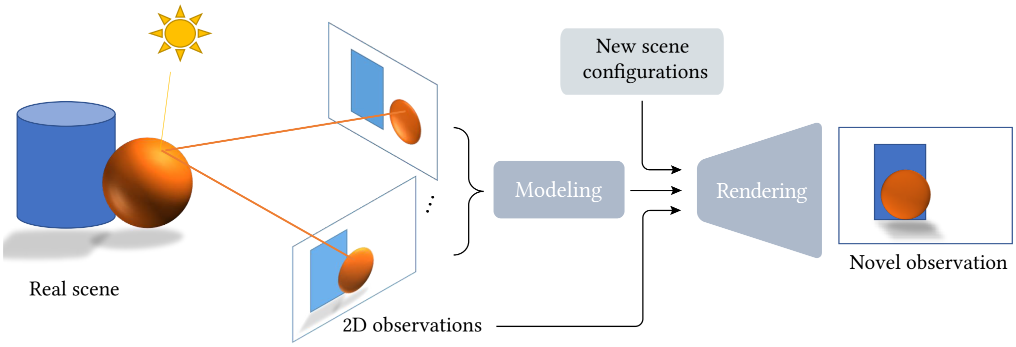


FIGURE 1.1: *The image-based rendering (IBR) pipeline takes a dense or sparse set of observations from real-world scenes as input and models the underlying structure of the scene (illumination, geometry, appearance, etc.) using an implicit or explicit 2D/3D representation. The IBR process ultimately enables rendering novel observations of the scene by re-configuring the scene parameters (e.g., the camera or light source position) inferred through the modeling process.*

scene content, so the rendering time can be accelerated by decoupling from scene complexity and simply re-sampling the captured images. The source of the input images to the IBR can be actual photographs, virtual content, or even a mixture of both. Light field (LF) rendering is one of the earliest forms of IBR, where an LF in a given scene captures information, including both the intensity and the direction of the light rays. By extracting appropriate 2D slices from the 4D LF, new views of the scene can be created. IBR was initially developed to change camera viewpoints from those that were captured; however, it has evolved beyond just visualizing scenes from different angles, and researchers are now interested in extracting and modifying other components of a scene, such as lighting and material properties of objects. The ability to retrieve the parameters of a real scene allows for the seamless integration of real and digital content, especially in augmented reality and visual effects applications. Additionally, IBR enables more straightforward model acquisition from photographs, which can be used to replace traditional geometric models in the modeling of complex scenes. This can be useful in creating scene content for gaming environments using available photos of real-world objects and environments or in helping an interior designer avoid the time-consuming process of designing every piece of furniture from scratch by working with a 3D model created from a collection of photos.

Nonetheless, key challenges must be addressed to regard an IBR approach as successful. The primary challenge is achieving high-fidelity reconstruction. Recent IBR representations [Tewari et al., 2020; Tewari et al., 2022] have demonstrated significant success in generating novel views of scenes with complex occlusion patterns and detailed structures. However, many of these representations are suitable for opaque surfaces and need help reproducing the correct view dependency effects for objects with different material appearances, such as transparent objects. Though the reconstruction provided by the current techniques is stable over time and contains few artifacts, there is no guarantee that the generated images will retain the quality of captured images and that the texture quality will not be degraded. The next challenge concerns the practical IBR, ensuring both the acquisition process and the rendering step are *efficient* in terms of time and computation. Accurate reconstruction of the scene primarily requires dense sampling, which, unfortunately, results in excessive storage, capturing, and processing time. Recent IBR approaches can function with sparse and unconstrained input images, allowing for easy, low-cost acquisition; yet, they struggle to provide fast optimization and real-time rendering. Real-time rendering and interaction, however, are essential for virtual and augmented reality applications in gaming, entertainment, healthcare, simulation, education, and other areas where hardware resources and memory bandwidth are often limited. Unraveling other graphic attributes to enable additional applications is another desired feature. With modern deep-learning-based techniques, it is now feasible to re-light the scene by changing the position of the light source or editing material properties, or even re-timing a dynamic sequence along with changing the viewpoint. However, there is still a lack of a unified presentation that can offer all these functions at once. Another concern is related to the scalability of an IBR method. It is highly desirable that the output of an IBR process can scale to any screen resolution or type, depending on whether it is a cell phone, a TV screen, or a head-mounted display (HMD), while remaining computationally efficient.

The approaches presented in this thesis aim to extend the scope of IBR algorithms by alleviating some of the remaining challenges. The first step towards achieving this goal is designing a neural network (NN) that already knows the "basic rules" of graphics (lighting, 3D projection, occlusion) and the "basic laws" of physical-based rendering (volume rendering, eikonal light transport) in a compact and differentiable form. Through the custom adaptation of NN and reformulation of the IBR representation, this thesis becomes able to (*i*) provide an interactive joint space, time, and light exploration in the captured real scenes and (*ii*) correctly reproduce the view dependency effect resulting from the interaction between the transparent objects and the surrounding environment. This way, IBR will gain greater versatility and provide more visually engaging experiences in virtual reality applications.

In the second step, inspired by the fact that humans are capable of spotting distortions in the images even without seeing their reference, this thesis develops a visibility metric that can locate artifacts in the reconstructed IBR images without the need to observe the reference image. Such a metric can contribute to the design of an effective IBR acquisition pipeline that will, in turn, minimize the compromise between reconstruction quality and capturing density by guiding the acquisition device to capture densely in regions where accurate reconstruction is highly challenging. Moreover, the human perception system is not always susceptible to visual errors, depending on the location, magnitude, and structure of errors in an image, hence allowing for a degree of imperfection in the rendered images, yet it is invisible to the human eye. Recognition of this insight enables us to develop efficient rendering techniques for improving visual content in virtual reality displays, where it often faces a trade-off between computational efficiency and perceptual quality.

1.2 Contributions

This thesis uses two tools to enable more efficient and versatile IBR: Implicit scene representation and image quality evaluation.

Implicit scene representation Traditional photorealistic rendering of real-world scenes proves tedious and challenging due to the need to reconstruct all physical parameters describing the rendered scenes. Recently, implicit scene representations have become a viable alternative to this task, where the entire scene is encoded into the parameters of a neural network. This compact scene representation is learned based on observations provided as unordered or structured sets of multi-view images. Inspired by this representation, Chapter 3 (based on Bemana et al. [2020]) presents X-Field,

- a compact solution for providing high-quality interpolation within a sparse (structured) set of light-view-time images,
- an efficient representation that can scale to high-resolution images and still provide real-time rendering, and
- a learning-based approach that does not require a large training dataset.

Despite recent advances in scene representations, existing approaches cannot properly reconstruct novel views of transparent objects with complex refraction and require special treatment. This problem is tackled by lifting the assumption that light rays are traversing in straight lines and adapting a physically correct approach to bend the light rays when they intersect with a refractive object in the scene. Specifically, Chapter 4 (based on Bemana et al. [2022]) integrates the physical laws of the eikonal light transport [Ihrke et al., 2007] with a state-of-the-art novel view synthesis method (NeRF [Mildenhall et al., 2020]) and presents

- a high-quality novel view reconstruction of refractive objects which does not require any shape prior and can work with unconstrained capturing setup and scene configuration,
- an implicit representation that can model refractive objects with a spatially varying index of refraction, and
- a volume rendering formulation for the curved light rays using Ordinary Differential Equation (ODE).

Image quality evaluation The development of an image quality metric can play an important role in accelerating the performance of IBR techniques. Commonly used image quality evaluations either focus on providing a single score per image or require a reference image to access the quality [Wang et al., 2004b; Zhang et al., 2018]. However, having a visibility metric with localized error prediction is essential for quality control, especially in IBR, where the artifacts are often localized. Moreover, the reference images in many applications are not readily accessible; they might be impossible to compute or unavailable. Therefore, it becomes more sensible not to always rely on reference pairs. To this end, Chapter 5 (based on Bemana et al. [2019]) proposes

- a per-pixel no-reference quality metric for identifying IBR artifacts,
- a training strategy to avoid false positive and negative predictions, and
- two applications to validate the proposed non-reference metric in light-field production.

A critical requirement for the faithful reconstruction of virtual 3D content is the reproduction of accommodation cues. Recent studies have shown that multilayer displays, such as light-field displays, are promising solutions for HMDs to provide near-correct accommodation cues [Hua, 2017]; however, a bigger challenge is involved in rendering multiple images in real-time for such advanced HMDs. In this regard, this thesis devises a perceptual model to locally determine where a low-cost decomposition strategy, namely, linear blending (LB) [MacKenzie et al., 2010], can be applied instead of the costly light-field rendering (LFS) [Lee et al., 2016] without sacrificing visual quality. Particularly Chapter 6 (based on Yu et al. [2019]) develops

- a perception-based hybrid decomposition method that combines the advantages of the above strategies and achieves both real-time performance and high-fidelity results,
- a gaze-dependent viewpoint sampling of LFS to improve reconstruction quality,
- a series of targeted perceptual experiments that measure the differences in visual quality between LB and LFS for different spatial frequencies, luminance contrasts, depth configurations, and eccentricities,
- a domain-specific calibration of the structural similarity index (SSIM) for predicting the visible differences between LB and LFS, which provides generalized perceptual insights beyond the scope of the perceptual experiments, and
- a unified optimization framework for the LB and LFS decompositions and an efficient adaptation of the simultaneous algebraic reconstruction technique (SART) to CUDA for real-time decompositions.

The author of this thesis is the main contributor of the work presented in each chapter; however, is not engaged in some parts in Chapter 6, including the implementation of ray-tracing and eye-tracking, the preparation of the scenes, and running perceptual experiments.

1.3 Outline

This thesis is organized as follows. Chapter 2 reviews and discusses relevant work. Chapter 3 introduces the proposed compact solution for joint light-view-time image interpolation. Chapter 4 presents the eikonal-based approach to correctly synthesizing the novel view of refractive objects. Moving on to the image quality metric in Chapter 5, where a no-reference visibility metric is designed for detecting artifacts in IBR images. Finally, Chapter 6 discusses the proposed perception-driven hybrid decomposition technique for multilayer accommodative displays. The conclusion of this thesis and discussion of promising research directions are provided in Chapter 7.

Chapter 2

Previous Work

This chapter reviews the previous work relevant to this thesis. The existing methods on image-based interpolation and view synthesis for transparent objects are discussed in Section 2.1 and Section 2.2, respectively. Then previous work regarding the objective image quality metrics are covered in Section 2.3. Finally, the topics related to multi-layered displays are reviewed in Section 2.4.

2.1 Image-Based Interpolation

Image-based interpolation is a long-standing problem in the vision and graphic community and spans many dimensions. This section reviews previous techniques to interpolate across discrete sampled observations in view (light fields), time (video), space-time, and illumination (reflectance fields). Table 2.1 summarizes this body of work along multiple axes.

2.1.1 View

View interpolation, aka novel view synthesis, refers to the process of generating novel viewpoints of a scene given a set of existing images or videos. This section reviews existing traditional approaches and recent learning-based methods for view interpolation.

Traditional approaches The concept of view interpolation was first introduced by Chen and Williams [1993]. It involves creating novel views by warping input images with pixels correspondence computed from image range data. Levoy and Hanrahan [1996], as well as Gortler et al. [1996], were also the first to formalize the concept of the light field (LF) and to devise acquisition hardware to capture it. Later view interpolation solutions, such as Unstructured Lumigraph Rendering (ULR) [Buehler et al., 2001; Chaurasia et al., 2013], involves creating proxy geometry to warp [Mark et al., 1997] multiple observations into a novel view and blend them with specific weights. Avoiding the difficulty of reconstructing geometry or 3D volumes has been addressed for LFs in [Du et al., 2014; Kellnhofer et al., 2017]. More recent works have used per-view geometry [Hedman et al., 2016b] and learned ULR blending weights [Hedman et al., 2018] to allow sparse input and view-dependent shading.

LF methods come first in Table 2.1, where they are checked "view" as they generalize across an observer's position and orientation. Depending on resources, a capture setup can be considered simple (cell phone) or more involved (light stage) as denoted in the "easy capture" column in Table 2.1. An important distinction is that a capturing can be *dense* or *sparse*, denoted as "Sparse" in Table 2.1. Sparsity depends less on the number of images but more on the difference between captured images. Very similar view positions [Kalantari et al., 2016] as for a Lytro camera can be considered dense, while 34 views on a sphere [Lombardi et al., 2019a] or 40 lights on a hemisphere [Malzbender et al., 2001] are sparse. The focus of this thesis is on wide camera baselines, with typically $M \times N$ cameras spaced by 5–10 cm [Flynn et al., 2019], and respectively a large pixel disparity ranging up to 250 pixels [Dabała et al., 2016; Mildenhall et al., 2019], where M and N are single-digit numbers, e. g., 3×3, 5×5 or even 2×1.

Multi-plane image representation With the advent of neural networks, Flynn et al. [2016] proposed the first method applying deep neural networks to the problem of view synthesis for a set of real-world images. They decompose the scene into multiple depth planes of the output view and construct a view-dependent plane sweep volume (PSV) to render novel views. Kalantari et al. [2016] adopt a similar idea to learn to synthesize novel views for LF data. They indirectly learn depth maps without depth supervision to interpolate between views in a Lytro camera. Instead of using proxy geometry, Penner and Zhang [2017a] have suggested using a volumetric occupancy representation. By learning how neighboring input views contribute to the output view, the multi-plane image (MPI) representation [Zhou et al., 2018] can be built, which enables high-quality local LF fusion [Mildenhall et al., 2019]. Inferring a volumetric/MPI representation can be facilitated with learned gradient descent [Flynn et al., 2019], where the gradient components directly encode visibility and effectively inform the NN on the occlusion relations in the scene. MPI techniques avoid the problem of explicit depth reconstruction and allow for softer, more pleasant results. A drawback in deployment is the massive volumetric data, the difficulty of distributing occupancy therein, and finally, the bandwidth requirements of volume rendering itself.

This thesis also involves a learning route and uses a NN to represent the scene implicitly (Chapter 3), and the deployment only requires a few additional kilobytes of NN parameters on top of the input image data and rendering in real-time. From yet another angle, ULR-inspired IBR creates an LF (view-dependent appearance) on the surface of a proxy geometry, i. e., a *surface light field*. Chen et al. [2018a], using an MLP, as well as Thies et al. [2019], using a CNN, have proposed to represent this information using a NN defined in the texture space of a proxy object. While inspired by the mechanics of sparse IBR, results are typically demonstrated for rather dense observations. In contrast, this thesis does not assume any proxy to be given but jointly represents the appearance and the geometry used to warp over many dimensions in a single NN, trained from sparse sets of images (Chapter 3).

Flow-based methods The Appearance Flow work of Zhou et al. [2016] suggests combining the idea of warping pixels with learning how to warp. While Zhou et al. [2016] typically consider a single input view, Sun et al. [2018b] employ multiple views to improve warped view quality. Both works use an implicit representation of the warp field, i. e., a NN that for every pixel in one view predicts from where to copy its value in the new view. While those techniques worked best for fixed camera positions that are used in training, Chen et al. [2019b] introduces an implicit NN of per-pixel depth that enables an arbitrary view interpolation. All these methods require extensive training for specific classes of scenes, such as cars, chairs, or urban city views.

This thesis takes this line of work further by constructing an implicit NN representation that generalizes jointly over complete geometry, motion, and illumination changes (Chapter 3). The task, on the one hand, becomes simpler as it does not need to generalize across different scenes, yet on the other hand, it is also harder as it requires generalizing across many more dimensions and provides state-of-the-art visual quality.

Implicit volumetric representation Another step of abstraction is voxel representation [Sitzmann et al., 2019a]. Instead of storing opacity and appearance in a volume, abstract "persistent" features are derived, which are then projected using learned ray-marching. The volume was learned per scene. Lombardi et al. [2019b] propose a voxel grid with an interpolation that is optimized using a CNN and encodes both dynamic geometry and appearance. Learned warping is performed to reduce memory requirements and improve the resolvable details. Recently, implicit scene representation [Sitzmann et al., 2019b], and [Saito et al., 2019] has become a promising approach for high-quality novel view synthesis. Mildenhall et al. [2020] introduce a volumetric opacity representation called NeRF that encodes geometry and appearance using a multi-layered perceptron (MLP) trained on a large set of multiple-view RGB images and proves to be extremely successful in novel view synthesis tasks. Despite producing high-quality results, MLP-based approaches are slow in rendering and optimization, especially for high-resolution images. A speed-up can be achieved with the recent grid-based structure (with no neural components) [Yu et al., 2021a; Yu et al., 2021b] or multiresolution hash table augmented with a shallow neural network [Müller et al., 2022], or multiple compact low-rank tensor components [Chen et al., 2022], which they directly optimize from the input images using gradients methods. However, these representations are usually less compact, resulting in large storage requirements, especially for rendering large scenes or high-resolution images. Such approaches are called "implicit" in Table 2.1 when the NN replaces the pixel basis, i. e., the network provides a high-dimensional `getPixel(x)`. These approaches use an MLP that can be queried for occupancy [Chen and Zhang, 2019; Sitzmann et al., 2019b; Saito et al., 2019], color [Oechsle et al., 2019; Sitzmann et al., 2019b; Mildenhall et al., 2020] etc. at different 3D positions along a ray for one pixel. The X-Field representation introduced in Chapter 3 makes two changes to this design. First, it predicts texture coordinates rather than appearance. These drive a spatial transformer [Jaderberg et al., 2015] that can copy details from the input images without representing them and do so at high speed (20 fps). Second, it trains a 2D CNN instead of a 3D MLP that, for a given X-Field coordinate, will directly return a complete 2D per-pixel depth and correspondence map. For an X-Field problem, this is more efficient than ray-marching and evaluating a complex MLP at every step [Tewari et al., 2022].

While implicit representations have so far been demonstrated to provide a certain level of fidelity when generalized across a class of simpler shapes (cars, chairs, etc.). This thesis makes the task simpler and generalizes less while producing quality to compete with the state-of-the-art view, time and light interpolation methods from computer graphics. Inspired by [Nguyen Phuoc et al., 2018], some work [Sitzmann et al., 2019a; Nguyen Phuoc et al., 2019; Sitzmann et al., 2019b] learns the differentiable tomographic rendering step, while other work has shown how it can be differentiated directly [Henzler et al., 2019a; Lombardi et al., 2019a; Mildenhall et al., 2020]. This thesis avoids tomography and works with differentiable warping [Jaderberg et al., 2015] with consistency handling inspired by unsupervised depth reconstruction [Godard et al., 2017; Zhou et al., 2017]. Avoiding volumetric representations allows for real-time playback while at the same time generalizing from view to other dimensions such as time and light.

2.1.2 Time (Video)

Videos comprise discrete observations of a scene over time and hence are a sparse capture of the visual world. To get smooth interpolation, e. g., , for slow-motion (individual frames), motion blur (averaging multiple frames) images need to be interpolated [Mahajan et al., 2009], potentially using NNs [Sun et al., 2018a; Jiang et al., 2018; Bao et al., 2019b; Bao et al., 2019a; Wang et al., 2018a]. More exotic domains of video re-timing, which involve annotation of a fraction of frames and one-off NN training, include the visual aspect in sync with spoken language [Fried and Agrawala, 2019]. Although the recent techniques [Reda et al., 2022; Sim et al., 2021] have shown great success in handling scenes with large uniform motion and the presence of complex occlusions, the computational cost for these methods is relatively high, hindering a real-time performance, especially for a high-resolution input. One can take an existing frame interpolation framework and perform multiple interpolation steps to reach any point within the view-time-light cube. However, even with a fast method like Jiang et al. [2018], this approach becomes inefficient in terms of run time, requiring up to seven steps for interpolation within the view-time-light cube.

2.1.3 Space-Time

Warping can be applied to space or time, as well as to both jointly [Manning and Dyer, 1999], resulting in LF video [Wang and Yang, 2005; Lipski et al., 2010; Wang et al., 2017; Zitnick et al., 2004]. Recent work has extended deep novel-view methods into the time domain [Lombardi et al., 2019a]. They also use warping for a very different purpose: deforming a pixel-basis 3D representation over time to avoid storing individual frames (motion compensation). Both methods of Sitzmann et al. [2019a] and Lombardi et al. [2019a] are limited by the spatial 3D resolution of volume texture and the need to process it, while this thesis works with 2D depth and color maps only (Chapter 3). As they learn the tomographic operator, this limit in resolution is not a classic Nyquist limit, e. g., sharp edges can be handled, but results typically are on isolated, dominantly convex objects. Ultimately, this thesis does not claim depth maps to be superior to volumes per se, instead suggests that 3D volumes have their strength for seeing objects from all views (at the expense of resolution), whereas using images is more for observing scenes from a "funnel" of views but at high 2D resolution. No work yet is able to combine high resolution and arbitrary views, not to mention time. Recent implicit-based representation methods [Park et al., 2021; Pumarola et al., 2021] deal with dynamic content by jointly learning a canonical NeRF volume and a deformation field or a dense scene flow fields [Li et al., 2022b; Gao et al., 2021] between the scene at a particular moment in time and the scene in canonical space. Such deformation can also be implemented as ray bending, where straight rays are deformed non-rigidly [Tretschk et al., 2021a]. While the input to these methods is merely a monocular video with only one view observation at each time stamp, it often contains a dense observation in the time domain. Moreover, the training step usually takes several days, and rendering is still far from real-time.

2.1.4 Light (Reflectance Fields)

While an LF is specific to one illumination, a *reflectance field* (RF) [Debevec et al., 2000] is a generalization additionally allowing for relighting, often just for a fixed view. Dense sampling for individually controlled directional lights can be performed using Light Stage [Debevec et al., 2000], which leads to hundreds of captured images. The

Method	Citation	Scene	View	Time	Light	Sparse	Unstructured	Real-time	Easy capture	Learned	Implicit	Diff. render.	Remarks
		Generalize				Interface				Implem.			
Unstructured Lumigraph	[Buehler et al., 2001]	✓	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	IBR
Inside Out	[Hedman et al., 2016b]	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	IBR; SfM; Per-view geometry
LF View Interp.	[Kalantari et al., 2016]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	Lytro; Learned disparity and fusion
Soft 3D	[Penner and Zhang, 2017a]	✓	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	MPI
Deep Blending	[Hedman et al., 2018]	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗	IBR; SfM; Learned fusion
Deep Surface LFs	[Chen et al., 2018a]	✗	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	Texture; Lumitexel; MLPs
Local LF Fusion	[Mildenhall et al., 2019]	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗	MPI
DeepView	[Flynn et al., 2019]	✗	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	MPI
Neural Textures	[Thies et al., 2019]	✗	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓	Texture; Lumitexel; CNNs
DeepVoxels	[Sitzmann et al., 2019a]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✓	✓	3D CNN
HoloGAN	[Nguyen Phuoc et al., 2019]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	Adversarial; 3D representation
Appearance Flow	[Zhou et al., 2016]	✓[1]	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	App. Flow; Fixed views
Multi-view App. Flow	[Sun et al., 2018b]	✓[2]	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	App. Flow; Learned fusion; Fixed views
Spatial Trans. Net IBR	[Chen et al., 2019b]	✓[3]	✓	✗	✗	✗	✓	✓	✗	✓	✓	✓	App. Flow; Per-view geometry; Free views
NeRF	[Mildenhall et al., 2020]	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	✓	MLPs, ray-marching
Plenoxels	[Yu et al., 2021a]	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	Voxel girds; no neural components
TensoRF	[Chen et al., 2022]	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	Compact low-rank tensors
Instant-NGP	[Müller et al., 2022]	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	Multiresolution hash table; Shallow MLP
Moving Gradients	[Mahajan et al., 2009]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	Gradient domain
Super SlowMo	[Jiang et al., 2018]	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	Occlusions: Learns visibility maps
Video-to-video	[Wang et al., 2018a]	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	Adversarial; Segmented content editing
Puppet Dubbing	[Fried and Agrawala, 2019]	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	Visual and sound sync.
Depth-aware Frame Int.	[Bao et al., 2019a]	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	Occlusions: Learns depth maps
MEMC-Net	[Bao et al., 2019b]	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	Occlusions: Learns visibility maps
XVFI	[Sim et al., 2021]	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	Learns occlusions maps; 1000-fps dataset
FILM	[Reda et al., 2022]	✓	✗	✓	✗	✓	✗	✗	✓	✓	✓	✗	Implicit occlusions handling
Layered Representation	[Zitnick et al., 2004]	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	MVS reconst.; Layered Depth Images
Video Array	[Wilburn et al., 2005]	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	Optical flow
Virtual Video	[Lipski et al., 2010]	✓	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗	Structure from Motion (SfM)
Hybrid Imaging	[Wang et al., 2017]	✓	✓	✓	✗	✓[4]	✗	✓	✓	✓	✗	✗	Lytro+DSLR camera system
Neural Volumes	[Lombardi et al., 2019a]	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓	✓	3D CNN; LightStage; Fixed time (video)
Scene Represent. Net	[Sitzmann et al., 2019b]	✓[5]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	3D MLP
PIFu	[Saito et al., 2019]	✓[6]	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	3D MLP
D-NeRF	[Pumarola et al., 2021]	✗	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	MLPs; Deformation field
NR-NeRF	[Tretschk et al., 2021a]	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	MLPs; Non-rigid deformation
Neural 3D Video	[Li et al., 2022b]	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	MLPs; 3D motion flow fields
Polynomial Textures	[Malzbender et al., 2001]	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	LightStage
Neural Relighting	[Ren et al., 2015]	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓	✗	MLP; LightStage and hand-held lighting
Sparse Sample Relighting	[Xu et al., 2018]	✓	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗	Optimized light positions
Deep Reflectance Fields	[Meka et al., 2019]	✓[7]	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	LightStage; Moving Performers
Total relighting	[Pandey et al., 2021]	✓	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓	LightStage; Fixed-view portrait relighting
Lumos	[Yeh et al., 2022]	✓	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓	Portrait relighting; Synthetic LightStage
Sparse Sample View Synth.	[Xu et al., 2019]	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗	Optimized lights as in [Xu et al., 2018]
Multi-view Relighting	[Philip et al., 2019]	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	Geometry proxy; Auxiliary 2D buffers
The Relightables	[Guo et al., 2019]	✓	✓	✗	✓	✗	✗	✓	✗	✓	✗	✗	LightStage; Depth sensor data
Deep Relightable Textures	[Meka et al., 2020]	✓[8]	✓	✗	✓	✗	✓	✗	✗	✓	✓	✓	LightStage; Neural rendering
NeRV	[Srinivasan et al., 2021]	✗	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	Known illumination; Decomposed BRDF
NeRD	[Boss et al., 2021]	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	MLPs; Decomposed BRDF
Relighting4D	[Chen and Liu, 2022]	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	Full-body relighting; Monocular video
RANA	[Iqbal et al., 2022]	✓	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	Articulated human relighting
Ours		✗	✓	✓	✓	✓	✗[9]	✓	✓	✓	✓	✓	

TABLE 2.1: *Comparison of space, time, and illumination interpolation methods* (rows) *with respect to capabilities* (columns)*, with an emphasis on deep methods.* (*[1−3,5]Similar-class scenes demonstrated, e. g., cars, chairs, urban city views; [4]LF sparse in time; [6]Clothed humans demonstrated; [7]Human faces shown; [8]Only for human performance capture, but can generalize to unseen performance; [9]Only structured grids are shown. Should support unstructured, as long as the transformation between views is known.*)

number of images can be reduced by employing specially designed illumination patterns [Fuchs et al., 2007; Peers et al., 2009; Reddy et al., 2012] to exploit various forms of coherence in the light transport function. For interpolation, the signal is frequently separated, such as into highlights, reflectance, or shadows [Chen and Lensch, 2005]. One method in this thesis (Chapter 3) also found such a separation helpful. Angular coherence in incoming lighting leads to an efficient reflectance field representation as polynomial texture maps [Malzbender et al., 2001], which can be further improved by neural networks whose expressive power enables one to capture non-linear spatial coherence [Ren et al., 2015], or generalize across views [Maximov et al., 2019]. Xu et al. [2018] directly regresses images of illumination from an arbitrary light direction when given five images from specific other light directions. The innovation is in optimizing what should be input at test time. Still, the setup requires custom capture dome equipment, as well as input images taken from those five very specific directions. For scenes captured under controlled illumination for multiple sparse views, generalization across views can be achieved by concatenation with a view synthesis method [Xu et al., 2019]. While the results are compelling on synthetic scenes, the method exhibits difficulties in handling complex or non-convex geometry, as well as high-frequency details such as specularities and shadows [Meka et al., 2019]. An approximate geometry proxy and extensive training over rendered scenes might compensate for inaccuracies in derived shadows and overall relighting quality [Philip et al., 2019]. Specialized systems for relighting human faces and characters remove many such limitations, including fixed view and static scene assumptions, using advanced Light Stage hardware that enables capturing massive data for CNN training [Meka et al., 2019] and complex optimizations that are additionally fed with multiple depth sensor data [Guo et al., 2019]. As only two images for an arbitrary face or character under spherical color gradient lighting are required at the test time, real-time dynamic performance capturing is possible. Free-viewpoint rendering of dynamic performers can also be achieved through CNN-based, LF-style interpolation in Meka et al. [2020], whereas Guo et al. [2019] capture complete 3D models with textures and can easily change viewpoint as well. The custom capture hardware required in these relighting approaches hinders their use in casual photography. Pandey et al. [2021] enables changing the lighting condition of a single portrait image but still relies on Light Stage data for training supervision. Yeh et al. [2022] overcome this limitation by utilizing synthetic custom data and a synthetic to real domain adaptation module, resulting in temporally consistent video portrait relighting. This thesis employs specific-scene training and removes the requirements for massive training data and costly capturing hardware [Meka et al., 2020; Guo et al., 2019]. At the same time, it enables the real-time rendering of animated scenes under interpolated dynamic lighting and view position (Chapter 3).

Recent techniques [Boss et al., 2021; Srinivasan et al., 2021] adapt the NeRF representation [Mildenhall et al., 2020] and employ a physically-based rendering to extract the volume density of a scene along with BRDF material properties of the objects, which then allows rendering novel views of the object under arbitrary illumination. However, these techniques primarily focus on relighting static objects and do not effectively incorporate both illumination and scene dynamics. Relighting full-body dynamic avatars is made possible by simultaneously estimating geometry, texture, and environment lighting from a short video clip of the person [Chen and Liu, 2022; Iqbal et al., 2022].

2.2 View Synthesis for Transparent Objects

As the focus of this thesis in Chapter 4 is the novel-view synthesis for refractive transparent objects, Section 2.2.1 first discusses this problem with an emphasis on recent neural rendering solutions that can handle specular effects. Section 2.2.2 also overviews the image-based modeling of transparent objects, which is a more general setup than required in this thesis, but still, some similarities can be found. Finally, the physics-based eikonal rendering is discussed (Section 2.2.3).

2.2.1 Reproducing Specular Effects

Complementary to Section 2.1.1, this section reviews solutions for novel-view synthesis in static scenes, where 3D representations such as multi-layer perceptron (MLP), voxel grids, and multi-plane images (MPI) are disentangled from material properties and image formation processes so that classic physics-based rendering can be employed to simulate complex lighting effects such as specular reflection and refraction.

Neural radiance fields (NeRF) Using the NeRF MLP-based representation [Mildenhall et al., 2020], the view-dependent RGB color and view-independent density are learned as sharp functions in space and smooth functions in angle, where the density determines the contribution of each location to the color integrated along any ray traversing the NeRF volume. In the case of near-mirror or near-glass reflection/refraction, appearance cannot be described as a smooth function of angle anymore [Guo et al., 2021]. As a result, low-frequency, view-dependent effects such as highlight positions can be partially reproduced; however, the scene seen in mirror reflections or through transparent objects appears notoriously blurry or with ghosting artifacts [Ichnowski et al., 2021]. Some solutions exist [Zhang et al., 2021b; Boss et al., 2021] to disentangle normal vectors and spatially-varying reflectance by manipulating the NeRF density representation. Still, highly specular surfaces with a clearly visible reflected environment cannot be reproduced. This thesis (Chapter 4) does not fully rely on the NeRF geometry and uses the diffuse scene only as a backdrop. Inspired by traditional image-based rendering [Sinha et al., 2012; Xu et al., 2021] for scenes with planar reflections, Guo et al. [2021] introduce NeRFReN, where an additional NeRF structure is proposed that renders a reflected image and composes it additively with the traditional NeRF rendering. MirrorNeRF [Wang et al., 2021b] employs a catadioptric imaging system based on an array of hemispherical mirrors, enabling a single-shot portrait reconstruction and rendering. An implicit representation of a continuous displacement field is learned per mirror to warp every sample point from rays emitted by a given mirror to a common reference NeRF space. Only the sample point position is warped, but its viewing direction is not changed, while in this thesis, view directions are bent within a novel NeRF structure. A similar concept of warping is employed to accommodate non-rigid object deformations to a canonical reference NeRF [Pumarola et al., 2020; Tretschk et al., 2021b], but again straight rays are considered in rendering. Recently, Huang et al. [2021] introduce high dynamic range HDR-NeRF that relies on RAW-captured images that might further improve the quality of specular effect rendering.

Other volumetric representations Sharper mirror reflection and transparency effects can be obtained using an extension of the multiplane image (MPI) representation, where for every pixel in a stack of parallel semi-transparent planes, directional information using learned basis functions is stored [Wizadwongsa et al., 2021]. Similarly, as for other MPI-based and neural light fields [Attal et al., 2021] methods, only narrow baselines are supported. Signed distance fields (SDF), possibly encoded into an MLP, can be used to represent surface geometry and recover non-spatially-varying reflectance using spherical Gaussians that, in turn, enables good quality reflections [Zhang et al., 2021a]. In an alternative point-based representation [Kolos et al., 2020], where each point is associated with a learnable photometric, geometric, and transparency descriptor, a relatively sharp depiction of semi-transparent objects is achieved when the non-distorted background is also known. However, specular effects are explicitly excluded from the training data.

In all those solutions, an important limiting factor is a straight light path assumption in the rendering formulation that neglects light reflection and refraction effects for complex geometry settings. The method in this thesis (Chapter 4) additionally reconstructs a volumetric index of refraction (IoR) field along with simulating the laws of physics associated with refractive effects, which enables explaining the input RGB images at the learning time and consequently provides a meaningful synthesis of novel views at the test time.

2.2.2 Transparent Surface Reconstruction

The visual appearance of transparent objects is strongly affected by their absorptive, refractive, and reflective properties and varies with background and illumination, which makes image-based reconstruction of such surfaces a difficult, fundamentally under-constrained problem that requires dedicated solutions. This section first discusses more foundational work that formulates theoretical conditions to make such reconstruction computationally tractable. Then, the environment matting techniques are discussed. These techniques are trying to solve a more focused problem of background deformation by a transparent object so that such an object can be composited onto different backgrounds. Finally, this section presents the practical solutions for reconstructing a complete transparent object geometry, e. g., in the form of explicit meshed or point-based and implicit NeRF models that are then suitable for arbitrary re-rendering. Typically such reconstruction is performed in controlled environments, often with the use of structured illumination. In contrast, other applications such as robotics require such reconstruction in the wild, in potentially cluttered scenes. Those two scenarios are briefly discussed with an emphasis on recent machine learning solutions, and the existing surveys [Ihrke et al., 2008; Ihrke et al., 2010] discuss a more comprehensive treatment.

Theoretical foundations Kutulakos and Steger [2008] investigate two-interface refractive light interaction with a surface, and for every pixel, recover multiple 3D points so that a ray exiting the surface can be reconstructed. They show that by using a three-viewpoint setup, they can reconstruct the underlying general geometry. Importantly, they enumerate a tractable number of light bounces that can be reconstructed along a light path for various acquisition setups. Importantly, they also demonstrate that by purely geometric means, it is impossible to reconstruct individual light paths when light bounces more than two times. The two-refraction case, where the correspondence between the incident and exit is found, has been investigated in follow-up work [Tsai et al., 2015].

Environment matting A deformation pattern in the light transport is reconstructed for a transparent object that is imposed on differently structured backgrounds using a sequence of images, where additionally undeformed background must be known [Zongker et al., 1999; Chuang et al., 2000; Peers and Dutré, 2003], or may remain unknown [Matusik et al., 2002; Wexler et al., 2002]. Khan et al. [2006] demonstrate that even significant departs of such a deformation pattern from physical refractive processes can be tolerated by human perception to make realistically-looking compositing of a transparent object over any background. Even a simple user markup with deformed/non-deformed stroke pairs might be sufficient to derive such a plausible deformation pattern [Yeung et al., 2011]. Along a similar line, Chen et al. [2019a] propose a massively-trained neural network that, at test time, derives a transparent object mask, an attenuation mask, and a realistically-looking deformation pattern all from a single photograph so that the transparent object can be composited onto any background.

Reconstruction in controlled environments Various dedicated setups that rely on light-field background displays [Wetzstein et al., 2011b], transmission imaging [Kim et al., 2017b], and X-ray computational tomography (CT) scanners [Stets et al., 2017] have been used for transparent object geometry reconstruction. In intrusive setups, which require immersing transparent objects into a liquid with matching IoR, straight light paths can be assumed that greatly simplifies CT reconstruction [Trifonov et al., 2006a] or range scanning when fluorescent liquid is employed [Hullin et al., 2008]. Inspired by environment matting, Wu et al. [2018] and Lyu et al. [2020] place a transparent object on a turntable in front of a coded background and capture its multiple views from a static camera position. Wu et al. [2018] derive the correspondence between the incident (camera) and exit rays that reach the background, which additionally requires rotating the background and finally consolidating the resulting point clouds into a clean geometric model. Lyu et al. [2020] perform coarse-to-fine mesh optimization, which is driven by differentiable tracing of the refractive two-bounce light path so that the distorted refractive pattern and object silhouettes match the captured photographs.

In-the-wild reconstruction Li et al. [2020] employ a cell phone to capture a small number of views along with segmented transparent object masks and a known environment map, which are provided as the input for their method. They propose an in-network differentiable rendering layer with a physical image formation model under an assumption of two-bounce refractive light paths to refine associated normal vectors so that a point-cloud model can be reconstructed that, in turn, explains the correspondence between the input images and environment map. Sajjan et al. [2020] show that by employing an RGB-D camera, the segmentation task is greatly simplified. At the same time, complex reflective and refractive light patterns enable a neural network to infer surface normals even from a single input image so that potentially unreliable depth information is further refined. Similar goals can be achieved using even a single RGB image and a massively trained encoder-decoder network [Stets et al., 2019]. As pointed out in Lyu et al. [2020], the domain gap can still be expected, as these networks [Stets et al., 2019; Li et al., 2020; Sajjan et al., 2020] are trained mostly on synthetic data. The most successful solutions for transparent object segmentation in RGB images rely on CNN [Khaing and Masayuki, 2019] or transformer [Xie et al., 2021] networks that, in turn, require training with large annotated datasets. Dex-NeRF [Ichnowski et al., 2021] does not require any prior dataset and derives a transparent object depth by searching along the ray traversing NeRF's volume for the first sample density whose value is larger than a given threshold. Multiple camera views are required for geometry reconstruction, where specular reflections from multiple light sources further improve the learned geometry quality.

2.2.3 Eikonal Rendering

Light propagation in media with varying IoR has been modeled based on formulations derived from the eikonal and transport equations. In Berger et al. [1990b], Berger et al. [1990a], and Musgrave [1990], mirage rendering is proposed by tracing rays through discrete atmosphere layers so that the IoR increases with elevation. Stam and Languénou [1996] extend this discrete formulation to media with continuously varying IoR by introducing the eikonal equation to rendering applications. Gutierrez et al. [2005] revisit mirage and other atmospheric effects rendering using such continuous formulation. Ihrke et al. [2007] derive from the eikonal equation a wavefront tracing technique to precompute irradiance distribution in a volume that enables the efficient rendering of media with non-homogeneous IoR. This thesis rather deals with an inverse rendering problem and aims to learn a continuous 3D IoR field. The eikonal equation provides a principled connection between gradients in the learned IoR and light propagation along curved trajectories that explains the input 2D images. In seismological applications, a factored eikonal formulation is used to train a network to predict travel time between any source-receiver pair in a continuous 3D space with non-homogeneous seismic velocities [Smith et al., 2021]. Training data involves massively sampled travel time measures between different points in the 3D space, while this thesis employs 2D images with refraction patterns. In interferometric tomography [Sweeney and Vest, 1973; Liu and Yang, 1989; Tian et al., 2011] 3D IoR field is reconstructed for investigating physical parameters such as temperatures or densities, e. g., in high-speed aerodynamic flows. Such techniques typically capitalize on phase modulation that changes with the optical path length of straight, rather than curved, light rays passing through transparent media and can be extracted from interferograms.

2.3 Image Quality Assessment

This section first discusses the objective image quality metrics (Section 2.3.1), with special emphasis on those that do not require a clean reference image (Section 2.3.2). Then, Section 2.3.3 briefly characterizes IBR-specific artifacts, as well as metrics specialized in their detection, which is the key focus of Chapter 5 in this thesis.

2.3.1 Image Metrics

Some applications and functions may only require a *quality* score while others need a *visibility* map [Chandler, 2013].

Image quality metrics (IQMs) evaluate the distortion magnitude and are typically trained on the mean-opinion score (MOS) data [Sheikh et al., 2006; Ponomarenko et al., 2009] that labels the entire image as a single quality score. The most commonly used IQMs such as PSNR, SSIM, MS-SSIM [Wang and Bovik, 2006], FSIM [Zhang et al., 2011], and CIELAB [Zhang and Wandell, 1997] are *full-reference* (FR) metrics that take as input the reference and distorted images and compute local differences that are pooled into a global, single quality score. Recently, it has been demonstrated that CNN-based FR-IQMs achieved the best performance in predicting MOS data [Amirshahi et al., 2016; Bosse et al., 2018]. Zhang et al. [2018] show that the distance between the features extracted from a pre-trained classifier such as VGG [Simonyan and Zisserman, 2014] can be used as a perceptual measure for IQM. They also employ crowdsourcing and create a large-scale patch-based dataset in two perceptual experiments: (1) two-alternative forced choice (2AFC) on distortion strength and (2) “same/not same" near-threshold distortion visibility. Then, they train different network architectures and report, in each case, a much better performance than traditional FR-IQMs in predicting their data from both experiments. A higher correlation with human judgment has been achieved by computing an SSIM-like texture and structure similarity measure in VGG feature space rather than image space [Ding et al., 2020; Ding et al., 2021].

Visibility metrics (VMs) predict the distortion perceptibility for every pixel in the form of visibility maps. VMs are specifically tuned for detecting near-threshold distortions, which is required in many graphics and vision applications that cannot tolerate any perceivable quality reduction and require local information on the distortion positions. To decide on the visibility of such near-threshold distortions, models of human vision are often employed, where the most prominent FR-VMs examples include: VDM [Lubin, 1995], VDP [Daly, 1992], and HDR-VDP-2 [Mantiuk et al., 2011a]. By predicting how the human visual system responds to temporal changes as well as the visual field, FovVideoVDP [Mantiuk et al., 2021] extends the VM prediction for videos in addition to images. In the specific task of predicting selected rendering and compression artifacts, the best performance has been achieved using machine learning [Čadík et al., 2013] and CNN-based techniques [Wolski et al., 2018a; Patney and Lefohn, 2018]. Recently, Andersson et al. [2020] presented FLIP, a perceptual VM which focuses particularly on differences between rendered images and ground truths.

2.3.2 No-Reference Metrics

This thesis focuses on the VMs due to the locality of their prediction; however, it is interested in a more challenging *no-reference* setup, where the reference image is unavailable. This section discusses the most successful and recent NR-IQMs that rely on machine learning techniques, and more comprehensive metric discussions can be found in surveys Chandler [2013] and Kim et al. [2017a]. Early machine learning techniques employed predefined features such as SIFT and HOG [Narwaria and Lin, 2010; Moorthy and Bovik, 2010; Saad et al., 2012; Tang et al., 2011], and measured their distortions with respect to natural image statistics [Wang and Bovik, 2006]. Recently, CNN architectures have been applied to such feature learning as well as the MOS regression at the same time [Bianco et al., 2016; Kang et al., 2014; Bosse et al., 2018; Talebi and Milanfar, 2018]. To compensate for a low number of MOS-labeled images, such solutions typically rely on patches, where they assign the same MOS score for all patches that belong to a given image [Kim et al., 2017a]. CNN-based models often have a fixed-size input requirement and lack the ability to fully exploit information across all regions of an image. To address these limitations, the vision transformer, which has been successful in vision tasks [Dosovitskiy et al., 2020; Zamir et al., 2022], has been adopted for no-reference tasks [Yang et al., 2022; Ke et al., 2021; You and Korhonen, 2021] and has shown to be superior to CNN-based models by capturing more global feature. Such practice is justified for specific classes of distortions that affect the whole image uniformly, which might be the case for certain types of image noise or compression artifacts. Still, it might confuse the network in case of localized distortions such as those occurring in IBR.

To compensate for the lack of true local reference images, Bosse et al. [2018] learn the importance of local patches. Still, their key motivation is not deriving the localized VM but rather estimating relative patch weights in the aggregated MOS rating. The work done by Lin and Wang [2018] employs a quality-aware generative network to hallucinate the reference image in an adversarial learning setup, which is further refined by an IQM-discriminator that is trained on ground truth references. Their hallucination-guided quality regression network is fed with the difference between the hallucinated and distorted images, as well as the distorted image itself, to predict the MOS value. The quality-aware generative network, hallucination-guided quality regression network, and the IQM-discriminator are jointly optimized in an end-to-end manner. Kim and Lee [2017] apply state-of-the-art FR-IQMs such as SSIM to generate proxy scores on patches as the ground truth to pre-train the model and then fine-tune their target NR-IQM. At the intermediate stages, the regression network considers mean values and the standard deviations of per-patch 100-element feature vectors, which are then pooled to a per-image quality score. This thesis also employs state-of-the-art FR-IQMs to perform an initial per-patch distortion annotation and strikes the required balance between different error magnitudes in the training data, which is essential for meaningful training and shift-invariant properties of an NR-VM. The research on NR-VMs is extremely sparse, presumably due to limited access to locally labeled images [Herzog et al., 2012; Čadík et al., 2013; Wolski et al., 2018a]. A notable exception is the work of Herzog et al. [2012], who employs a support vector machine (SVM) to predict per-pixel distortions for selected rendering artifacts (they do not consider IBR) and achieve performance comparable to FR-VMs. Here, this thesis attempts to demonstrate that time-consuming manual per-pixel distortion labeling is not strictly required.

In cases where training data is easy to produce–such as uniform distortions like noise, JPEG, etc.–and no perceptual calibration is required, supervised training has been employed to detect aliasing artifacts [Patney and Lefohn, 2018]. In contrast, the method in this thesis (Chapter 5) deals with limited training data because only very few ground truth images are available for IBR and require perceptual calibration.

Vogels et al. [2018] have proposed a method to denoise path-traced images. To steer the amount of denoising, they also train a neural network to predict distortion in terms of MC variance, which is as unknown as the pixel value to be MC-estimated itself. Similar to their works, this thesis employs an NR metric to steer adaptation which is controlling capture hardware in this case (Chapter 5).

Their task is different as they predict SSIM error from a pair of images, where one is noisy and the other is denoised. This restricts the distortions to the difference between denoised and reference, which are smaller than IBR artifacts and also do not need to be perceptually calibrated. The fact that images with MC noise can be generated in arbitrary amounts also underlines what is the focus of this thesis: coping with limited training data.

2.3.3 Artifacts in Image-Based Rendering

Image-based rendering for structured or unstructured light fields (LFs) of real-world scenes involves several computational steps, such as depth reconstruction, neighboring view-image warping, warped view-image blending, and disocclusion hole in-painting. Each of these steps is prone to inaccuracies that manifest themselves as IBR-specific artifacts such as object shifting (incorrect depth), crumbling, distorted edges (depth discontinuities, e. g., due to compression), popping (fluctuations in depth), ghosting (depth inaccuracy, view blending), stretching, blurry or black regions (in-painting) [Tian et al., 2018]. Specialized IBR quality metrics often leave one view as the reference [Waechter et al., 2017; Conze et al., 2012; Solh et al., 2011; Bosc et al., 2011] or search for matching image blocks after their registration [Battisti et al., 2015; Gu et al., 2017], and then employ customized FR-IQMs. NR-IQMs typically focus on detecting selected distortion types such as blurring and ghosting [Berger et al., 2010], ghosting and popping [Guthe et al., 2016], blurring, stretching, and black holes [Tian et al., 2018], and aggregation into one final scalar score. Perceptual

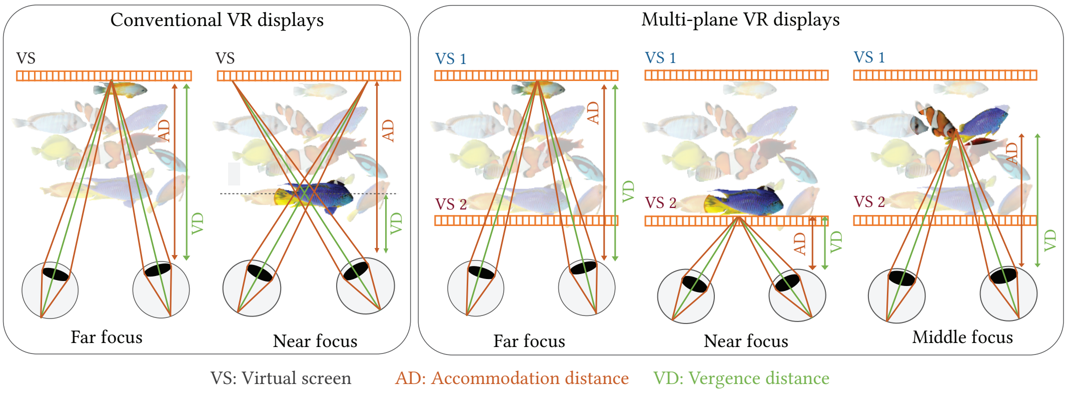


FIGURE 2.1: *When looking at objects in the real world, two mechanisms consistently occur: vergence and accommodation. Vergence refers to the binocular eye movement that directs both eyes on the target, while accommodation involves the adjustment of the eye lens to focus and produce clear images. In conventional VR displays (left), vergence and accommodation only match at the virtual screen plane, and when the user looks at objects at different depths, the accommodation cue is triggered incorrectly. Multi-plane displays (right) offer a solution to this issue. Studies [Hua, 2017] demonstrate that these displays are capable of projecting 3D images and triggering eye accommodation across display planes, enabling users to focus on objects at any depth seamlessly.*

experiments have been performed to understand how the observers rate the severity of different artifacts as a function of rendering parameters such as the number of blended views and viewing angles [Vangorp et al., 2011]. A skillful pre-processing of depth (e. g., depth blurring in uncertain regions) and choice of particular algorithmic solutions can substantially suppress artifacts [Hedman et al., 2016a; Serrano et al., 2019], eventually using a neural network trained to predict blending weights to combine the warped images [Hedman et al., 2018]. More objectionable distortion types can be traded-off with those more visually appealing (e. g., blurry depth that is more consistent but further from the ground truth). Instead of focusing on selected distortion types, Ling and Le Callet [2018] propose to learn a dictionary based on manually labeled data. The features extracted from an image allow predicting a MOS value using support vector machine regression. Data labeling can be time-consuming, as Ling et al. [2019] create artificial training data to simulate occlusion problems. A Generative Adversarial Network (GAN) discriminator [Goodfellow et al., 2014], targeted to identify in-painted image regions, is used to predict a quality score.

All the discussed work on IBR quality evaluation essentially focuses on providing a single score per image, which also serves as a metric for performance evaluation. While some FR-IQMs generate viable per-pixel VMs at intermediate stages [Conze et al., 2012; Solh et al., 2011], their accuracy is not formally evaluated. The same holds for the NR-IQM [Ling and Le Callet, 2018]. This thesis differs from all previous work by pursuing the NR-VM setup to detect local IBR distortions using CNN-based techniques.

2.4 Multi-Layered Displays

This section gives an overview of near-eye displays supporting accommodation cues (Section 2.4.1) and image decomposition algorithms targeted for such displays (Section 2.4.2), as well as selected aspects of foveated rendering (Section 2.4.3) that are central to this thesis.

2.4.1 Accommodative Displays

In order to support correct accommodation cues in near-eye displays, various methods have been proposed.

Multi-plane displays These display project images on different depth planes and form near-correct 3D volumetric images (Figure 2.1). The system architecture can be classified into two categories: systems based on time-multiplexing with switchable lenses [Love et al., 2009; Hu and Hua, 2014] or systems based on beam splitters and multiple physical screens [Akeley et al., 2004; MacKenzie et al., 2010]. Time-multiplexing systems can be designed in smaller form factors, but the requirement for high-refresh-rate screens and fast tunable-focus devices is a major obstacle. Although multi-screen systems have a major drawback in their large form factor, they offer a larger FOV than time-multiplexing systems. Another major obstacle of both architectures is the requirement for eye tracking since the images are generated for a fixed viewing position. Recently, focal surface displays have been developed to represent continuous 3D imagery [Matsuda et al., 2017]. They eliminated the need for eye tracking in the case of single-plane generation, but they are computationally demanding and based on expensive LCoS SLMs. Another approach to avoid eye-tracking is to perform per-region optimization at multiple gaze points, but it requires costly optimization and precise calibration of the eye rotation axis [Lee et al., 2017]. Since eye-tracking is essential in practical multi-plane system settings, this thesis exploits the eye-tracking system further to develop foveated rendering strategy.

Light-field displays The light-field display controls the 4D ray space of the light generated by the display to produce the motion parallax and vergence cues. Recently, light-field displays supporting focus cues have been proposed based on microlenses [Lanman and Luebke, 2013; Hua and Javidi, 2014]. However, those designs have an intrinsic trade-off between angular and spatial resolution. Light-field displays based on multi-layered architecture [Maimone and Fuchs, 2013; Huang et al., 2015a; Moon et al., 2017] have been demonstrated as an efficient platform for providing focus cues.

Other methods Holographic displays can project a replica of real-world scenes and provide accurate focus cues [Yeom et al., 2015]. However, the limited pixel size and resolution of digital wavefront modulators impose a significant trade-off between the eyebox size and FOV [Maimone et al., 2017]. Another approach is to change the depth of the 2D image plane dynamically with focus-tunable devices [Aksit et al., 2017; Dunn et al., 2017]. Although viewers can observe the images with correct accommodation cues, the requirement for a dynamic system may lead to latency issues. Instead of generating complete focus cues, the vergence-accommodation conflict also can be alleviated by projecting all-in-focus images [Konrad et al., 2017]. However, this method has a trade-off between the spatial resolution and the reproducible focus range. Recently, it is also demonstrated that proper rendering of chromatic aberration can effectively trigger accommodation without changing optical focus cues [Cholewiak et al., 2017].

In this thesis, the rendering strategy is mainly built on the principle of additive light-field displays with accommodation cues [Moon et al., 2017].

2.4.2 Decomposition Algorithms

Light-field displays In multi-layered light-field displays, the light fields are parameterized by a group of pixels on multiple layers. For multiplicative displays, the optimization system is described in tensor form and solved by various factorization algorithms [Wetzstein et al., 2011a; Huang et al., 2015b]. Additive light-field displays based on the polarization LCDs [Lanman et al., 2011] or incoherent summation of pixel intensities reflected from holographic optical elements [Lee et al., 2017] have also been proposed. For those architectures, LFS is formulated with a linear least-squares error problem and solved with the simultaneous algebraic reconstruction technique (SART) for online calculation [Andersen, 1984] or the trust-region method [Coleman and Li, 1996] for offline calculation. In LFS, the generation of target light fields requires high computational cost, and real-time performance is only possible by reducing the number of iterations [Lanman et al., 2011; Huang et al., 2015b]. To enhance the rendering speed, an adaptive sampling strategy was proposed [Heide et al., 2013], but the performance improvement was only demonstrated for offline rendering scenarios. This thesis saves the computational cost of generating the target light fields and decomposition through selective rendering and optimization.

Multi-plane displays In multi-plane displays, the linear blending rule assigns pixel values proportional to the distance between a target point and display planes [Akeley et al., 2004]. Although it can effectively trigger accommodation [MacKenzie et al., 2010], occlusion boundaries and non-Lambertian surfaces are imperfectly rendered in LB due to the simple consideration of a single image and depth map. In order to correctly generate artifact-free scenes, the retinal optimization (RO) [Narain et al., 2015; Mercier et al., 2017], which optimizes a focal stack, has been proposed. However, the target focal stack, in fact, implicitly contains the 4D light-field information [Levin and Durand, 2010]. Therefore, LFS that optimizes the 4D light fields also can be employed in multi-plane display architecture. This thesis in Chapter 6 is based on LFS since the implementations of current LFS algorithms are demonstrated to be more efficient than RO. It also revisits LFS in the context of gaze-contingent rendering to improve the perceived image quality and reduce computational costs.

2.4.3 Foveated Rendering

Gaze-contingent techniques have been used to improve image quality in various applications such as tone reproduction [Jacobs et al., 2015], depth-of-field modeling [Mauderer et al., 2014], disparity manipulation [Kellnhofer et al., 2016a], and viewing comfort improvement [Duchowski et al., 2014] in stereoscopic displays. Foveated rendering uses gaze information to improve rendering efficiency by reducing quality for the periphery. This is usually achieved by reducing the density of rendered image samples with increasing eccentricity [Guenter et al., 2012; Swafford et al., 2016; Patney et al., 2016]. In accommodative light field displays, Sun et al. [2017] propose a foveated rendering solution, which accounts for depth information and the current state of the accommodation to choose optimal ray directions in the OptiX renderer. In this thesis, the ray selection is dictated by choice of local decomposition technique for multi-plane displays and supported by an analysis of local luminance contrast and visibility of artifacts caused by the LB.

Chapter 3

Implicit View, Light, and Time Image Interpolation

This chapter introduces X-Field —a set of 2D images taken across different view, time or illumination conditions, i. e., video, light field, reflectance fields or combinations thereof—by learning a neural network (NN) to map their view, time or light coordinates to 2D images. Executing this NN at new coordinates results in joint view, time, or light interpolation. The key idea to make this workable is a NN that already knows the “basic tricks” of graphics (lighting, 3D projection, occlusion) in a hard-coded and differentiable form. The NN represents the input to that rendering as an implicit map that for any view, time, or light coordinate and for any pixel can quantify how it will move if view, time, or light coordinates change (Jacobian of pixel position with respect to view, time, illumination, etc.). The proposed X-Field representation is trained for one scene within minutes, leading to a compact set of trainable parameters and hence real-time navigation in view, time and illumination.

3.1 Introduction

Current and future sensors capture images of one scene from different points (video), from different angles (light fields), under varying illumination (reflectance fields), or subject to many other possible changes. In theory, this information will allow for exploring time, view, or light changes in Virtual Reality (VR). Regrettably, in practice, sampling this data densely leads to excessive storage, capture, and processing requirements. In higher dimensions—, here it is demonstrated as 5D—, the demands of dense regular sampling (cubature) increase exponentially. Alternatively, sparse and

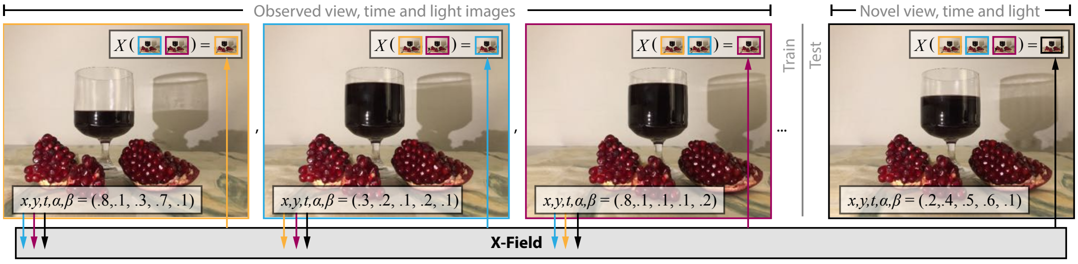

Observed view, time and light images
Novel view, time and light
$X(\ ,\ ) =$
$x,y,t,\alpha,\beta = (.8,.1, .3, .7, .1)$
$x,y,t,\alpha,\beta = (.3, .2, .1, .2, .1)$
$x,y,t,\alpha,\beta = (.8,.1, .1, .1, .2)$
$x,y,t,\alpha,\beta = (.2,.4, .5, .6, .1)$
Train
Test
X-Field

FIGURE 3.1: *This chapter presents a method to interpolate view, light, and time in a set of 2D images labeled with coordinates (X-Field) where a neural network (NN) is trained to regress each image from all others. The first (yellow) image is the NN output (yellow up arrow) when the blue and purple observed images and their coordinates x, y, t, α, β are input (yellow down arrows). The blue and purple observations form additional constraints, visualized as colored boxes. Provided with an unobserved coordinate (black up arrow) the NN produces, from the observed images and coordinates (black down arrow), a novel high-quality 2D image in real-time.*

irregular sampling overcomes these limitations but requires faithful interpolation across time, view, and light. This chapter suggests taking an abstract view of all those dimensions and simply denoting any set of images conditioned on parameters as an "X-Field", where X could stand for any combination of time, view, light, or other dimensions like the color spectrum. This chapter will demonstrate how the right neural network (NN) becomes a universal, compact, and interpolatable X-Field representation. While NNs have been suggested to estimate depth or correspondence across space, time, or light, this thesis, for the first time, suggests representing the complete X-Field implicitly [Niemeyer et al., 2019; Chen and Zhang, 2019; Oechsle et al., 2019; Sitzmann et al., 2019b], i. e., as a trainable architecture that implements a high-dimensional `getPixel`. The main idea of this chapter is shown in Figure 3.1: from sparse image observations with varying conditions and coordinates, a mapping is learned to take the space, time, or light coordinate and generate the observed sample image as an output. Importantly, when given a non-observed coordinate, the output is a faithfully interpolated image. The key to making this work is the right training and a suitable network structure involving a very primitive (but differentiable) rendering (projection and lighting) step. The proposed architecture is trained for one specific X-Field to generalize across its parameters, but not across scenes. However, per-scene training is fast (minutes), and decoding occurs at high frame rates (ca. 20 Hz) and high resolution (1024×1024). In a typical use case of the VR exploration of an X-Field, the architecture parameters only require a few additional kilobytes on top of the image samples. The results of the proposed method are compared to several other state-of-the-art interpolation baselines (NN and classic, specific to certain domains and general) as well as to ablations of the method itself.

3.2 Background

This chapter is motivated by two main observations: First, representing information using NNs leads to interpolation. Second, this property is retained if the network contains more useful layers, such as a differentiable rendering step. Both will be discussed next:

Deep representations help interpolation. It is well-known that deep representations suit interpolation of 2D images [Radford et al., 2015; Reed et al., 2015; White, 2016], audio [Engel et al., 2017], or 3D shape [Dosovitskiy et al., 2015] much better than the pixel basis. Consider the blue and orange bumps in Figure 3.2 (a); these are observed. They represent flat-land functions of appearance (vertical axis), depending on some abstract domain (horizontal axis) that later will become space, time, reflectance, etc., in an X-Field. The aim is to interpolate something similar to the unobserved violet bump in the middle. Linear interpolation in the pixel basis (solid lines) will fade both in, resulting in two flat copies. Visually this would be unappealing and distracting ghosting. This difference is also seen in the continuous setting of Figure 3.2 (b) that can be compared to the reference in Figure 3.2 (c). When representing the bumps as NNs to map coordinates to color (dotted lines), note that they are slightly worse than the pixel basis and might not match the NNs. However, the interpolated, unobserved result is much closer to the reference, and this is what matters in X-Field interpolation. To benefit from interpolation, typically, substantial effort is made to construct latent codes from images, such as auto-encoders [Hinton and Salakhutdinov, 2006], variational auto-encoders [Kingma and Welling, 2013], or adversarial networks [Goodfellow et al., 2014]. However, this step is not always

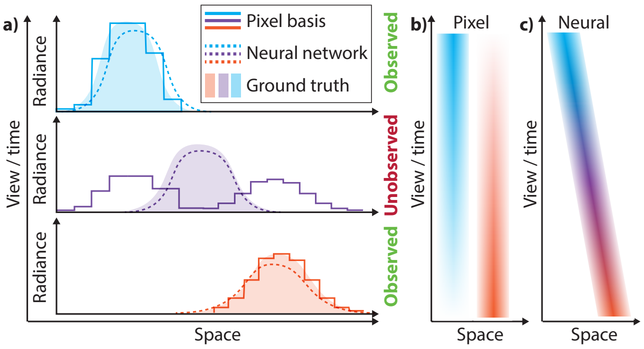


FIGURE 3.2: *NN and pixel interpolation: **a)** Flatland interpolation in the pixel **(lines)** and the NN representations **(dotted lines)** compared to a reference **(solid)** for a 1D field (vertical axis angle; horizontal axis space). The top and bottom are observed, and the middle is unobserved, i. e., interpolated. **b,c)** Comparing the continuous interpolation in the pixel and the NN representation visualized as a (generalized) epi-polar image. Note that the NN leads to smooth interpolation, while the pixel representation causes undesired fade-in/fade-out transitions.*

required in the common graphics task of image (generalized) interpolation. In the problem this chapter is dealing with, the latent space is given as beautifully laid-out space-time X-Field coordinates and only need to learn to decode these into images.

(Differentiable) rendering is just another non-linearity. The second key insight is that the above property holds for any architecture as long as all units are differentiable. In particular, this allows for a primitive form of rendering (projection, shading, and occlusion units). These units do not even have learnable parameters. Their purpose, instead, is to relieve the NN from learning basic concepts like occlusion, perspective, etc. Figure 3.2 shows interpolation of colors over space. Consider regression of appearance using a multi-layer perceptron (MLP) [Oechsle et al., 2019; Sitzmann et al., 2019b; Chen and Zhang, 2019] or convolutional neural network (CNN). CNNs without the Coord-Conv trick [Liu et al., 2018] are particularly weak at such spatially-conditioned generation. But even with Coord-Conv, this complex function is unnecessarily hard to find and slow to fit. In contrast, methods that sample the observations using warping [Jaderberg et al., 2015] are much more effective in changing the view [Zhou et al., 2016]. Figure 3.3 shows a validation experiment that compares classic pixel-basis interpolation and neural interpolation of color and warping. Using a NN provides smooth epipolar lines, and using warping, adds the details.

Motivated by those observations, this chapter first introduces the function that the proposed method aims to learn (Section 3.3), followed by the architecture for implementing it (Section 3.4).

3.3 Objective

The X-Field is represented as a non-linear function:

$$L_{\text{out}}^{(\theta)}(\mathbf{x}) \in \mathcal{X} \rightarrow \mathbb{R}^{3\times n_{\text{p}}},$$

with trainable parameters θ to map from an n_{d}-dimensional X-Field coordinate $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{n_{\text{d}}}$ to 2D RGB images with n_{p} pixels. The X-Field dimension depends on the capture modality: A 4D example would be two spatial coordinates, one temporal dimension, and one light angle. Parametrization can also be as simple as scalar 1D

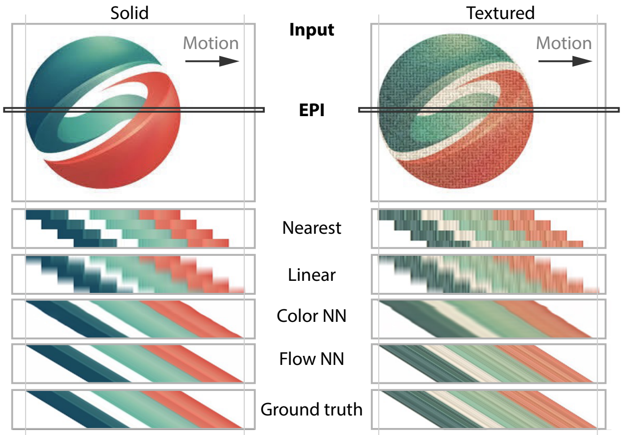


FIGURE 3.3: *Validation experiment: Different interpolation* ***(rows)****, for two variants* ***(columns)*** *of a right-moving SIGGRAPH Asia logo (a 1D X-Field). For each method, the same epipolar slice is shown (i. e., space on the horizontal axis; time on the vertical axis), which is marked in the input image. Nearest and linear sampling show either blur or step artifacts. A NN to interpolate solid color depending on time succeeds but lacks the capacity to reproduce textured details where the fine diagonal stripes are missing. A NN that instead interpolates flow captures the textured stripes.*

time for video interpolation. The symbol L_{out} is chosen as images are in units of radiance with a subscript to denote them as output.

The subset of *observed* X-Field coordinates is denoted as $\mathcal{Y} \subset \mathcal{X}$ for which an image $L_{\text{in}}(\mathbf{y})$ was captured at the known coordinate $\mathbf{y}$. Typically $|\mathcal{Y}|$ is sparse, i. e., small, like 3×3, 5×5 for view changes or even 2×1 for a stereo capture. This mapping L_{out} is found by optimizing for

$$\theta = \underset{\theta'}{\arg\min}\, \mathbb{E}_{\mathbf{y}\sim\mathcal{Y}} || L_{\text{out}}^{(\theta')}(\mathbf{y}) - L_{\text{in}}(\mathbf{y}) ||_1,$$

where $\mathbb{E}_{\mathbf{y}\sim\mathcal{Y}}$ is the expected value across all the discrete and sparse X-Field coordinates $\mathcal{Y}$. In prose, an architecture L_{out} is trained to map vectors $\mathbf{y}$ to captured images $L_{\text{in}}(\mathbf{y})$ in the hope of also getting plausible images $L_{\text{out}}(\mathbf{x})$ for unobserved vectors $\mathbf{x}$. It is targeted for interpolation; $\mathcal{X}$ is a convex combination of $\mathcal{Y}$ and does not extend beyond. Note that training never evaluates any X-Field coordinate $\mathbf{x}$ that is not in $\mathcal{Y}$, as the image $L_{\text{in}}(\mathbf{x})$ at that coordinate is not available.

3.4 Architecture

The L_{out} is designed using three main ideas. First, appearance is a acombination of appearance in observed images. Second, appearance is assumed to be a product of shading and albedo. Third, it is assumed that the unobserved shading and albedo at $\mathbf{x}$ are a warped version of the observed shading and albedo at $\mathbf{y}$. These assumptions do not strictly need to hold, in particular not for splitting albedo and shading: when they are not fulfilled, the NN just has a harder time capturing the relationship of coordinates and images. The pipeline of proposed approach L_{out}, depicted in Figure 3.4, implements this in four steps: decoupling shading and albedo (Section 3.4.1), interpolating images (Section 3.4.2) as a weighted combination of warped images (Section 3.4.3), representing flow using a NN (Section 3.4.4) and resolving inconsistencies (Section 3.4.5).

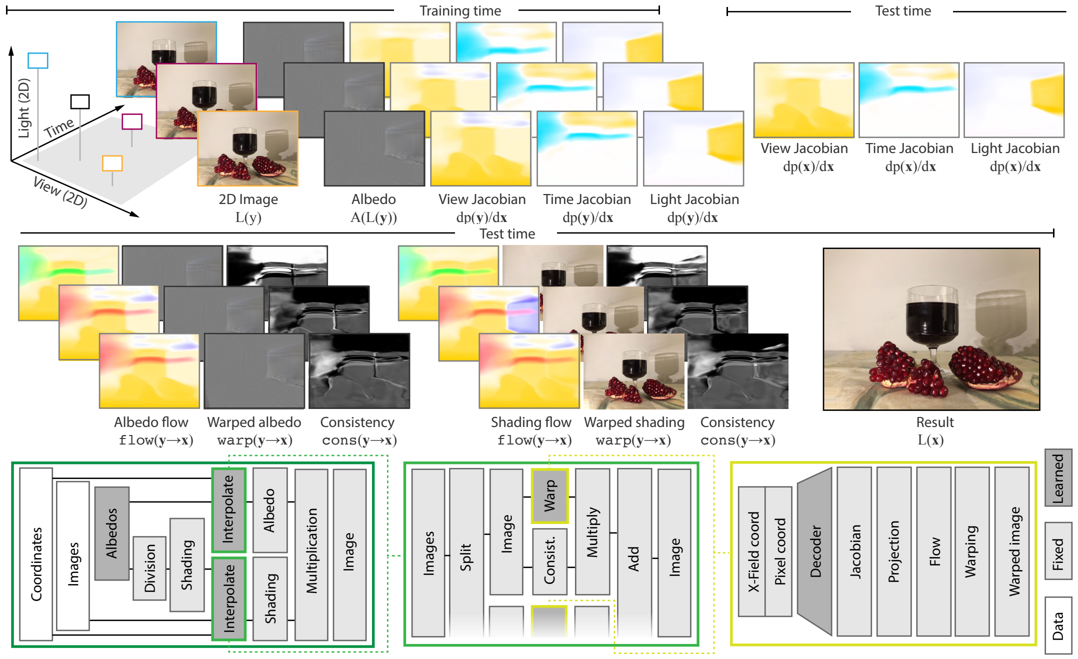


FIGURE 3.4: *Data flow for an example with three dimensions (one view, one light, one temporal) and three samples, denoted as colors, as in Figure 3.1 and stacked vertically in each column. In the first row, the 2×3 Jacobian matrix is always visualized as separate channels i. e., as three columns with two dimensions each. Values are 2D vectors, hence visualized as false colors. At test time, the Jacobians are evaluated at the output X-Field coordinate only; hence, only a single row is shown. In the second row, each observation is separately warped for shading and albedo, leading to 2×3 flow, result, and weight images. The last row shows the flow of information as a diagram.* Learned *is a tunable,* Fixed *a non-tunable step (i. e., without learnable parameters).* Data *denotes access to inputs.*

3.4.1 De-Light

De-lighting splits appearance into a combination of shading, which moves in one way in response to changes in X-Field coordinates, e. g., highlights move in response to view changes or shadows move with respect to light changes, and albedo, which is attached to the surface and will move with geometry, i. e., textures. To this end, every observed image is decomposed as $L_{\text{in}}(\mathbf{y}) = E(\mathbf{y}) \odot A(\mathbf{y})$, a per-pixel (Hadamard) product $\odot$ of a shading image E and an albedo image A. This is done by adding one parameter to θ for every observed pixel channel in E, and computing A from L_{in} by division as $E(\mathbf{y}) = L_{\text{in}}(\mathbf{y}) \odot A(\mathbf{y})^{-1}$. Both shading and albedo are interpolated independently:

$$L_{\text{out}}(\mathbf{x}) = \mathtt{int}(A(L_{\text{in}}(\mathbf{y})), \mathbf{y} \to \mathbf{x}) \odot \mathtt{int}(E(L_{\text{in}}(\mathbf{y})), \mathbf{y} \to \mathbf{x}) \qquad (3.1)$$

and recombined into new radiance at an unobserved location $\mathbf{x}$ by multiplication. The next part details the operator `int`, working the same way on both shading $E(L_{\text{in}})$ and albedo $A(L_{\text{in}})$.

3.4.2 Interpolation

Interpolation warps all observed images and merges the individual results. Both warp and merge are performed completely identically for shading E and albedo A,

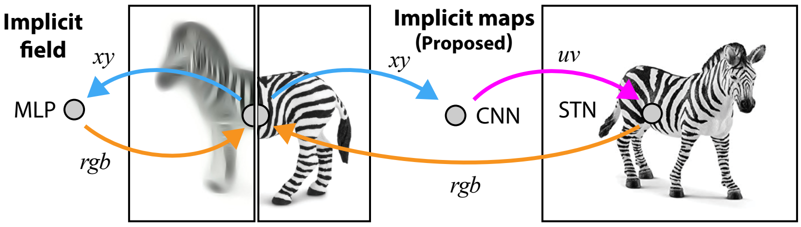


FIGURE 3.5: *Implicit maps: implicit fields **(left)** typically use an MLP to map 3D position to color, occupancy etc. The proposed method **(right)** adds an indirection and maps pixel position to texture coordinates to look up another image.*

which is neutrally denoted as I, as in:

$$\texttt{int}(I, \mathbf{y} \rightarrow \mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} (\texttt{cons}(\mathbf{y} \rightarrow \mathbf{x}) \odot \texttt{warp}(I(\mathbf{y}), \mathbf{y} \rightarrow \mathbf{x}))\,. \tag{3.2}$$

The result is a weighted combination of deformed images. Warping (Section 3.4.3) models how an image changes when X-Field coordinates change by deforming it, and a per-pixel weight is given to this result to handle flow consistency (Section 3.4.5).

3.4.3 Warping

Warping deforms an observed into an unobserved image, conditioned on the observed and the unobserved X-Field coordinates:

$$\texttt{warp}(I, \mathbf{y} \rightarrow \mathbf{x}) \in \mathcal{I} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{I}. \tag{3.3}$$

which utilizes a spatial transformer (STN) [Jaderberg et al., 2015] with bi-linear filtering, i. e., a component that computes all pixels in one image by reading them from another image according to a given flow map. STNs are differentiable, do not have any learnable parameters, and are efficient in executing at test time. The key question is, (Figure 3.5) from which position $\mathbf{q}$ should a pixel at position $\mathbf{p}$ read when the image at $\mathbf{x}$ is reconstructed from the one at $\mathbf{y}$?

To answer this question, let's look at the Jacobians of the mapping from X-Field coordinates to pixel positions. Here, Jacobians capture, for example, how a pixel moves in a certain view and light if time is changed, or its motion for one light, time and view coordinate if the light is moved, and so forth. Formally, for a specific pixel $\mathbf{p}$, the Jacobian is:

$$\texttt{flow}_{\partial}(\mathbf{x})[\mathbf{p}] = \frac{\partial \mathbf{p}(\mathbf{x})}{\partial \mathbf{x}} \in \mathcal{X} \rightarrow \mathbb{R}^{2 \times n_\mathrm{d}}, \tag{3.4}$$

where $[\cdot]$ denotes indexing into a discrete pixel array. This is a Jacobian matrix with size $2 \times n_\mathrm{d}$, which holds all partial derivatives of the two image pixel coordinate dimensions (horizontal and vertical) with respect to all n_d-dimensional X-Field coordinates. A Jacobian is only differential and does not yet define the finite position $\mathbf{q}$ to read for at a pixel position $\mathbf{p}$ as required by the STN. In order to find $\mathbf{q}$, the change in X-Field coordinate $\mathbf{y} \rightarrow \mathbf{x}$ is *projected* to 2D pixel motion using finite differences:

$$\texttt{flow}_{\Delta}(\mathbf{y} \rightarrow \mathbf{x})[\mathbf{p}] = \mathbf{p} + \Delta(\mathbf{y} \rightarrow \mathbf{x})\texttt{flow}_{\partial}(\mathbf{x})[\mathbf{p}] = \mathbf{q}. \tag{3.5}$$

Here, the finite delta in X-Field coordinates $(\mathbf{y} \rightarrow \mathbf{x})$, an n_d-dimensional vector, is multiplied with an $n_\mathrm{d} \times 2$ matrix, and added to the start position $\mathbf{p}$, producing an absolute pixel position $\mathbf{q}$ used by the STN to perform the warp. In other words, Equation (3.4) specifies how pixels move for an infinitesimal change of X-Field coordinates, while Equation (3.5) gives a finite pixel motion for a finite change of X-Field coordinates. In the next part, the learned representation of the Jacobian $\texttt{flow}_\partial$ is discussed, which is the core of the proposed approach.

3.4.4 Flow

Input to the flow computation is only the X-Field coordinate $\mathbf{x}$, and output is the Jacobian (Equation (3.4)). This function is implemented particularly using a CNN.

Implementation The implementation starts with a fully connected operation that transforms the coordinate $\mathbf{x}$ into a 2×2 image with 128 channels. The Coord-Conv [Liu et al., 2018] information (the complete $\mathbf{x}$ at every pixel) is added at that stage. This is followed by as many steps as it takes to arrive at the output resolution, reducing the number of channels to produce at n_d output channels. For some input, it can be acceptable to produce a flow map at a resolution lower than the image resolution and warp high-resolution images using low-resolution flow, which preserves details in color, but not in motion.

Compression Changes in some X-Field dimension can only change the pixel coordinates in a limited way. One example is the view: all changes of pixel motion with respect to known camera motion can be explained by disparity [Forsyth and Ponce, 2002]. So instead of modeling a full 2D motion to depend on all view parameters, only a per-pixel disparity is generated, and the flow Jacobian is computed from disparity in closed form using reprojection. For the dataset used in this chapter, this is only applicable to view changes, as no such constraints are in place for derivatives of time or light.

Discussion It should also be noted that no pixel-basis RGB observation $L_\mathrm{in}(\mathbf{y})$ ever is input to $\texttt{flow}_\partial$, and hence, all geometric structure is encoded in the network. Recalling Section 3.2, this can be regarded as both a burden and a requirement to achieve the desired interpolation property: if the geometry NN can explain the observations at a few $\mathbf{y}$, it can explain their interpolation at all $\mathbf{x}$. This also justifies why $\texttt{flow}_\partial$ is a NN, and directly learning a pixel-basis depth-motion map is not necessary; otherwise, it would not be interpolatable. An apparent alternative would be to learn $\texttt{flow}'_\partial(\mathbf{x},\mathbf{y})$ to depend on both $\mathbf{y}$ and $\mathbf{x}$, so as not to use a Jacobian but allow any mapping. Regrettably, this does not result in interpolation. Consider a 1D view alone: Using $\texttt{flow}_\partial(\mathbf{x})$ has to commit to one value that just minimizes image error after soft blending. If a hypothetical $\texttt{flow}'_\partial(\mathbf{x},\mathbf{y})$ can pick any different value for every pair $\mathbf{x}$ and $\mathbf{y}$, it will do so without incentive for a solution that is valid in between them. Finally, it should be noted that using skip connections is not applicable to the proposed setting, as the decoder input is a mere three numbers without any spatial meaning.

3.4.5 Consistency

To combine all observed images that are warped to the unobserved X-Field coordinate, each pixel in an image is weighted by its *flow consistency*. For a pixel $\mathbf{q}$ to contribute to an image at pixel $\mathbf{p}$, the flow at $\mathbf{q}$ has to map back to $\mathbf{p}$. If not, evidence for not being an occlusion is missing, and the pixel needs to be weighted down. Formally, consistency of one pixel $\mathbf{p}$ when warped to coordinate $\mathbf{x}$ from $\mathbf{y}$ is the partition of unity of a weight function:

$$\texttt{cons}(\mathbf{y} \rightarrow \mathbf{x})[\mathbf{p}] = w(\mathbf{y} \rightarrow \mathbf{x})[\mathbf{p}](\sum_{\mathbf{y}' \in \mathcal{Y}} w(\mathbf{y}' \rightarrow \mathbf{x})[\mathbf{p}])^{-1}. \tag{3.6}$$

The weights w are smoothly decreasing functions of the 1-norm of the delta of the pixel position $\mathbf{p}$ and the backward flow at the position $\mathbf{q}$ where $\mathbf{p}$ was warped to:

$$w(\mathbf{y} \rightarrow \mathbf{x})[\mathbf{p}] = \exp(-\sigma|\mathbf{p} - \texttt{flow}_\Delta(\mathbf{x} \rightarrow \mathbf{y})[\mathbf{q}])|_1). \tag{3.7}$$

Here $\sigma = 10$ is a bandwidth parameter chosen manually. No benefit was observed when making σ a vector or learning it.

Discussion In other work, consistency has been used in a loss, asking for consistent flow for unsupervised depth [Godard et al., 2017; Zhou et al., 2017] and motion [Zou et al., 2018] estimation. The proposed approach does not have consistency in the loss during training but inserts it into the image compositing of the architecture, i. e., also to be applied at test time. In other approaches—that aim to produce depth, not images—consistency is not used at test time. The predicted flow can be inconsistent: for very sparse images such as three views, many occlusions occur, leading to inconsistencies. Also, flow due to, e. g., caustics or shadows, probably has a fundamentally different structure compared to multi-view flow, which has not been explored in the literature. The graphics question answered here is, however, what to do with inconsistencies. To this end, instead of a consistency loss, the proposed architecture applies multiple flows such that the combined result is plausible when weighing down inconsistencies. In the worst case, no flow is consistent with any other, and w has similar but small values for large `cons`, which lead to equal weights after normalization, i. e., linear blending.

3.5 Results

This section provides a comparison of the method to other works (Section 3.5.1), an evaluation of scalability (Section 3.5.2), and a discussion of applications (Section 3.5.3).

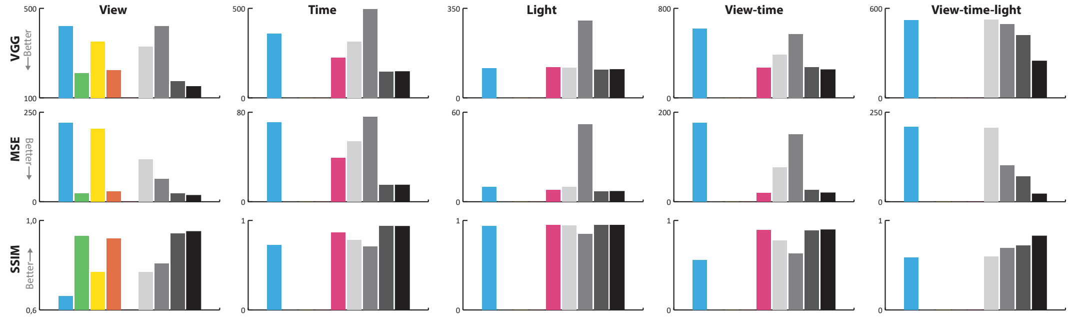
3.5.1 Comparison

The method is compared to other *methods*, following a specific *protocol* and by different *metrics* to be explained now:

Methods The following methods are considered: PROPOSED, BLENDING, WARPING, KALANTARIETAL, Local light-field fusion (LLFF), SUPERSLOWMO, and three ablations of the method: NOCORDCONV, as well as NOWARPING and NOCONSISTENCY. Linear BLENDING is not a serious method, but documents the sparsity: plagued by ghosting for small baselines, as the baseline/sparsity poses a difficult interpolation task, far from linear. It is applicable to all dimensions. WARPING and SUPERSLOWMO first estimate the correspondence in image pairs [Sun et al., 2018a] or light field data [Dabała et al., 2016] and later apply warping [Mark et al., 1997] with ULR-style weights [Buehler et al., 2001]. Note how ULR weighting accounts for occlusion. Warping is applicable to time ([Jiang et al., 2018]) and view interpolation ([Dabała et al., 2016]). KALANTARIETAL and LLFF are the publicly available implementations

TABLE 3.1: *Results of different methods* ***(columns)*** *for different dimensions* ***(rows)*** *according to different metrics. Below is the same data as the diagrams. Colors encode methods. The best method according to one metric for one class of X-Field is denoted in bold font (for L_2 and VGG, less is better, while for SSIM, more is better).* [1]*For view-time interpolation, combined with LLFF.*

View	Time	Light	● LINEAR			● WARPING			● KALANTARI			● LLFF			● SUSLOMO[1]			● NOWARP			● NOCC			● NOCONS			● PROPOSED		
			VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM
✓			421	221	.662	210	2.28	0.929	351	20.39	.769	223	2.78	.919	—	—	—	330	11.78	.768	421	6.45	.806	175	2.25	.941	**151**	**1.79**	**.951**
	✓		359	71	.723	—	—	—	—	—	—	—	—	—	224	3.90	.867	315	5.43	.778	497	7.63	.706	147	1.45	.935	**147**	**1.46**	**.935**
		✓	116	9	.940	—	—	—	—	—	—	—	—	—	120	.784	.947	119	0.95	.941	302	5.25	.848	111	0.68	.948	**111**	**0.66**	**.948**
✓	✓		620	176	.558	—	—	—	—	—	—	—	—	—	269	1.99	.892	388	7.67	.775	571	14.97	.632	273	2.61	.888	**252**	**2.00**	**.896**
✓	✓	✓	522	209	.584	—	—	—	—	—	—	—	—	—	—	—	—	523	20.60	.595	493	10.10	.692	419	7.09	.719	**247**	**2.19**	**.827**



of Kalantari et al. [2016] and Mildenhall et al. [2019]. Both are applicable to and tested on light fields, i. e., view interpolation only. To evaluate other works in higher dimensions, a hypothetical combination is introduced, such as first using LLFF for view interpolation followed by SUPERSLOWMO for time interpolation. Finally, three ablations of the method are presented. The first, NOCORDCONV, regresses without Coord-Conv, i. e., will produce spatially invariant fields. The second, NOWARPING, uses direct regression of color values without warping. The third, NOCONSISTENCY, does not perform occlusion reasoning but averages directly. These are applicable to all dimensions.

Protocol Success is quantified as the expected ability of a method to predict a set of held-out LF observed coordinates $\mathcal{H}$ when trained on $\mathcal{Y} - \mathcal{H}$, i. e., $\mathbb{E}_{\mathbf{h}\sim\mathcal{H}} L_{\text{out}}(\mathbf{h}) \ominus_{\text{m}} L_{\text{in}}(\mathbf{h})$, where $\ominus_{\text{m}}$ is one of the metrics to be defined below. For dense LF, the held-out protocol follows Kalantari et al. [2016]: four corner views as an input. Sparse LF interpolation is on 5×5, holding out the center one. For time interpolation, a triplet is used, i. e., the network is trained on past and future frames, while the middle one is withheld.

Metrics The L_2, SSIM, and VGG [Zhang et al., 2018] metrics are utilized for comparing the predicted to the held-out view.

Data The evaluation set consists of the publicly available LF data from Levoy and Hanrahan [1996], Penner and Zhang [2017a], Dabała et al. [2016], and Kalantari et al. [2016], LF video data from Sabater et al. [2017], sequences from Butler et al. [2012], relighting data from Xu et al. [2018] as well as custom captured reflectance field video. For aggregate statistics, five LFs, three videos, and one view-time-light X-Field are used. New X-Fields data are also proposed by this thesis which is captured using a minimalist setup: a pair of mobile phones. The first phone takes the photo; the second one provides the light source. Both are moved with one, two, or three degrees of

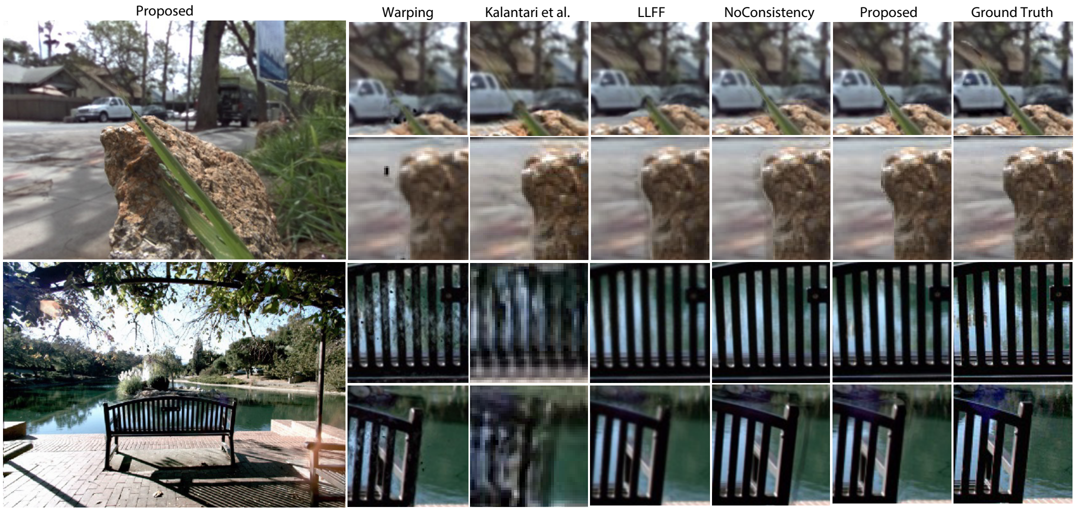


FIGURE 3.6: *Comparison of the proposed approach for view interpolation to other methods for two scenes **(rows)**. The top scene, from Kalantari et al. [2016] is a dense LF; the one below, from Penner and Zhang [2017a], is sparse. Columns show, left to right,* PROPOSED *at the position of the withheld reference, the results from* (WARPING, KALANTARIETAL, LLFF, *and* NOCONSISTENCY *and* PROPOSED), *as well as the ground truth as insets.*

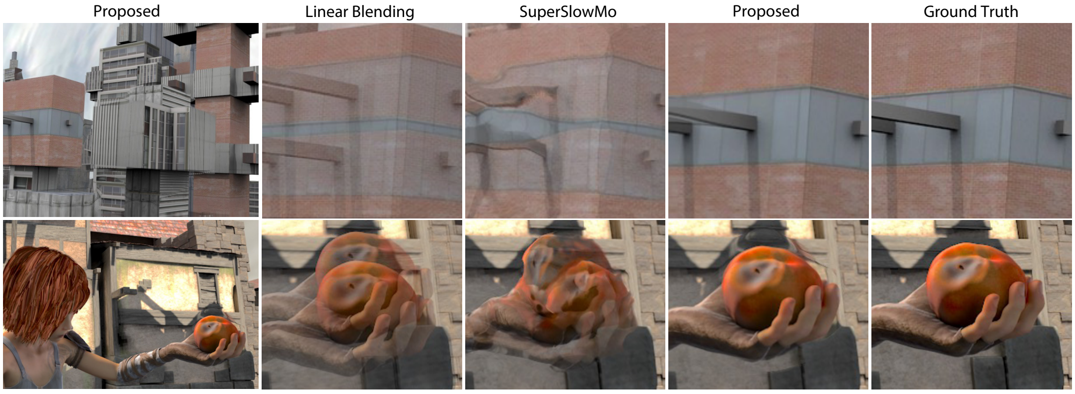


FIGURE 3.7: *Temporal interpolation for two scenes **(rows)** using different methods **(columns)**. See Section 3.5.1 for a discussion.*

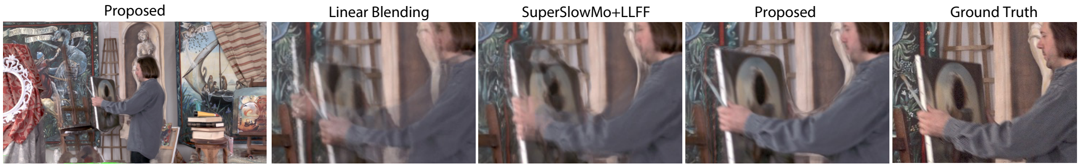


FIGURE 3.8: *Results for view-time interpolation. The input was a 2×2×2 X-Field: 2×2 sparse view observations with two frames.*

freedom, depending on the scene. All animation is produced by stop motion. Several X-Fields are captured, but only one has additional reference views to quantify quality.

Results Table 3.1 summarizes the outcome of the main comparison. It can be seen that the proposed method provides the best quality in all tasks according to all metrics in all domains. For example, images corresponding to the plots in Table 3.1, please see Figure 3.6 for interpolation in space, Figure 3.7 and Figure 3.9 for time, Figure 3.8 for space-time, Figure 3.10 and Figure 3.11 for light and Figure 3.12 for view-time-light results. Each figure shows the input view and multiple insets that visualize the results from all competing methods. Figure 3.6 shows results for view interpolation. Here,

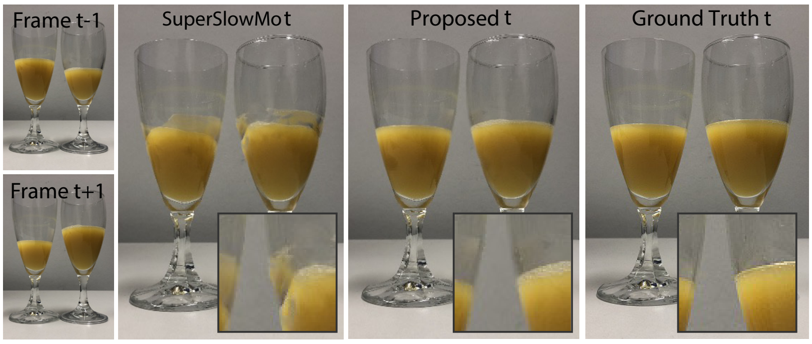


FIGURE 3.9: *Interpolation of two frames **(shown left)** compared to a reference using the proposed approach and state-of-the-art SuperSlowMo [Jiang et al., 2018].*

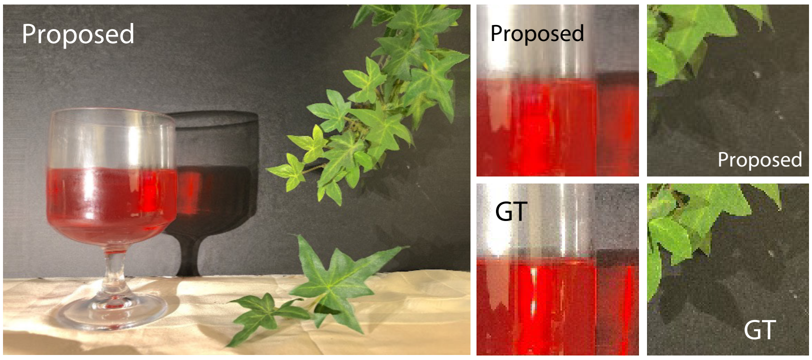


FIGURE 3.10: *Interpolation in the light dimension. Note that the interpolated image is plausible, even in the presence of cast shadows or caustics and transparency, maybe at the slight expense of blurring highlights and ghosting shadows.*

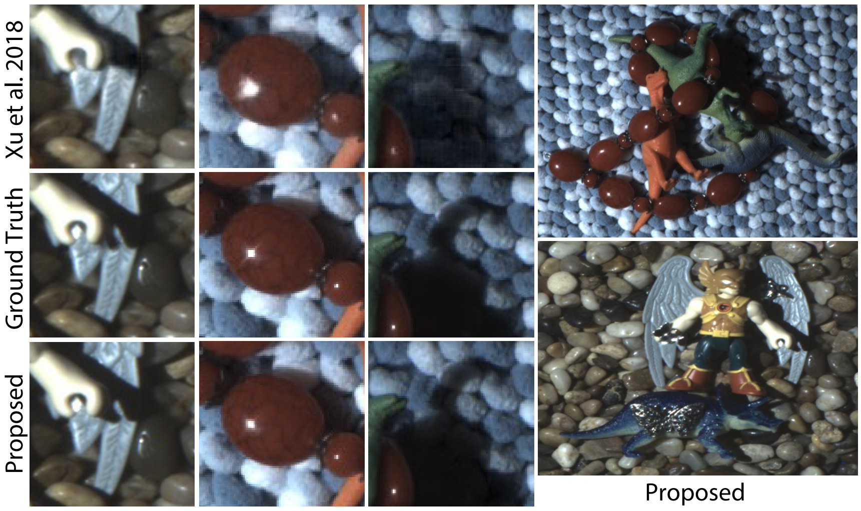


FIGURE 3.11: *Comparison between Xu et al. [2018] **(top)**, the GT **(middle)** and the proposed approach **(bottom)** for a 10 degree baseline.*

WARPING produces crisp images but pixel-level outliers that are distracting in motion, e. g., for the bench. KALANTARIETAL and LLFF do not capture the tip of the grass (top row). Instead, ghosted copies are observed. KALANTARIETAL is not supposed to work for larger baselines [Kalantari et al., 2016] and is only shown for completeness on the bench scene. LLFF produces slightly blurrier results for the sparse bench scene. The NOCONSISTENCY shows the tip of the grass but on top of ghosting. PROPOSED has details and plausible motion and is generally most similar to the ground truth. The temporal interpolation comparison in Figure 3.7 indicates similar conclusions: BLENDING is not a usable option; not handling occlusion, also in time,

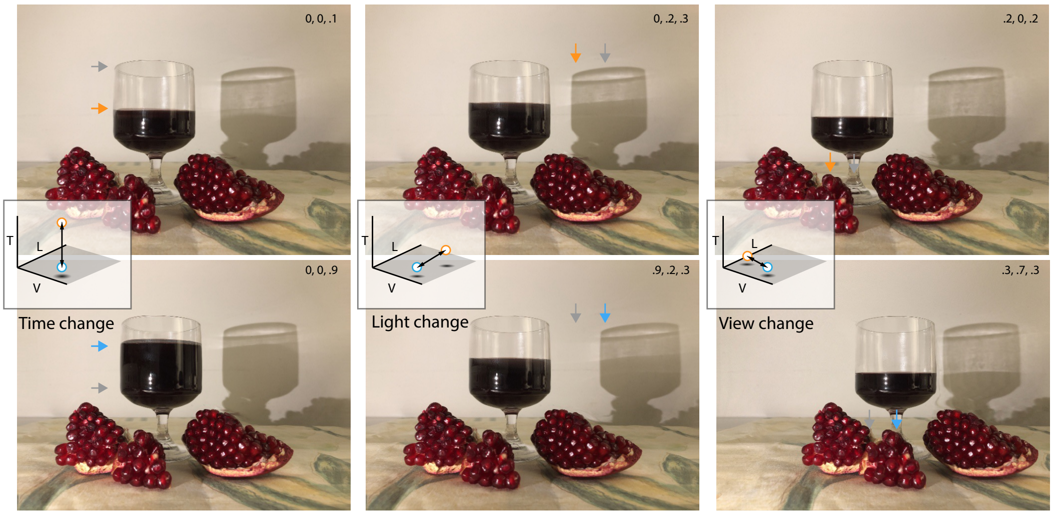


FIGURE 3.12: *Exploring a view-time-light X-Field. Each column shows a change in the dominant X-Field dimension. The input was a 3×3×3 X-Field. All images are at unobserved intermediate coordinates. Colored arrows indicate how image features have moved in response.*

TABLE 3.2: *Relighting comparison to Xu et al. [2018] for different baselines.*

Method	20° Baseline			30° Baseline			45° Baseline		
	VGG	MSE	SSIM	VGG	MSE	SSIM	VGG	MSE	SSIM
○ XUETAL	192	.1424	.954	194	.1561	.950	196	.1580	.950
● PROPOSED	93	.0335	.989	134	.0718	.970	169	.1220	.958

creates ghosting due to overlap. SUPERSLOWMO fails for both scenes as the motion is large. The motion size can be seen from the linear blending. Ultimately, PROPOSED is similar to the ground truth. Interpolation between triplets of images can represent strong, non-rigid changes involving transparency, scattering, etc (Figure 3.9). When interpolating across view and time as in Figure 3.8, ghosting effects get stronger for others as images get increasingly different. Interpolation across light is seen in Figure 3.10. For light interpolation, the method of Xu et al. [2018] is an extension of the ablation DIRECT by an additional optimization over sample placement when assuming a capture dome. The results of the proposed method are compared to other methods when trained on their data. Please note that their method cannot be applied to the proposed data used as it requires a custom capture setup. Figure 3.11 shows a comparison from interpolating across a neighborhood of 3×3 images out of the 541 dome images, covering a baseline of approximately 20 degrees. The direct regression blurs both the shadows and the highlights, while the method deforms the image, retaining sharpness. Table 3.2 quantifies this result as the average across their test images "Dinosaur", "Jewel" and "Angel". Besides the 10-degree column corresponding to Figure 3.11, other baselines are also included. It can be observed that for wider baselines [Xu et al., 2018], both methods converge in quality. PROPOSED can have difficulties where deformations are not fully rigid, as seen for faces, but compensates for this to produce plausible images. The proposed approach can both numerically and visually produce state-of-the-art interpolation in view and time in high spatial resolution and at high frame rates. Next, the dependency of this success on different factors is explored.

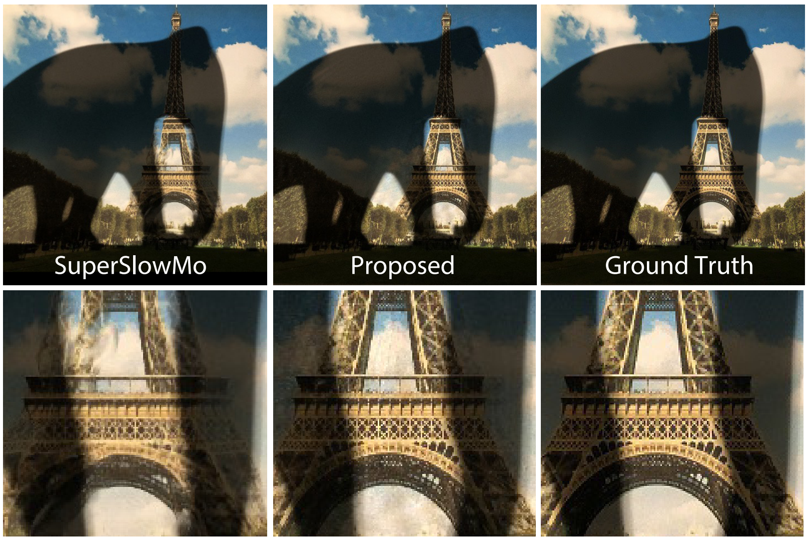


FIGURE 3.13: *Splitting albedo and shading: When the elephant's shadow meets a texture of the Eiffel Tower unprepared, a single-layer method such as* SUPERSLOWMO *cannot find a unique flow and produces artifacts. the proposed method leaves both shadow and texture structures mostly intact.*

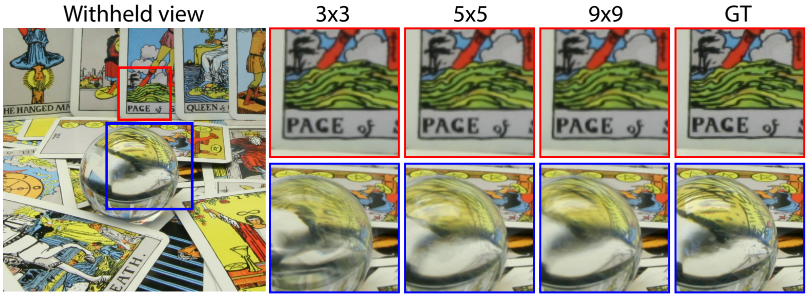


FIGURE 3.14: *Visual quality of the proposed approach as a function of increasing* ***(left to right)*** *training set size for view interpolation.*

3.5.2 Evaluation

The approach is evaluated in terms of scalability with training effort and observation sparsity, speed, and detail reproduction. These tests are performed on the view interpolation only.

Analysis of albedo splitting Figure 3.13 shows an example of a scene that benefits from albedo splitting for a light interpolation. It shows that splitting albedo and shading is critical for shadows cast on textured surfaces.

Observation sparsity The proposed method is also evaluated when interpolating extremely sparse data. Table 3.3 shows the interpolation quality of the method based on the number of training exemplars, also seen in Figure 3.14.

TABLE 3.3: *Error for the* Crystal Ball *scene with resolution 512×512 using different metrics* ***(columns)*** *for different view counts* ***(rows).***

LF	VGG19	L2	SSIM
3×3	140	.005	.90
5×5	119	.003	.93
9×9	102	.002	.95

Speed At deployment, the proposed design requires no more than taking a couple of numbers and passing them through a decoder for each observation, followed by warping and weighting. The end speed for view navigation is

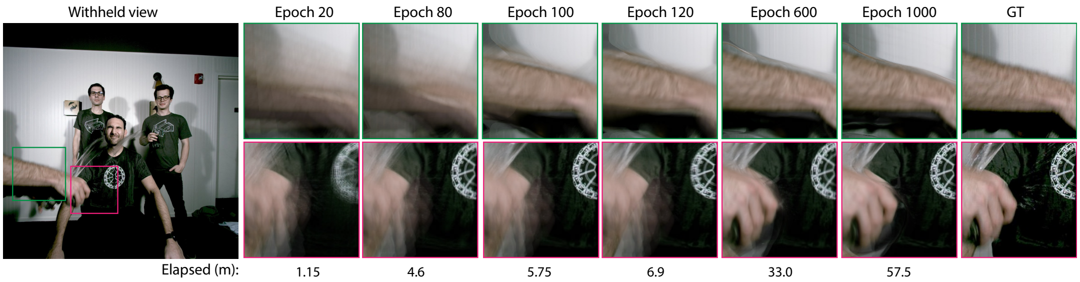


FIGURE 3.15: *Progression of visual fidelity for different training efforts (horizontal axis) for two insets (vertical axis) in one scene. After 500 epochs (ca. 30 minutes), the result is usable, and it converges after 1000 epochs (ca. 1h). Note that epochs are short as the training data is an LF array with a size of 5×5.*

around 20 Hz (on average 46 ms per frame) at 1024×1024 for a 5×5 LF on an Nvidia 1080Ti with 12 GB RAM.

Training effort The proposed approach needs to be trained again for every LF. Typical training time is listed in Table 3.4. Figure 3.15 shows the progression of interpolation quality over learning time. It can be seen that even after little training, results can be acceptable. Overall, in the proposed approach, training of the NN requires a workable amount of time compared to approaches trained in the order of many hours or days.

TABLE 3.4: *Training time (minutes) and network parameters for different resolutions for a 5×5 LF array and spatial interpolation.*

	512^2	1024^2	1764^2
Time	28	60	172
Params	482 k	492 k	492 k

Smoothness The depth and flow map produced by the method are smooth in view and time and may lack detail. It would be easy to add skip connections to get the details from the appearance. Regrettably, this would only work on the input image, which needs to be withheld at training and is unknown at test time. An example of this is seen in Figure 3.16. This smoothness is one source of artifacts. Overcoming this, e. g., using an adversarial design, is left to future work.

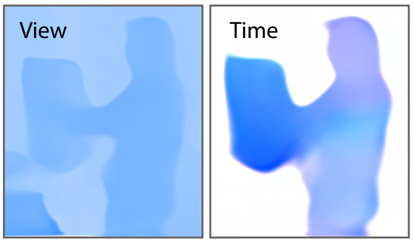


FIGURE 3.16: *Correspondence depth and flow maps for Figure 3.8.*

Coherence The method might miss details or be over-smooth, but it is coherent, as first, it does not regress colors that flicker, only texture coordinates; second, Jacobians are multiplied with view differentials in a linear operation, and hence smooth; third, as the NNs used to produce Jacobians are smooth functions and, finally, soft occlusion is smooth.

3.5.3 Applications

Figure 3.17 demonstrates motion blur (time interpolation), depth-of-field (view interpolation), and both (interpolating both).

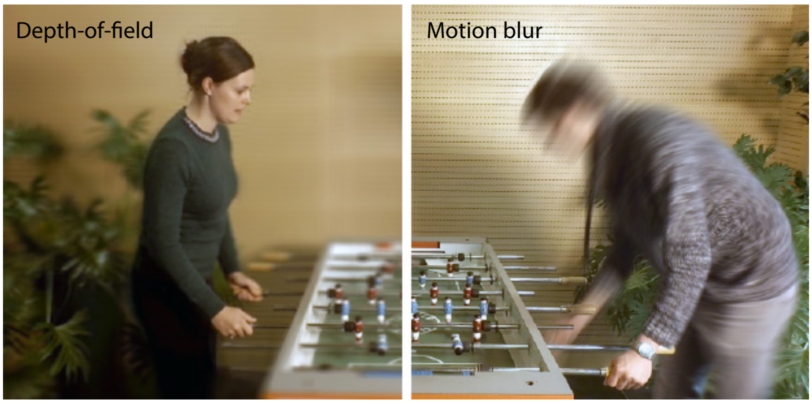


FIGURE 3.17: *Two LF video-enabled effects, computed using view interpolation: Depth-of-field* ***(left)*** *and motion blur* ***(right)****. For both, many images at X-Field coordinates are generated and averaged to cover a lens resp. shutter.*

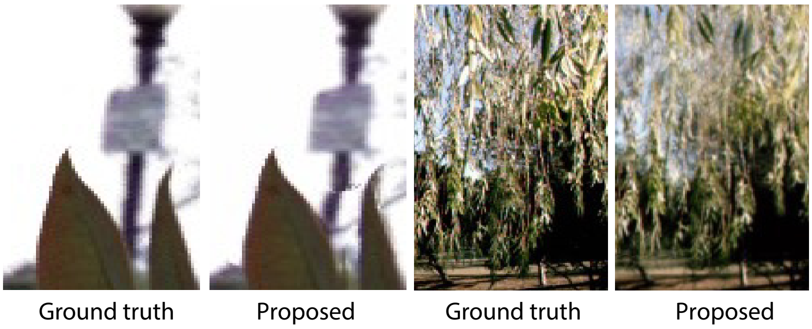


FIGURE 3.18: *Two failure cases of the proposed method, documenting, left, insufficient data (the lamp post is only visible in one view and happens to become attached to the foreground leaf) and right, insufficiency of the capacity (the depth structure of the twigs is too complex to be represented by the proposed architecture).*

3.6 Discussion and Limitations

The success of the proposed method largely depends on three factors, Data, Model, and Capacity, which will be discussed in the next part.

Data The method is trained using very sparse observations, often only a dozen images. It is clear that information not present in any image will not be reconstructed. Even parts observed in only one image can be problematic (Figure 3.18, left). A classic example is an occlusion: if only three different views are available and two occlude an area that is not occluded in one view, this area will be filled in. However, this fill-in will occur in X-Field Jacobian domain. Hence, disoccluded pixels will change their position similar to their spatial neighbors. Artifacts manifest as rubber-like stretches between the disoccluding and the occluding object. The chair example from Figure 3.19 shows artifacts resulting from a lack of data. Similarly, the foam in Figure 3.19 is stochastic and different in every image, and hence unable to form fine-scale correspondence. The consistency weighting typically removes them. Future work might overcome this limitation by training on more than one scene.

Model The proposed approach combines a primitive, hard-wired image formation model with a learned scene representation. As long as the data roughly follows this model, this is a winning combination. Scenes that are entirely beyond the model's scope might fail and will do so independently of the amount of data or the

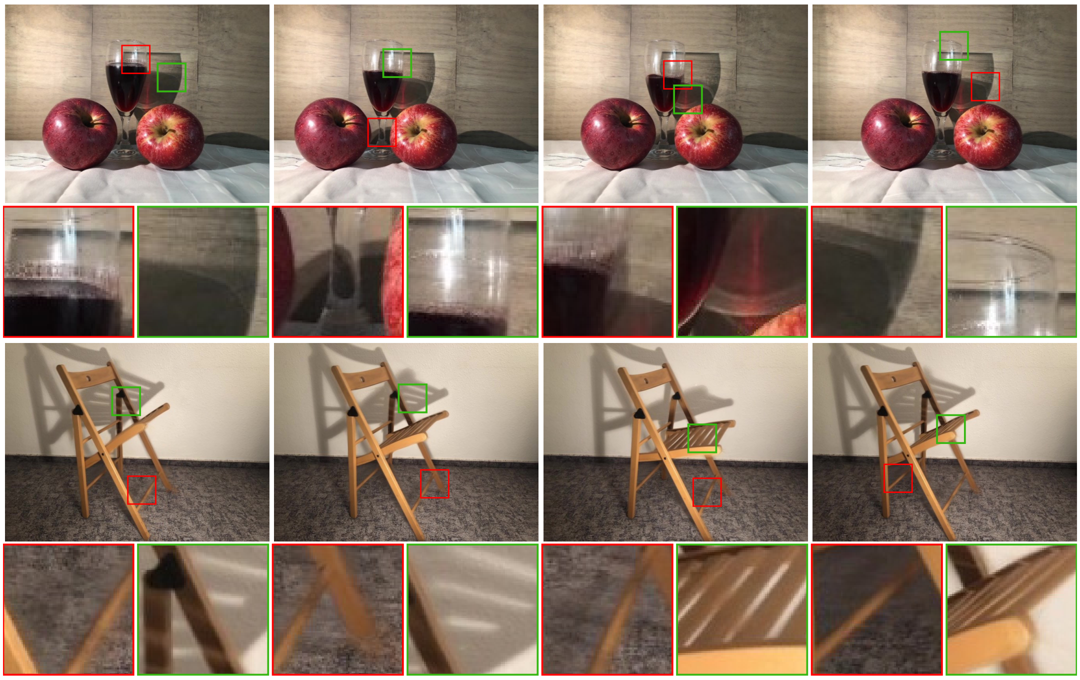


FIGURE 3.19: *The interpolation results of proposed method for two scenes ("Apple" and "Chair"): Insets in red identify regions where artifacts appeared, and insets in green indicate challenging examples that the method interpolated successfully. Artifacts mainly happened due to lack of training data; in the apple scene* ***(top)****, a 3×3×3 X-Field capture, the caustics in the shadow shows in only one view, and the foam is stochastic and different at each level of the liquid. In these regions, appearance does not properly interpolate but fades in and out, leading to ghosting or blurring. In the chair scene* ***(bottom)****, which is a 5×5×5 X-Field, the texture on the carpet beneath the chair gets blurry as this part of the carpet becomes visible only in one view due to occlusion caused by the chair, and its shadow. However, the method could handle soft shadow casting on a textured background or when there is a moving shadow of a complex object occluded with the object itself.*

representation capacity. The key assumption of the proposed design is that changes are explained by the flow. This is not a reasonable assumption with dominant transparency [Kopf et al., 2013]. Changes in brightness due to casual capture with auto-exposure can cause variations that the proposed deformation model fails to explain. In an X-Field non-unique flow is common: after one bounce, multiple indirect shadows might overlap and move differently. This is addressed by processing the signal so a unique flow becomes more applicable: by splitting shading and albedo, by representing the full X-Field Jacobian, by learning a non-linear inverse flow instead of linearly interpolating a forward flow, etc. Finally, if all flows are wrong, consistency weighting degenerates to linear blending. Future work could learn layered flow [Sun et al., 2012].

Capacity Finally, even if all data is available, the model is perfect, and the model assumptions are fulfilled, the NN needs to have the capacity to represent the input to the model. Naturally, any finite model can only be an approximation, and hence, the flow and, consequently, shape, illumination, and motion become smooth. The NN allows for some level of sharpness via non-linearities as in other implicit representations [Niemeyer et al., 2019; Oechsle et al., 2019; Chen and Zhang, 2019] but the amount of information is finite (Figure 3.18, right). Capturing sharp silhouettes is clearly possible, but to represent a scene with stochastic variation, stochasticity should be inserted [Karras et al., 2019] in combination with a style loss.

3.7 Conclusion

This chapter represents the X-Field as a NN that produces images conditioned on view, time, and light coordinates. The interpolation is high-quality and high-performance, outperforming several competitors for dynamic changes of advanced light transport (all BRDFs, (soft) shadows, GI, caustics, reflections, transparency), as well as fine spatial details (plant structures), both for single objects (still-life scenes) and entire scenes (tabletop soccer, parks). The particular structure of a network that combines a learnable view-time geometry model, combined with warping and reasoning on consistency has been shown to perform better than direct regression of color or warping without handling occlusion and state-of-the-art domain-adapted solutions. It is worth mentioning that this success becomes mainly possible because a general task is changed to a much simpler one: instead of interpolating all possible combinations of images, the method only interpolates within a fixed set. Strong generalization is a useful and exciting scientific goal, in particular from an AI perspective. But, depending on the use case, it might not be required in applied graphics: With the proposed approach, after 20 minutes of pre-calculation, one can deploy an X-Field in a VR application to play back at interactive rates. A user enjoying this high-quality visual experience might not ask if the same network could generalize to a different scene or not.

In future work, other data, such as data from Lightstages or sparse and unstructured capture, as well as extrapolation, should be explored. It is also desirable to reduce training time further (eventually using learned gradient descent [Flynn et al., 2019]) and explore interpolation along other domains such as wavelength or spatial audio [Engel et al., 2017], as well as reconstruction from even sparser observations.

Chapter 4

Novel View Synthesis for Refractive Objects

This chapter tackles the problem of generating novel-view images from collections of 2D images showing refractive and reflective objects. Current solutions assume opaque or transparent light transport along straight paths following the emission-absorption model. Instead, the proposed method in this chapter optimizes for a field of 3D-varying Index of Refraction (IoR) and trace light through it that bends toward the spatial gradients of said IoR according to the laws of *eikonal* light transport.

4.1 Introduction

Given images with different views of a refractive object, it is a challenging task to synthesize a novel view. The issue is that the refractive object takes its appearance from the surroundings by bending and internally reflecting the rays of light that travel through the object. By fully digitizing the object and its surroundings, one can synthesize novel views [Trifonov et al., 2006b; Hullin et al., 2008; Ihrke et al., 2010; Stets et al., 2017], but this approach requires a lot more information than a simple set of images. For instance, a dedicated hardware setup is required to digitize a transparent object [Ihrke et al., 2010; Stets et al., 2017; Lyu et al., 2020]. Deep learning offers an alternative approach where a Neural Network (NN) is trained to estimate the shape of such objects in more arbitrary surroundings [Stets et al., 2019; Sajjan et al., 2020; Li et al., 2020]. A deep learning technique based on a synthetic dataset, however, often returns a faulty estimate when presented with an image significantly different from those in the training data [Lyu et al., 2020].

One way to avoid the difficulties in representing a wide enough range of transparent object appearances in one synthetic dataset is to learn the radiance field of a given object based on a set of images capturing its appearance as observed from different directions [Lombardi et al., 2019b; Mildenhall et al., 2020]. This is useful for locating

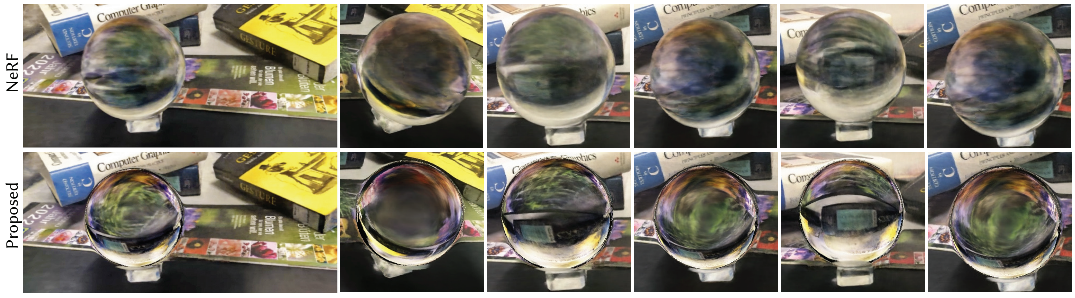


FIGURE 4.1: *Novel-view synthesis using Neural Radiance Fields (NeRF) (top) as well as the proposed eikonal approach in this chapter (bottom) for a real scene containing a refractive object.*

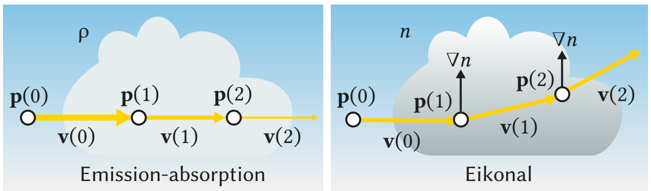


FIGURE 4.2: *Emission-absorption (left) and eikonal light transport (right). Light is the yellow arrow, its thickness indicates strength. Here, three discrete steps are considered, where in emission-absorption, the direction remains unaltered. It changes in the eikonal formulation according to the gradient of the IoR, n, a vertical gradient here. In the eikonal case, strength remains unaffected.*

and estimating the distance to transparent objects [Ichnowski et al., 2021]. However, since the neural radiance field approaches do not consider refraction, they can not be used out of the box for refractive Novel-View Synthesis (NVS).

To enable this, this chapter devises a learning-based method that optimizes for the field of 3D spatially-varying IoR given a set of 2D images picturing a refractive object. Existing solutions to learn 3D fields capturing scene geometry are based on opaque or transparent light transport along straight paths. In the presence of transparent objects, however, light bends, i. e., it changes its direction. The precise way in which light paths are curved depends on a certain *eikonal equation* operating on spatial gradients of the IoR field, which can be solved – and differentiated over in learning – in practice with the appropriate formulation. The resulting method allows for novel-view synthesis (Figure 4.1) in 3D scenes with complex objects involving strong refractive and internal reflection effects.

4.2 Light Transport ODE

This section discusses three approaches to model the interaction of light and matter as Ordinary Differential Equations (ODEs): a complete model (Section 4.2.1), an emission-absorption-only model (Section 4.2.2) and an eikonal-only model (Section 4.2.3). The complete one handles refractive and non-refractive scenes but was only applied to synthetic scenes in the literature. The emission-absorption one can be used for inverse rendering but excludes refraction. The eikonal one, in combination with the emission-absorption one, can handle refractive transparency in practical inverse rendering.

4.2.1 Complete Model

When light travels through a scene, it changes its radiance L due to absorption and emission as described by the (refractive) Radiative Transfer Equation (RTE) [Preisendorfer, 1957]

$$n(s)^2 \frac{\mathrm{d}(L/n^2)}{\mathrm{d}s} = -\sigma(s)L(s) + q(s), \tag{4.1}$$

where n is the IoR and $s \in [0, \infty)$ is the distance along a (curved) light path, σ is the extinction coefficient, and q/σ is the source function (which includes in-scattering) [Chandrasekhar, 1950]. The quantity L/n^2 is sometimes referred to as basic radiance. For a spatially varying n, light also changes its position $\mathbf{p}$ and direction $\mathbf{v}$ due to refraction according to the laws of eikonal light transport [Stam and Languénou, 1996; Gutierrez et al., 2005; Ihrke et al., 2007], see Figure 4.2. This can be described using

Hamilton's equations for ray tracing [Ihrke et al., 2007]:

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}s} = \frac{\mathbf{v}(s)}{n(s)} \qquad \text{and} \qquad \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}s} = \nabla n(s), \tag{4.2}$$

where $\mathbf{v}$ is not unit length but normalized by n. This model has been used in a virtual setting to render advanced visual phenomena, including refraction, total internal reflection, and scattering [Gutierrez et al., 2005; Ihrke et al., 2007; Ament et al., 2014; Pediredla et al., 2020]. Unfortunately, this is an ideal model that has not been demonstrated to be tractably used for NVS directly. A simplification can be made by ignoring refraction and introducing a different, also simplified model that allows NVS for refraction.

4.2.2 Emission-Absorption-Only Model

In NeRF [Mildenhall et al., 2020], radiance remains subject to emission and absorption

$$\frac{\mathrm{d}L}{\mathrm{d}s} = -\sigma(s)L(s) + q(s), \tag{4.3}$$

but travels along a constant direction $\mathbf{v}$ and the change of direction is assumed zero (Figure 4.2-left):

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}s} = \mathbf{v} \qquad \text{and} \qquad \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}s} = 0. \tag{4.4}$$

This is classic ray-marching along straight rays [Max, 1995].

4.2.3 Eikonal-Only Model

Complementary and finally, a simplified light transport is considered, which does not emit or absorb,

$$\frac{\mathrm{d}L/n^2}{\mathrm{d}s} = 0, \tag{4.5}$$

but changes direction as per the eikonal equation (Figure 4.2-right):

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}s} = \frac{\mathbf{v}(s)}{n} \qquad \text{and} \qquad \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}s} = \nabla n(s). \tag{4.6}$$

4.2.4 Solving

Concisely, all three variants can be formulated as position-motion-radiance state vector and its derivative:

$$\mathbf{z}(s) = (\mathbf{p}, \mathbf{v}, L) \qquad \text{and} \qquad \mathbf{z}'(s) = \frac{\mathrm{d}\mathbf{z}(s)}{\mathrm{d}s}. \tag{4.7}$$

For all three approaches, coupled ODEs

$$\mathbf{z}(s_1) = \mathbf{z}(s_0) + \int_{s_0}^{s_1} \mathbf{z}'(s)\mathrm{d}s = \texttt{odeSolve}(s_0, s_1, \mathbf{z}, \mathbf{z}') \tag{4.8}$$

need to be solved to compute the final state given the initial state as well as the IoR, emission and absorption fields.

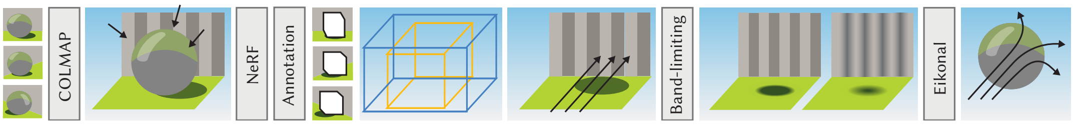


FIGURE 4.3: *Overview of proposed pipeline enabling final Eikonal training: It starts by establishing camera correspondence over all input images using COLMAP [Schönberger and Frahm, 2016]. Then the NeRF representation is utilized to explain the scene using emission-absorption and straight rays. In a semi-automated process, a 3D box region encompassing the refractive is identified. This refractive volume is not properly explained and is excluded from the NeRF fit. The view-independent part of this fit is discretized into a 3D grid which then enables the final progressive training using eikonal equations and curved rays in the last step.*

Typically, numerical integration such as Euler solvers [Hairer and Wanner, 1996] is used to solve for the state. Working backward, to compute gradients of the emission or absorption is done by automatic differentiation of forward Euler solvers [Mildenhall et al., 2020; Henzler et al., 2019b]. Unfortunately, this requires memory in the order of the number of steps a solver takes. When also accounting for IoR with many small steps, this can quickly become prohibitive. Instead, the adjoint formulation [Pontryagin, 1987] is adapted from NeuralODE [Chen et al., 2018b; Stam, 2020] which uses constant memory also in backward mode to perform `odeSolve`.

4.3 Pipeline

The proposed approach has two main steps: First (Section 4.3.1), reconstructing the opaque scene using a non-eikonal emission-absorption model with straight rays (Section 4.2.2) and, second (Section 4.3.2), modeling the remaining refractive part using an eikonal formulation (Section 4.2.3).

The result of the first step is an input to the second step, i. e., a non-refractive 3D explanation of the world is first trained which becomes the input to a second training that 3D-bends rays inside a fixed non-refractive world so that 2D input images can be explained (Figure 4.3).

4.3.1 Non-Eikonal Step

In this step, assuming straight rays, a NeRF model of emission ($\bar{q}$) and absorption ($\bar{\sigma}$) is trained. This is used to represent the background and to find the 3D region not properly explained by the model. A multi-scale version of this model is learned, which will be used in the next step.

Registration In the first step, the camera matrices are calculated to transform the camera space of each input image into one reference view using COLMAP [Schönberger and Frahm, 2016]. Hence, the 3D ray corresponding to every 2D pixel is known.

Diffuse-opaque init Given this information an off-the-shelf NeRF is learned that describes emission and absorption as two Multi-Layered Perceptron (MLP)s that fit continuous functions $\bar{q}(\mathbf{p}, \omega) \in \mathbb{R}^3 \times \Omega \mapsto \mathbb{R}^3$ and $\bar{\sigma}(\mathbf{p}) \in \mathbb{R}^3 \times \Omega \mapsto \mathbb{R}$ mapping position and direction to RGB color or scalar opacity. Let θ and ϕ denote the MLP parameters of the emission and absorption models resulting from this optimization.

Masking The model above of $\bar{q}$ and $\bar{\sigma}$ will not be reliable for refractive objects. Hence, these parts of 3D space need to be eliminated and explained by the proposed eikonal approach. The parts that are non-refractive will be input into this step. The refractive part of the scene assumes to be bounded by a 3D box $\Pi \in \mathbb{R}^{3\times2}$ that exclusively contains refractive objects. This results in a *masked* emission model q, respectively a masked σ:

$$q(\mathbf{p},\omega) \text{ resp. } \sigma(\mathbf{p}) = \begin{cases} 0 & \text{for } \mathbf{p} \in \Pi \\ \bar{q}(\mathbf{p},\omega) \text{ resp. } \bar{\sigma}(\mathbf{p}) & \text{otherwise.} \end{cases} \tag{4.9}$$

The bounding box Π is estimated as follows: For a given set of views (the 10 percent of the training images uniformly distributed around the refractive object), the user selects a few points on the horizontal and vertical extent of the refractive object in the scene. Given collected 2D points from the images, the depth map computed from the NeRF model is used to compute their corresponding 3D locations. Then, the 0.02 and 0.98 percentiles of all points along each spatial dimension are calculated, and they are multiplied by a constant value of 1.2 to make sure the box encompasses the entire object. The parameters of Π are given by the minimum and maximum coordinate values of the points.

Progressive grids Solving for the eikonal directly given σ and q is challenging. The problem is that when rays bend a lot, it becomes harder to find correspondences between input images and background. Moreover, the bending depends on the spatial gradient of the IoR rather than the IoR directly, which is an operation known to be numerically demanding to optimize over. Addressing this challenge, the proposed method in this chapter instead learns eikonal transport using different progressively finer versions of the emission and absorption models. This is inspired by progressive spatial encodings [Park et al., 2021], but instead of blurring the periodic spatial functions, the radiance function itself is blurred. It is not obvious how to make a coarser version of q or σ, which are MLPs. In particular, the preliminary experiments using slower-varying or fewer spatial encodings did not result in the desired band-limiting. Instead, the regular grids are considered. These are typically struggling to resolve fine details or to work in 5D, but fortunately, this is not required here. Hence, the masked emission and absorption solution are sampled to a 3D grid P and Q, and the Q is averaged over the angular domain:

$$\begin{aligned} Q_i(\mathbf{p}) &= \mathbb{E}_{\mathbf{y}}[\mathbb{E}_\omega[q(\mathbf{y},\omega)\kappa_i(|\mathbf{p}-\mathbf{y}|)]] && (4.10)\\ P_i(\mathbf{p}) &= \mathbb{E}_{\mathbf{y}}[\sigma(\mathbf{y})\kappa_i(|\mathbf{p}-\mathbf{y}|)], && (4.11)\end{aligned}$$

where κ_i is a Gaussian kernel of increasing frequency bandwidth for increasing levels i. In the experiments throughout this chapter, a grid size of 128^3 is considered, and the values inside the grid are interpolated with a trilinear interpolation scheme.

4.3.2 Eikonal Step

At this step, given the hierarchy of grids describing the emission and absorption in the scene for all locations $\mathbf{p} \notin \Pi$ outside the refractive box, an IoR field is defined on $\mathbf{p} \in \Pi$ to explain both the non-refractive 3D grids and the 2D images.

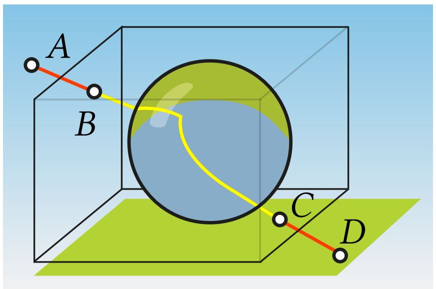


FIGURE 4.4: *Enter and exit.*

Masked traversal Figure 4.4 shows a red ray starting from point A and traversing the world outside Π, which is hit at point B. The emission and absorption models $\bar{q}$ and $\bar{\sigma}$ are used to trace the straight ray from A to B (Section 4.2.2). A key concept is to *enter and exit* the refractive representations in a masked traversal, as well as training with masked rays and progressively. Starting at B, eikonal ray-marching curves out the yellow path (Section 4.2.3) according to an IoR model that maps spatial position to IoR: $n(\mathbf{p}) \in \mathbb{R}^3 \mapsto \mathbb{R}$. When this ray leaves Π at C, it continues with emission and absorption on a straight path, eventually receiving a contribution at D or other points along the straight line. Training of n – that is also an MLP whose parameters are denoted as ψ – proceeds similar to NeRF, but instead of marching geometrically, an ODE in position-motion-radiance space is solved (and back-propagated through). Recall that the $\mathbf{z}$ denotes a position-motion-radiance vector. The dot notation is not utilized to pick an element in the vector so that $\mathbf{z}.\mathbf{p}$ denotes the position and $\mathbf{z}.L$ the radiance, for example. In the mixed refractive/non-refractive case, the state ODE is

$$\mathbf{z}'_\psi(s) = \left\{ \begin{array}{ll} \text{Eqs. 4.5 and 4.6 s.t. } n_\psi & \text{if } \mathbf{z}_\psi(s).\mathbf{p} \in \Pi \\ \text{Eqs. 4.3 and 4.4 s.t. } q_\theta \text{ and } \sigma_\phi & \text{otherwise.} \end{array} \right\}, \tag{4.12}$$

so the state change is non-eikonal outside the box and eikonal inside. It is made to depend on ψ, but not on θ and ϕ, as these are fixed both in the forward and backward pass of this step.

Let $\mathbf{z}_i$ denote the state of a ray through pixel i. It is then solved using

$$\psi^\star = \arg\min_\psi \mathbb{E}_i[|\texttt{odeSolve}(s_0, s_1, \mathbf{z}, \mathbf{z}', \psi).L - \mathbf{z}_i.L|], \tag{4.13}$$

where ψ is an extra argument for `odeSolve` with parameters that condition $\mathbf{z}'$. As a ray cannot change direction outside Π, the condition in Equation (4.12) can be handled by loop splitting in practice: First, the ray is traced straight, then traced eikonal, and then it is traced straight once more, eliminating the conditional statement in Equation (4.12).

Masked rays Since the considered MLP for estimating the IoR is only evaluated inside the bounding box, the eikonal training starts by making sure a batch contains only the rays that are hitting the box.

Progression This step starts by finding an IoR field that explains a coarse version of the emission-absorption grid. When the change of error falls below a threshold, it is switched one level up to a finer grid (Figure 4.5). The number of parameters in the MLP to represent the IoR is the same at all levels. Note the final images are rendered using the full NeRF model instead of a grid.

Interior radiance field Non-transparent objects might be present in the interior of the transparent object that is located in Π. To explain these, another NeRF is trained for the radiance in Π. The IoR field in Π is available from the eikonal step (Section 4.3.2), and it is fixed together with the opaque NeRF (Section 4.3.1). The ray paths only bend when encountering the transparent object in Π. Conclusively,

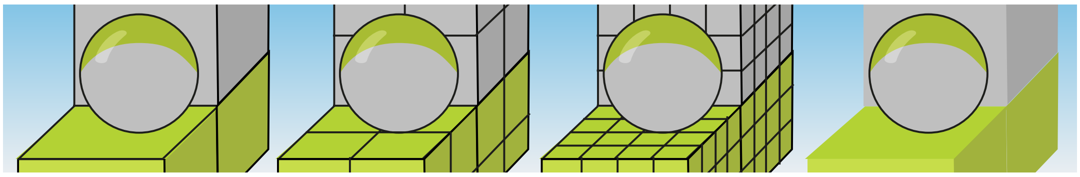


FIGURE 4.5: *The eikonal solution is fit to progressively finer discrete-grid approximations of the opaque NeRF solution and, finally, the continuous field.*

the proposed solution consists of the opaque NeRF, the MLP for the IoR field, and a NeRF for the interior of the transparent object. Together, these have been trained sequentially to explain the input images.

4.4 Implementation Details

The NeRF implementation follows Mildenhall et al. [2020], and the second MLP to represent the IoR field is a 6-layer MLP with 64 hidden dimensions with a skip connection that concatenates the input to the third layer's activation. Similar to NeRF, positional encoding with five frequencies is applied to the input. For stable training, as suggested by Chen et al. [2018b], the Softplus activation with $\beta = 5$ is used for all layers instead of a non-smooth function like ReLU, and all layers are initialized with the Xavier uniform. In the non-eikonal step (training NeRF), the same setting is used for the training as described by Mildenhall et al. [2020], and the optimization is done for 150k iterations. This takes around 12 hours to converge on a single NVIDIA 1080Ti with 12 GB RAM. For the Eikonal step, a batch size of 1024 rays is considered, and the entire space is traversed with 128 ODE steps. In this step, the training takes around 5 hours for 5k iterations. The Neural ODE PyTorch implementation Chen et al., 2018b; Stam, 2020 is employed to backpropagate through the ODE with the adjoint method. In the progression part, the 3D grid is filtered with a Gaussian kernel with a normalized frequency bandwidth of 0.08 cycles per sample, and it is doubled for every 1k iterations. The last step takes a single MLP similar to the NeRF fine network [Mildenhall et al., 2020], but with 128 hidden dimensions to represent the interior radiance field. As there is no hierarchical volume sampling involved, 512 steps are considered along the ray to properly sample both interior and exterior radiance fields, which takes around 12 hours to optimize over 10k iterations. With the complete model of the proposed method, it takes around 85 seconds to render a frame of 672×504 resolution.

4.5 Results

The goal of this chapter is NVS with plausible coherence in scenes with transparent objects. The result of the proposed approach is compared with existing methods using standard Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) metrics. It is also further evaluated with a user study.

Scenes This thesis proposes four real scenes that include refractive objects with unknown geometry: BALL, GLASS, PEN and WINEGLASS. An iPhone 8 camera is used to capture 96, 97, 105, and 102 views, respectively, for each scene, and 1/10 of all views are held out for the test set. All images are down-sampled to the resolution 672×504 pixels.

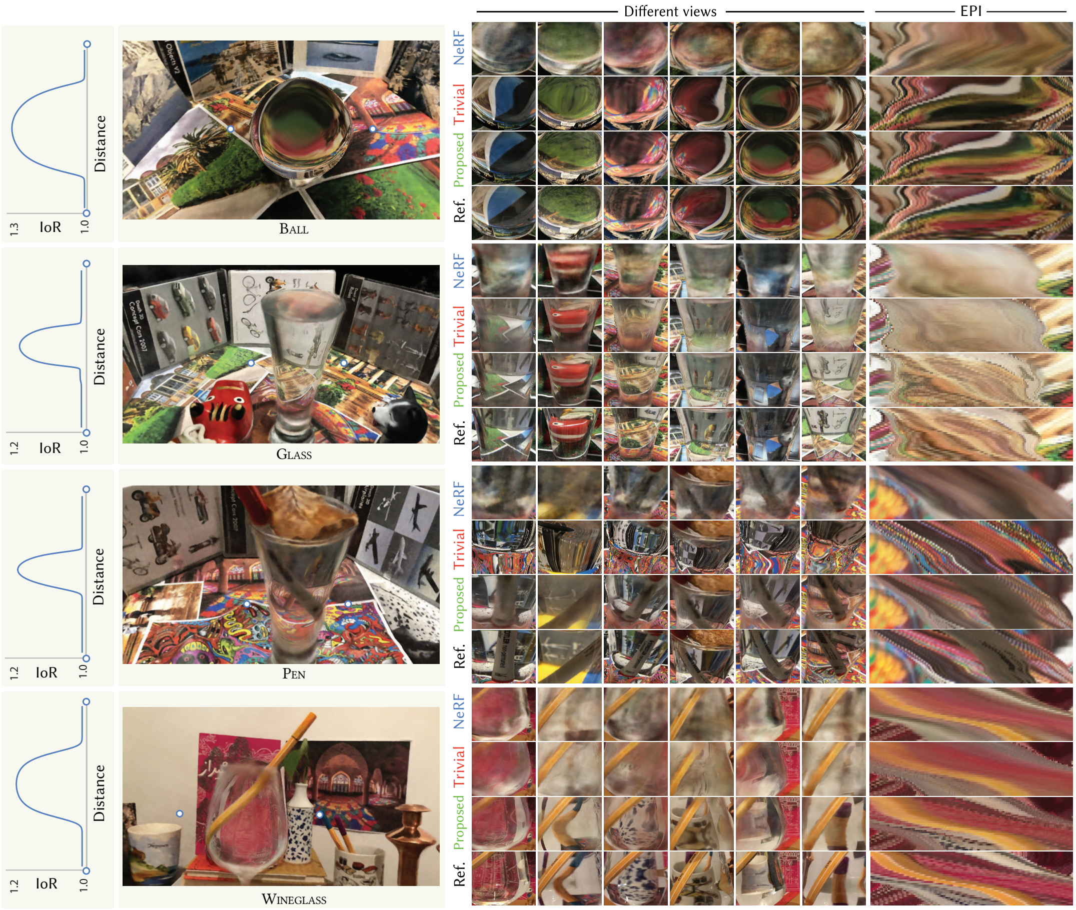


FIGURE 4.6: *The left block shows the cross-section of recovered IoR by the proposed method for a scanline between the white dots shown in the reconstructed test view in the second block. The third block shows insets taken from novel views produced by three different methods (rows) for different viewpoints (columns). The right block shows a pseudo-epipolar view using a continuous camera trajectory, again for all methods.*

Methods The proposed method `Proposed` is compared with `NeRF` and another method, namely `Trivial`.

In `Trivial`, the IoR field is simply reconstructed using the density field of refractive objects recovered by NeRF. This is done first by executing the NeRF model for a discrete set of samples along the rays coming from the input camera poses and crossing the bounding box Π, and setting the density to zero for the samples outside the box. Then, for each ray, both the front and back surface position of the refractive object are estimated by forward and backward ray marching until an opacity threshold is reached (similar to how the depth maps are computed in NeRF). For the samples that fall between the intersections, a constant IoR value is assigned (1.5 in the case of glass and 1.33 in the case of thin glass filled with water), and IoR value of 1.0 is considered for the regions outside. Then an MLP is employed to map each 3D point inside the box Π to its calculated IoR.

Qualitative comparisons Figure 4.6 facilitates a visual comparison of `Proposed` with `NeRF` and `Trivial`. The insets show novel view reconstructions of different viewpoints for all methods. Please refer to the supplemental video for an animated version of these results. `NeRF` tends to "fake" refraction by considering diffuse content on the surface of the transparent object and assigning a view-dependent color for

TABLE 4.1: *Quantitative comparison according to different metrics for each method (row) and different scenes (columns). The numbers in the User column say how often in the conducted user study, the method was considered closer to the reference than ours. As these numbers are significantly ($p < 0.01$) smaller than the chance level 50%, the proposed method was for all scenes considered closest to the reference in the majority of the comparisons shown to the users.*

	BALL				GLASS				PEN				WINEGLASS			
	PSNR	SSIM	LPIPS	User	PSNR	SSIM	LPIPS	User	PSNR	SSIM	LPIPS	User	PSNR	SSIM	LPIPS	User
`NeRF`	**27.384**	0.945	0.042	0.27%	**27.146**	**0.924**	0.066	3.83%	**27.749**	0.933	0.059	9.58%	**29.011**	**0.947**	0.045	24.93%
`Trivial`	24.373	0.933	0.034	9.31%	25.930	0.914	0.059	7.39%	23.070	0.912	0.060	37.26%	26.739	0.935	0.052	1.33%
`Proposed`	26.720	**0.951**	**0.023**		26.525	0.922	**0.050**		27.803	**0.935**	**0.047**		27.789	0.940	**0.042**	

each point on the surface. Under the condition of extreme view changes, as can be seen in all scenes, `NeRF` fails to properly reproduce the color, and it tends to average all observations leading to a blurry result. `NeRF` also seems to struggle with the reconstruction of an occluder inside the transparent objects, although multi-view consistency holds for the object inside. In the PEN scene, `NeRF` failed to assign transparent content on the surface of the glass in order to properly reconstruct the pen inside.

`Trivial` assumes a constant IoR field inside the entire refractive object, and in case of spatially varying IoR, the refraction tends to be wrong for some regions (e. g., towards the top and the bottom of the glass in the GLASS and the PEN scenes). `Trivial` performs better on the BALL scene as the crystal ball has a constant IoR inside. However, due to the mere fact that the NeRF density field for the refractive object is not always valid, the estimated IoR of `Trivial` might not be very accurate, and the refracted background becomes misplaced in some regions. In contrast, `Proposed` reproduces sharper details and aligns better with the reference. Moreover, in order to assess the temporal consistency of each method, in the right block, the corresponding pseudo-epipolar image is also shown, which is created by stacking a selected scanline for 30 subsequent video frames using a continuous camera trajectory. A good optical flow continuity can be observed between the stacked scanlines for all methods, but clearly, the flow fidelity with respect to the reference is best for `Proposed`. `NeRF` and `Trivial` feature significant blur that is also visible in the insets in the middle column.

User study Unfortunately, no method exists to quantify the main aim of this work, plausible refractive and reflective flow. To quantify the coherency, a small user study is performed. A reference photograph and two images produced by `Proposed` and either `NeRF` or `Trivial` (selected randomly) were shown to 73 participants, 10 image triplets for each scene. The participant then had to indicate which one is visually closer to the reference photograph in a Two-alternative Forced Choice (2AFC) experiment. All three images are presented simultaneously without any time limit; the position of reference is fixed while randomized for the other two images. Five different views for each of the four scenes are selected, and the participant selections are aggregated over those views. For each scene, Tab. 4.1 reports how often a competitor is selected; hence less is better, while the chance level is 50%. All outcomes are significant at the $p < 0.01$ level for a binomial test at $N = 73$.

Quantitative comparisons Tab. 4.1 presents the quantitative results of the user study (in the "User" columns), and the different metrics averaged over the test set. It can be seen that `NeRF` has consistently the highest PSNR, which is a metric relatively insensitive to blur or structure preservation. When it comes to SSIM, which is a metric more aware of the structures, it comes to a draw. At the most advanced metric,

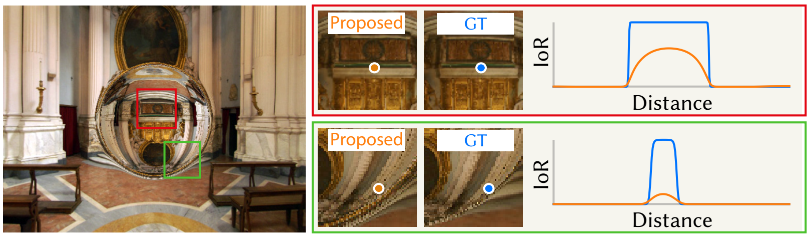


FIGURE 4.7: *IoR cross-section of the proposed method (orange) and ground truth (blue) for a scanline along the pixel marked with a dot in each inset.*

LPIPS – which is based on human image artifact perception and better tolerates small spatial misalignment with respect to the reference – `Proposed` always wins. `Trivial` is sometimes better than other methods but never wins. The participants of the user study almost consistently indicate that `Proposed` leads to less perceived differences with respect to the reference views. A relatively high score of `Trivial` for PEN can be attributed to the background sharpness and its color saturation that could have an appeal to some of the experiment participants, who somehow neglected strong background distortions, and the lack of pen in the water, as also visible in Figure 4.6. Also, a relatively high score of `NeRF` for WINEGLASS can be attributed to some views, where the background contained less high-frequency details, so that blur typical for this method was less perceivable.

Reference comparison. While the proposed method makes use of an IoR, it is not forced to use actual physical values. For a scene with a known IoR (Figure 4.7), the method can reconstruct the image faithfully, while a cross-section shows the IoR is indeed quite different from the reference IoR. It again needs to be mentioned that the method is suitable for NVS, not for the reconstruction of the 3D structure.

4.6 Conclusion

Given a set of 2D images containing refractive materials, this chapter explored the problem of optimizing for the field of 3D-spatially varying IoR with the purpose of NVS. Existing solutions that learn 3D fields for NVS are based on opaque or transparent light transport along straight paths. As opposed to this, the proposed method in this chapter models the bending of light according to the eikonal equation from geometric optics. This enables perceptually better NVS in 3D scenes with complex objects exhibiting strong refractive effects. This chapter is subject to several assumptions. The eikonal equation deals with refraction and total internal reflection but is not separated into partial reflection and refraction. Partial reflection and refraction in continuously varying media is difficult even in forward simulation and left for future work. Moreover, the optimization is done sequentially, where the diffuse world is first learned, followed by the transparent objects in a second pass that relies on a user marking the bounding box of the specular object to aid the task. Ideally, this should be done jointly and in a fully automated way. As a consequence, it is assumed that the diffuse world is sufficiently observed, making it unable to reconstruct parts exclusively revealed in the refraction. Despite simplifying assumptions, this chapter, by means of eikonal light transport for the first time, included refraction and total internal reflection in a model that learns 3D fields from images of transparent objects to accomplish the synthesis of novel views.

Chapter 5

A No-Reference Metric for Predicting IBR Artifacts

Image metrics predict the perceived per-pixel difference between a reference image and its degraded (e. g., re-rendered) version. In several important applications, the reference image is not available, and image metrics cannot be applied. This chapter devises a neural network architecture and training procedure that allows predicting the MSE, SSIM or VGG16 image difference from the distorted image alone while the reference is not observed. This is enabled by two insights: The first is to inject sufficiently many un-distorted natural image patches, which can be found in arbitrary amounts and are known to have no perceivable difference to themselves. This avoids false positives. The second is to balance the learning, where it is carefully made sure that all image errors are equally likely, avoiding false negatives. Surprisingly, the resulting no-reference metric, subjectively, can even perform better than the reference-based one, as it had to become robust against misalignment.

5.1 Introduction

Computer vision or graphics experts easily recognize image artifacts that might be highly domain-specific. An image-based rendering (IBR) specialist will quickly notice where depth estimation failed, where transparency was not handled, or where a highlight did not move correctly. Similarly, in computer graphics, artifacts resulting from Monte Carlo noise in image synthesis when producing a feature film, or shadow bias [Williams, 1978] in a computer game are easily spotted by domain experts. The assessment typically is not limited to detection but importantly includes judging magnitude as well as spatial locality. The importance of interacting with errors can be seen from photographs with spatially annotated over- and under-expose artifacts, as done, for instance, by Coleman [2012]. Remarkably, all this is not achieved by comparing an image to a reference but by experience and intuition built from knowing

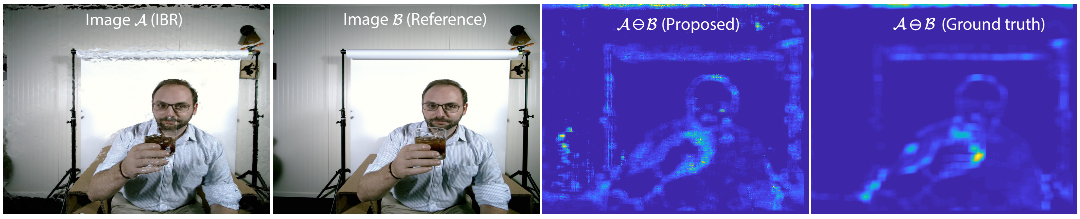


Image $\mathcal{A}$ (IBR)
Image $\mathcal{B}$ (Reference)
$\mathcal{A} \ominus \mathcal{B}$ (Proposed)
$\mathcal{A} \ominus \mathcal{B}$ (Ground truth)

FIGURE 5.1: *Given an image $\mathcal{A}$ (first) that is a version of a reference $\mathcal{B}$ (second) distorted by IBR artifacts, the proposed method in this chapter predicts their per-pixel difference map $\mathcal{A} \ominus \mathcal{B}$ (third) without observing $\mathcal{B}$. The fourth image shows the ground truth difference $\mathcal{A} \ominus \mathcal{B}$. Here, it is shown for MSE, but other metrics, such as SSIM or VGG16 are also possible.*

what natural images look like and how images with artifacts differ. Is it possible for a machine to perform such a task?

More formally, this chapter faces the challenge illustrated in Figure 5.1. Given an image $\mathcal{A}$ that is a distorted version of a reference $\mathcal{B}$, the aim is to predict their difference $\mathcal{A} \ominus \mathcal{B}$ without access to $\mathcal{B}$. The ground truth metric response could simply be the mean square error (MSE as used in Figure 5.1), a more perceptual metric like SSIM [Wang et al., 2004b] or even VGG-16 activation differences that are effective as an image metric [Simonyan and Zisserman, 2014; Zhang et al., 2018]. More particularly, it can go beyond the typical mean opinion scores [Talebi and Milanfar, 2018] given to uniform distortions such as noise or JPEG compression and can seek to produce localized distortion visibility maps without accessing the reference.

This chapter chooses to study one specific form of artifacts that arise in image-based rendering (IBR) [McMillan and Bishop, 1995; Gortler et al., 1996], in particular when employed for novel-view synthesis from sparse light fields (LFs) [Levoy and Hanrahan, 1996]. It is important in virtual reality and movie production, where LFs are used to provide head motion parallax and special effects. Moreover, having a localized error prediction is also important for quality control. In IBR, artifacts are very localized (e. g., around certain depth edges), and creating opinion scoring or even spatial-angular annotated datasets of LF artifacts in a size that is sufficient for machine learning appears to be a daunting task. The proposed method proceeds without all of this.

Addressing this challenge, the method in this chapter makes use of convolutional neural networks. It will show how learning this mapping right away will result in many false positives or false negatives. Instead, two important ingredients come together in the method. First, as the number of images containing artifacts is typically limited, the training data is augmented with natural images that are free from artifacts. Second, the right balance between natural and distorted training data should be considered.

Not requiring a reference is useful whenever the original is inaccessible (lost, impossible to compute, unavailable, undefined). Furthermore, two applications of the presented no-reference metric are demonstrated in light field production. In the first application, a sparse light field is first captured, followed by an interpolation of the intermediate views. If the metric indicates those intermediate views have errors, those views will be recaptured. This allows for acquiring a higher-quality light field in a much shorter time compared to dense LF capturing. In the second application, metric prediction is employed to identify the local distortions as a guide for interactive depth adjustment.

5.2 Overview

Test-time input to the proposed method is a single distorted RGB image $\mathcal{A}$. While the considered distortions are always IBR artifacts resulting from a specific depth reconstruction and specific IBR method, the interna of how this image is generated (e. g., the depth map) is transparent, and only the outcome result is needed. Withheld is the reference RGB image $\mathcal{B}$. In the case of IBR, such a distorted-undistorted pair is typically produced by rendering a known image from other known views.

The output of the method is a single-channel image that predicts a given difference metric response $\mathcal{A} \ominus \mathcal{B}$, where the $\ominus$ operator depends on the choice of the specific metric, e. g., MSE, SSIM [Wang et al., 2004b], or VGG16 [Simonyan and Zisserman, 2014]. High values are produced where the images are different and small values

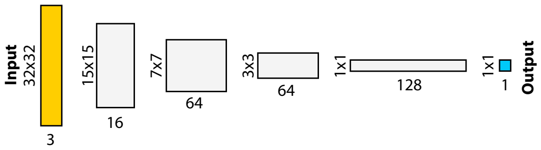


FIGURE 5.2: *The suggested architecture consumes* 32 × 32 *patches (yellow left) and applies a cascade of* 3 × 3 *convolutions, followed by non-linearities (ReLU). Spatial resolution is reduced* (height) *and feature count increases* (width) *before a final prediction of the metric response is produced* (blue, right).

where they are similar. This output is accurate if it has few false positives or negatives. False positives correspond to predicting a perceived difference where there are no artifacts, and false negatives correspond to visible artifacts the metric fails to report. Note that two forms of approximations are made here: the first is the error that the metric itself makes when comparing two images relative to human judgment. The second is the error that the method has with respect to a prediction. Ultimately, the method is a prediction of a prediction but, surprisingly, can perform better than one prediction alone.

5.2.1 Training Data

The training data used in this chapter comprises existing metric responses $\mathcal{A} \ominus \mathcal{B}$ to the distorted image $\mathcal{A}$ and the clean reference image $\mathcal{B}$. Strictly speaking, learning does not even observe the reference image $\mathcal{B}$, but in practice, it is required to compute the metric response $\mathcal{A} \ominus \mathcal{B}$.

The training dataset consists of the captured LF images of 42 different scenes, which come from the Stanford LF repository [1], the Fraunhofer IIS light field dataset [Dabała et al., 2016], Google Research work [Penner and Zhang, 2017b], and Technicolor [Sabater et al., 2017] as well as from newly captured LF images provided by this thesis. All 4D LF datasets comprise conventional 2D images in a resolution up to 2k×2k, taken from a range of sparse viewpoints, such as in a 3×3 camera array with known camera positions. For each LF viewpoint, first, the depth is extracted using a light field depth estimation technique [Dabała et al., 2016], then the images are warped [Mark et al., 1997] to that view. For each LF, four corner views are used to generate novel-view images at the positions of the remaining views. Each warped view corresponds to one original view where the response of a full-reference metric is computed to this pair. With approx. 9 views per LF and 42 LFs in total, this amounts to only 210 unique images, i. e., a comparatively low number for a training task. Six scenes were used for testing and the rest for training. The same split is also applied later for the user study. The test scenes are totally different from the training scenes, which is important as the number of scenes in the training set is small, and generalization across them is an additional challenge. The natural images used in the training and test dataset are sourced from the Inria Holidays image dataset [Jegou et al., 2008], which has a comparable resolution to the LF images.

The proposed method in this chapter is independent of the actual underlying metric $\ominus$ that is used for the prediction of the difference. This response will be denoted neutrally as $\mathcal{A} \ominus \mathcal{B}$. Three metrics are explored in this chapter: MSE, SSIM, and VGG16. MSE is defined as the average per-pixel RGB difference vector length squared. The SSIM metric uses the original implementation [Wang et al., 2004b]. VGG16 [Zhang et al., 2018] transforms both $\mathcal{A}$ and $\mathcal{B}$ into the VGG16 feature space and picks the activations at layer five, which is 512-dimensional. The L_2 difference of

[1] http://lightfield.stanford.edu/lfs.html

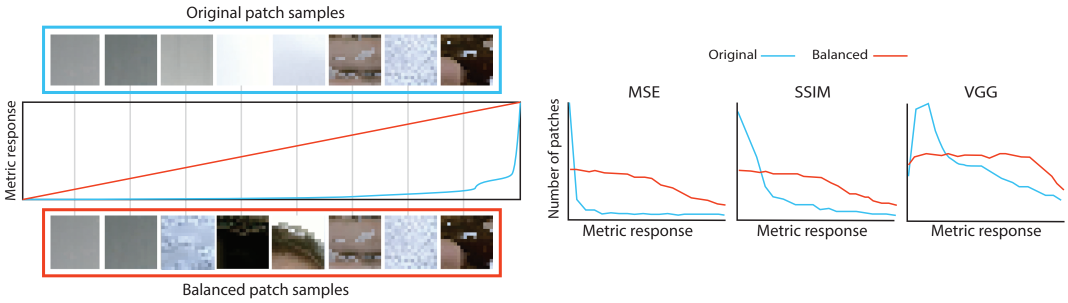


FIGURE 5.3: *When sampling uniformly from IBR patches, the error distribution is skewed towards low errors* (blue)*. The suggested balancing* (red) *adjusts the samples to have a uniform range of errors. The three lower plots show the actual distribution before and after balancing for different metric responses.*

these two vectors is used as the metric response. Each metric is normalized with the 95th percentile of their responses across the training dataset to fall between 0 and 1.

5.2.2 Architecture

The presented method is built upon a simple encoder P [Ronneberger et al., 2015] that has learnable parameters Θ and predicts the error map $P(\mathcal{A}|\Theta)$ by observing $\mathcal{A}$ (Figure 5.2). The network comprises 5 layers (32×32 patch size) with the total number of $|\Theta| = 175,537$ learnable parameters and is trained on all patches of the training set in a sliding window fashion. The loss is the L_1 error of the predicted metric response, so $||P(\mathcal{A}|\Theta) - (\mathcal{A} \ominus \mathcal{B})||_1$. Note that the loss is always L_1, while the metric can be the L-norm-like MSE as well as SSIM or VGG16.

5.2.3 Balancing

So far, it has been explained why, and it will be shown from the ablation study that it is important to have natural patches, but the question is how many. With an unlimited number, the metric prediction simply always returns zero because natural patches have no error to themselves. The proposed solution starts with a half-half mix of distorted and clean patches. Regrettably, many of the distorted patches, which make 50 % of the total, also have small errors that are close to zero. These patches are exactly those for which IBR was successful, i. e., did not have any artifacts. Depending on the metric, this imbalance can be very strong, and in particular, for MSE, it is extremely heavy-tailed (Figure 5.3). To address this, the proposed method suggests balancing the error distribution for the distorted half when creating the training data as follows: First, all patches are sorted based on their metric response into a priority queue. Then, a random sample is uniformly drawn within the range of zero to the 95th percentile of the metric response distribution. For every sample i with value ξ_i, a patch j with the most similar metric response d_i is selected and will be added to the training dataset and removed from the queue. When the minimum difference $\xi_i - d_j$ is larger than a threshold ϵ, the chosen sample is rejected. This is repeated until a target patch count, such as 250 k, is reached.

5.3 Evaluation

5.3.1 Methods

Training Strategies Three different strategies are compared for training. The first is the proposed method, the other two are ablations. FULL is the proposed method

TABLE 5.1: *The error of the metric predictions on the test data for different variants of the method and different partitions (*All/Clean/Distorted*) of the training data* (columns) *on different metrics* (rows). *Winners per partition are marked in bold.*

Metric	Full			NoNatural			NoBalance		
	All	Cle.	Dist.	All	Cle.	Dist.	All	Cle.	Dist.
MSE	**.098**	.006	.189	.137	.092	**.182**	.102	**.003**	.201
SSIM	**.078**	.013	.143	.143	.159	**.127**	.080	**.012**	.149
VGG	**.085**	**.006**	.165	.207	.293	**.121**	.092	.008	.176

involving 50 % natural patches and a balancing of the other 50 % as described in Section 5.2.2. NoBalance is realized by a similar 50/50-split, but the network is trained on all distorted patches without the balancing. NoNatural adapts the balancing to take 100 % of the patches coming from IBR without adding the natural patches as described in Section 5.2.2. All training sets, albeit processed differently, have the same size of ca. .5 M patches.

Error As the goal is to predict the metric responses, the prediction error is the same as the loss, the absolute difference between the ground truth metric response and the prediction of that response. As these errors also come in arbitrarily different scales for different metrics, they are normalized with the global 95th percentile of the GT metric response across the balanced training dataset. The errors in metric prediction errors for split subsets are additionally reported to understand the false/true-positive and false/true-negative tendencies. In All, the error for the whole test dataset is computed. Additionally, two subsets of the test dataset are considered. The first subset is Clean, which includes only natural patches. The second one is Distorted which contains only IBR patches, including those that might also come out with very low or even with no error. Note that this is a partitioning of the test set and not of the training set.

5.3.2 Quantitative Results

This section discusses both the means and full error distributions of all training strategies for different partitions and different metrics.

Means The means of all methods are compared in Table 5.1. It can be seen that the proposed method (Full) has the smallest error across different metrics compared to both other variants (bold in column All). By looking into the partitioning, it can be noticed for the Distorted partition, the NoNatural strategy performs best. This is expected as training is done with all distorted patches, which comprise the maximal variety of distortion. This makes the resulting metric sensitive to all kinds of distortions. As a result, the probability of false negatives, i. e., claiming patches with an error to be fine, becomes low. It also appears that for the Clean partition, the NoBalance strategy performs best. This also is expected as in training, 50 % of data comprises natural (undistorted) patches, and due to the NoBalance strategy, small errors dominate in the distorted patches. This makes the resulting metric particularly sensitive to near-threshold distortions. In this case, the probability of false positives, i. e., reporting a high metric response for no-error patches, is low. All statements are true (significant, $p < .01$, t-test after testing for Gaussianity) across all metrics,

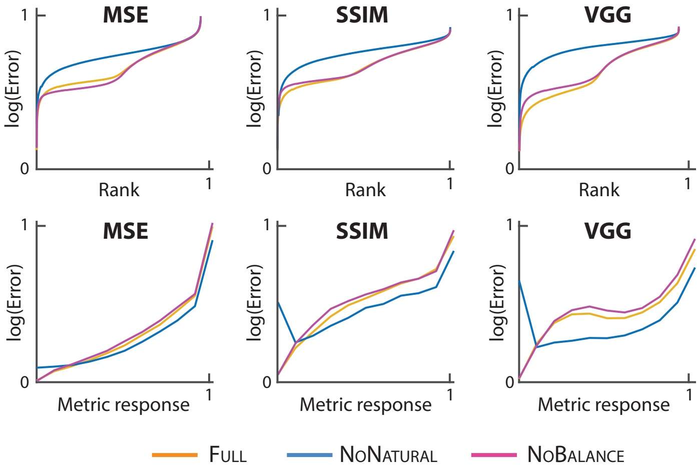


FIGURE 5.4: *Analysis of metric prediction error for different metrics and variants of the proposed method. The top plots show sorted error distributions. The bottom row plots show the correlation between metric response and metric prediction error. All vertical axis is log scale.*

indicating that the FULL approach is independent of the underlying metric. A positive exception is VGG, where the FULL approach even performs better than NOBALANCE on the CLEAN partition.

Distributions Figure 5.4 shows the distribution of errors for different metric predictions (top) and the correlation of the prediction error and metric response (bottom). In each plot, colors encode the variants of the proposed approach (NONATURAL, NOBALANCE, FULL). Each plot in the first row of Figure 5.4 shows the sorted error of the metric prediction in ascending order. The FULL approach performs better than other variants across the entire range, with the exception of MSE prediction for low errors. This indicates that the mean is a good characterization of the performance. In all cases, there is a sudden increase in the error that occurs around 50 % of the population, i. e., the error for the first half of the population seems to follow a different trend than the second half. This could be attributed to the patches where reference and input are (partially) not aligned, which make up roughly 50 % of the population as well. Unfortunately, there is no way to tell apart a misaligned patch that is judged by FR metrics as different with respect to a displaced reference. Hence, large errors are expected to become undetectable at some error level. The exception is the regime in MSE where the FULL approach is worse on low errors and slightly better on high errors, while it performs best on average in (Table 5.1). This can be difficult to comprehend due to the log scale of the vertical axis. Each plot in the second row in Figure 5.4 shows the error of the prediction on the vertical axis and the metric response on the horizontal axis as a connected scatter plot. The plots are in accordance with Table 5.1: The NONATURAL method, which performs best in predicting high metric responses, has a high error on patches with a small metric response (false positives). Symmetrically, the NOBALANCE method, which is the best at predicting low metric responses, produces high errors on patches with a high metric response (false negatives). FULL method is always a bit worse than one other method in one region (except at the unique point where both cross), but on average, performs best overall.

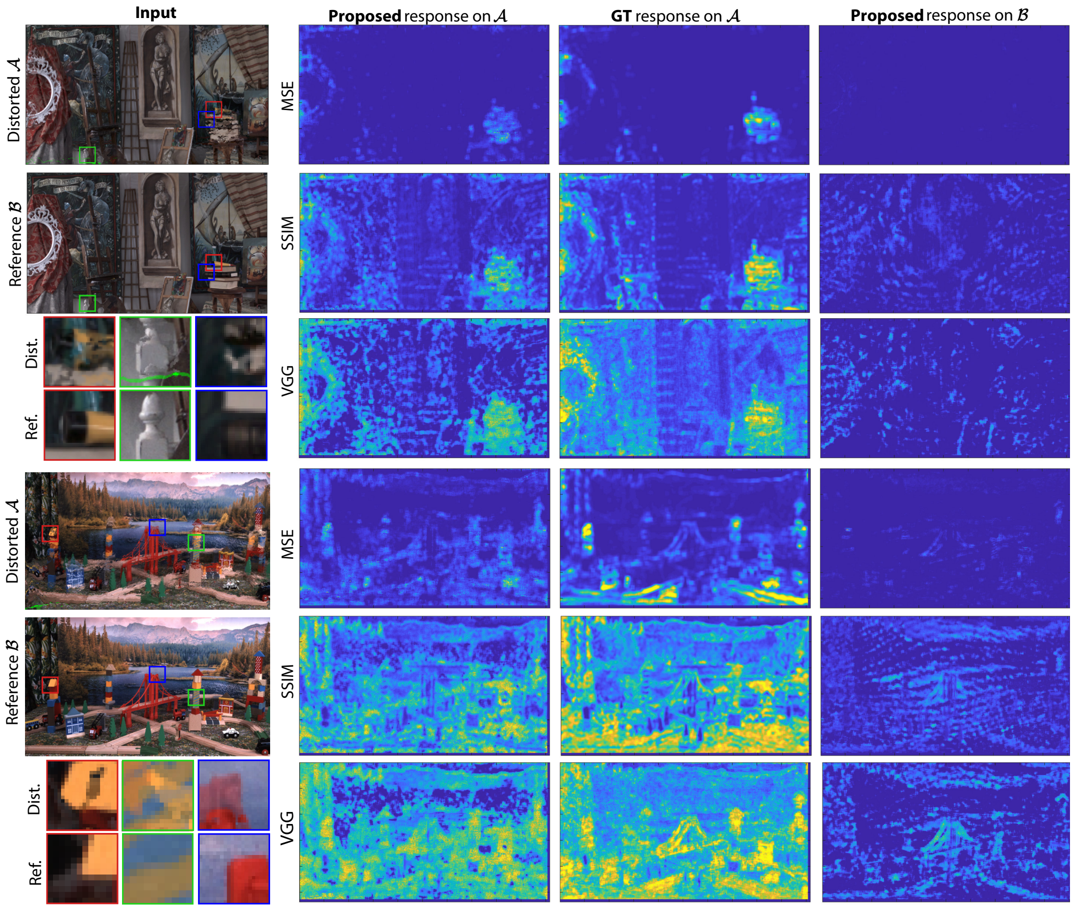


FIGURE 5.5: *Comparing the response to a pair of an image $\mathcal{A}$ and its distorted version $\mathcal{B}$* (first column). *The response of the proposed method* (second column) *is similar to the ground truth* (third column). *When executed on the clean reference* (fourth column), *only very few false positives are reported.*

5.3.3 Qualitative Results

Example metric outputs Figure 5.5 shows an analysis of the response of all metrics to two different LFs from the test set. The first column shows the distorted input $\mathcal{A}$ at the top, below the hidden reference $\mathcal{B}$, and below these three insets from both. The second column shows a predicted response of the proposed method $\mathcal{A} \ominus \mathcal{B}$ for different metrics: MSE on top, followed by SSIM and VGG. A false-color coding, where cold colors indicate a low response and warm colors indicate a high response, is used. The third column shows the GT response for the same. It is evident that there is a similarity between the prediction and the ground truth. While it slightly errs towards conservative, i. e., miss a few errors. How some of these errors are only false findings, i. e., a limitation of the metrics becomes apparent from the user study to follow. The last column shows a sanity check where the proposed metric is exposed to hidden reference image $\mathcal{B}$. The hidden reference obviously does not contain any error, and consequently, reporting one is a false positive. It shows the metric has a response in areas that are correct but look like IBR artifacts, but in most areas has no response. In summary, this indicates that the metric is able to localize and scale errors to a hidden reference in images with artifacts while avoiding producing a signal when facing clean images. It might appear that MSE has fewer false positives than SSIM or VGG when inspecting the last column; simply more deep blue, very close to perfect in the first row. However, such a trend is not supported by the numbers in Table 5.1 or the plots in Figure 5.4. The true reason for this impression might be

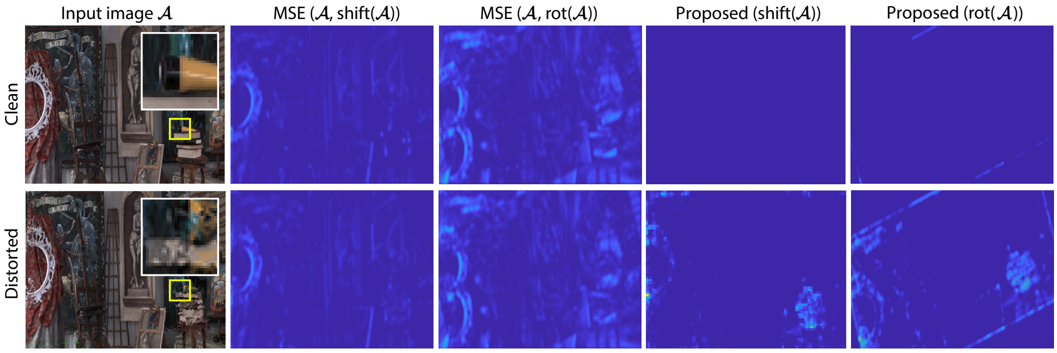

Input image $\mathcal{A}$
MSE ($\mathcal{A}$, shift($\mathcal{A}$))
MSE ($\mathcal{A}$, rot($\mathcal{A}$))
Proposed (shift($\mathcal{A}$))
Proposed (rot($\mathcal{A}$))
Clean
Distorted

FIGURE 5.6: *Transform-invariance of the proposed approach: When computing the distance between a clean input image $\mathcal{A}$* (first column, first row) *and a misaligned reference $\mathcal{B}$ (not shown here, 20-px shifted or 20 degrees rotated copy of $\mathcal{A}$), a common metric such as MSE will show a strong response* (first row, second and third columns). *Such a response is numerically correct but far from human assessment, which would be more similar to the response of proposed metric* (first row, fourth and fifth columns). *Symmetrically, repeating the experiment on a distorted input, the proposed approach correctly localizes the distortions around the books* (inset) *as if the reference had been aligned.*

that the SSIM and VGG response simply have a larger receptive field per-se: MSE is per-pixel while VGG is affected by up to 32 × 32 pixels. Even the ground truth response is dense (less deep blue). Consequently, the metric prediction, in case of error, also makes spatially more extended, denser mistakes.

Transformation-invariance Surprisingly, results produced by the presented method can turn out to be better than their own supervision, as the method is forced to come up with strategies to detect problems without seeing the reference. This makes it immune to a common issue of many image metrics: misalignment [Kellnhofer et al., 2016b]. Even a simple shift in image content will result in many false positives for classic metrics (Figure 5.6). An image that has merely been shifted is reported to be very different from a reference by all the metrics used in this chapter; however, it is less different from the reference compared to the one with IBR artifacts. In contrast, the proposed metric does not care about transformation, but when IBR artifacts are added, they are detected. As the metric is oblivious to the ground truth, it is not subject to such a misconception. While not quantifiable, the result is arguably more similar to human judgment, as indicated by the user experiment in the next subsection.

5.3.4 User Study

A user experiment is done to validate that the predicted responses of the proposed metric spatially correlate with the visibility of artifacts to human subjects. The human responses are quantified by means of per-pixel annotations, which are painted on top of images showing IBR artifacts. Note that no user responses were used for training.

Methods Naïve users were asked to use a binary painting interface to mark errors in a rendered image for each of the six LFs of the test dataset in an open-ended session that took 15 minutes on average. The binary responses are then averaged into a continuous fraction (percentage) of users that detected the location of the artifacts.

Analysis Asking $N = 10$ users, the Pearson linear correlation R is computed (higher values are better; statements are highly significant as the correlation is computed

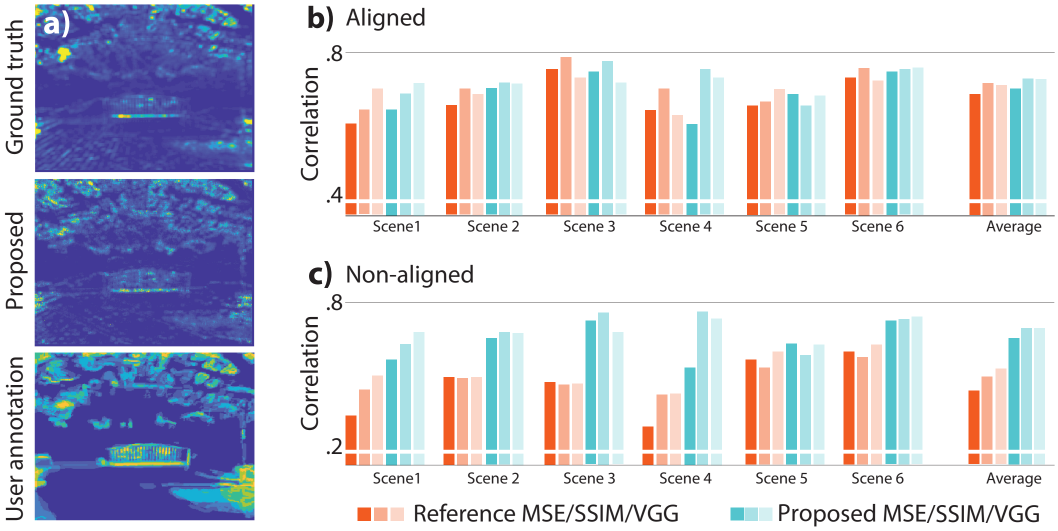


FIGURE 5.7: *Exemplary user study result* (a). *Correlation (significant, $p < .001$) of MSE/SSIM/VGG and user responses (red) compared to the predictions of the proposed metric for the three metrics (blue) for different scenes and as an average across scenes to the right* (b). *It can be seen that in the non-aligned conditions, these differences get stronger* (c).

on a high number of image pixels) and reported in Figure 5.7-b. It appears that for many scenes, as well as for the average across scenes, the proposed method has a higher correlation with user annotation than the metric it was supervised on. This could be due to the fact that the network had learned to become independent of a reference, similar robustness that the HVS employs. There is no clear trend on which of the metric response predictions correlates the most with the user annotations. The differences between scenes, however, seem more pronounced. The experiment is repeated with a non-aligned reference (shifted a mere 20 px to the right), and the correlations are reported in Figure 5.7-c. It can be observed that the proposed metric shows higher correlations for all metrics across different scenes, indicating more robustness to alignment issues when predicting user responses.

Perceptualization Finally, a linear correlation R is computed by fitting a model $x_i = a \cdot y_i + b$, where x_i is the user response and y_i is the response of the proposed metric for pixel i. This allows a "perceptualization" of the metrics response. Fitting multiple models a, b in a leave-one-out protocol to 5 of all 6 scenes produces an average error of .05/.04/.02 for MSE/SSIM/VGG, respectively, indicating that this perceptualization generalizes to some extent.

5.3.5 Other Architectures

Alternative architectures are explored with or without balancing. A simple solution would be to use a supervised image translation network such as Pix2Pix [Isola et al., 2017] to map from entire IBR images to the metric response. Unfortunately, training these on the training data used in this chapter converges to a flat response of zero, as artifacts are too rare and subtle to be picked without the suggested balancing. Future work could investigate combining the balancing with other architectures.

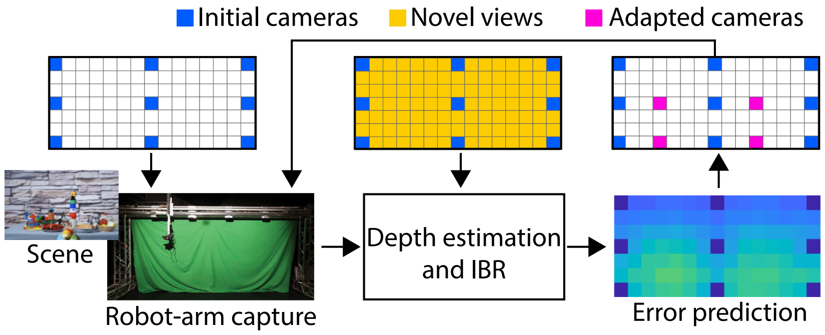


FIGURE 5.8: *Proposed pipeline for adaptive LF sampling by bounding the reconstruction error predicted by the proposed metric.*

5.4 Applications

In this part, two practical applications of an NR-IQM will be demonstrated in light field production. The first is accelerating automated adaptive LF capture (Section 5.4.1), and the second employs the proposed NR-IQM as feedback in an interactive depth manipulation system (Section 5.4.2).

5.4.1 Adaptive Light Field Capturing

Capturing a dense set of input view images results in a high-quality reconstruction but remains a time-consuming process or may require a bulky setup. The main observation is that not all input view images contribute equally to the reconstruction of novel-view images. The proposed metric helps identify and capture these. Images from views dominated by planar diffuse surfaces can reliably be predicted from images taken from other views showing this very same surface. Hence, dense capturing from these views is needed and thus not efficient. In contrast, occlusions and specularity can be more challenging because it must be ensured that each scene element is visible in at least two camera views (when using multi-view stereo) to compute depth. Sparse capturing from these views would sacrifice the reconstruction quality. To both of these ends, this chapter introduces an adaptive capturing mechanism as illustrated in Figure 5.8 to capture an image for a view only if it cannot be extrapolated from other views.

Setup This chapter studies adaptive capturing by means of a large-scale translation stage equipped with a digital camera. The position of the camera can be controlled with a precision of 80 μm in the horizontal and 50 μm in the vertical direction. This allows for a very dense capture of the scene. While this takes longer to capture, it serves as a unique baseline as it is possible to compare the metric prediction to the actual error present.

Procedure First, a sparse set of images are captured, and the depth maps are estimated for all views. Then, a set of intermediate views are rendered using a DIBR method [Dabała et al., 2016], and the reconstruction error is measured for each rendered view. All pixels are simply averaged in each view image, producing a single scalar value. The capturing grid is then subdivided into smaller regions where the average predicted reconstruction error is larger than a given threshold. This process is repeated until the desired quality is achieved. By this approach, the number of captured views can be substantially reduced, and the scene is recaptured only at

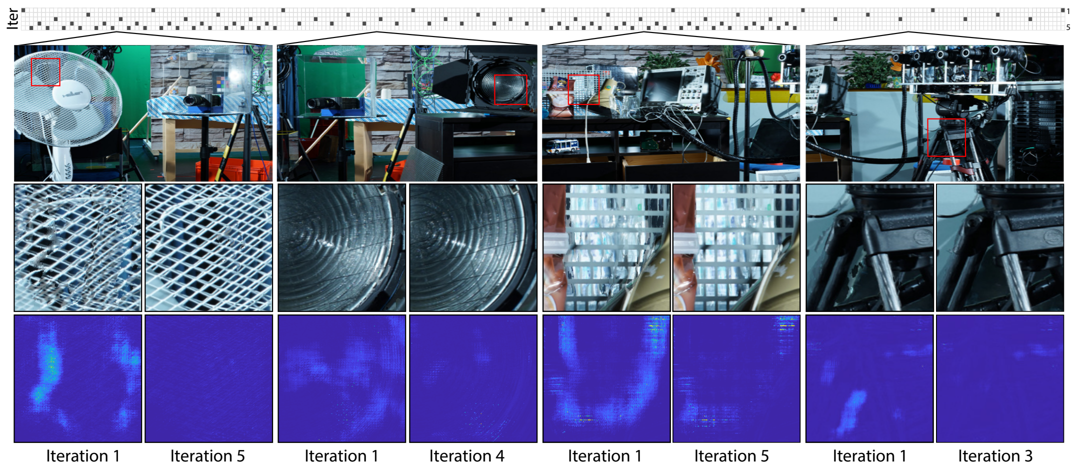


FIGURE 5.10: *Adaptive panoramic light field capturing: The top row shows a grid indicating the camera placement at different iterations. The second row shows the selected rendered views based on the keyframes that are captured. The insets in the third row show the marked patches from the rendered views in the first iteration and in the iteration where the desired quality is achieved. The fourth row shows the metric predictions for the corresponding patches in each iteration.*

locations where reconstruction is poor. Predicting the reconstruction error of the novel view is the key to making such an approach work. Classic full-reference image quality metrics require a dense capture to provide reference images to compute the error, which is not practical as the goal in this thesis is to reduce the number of captured images in the first place. In contrast, the proposed no-reference metric in this chapter can measure the error in the novel view images without providing their reference images, resulting in an efficient approach.

Evaluation To evaluate the effectiveness of the metric in this application, two LF sequences are captured and adapted according to the MSE metric.

Array The scene is captured with an array of 7×15 images as shown in Figure 5.8 (left). Figure 5.9 shows the ground truth MSE (left) and the proposed metric prediction (right), where each grid element denotes a camera position. The dark blue grid elements indicate the camera positions where actual keyframes were captured while rendering has been performed for all remaining intermediate positions. As can be seen, the distribution of reconstruction error, as predicted by the metric, correlates well with the ground truth. Figure 5.8 (right) shows new camera locations that are required to reduce the true average reconstruction error below .004.

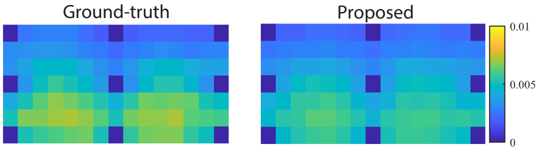


FIGURE 5.9: *Reconstruction error of intermediate novel views.* Left*: Ground truth MSE values,* right*: The proposed metric MSE prediction.*

Panoramic The proposed metric can potentially be beneficial for efficient panoramic (i. e., one-dimensional, linear) light field capturing. As it is shown in Figure 5.10, depending on the scene content, not all regions in the scene require equally dense camera placement. The metric successfully guides the capturing setup to take more photos in the regions with thin structures, substantial disocclusions, or specularites where accurate reconstruction is highly challenging. Overall, capturing 76 instead of

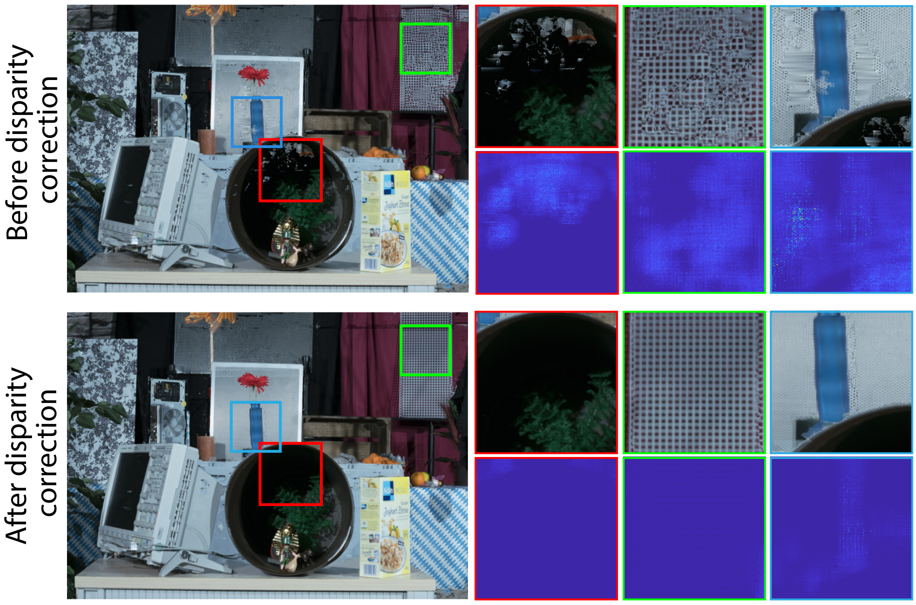


FIGURE 5.11: *Interactive depth adjustment. The marked patches show the regions in the rendered view where the proposed method predicts the MSE (top), and the bottom row shows the corresponding patches after applying the manual disparity refinement.*

720 images – a sparsity of 10.5 % – reduces the total capture time from 59 minutes to 4.9 minutes, i. e., by 91 %.

5.4.2 Interactive Depth Adjustments

Long acquisition times involved in capturing dense light fields make it a tedious and impractical task for some application fields. One such field is movie production, where the presence of highly dynamic scenes and time pressure discourages the use of dense light fields, and in such cases, only sparse light field capture using video camera arrays is seen as a convenient solution.

Unfortunately, automatic error-free light field reconstruction from a sparse capture is still an unsolved problem. To this end, there are ongoing research efforts to address the challenges, such as the estimation of disparity in the presence of homogeneous areas, repetitive structures, fine-grained objects, or specularities. In such cases, interactive disparity estimation improvement seems to be the most promising solution to achieve a high-quality view rendering [Wildeboer et al., 2011; Kap-Kee, 2015; Lin et al., 2012; Cao et al., 2011]. However, this requires detecting possible view rendering artifacts as fast as possible to reduce the post-processing time. As shown in the right-most image of the second row in Figure 5.10, spotting an artifact is not a trivial task and sometimes requires carefully scanning the view rendering result. The presented quality estimation metric can significantly simplify this process by allowing the automatic analysis of several rendered views. By observing the predicted visibility map, which identifies the local distortions, the user can quickly spot the problematic regions. Using a post-production software suite [2] to perform an interactive view rendering with only a small subset of cameras allows detecting the captured view responsible for the error. The inspection of the corresponding disparity map followed by an approach similar to Wildeboer et al. [2011] and Kap-Kee [2015] finally allows fixing the view rendering error. This is achieved by the manual creation of a geometry proxy in 3D space for objects whose disparity map could not

[2] https://www.iis.fraunhofer.de/realception

be computed automatically. The proxy is then used to bound the admissible depth values for subsequent disparity estimation.

The results of this procedure are illustrated in Figure 5.11. The contained repetitive structures are very challenging for automatic disparity estimation and consequently lead to many view rendering artifacts, as clearly indicated by the depicted error map. To solve these issues, a user has added proxy-based disparity constraints for the waste basket (and the contained figurine), the grid structure behind the flower, and the grid structure in the upper right corner of the image. By these means, a much better view rendering could be achieved, as shown in Figure 5.11. The proposed metric has reduced the time required to find those reconstruction errors, leaving more time for a user to correct them.

5.5 Conclusion

This chapter has demonstrated that with properly adjusted training data (prioritization and natural supervision), a CNN can learn how to predict the difference between an image to a hidden reference. The proposed approach is independent of the metric used and can reproduce MSE, SSIM, and VGG prediction. Other metrics such as HDR-VDP-2 [Mantiuk et al., 2011a] or the CNN-based metric of Wolski et al. [2018a] would likely be predictable in a similar fashion. Such a metric can be applied to several applications. As demonstrated, this includes adaptive light field sampling of complex scenes and interactive depth editing. Moreover, since, in contrast to any existing no-reference metric, the proposed approach provides a predicted error map, this opens the potential for many novel applications, such as interactive or automatic view rendering error correction. In future work, it would be interesting to overcome the limitations of the paired input, eventually using an adversarial [Goodfellow et al., 2014] design, and learn the prediction only from pairs and without the metric or only from pairs of undistorted-metric or distorted-metric.

Chapter 6

A Perception-Driven Decomposition for Multi-Layer Displays

Multi-focal plane and multi-layered light-field displays are promising solutions for addressing all visual cues observed in the real world. Unfortunately, these devices usually require expensive optimizations to compute a suitable decomposition of the input light field or focal stack to drive individual display layers. Although these methods provide near-correct image reconstruction, a significant computational cost prevents real-time applications. A simple alternative is a linear blending strategy which decomposes a single 2D image using depth information. While it provides real-time performance, it generates inaccurate results at occlusion boundaries and on glossy surfaces. This chapter proposes a perception-based hybrid decomposition technique that combines the advantages of the above strategies and achieves both real-time performance and high-fidelity results. The fundamental idea is to apply expensive optimizations only in regions where it is perceptually superior, e.g., depth discontinuities at the fovea, and fall back to less costly linear blending otherwise. This chapter presents a complete, perception-informed analysis and model that locally determines which of the two strategies should be applied. Later. a new synthesis method is proposed to perform image decomposition. The results are analyzed and validated in user experiments on a custom multi-plane display.

6.1 Introduction

In recent years, head-mounted displays (HMDs) have emerged as a major virtual (VR) and augmented reality (AR) technology, and currently, they have many potential applications in a diverse set of fields, including gaming, video, medicine, simulation, and aviation. Stereo HMDs can display 3D content with binocular disparity, which is one of the critical cues for stereopsis and depth perception of the brain. As the use of binocular disparity in HMDs has already been successfully commercialized, research efforts are recently getting directed toward enhancing 3D perception by introducing support for other types of cues. A critical requirement for a faithful reconstruction of virtual 3D content is the reproduction of correct accommodation cues, which allows a natural depth perception by triggering changes in the focal distance of the eye [Lambooij et al., 2009; Banks et al., 2016]. However, developing HMDs with correct accommodation cues is an extremely challenging task due to the limitations imposed by optics on the hardware design. Any improvement in this direction must satisfy the requirements of a consumer product, such as having a small form factor but usually, there is a trade-off between these requirements and the optical capabilities of the display, such as the field of view (FOV) and display resolution [Hua, 2017]. In addition to these hardware challenges, generating 3D content for such displays is another important issue since it requires efficient processing of a larger amount of data compared to 2D images. Furthermore, there is always a concern about compatibility with different display architectures [Kramida, 2016].

Recent studies have shown that multi-layer displays, such as multi-plane displays or light-field displays, are practical solutions for HMDs to provide near-correct accommodation cues [Hua, 2017; Kramida, 2016]. A crucial step of rendering in a multi-layered system is the decomposition of an input scene into layers for a correct 3D perception [Narain et al., 2015]. The most straightforward decomposition method is linear blending (LB), where the input is a single viewpoint image with a depth map [Akeley et al., 2004; MacKenzie et al., 2010]. Although this technique is computationally efficient, it usually fails at occlusion boundaries or non-Lambertian surfaces. To overcome this limitation, two approaches have been proposed: retinal optimization (RO) [Narain et al., 2015; Mercier et al., 2017] and light-field synthesis (LFS) [Huang et al., 2015b; Lee et al., 2016], which optimize the decomposition based on a focal stack and a 4D light field, respectively. The improved quality comes at a high computational cost of the optimization (5 Hz at 512×512 resolution as reported by Mercier et al. [2017]) and input generation. In addition, although these techniques perform better at occlusion boundaries [Zannoli et al., 2016], they may perform worse in driving the eye accommodation [Mercier et al., 2017].

In order to combine the desired features of different algorithms, the most promising solution would be designing a hybrid decomposition technique. Such an approach could select the decomposition method locally depending on the scene content in order to obtain the best perceptual quality possible. For real-time rendering applications, this type of hybrid decomposition has to be implemented efficiently. Thanks to the recent developments in GPU hardware, new cards introduce separate cores for massively computational tasks (e.g. recently announced Nvidia RTX platform), which encourages such content-dependent local optimizations to be performed in parallel to the traditional graphics pipeline. However, in order to propose a robust hybrid algorithm, a clear understanding of the perceptual quality differences among various decomposition methods is required. So far, there has been very little research comparing the visual quality of LB, LFS, and RO methods. In addition, the conditions which lead to the failure of the LB method at occlusion boundaries are not thoroughly investigated in previous works.

To address these issues, this chapter provides a perceptual evaluation of different decomposition methods and proposes a perception-driven hybrid decomposition technique. In the first part, as a preliminary step towards the hybrid decomposition, an improved gaze-contingent LFS method is introduced to generate the input viewpoints exclusively inside the pupil. This solution achieves similar results to RO but with a significantly lower amount of computational cost. Consequently, the RO method is skipped, and only the gaze-contingent LFS is considered. In the second part, a perceptual evaluation methodology is designed to determine for which multi-plane display configurations and scene content the inexpensive LB can be applied without a loss of visual quality and when the gaze-contingent LFS is necessary. In the evaluation analysis, only texture and occlusion boundaries are considered, as they are more responsible for driving accommodation [Mathews and Kruger, 1994; MacKenzie et al., 2010] and depth order perception [Zannoli et al., 2016]. Through a series of perceptual experiments, a detection threshold is derived, which then allows the establishment of the selection rule for the decomposition algorithm such that:

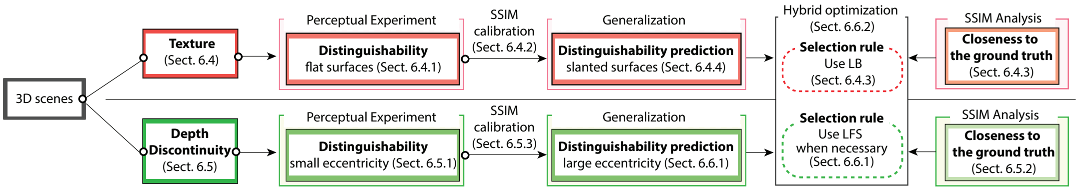


FIGURE 6.1: *Overview of the methodology in the proposed approach. First, the 3D scenes are analyzed based on texture and depth discontinuity. The limited cases are investigated with perceptual experiments and extended to the general cases through the prediction using custom-calibrated SSIM. The selection rules are then derived from the predicted distinguishability between LB and LFS and closeness to the ground truth obtained by SSIM analyses. Finally, the hybrid optimization framework is developed based on the selection rules.*

1. when LB and LFS are visually indistinguishable, LB should be selected, and
2. when LB and LFS are distinguishable, the method that yields result closer to the ground truth should be chosen.

Based on the selection rule, the proposed hybrid decomposition approach is described in order to combine linear blending and light-field synthesis methods. To further improve the performance, the foveal and peripheral vision characteristics are also taken into account. Consequently, this chapter proposes a content-dependent and gaze-contingent hybrid decomposition algorithm for multi-layered accommodative displays, which enables real-time rendering performance and high-quality reconstruction.

The main contributions of this chapter are:

- a gaze-dependent viewpoint sampling of LFS for enhanced reconstruction quality,
- a series of targeted perceptual experiments that measure the differences in the visual quality obtained by LB and LFS for various spatial frequencies, luminance contrasts, depth configurations, and eccentricities,
- a domain-specific structural similarity index (SSIM) calibration for visible difference prediction between the LB and LFS that generalizes perceptual insights beyond the scope of the perceptual experiments,
- a unified optimization framework for the LB and LFS decompositions,
- an efficient adaptation of the simultaneous algebraic reconstruction technique (SART) to CUDA for real-time decomposition.

6.2 Overview

The goal of this chapter is to develop the perceptual evaluation methods and the hybrid decomposition of a gaze-contingent LFS (Section 6.3) and LB. The overall pipeline is outlined in Figure 6.1. First, the 3D scenes are analyzed based on texture (Section 6.4) and depth discontinuity (Section 6.5), which are important for quality perception and driving accommodation. While it is ideal to evaluate all possible scenarios through perceptual experiments, as the parametric space of texture and depth discontinuity is vast, perceptual experiments are performed on distinguishability between LFS and LB in a limited parametric space. To explore the full space, a visual quality metric SSIM [Wang et al., 2004a] is calibrated to predict the experimental outcomes and predict distinguishability in general cases. The employment of SSIM is motivated by a recent study showing that advanced metrics such as SSIM and HDR-VDP [Mantiuk et al., 2011b] provide a similar and good prediction on a narrow, well-defined task after proper training and calibration with relevant perceptual data [Adhikarla et al., 2017]. Specifically, the perceptual experiments are conducted for flat textured surfaces in Section 6.4.1 and depth discontinuities at small eccentricities in Section 6.5.1. Through the calibrations of SSIM in Section 6.4.2 and Section 6.5.3, the distinguishability in general cases such as slanted textured surfaces in Section 6.4.4 or depth discontinuities at large eccentricities in Section 6.6.1 are predicted. For selecting a proper algorithm when LFS and LB are distinguishable, SSIM analysis is performed to find the algorithm closer to the ground truth in Section 6.4.3 and Section 6.5.2. Finally, the best decomposition algorithm is determined, which is LB for textured surfaces and LFS for depth discontinuities depending on depth difference, luminance contrast, and eccentricities. Since the transition between LFS and LB is required at depth discontinuity, the selection rule is developed in Section 6.6.1, and the hybrid optimization framework is proposed in Section 6.6.2.

6.3 Gaze-Contingent Light Field Synthesis

In order to propose the hybrid decomposition strategy, the existing decomposition methods are evaluated for multi-layer displays with respect to computational complexity and visual quality criteria. LB is a fast decomposition method, and it is suitable for regions where an accurate reconstruction is not required. On the other hand, when a high-quality reconstruction is required, the hybrid decomposition algorithm should select more complex methods such as LFS and RO. While LFS reconstructs a sparse set of light field views, RO reproduces a focal stack rendered from dense light fields inside the pupil. Although LFS is computationally more efficient than RO, a recent study shows that LFS suffers from contrast degradation, and RO might be a better alternative for preserving the contrast [Lee et al., 2017]. However, the loss of contrast in LFS might originate from using a wide eye box that is larger than the pupil size, where some of the viewpoints fall outside the observer's pupil [Huang et al., 2015b; Lee et al., 2017]. On the contrary, RO provides a higher level of contrast by rendering the dense light fields exclusively inside the pupil and further processing them to generate focal images at multiple depths.

Both LFS and RO might be used to produce high-quality outputs when required by a hybrid decomposition algorithm. But the issue of contrast degradation has to be addressed to get the benefit of LFS. To this end, a gaze-contingent viewpoint sampling approach is proposed to enhance LFS image quality compared to the implementation using a wide eye box. The approach aims to generate light-field viewpoints only inside the pupil, using the pupil position from an eye tracker. This solution effectively avoids contrast degradation in the LFS method. The gaze-contingent method requires the addition of an eye tracker device to the hardware, but as it is discussed in Section 2.4, this requirement applies to any practical decomposition method for multi-focal displays.

The quality of the proposed gaze-contingent LFS method is validated using simulated contrast curves of the reconstructed images from various decompositions (Figure 6.2). The contrast curves show the magnitude of luminance contrast for different spatial frequencies with respect to accommodation depth. In accommodative displays, the contrast of the images should be maximized at the object plane because a higher gradient of the contrast curve more effectively drives the accommodation

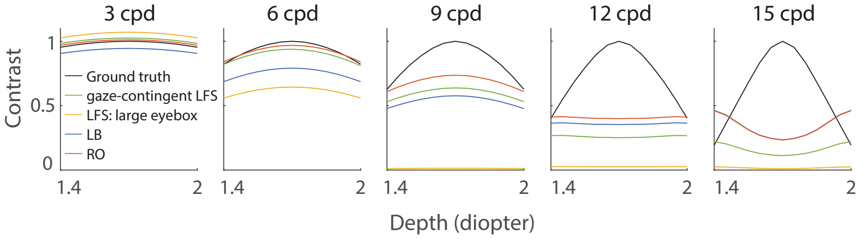


FIGURE 6.2: *Contrast curves for various optimization algorithms for various spatial frequencies. While LFS with a large eye box exhibits significant contrast reduction for high spatial frequencies, gaze-contingent LFS shows much higher contrast values over the entire frequency range, providing similar quality to LB or RO.*

toward the object plane [Ravikumar et al., 2011]. In order to obtain the contrast curves, first retinal images are generated at various focal depths between 1.4 D and 2 D. The 0.6 D gap is chosen since it is widely used to attain sufficient resolution for triggering accommodation at intermediate planes and minimizing the number of display planes [MacKenzie et al., 2010]. Then, the Fourier transform is used to extract the luminance values at the target spatial frequency. Finally, all values are normalized with the peak value of the contrast curve of the ground truth. A similar analysis has been performed by Lee et al. [2017]. The ground truth, gaze-contingent LFS, LFS with large eye box, LB, and RO are used for the analysis in this chapter. During the evaluations, the number of viewpoints is set to 13 inside a 4 mm-diameter pupil for gaze-contingent LFS. It is empirically found that using a larger number of views does not improve the image quality for gaze-contingent LFS. The large eye box case assumes 5 × 5 viewpoints inside an 8 × 8 mm eye box. The sinusoidal patterns of various spatial frequencies are projected in the middle plane between two display layers placed at 1.4 D and 2 D, respectively. The resolution is set to 15 cpd, which is the maximum resolution supported by the display. The analysis shows that LFS with a large eye box significantly degrades the quality beyond approximately 6 cpd. In contrast, the gaze-contingent LFS provides a quality comparable to RO or LB for 3–9 cpd, which is the critical range for driving accommodation [Mathews and Kruger, 1994; MacKenzie et al., 2010]. The noticeable deviations are observed for high spatial frequencies; however, all algorithms already fail to reproduce the correct contrast curve due to the limited frequency support of the display. The maximum reproducible frequency increases as the distance between the displays decreases [Narain et al., 2015]. Therefore, this suggests that the gaze-contingent LFS attains the quality offered by RO and is suitable for use with LB in a hybrid decomposition approach. From now on, the gaze-contingent LFS is simply referred to as LFS.

6.4 Effect of Texture on Decomposition

LB method performs poorly in regions that are affected by occlusion [Narain et al., 2015]. However, it preserves the contrast relatively well in other regions (see Figure 6.2). Therefore, the use of LB still can be a good option on textured regions except at those problematic regions around occlusion boundaries where LFS gives a better result than LB. The previous work shows that the difference between the two methods is noticeable when the content has high spatial frequencies [Narain et al., 2015]. Given this observation, a spatial frequency threshold is found in order to switch from one decomposition method to the other in order to get the optimal quality. The analysis in this chapter includes showing both flat and slanted surfaces with various

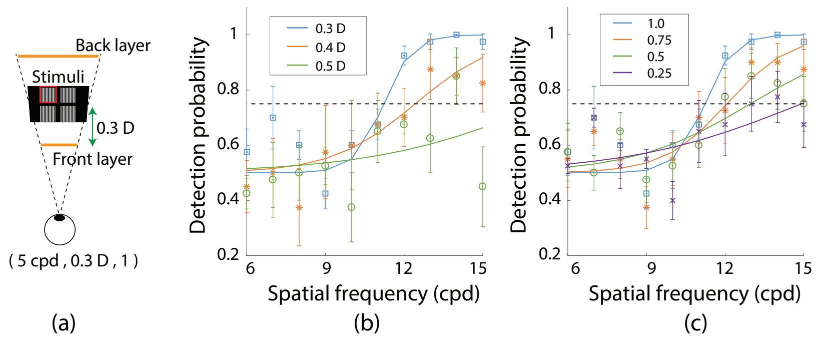


FIGURE 6.3: *The configuration of displays and stimuli (a). The probability of detecting the difference between LB and LFS methods for various depth (b) and contrast levels (c).*

slopes. First, a perceptual experiment is conducted to investigate the conditions when two algorithms are indistinguishable for an observer viewing flat surfaces on the prototype display. Then, a set of analyses is performed using an objective quality metric in order to generalize the observed findings to slanted surfaces. In addition to allowing the evaluation of different scene configurations, the use of an objective metric helps avoid any issues due to the lack of ground truth as well. Unfortunately, there is no domain-specific metric designed for such an evaluation. Therefore, the data obtained from conducted perceptual experiments are used to calibrate an existing full-reference quality metric. The Structural Similarity (SSIM) metric is employed for this purpose, as it is widely used for the objective evaluation of visual quality in other domains [Wang et al., 2004a].

6.4.1 Perceptual Experiment

In order to test the distinguishability between LB and LFS, a perceptual experiment is conducted on a two-plane prototype display in a monocular viewing setting. The stimuli consist of two pairs of flat sinusoidal patterns. One pair contains two identical patterns generated using only LFS, and the other pair contains two different patterns generated using LFS and LB. The experiment is done with the two-alternative forced choice (2AFC) procedure, and the participants are asked to select a pair of patterns that look different from each other. While two pairs are shown at the top and bottom positions, the order of patterns is completely randomized among trials. Then, using the number of correct responses for different combinations of Michelson contrast, stimuli depth, and spatial frequencies from 6 to 15 cpd, the probability of detection is computed. Figure 6.3 (a) shows representative stimuli used in the conducted experiment, where the LB stimulus has a red frame around the pattern. For that stimuli, the correct response is the top pair. In total, five participants took the experiment. All participants were naïve, paid, and have a normal or corrected-to-normal vision. The display resolution is 15 cpd, and the display separation is set to 0.6 D. Section 6.7.2 describes more details on the experimental setup.

The frequency that corresponds to 75% detection probability is taken as the detection threshold. It is computed by fitting a psychometric sigmoidal function to the collected data. The detection probabilities from the experiment and fitted sigmoids are shown in Figure 6.3 (b-c). Figure 6.3 (b) is obtained for various depths of the stimuli, while the Michelson contrast is fixed at 1. The depth is measured as the distance from the front display, where 0.3 D corresponds to the middle plane. The frequency threshold has the smallest value for the middle plane stimuli, where the reconstruction quality of decomposition algorithms is the lowest [MacKenzie et al.,

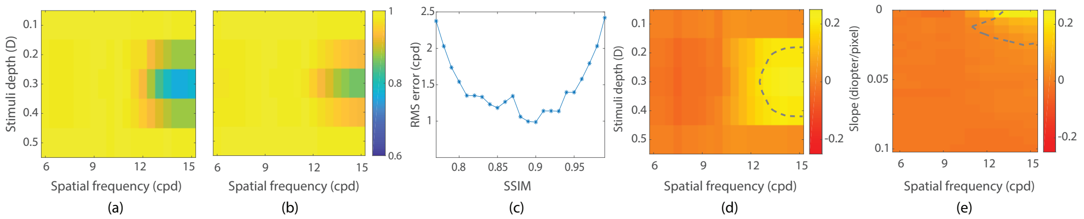


FIGURE 6.4: *The minimum SSIM values of the SSIM maps between the LB and LFS for (a) contrast = 1 and (b) contrast = 0.5. (c) The RMS error between the predicted cutoff frequencies and experimental results. (d) The SSIM value against the ground truth. The positive region: LB is closer to the ground truth. The negative region: LFS is closer to the ground truth. The yellow region bounded by the dashed line represents the conditions where the two methods are distinguishable. (e) The ground truth comparison for slanted surfaces.*

2010; Narain et al., 2015]. Figure 6.3 (c) shows the results for various contrasts, while the stimuli depth is fixed to 0.3 D. These results indicate that the frequency threshold increases as the contrast decreases.

6.4.2 Calibrating SSIM

The above experiment considers only a small subset of different texture and depth configurations that can occur in complex scenes. One option to investigate a wider range of stimuli is performing more extensive perceptual experiments. Instead, this chapter relies on image quality metrics which have been recently demonstrated to be successful in simulating the visibility of different artifacts when calibrated on a problem-specific dataset [Adhikarla et al., 2017; Wolski et al., 2018b]. Consequently, the SSIM metric is adapted for predicting distinguishability between LFS and LB methods and used in further investigation. An additional and critical benefit of such a strategy is that it allows comparing the decomposition techniques to ground-truth images. This is challenging using perceptual experiments due to the lack of a reference light-field display.

The proposed SSIM-based metric takes as an input two perceived images, simulated for a specific focus, and computes an SSIM map. The metric later takes the minimum value of the map as the dissimilarity measure between the two images. First, this procedure is used to simulate the previous experiment. To this end, the dissimilarity index is computed between LFS and LB methods for different combinations of luminance contrast, frequency, and depth, assuming that the observer focuses on the target object plane. Figure 6.4 (a) and (b) show the results for the stimuli when the Michelson contrast of the stimuli is fixed to 1 and 0.5. Smaller values of the maps indicate a larger difference between the results of the two methods. Similar to the result in Figure 6.3 (b), the transition behavior is observed around 12 cpd for the middle plane (0.3 D), and the transition point moves towards higher frequencies for stimuli closer to the display plane. Figure 6.4 (b) reveals that the SSIM values overall increase for lower luminance contrast, which is in agreement with Figure 6.3 (c).

To use the metric as a visibility predictor, an SSIM threshold is selected such that it corresponds to the visibility threshold. In other words, all the image regions for which the SSIM index is above the SSIM threshold should contain only invisible differences while the regions with smaller SSIM values contain visible differences. The optimal SSIM threshold is determined as the value which minimizes the RMS error between the predicted and measured frequency thresholds obtained in the experiment in Section 6.4.1. The lowest error was obtained for the SSIM threshold of 0.9. The rest of the evaluations in this chapter are based on this value.

6.4.3 Comparison with Ground Truth

The SSIM analysis is used to select the algorithm which is closer to the simulated ground truth. First, two SSIM maps are obtained in the same way as Figure 6.4 (a) by comparing LFS with ground truth and LB with ground truth. Then the pixel-wise difference between the SSIM map of LFS and LB is considered. The result is shown in Figure 6.4 (d). It is vivid that LB is better at reproducing the ground truth in the region inside the dashed half-circle, where LFS and LB are distinguishable according to the previous analysis. Outside this region, LFS performs better, particularly at low frequencies, but it is still acceptable to use LB due to indistinguishability. Interestingly, these results suggest that the computationally efficient LB provides higher fidelity reconstruction compared to the computationally expensive LFS on high spatial frequencies. Although previous study [Narain et al., 2015] and the conducted analysis (Figure 6.2) suggest that such high contrast reconstruction can lead to incorrect contrast curve, the eye accommodation is dominantly driven by 4-8 cpd and the failures of LB in reproducing contrast gradient at high frequencies are negligible [MacKenzie et al., 2010]. Furthermore, it should be noted that even LFS or RO fails to reproduce the correct contrast curves in such cases, as shown in Figure 6.2. Therefore, LB is selected as the best algorithm which provides high contrast in retinal images.

6.4.4 Generalization to Slanted Surfaces

In many studies, the quality of reconstruction has been tested on planar surfaces at a fixed depth [MacKenzie et al., 2010; Narain et al., 2015; Lee et al., 2017]. However, most 3D scenes contain various slanted surfaces. Hence, the analysis in this chapter is also extended to slanted surfaces with various slopes. At each spatial frequency, slanted surfaces up to the maximum slope of 0.1 D/pixel are created. In the designed display prototype with the 0.6 D separation, this maximum slope corresponds to a 6-pixel-wide slanted surface extending from the front display to the back display. Since a fewer number of pixels cannot fully represent one cycle of the minimum spatial frequency, the steeper surfaces are regarded as occlusion boundaries. The previous analysis of flat surfaces compared the focal images at the target stimulus plane. In the presence of a slanted surface, however, the reconstruction quality should be checked at every possible focal state. Therefore, seven focal images are computed between two layers with a step size of 0.1 D. For each focal image of each algorithm, the minimum SSIM value is found in the comparison against the ground truth focal image. Among all focal depths, based on the minimum SSIM, the worst case is found. Then, the difference between the SSIM map of LFS and LB is measured to compute the closeness to the ground truth, as shown in Figure 6.4 (e). The border of the distinguishable region is indicated with the gray dashed line. Similar to the flat surfaces, two methods are distinguishable for high spatial frequency texture at low slopes. Inside this distinguishable region, LB still performs better than LFS.

In summary, the conducted analyses reveal that for flat and slanted surfaces with sinusoidal patterns as textures, LB and LFS methods are distinguishable only for high spatial frequency textures, and LB provides higher fidelity reconstruction when they are distinguishable. Since this holds for foveal vision, it is evident that the same algorithm holds for peripheral vision because contrast sensitivity declines in the peripheral region.

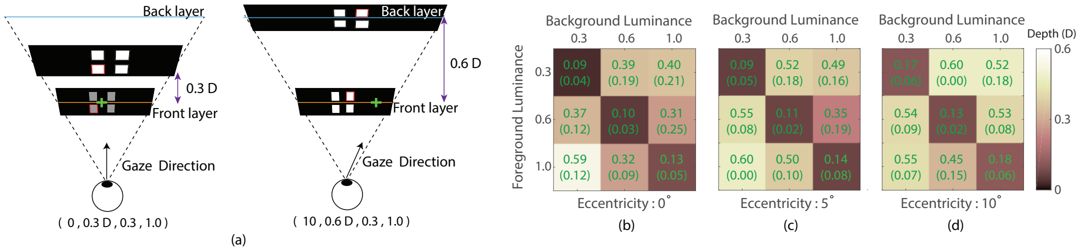


FIGURE 6.5: *Artifact perception at depth discontinuity. (a) The configuration of the perceptual experiment. The values in parentheses represent the eccentricity, depth difference, foreground luminance, and background luminance. (b–d) The experimental results on the depth difference threshold at which LB and LFS are distinguishable. The mean values are shown with the standard deviations in parentheses.*

6.5 Effect of Depth Discontinuity on Decomposition

Another factor that affects the decomposition quality is the depth difference between two surfaces with an occlusion boundary. This part of the chapter investigates the distinguishability between LFS and LB as a function of depth difference, luminance contrast, and eccentricity. Contrary to the analysis on spatial frequency in Section 6.4, the LFS is always closer to the ground truth compared to LB, but it is important to clearly identify the conditions in which LB can still be employed without causing any visible loss of quality.

6.5.1 Perceptual Experiment

The perceptual experiment in this part follows the one in Section 6.4.1 where the 2AFC procedure is used, and the participants are asked to select the pair consisting of different patterns. For each luminance contrast and eccentricity, the Quest procedure is employed to find the threshold of depth difference at which LB becomes distinguishable [Watson and Pelli, 1983]. The depth difference ranges from 0.05 D to 0.6 D with a 0.05 step size. Two representative stimuli are illustrated in Figure 6.5 (a). While the foreground objects are fixed on the front display, only the depth of occluded objects is altered. For the experiments at higher eccentricities, the gaze direction is guided by a target green cross, and the observers' gaze position is monitored using the eye tracker. In order to avoid incorrect measurements due to accidental glances, the stimulus is hidden when the gaze position slightly deviates from the target cross. The whole set of stimuli spans 3° of visual angles. In order to avoid image degradation due to the aberration near the boundaries of the display, the position of the stimuli is fixed at the center of the display, and the position of the target cross is changed to control stimulus eccentricity.

The results of this experiment are shown in Figure 6.5 (b–d). An increase in depth difference thresholds with respect to eccentricity can be seen as expected. This implies that the human visual system (HVS) is less sensitive to the incorrect edges generated by LB in the peripheral visual field, and it provides the flexibility of using LB instead of LFS at edges located in the periphery to improve performance. Another observation is that the difference between LB and LFS decompositions is highly distinguishable at low luminance contrast edges, which is a finding that is in the opposite direction of the analysis on texture, where the difference between LB and LFS is reduced with the luminance contrast reduction as shown in Figure 6.4 (a,b). Notice that at the occlusion boundaries, the mixed signals from the focused and defocused image regions are perceived, which is not the case for local texture perception. The following section further analyzes this interesting trend.

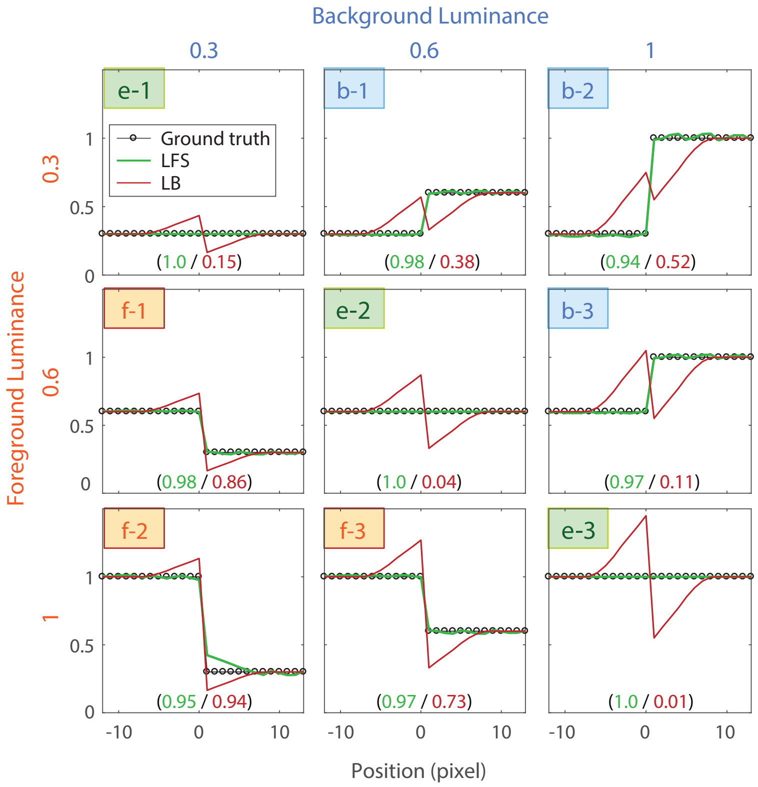


FIGURE 6.6: *1D luminance profiles at the edge between the front and back planes separated with the 0.6 D depth difference for LB and LFS decompositions. The right half of each plot corresponds to the back plane, and the left half to the front plane, which is also the focus plane. At the bottom of each case, the SSIM value pairs of (LFS vs. ground truth / LB vs. ground truth) are shown in the parentheses. These SSIM values clearly indicate that LFS surpasses LB in all cases.*

6.5.2 Analysis of Edge Profiles

In order to clarify the occlusion perception, the conducted analysis in this part investigates 1D luminance profiles (Figure 6.6) that are produced at the fovea by LB and LFS methods while observing a depth edge between the front and back planes. It is assumed that the eye is always focused on the front plane, which leads to the strongest artifacts [Narain et al., 2015].

The E-1—E-3 types in Figure 6.6 show the depth discontinuity where the luminance values are the same for the front and back planes. In such conditions, the artifact patterns in the LB decomposition can be attributed to an interaction of two factors: optical blur in the back plane and luminance additivity in the two-plane display. As the energy of the blurred signal increases with the back-plane luminance, the artifact's absolute magnitude is larger in the E-3 than in the E-1 case. However, the artifact detectability, akin to Weber's law, depends on its luminance contrast with respect to the uniform background; thus, the E-1—E-3 types have similar thresholds (Figure 6.5 (b)). In general, the eye sensitivity for this type of artifact is relatively high, as the contrast detection thresholds at uniform backgrounds are relatively low [Legge and Foley, 1980]. The artifact contrast increases with depth discontinuity, so that it can easily be detected even for small depth differences (Figure 6.5 (b)).

The F-1—F-3 types in Figure 6.6 show the depth discontinuity where the front-plane luminance values are higher than their back counterpart. Similar artifact patterns as in the E-1—E-3 types are created, but this time they are imposed on contrast edges that act as contrast maskers [Legge and Foley, 1980]. Effectively

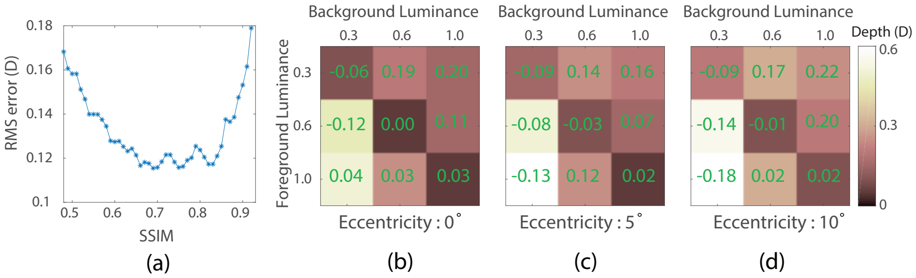


FIGURE 6.7: *SSIM calibration. (a) The RMS error between the predicted depth thresholds and experimental results. (b–d) the predicted depth difference thresholds from SSIM for various eccentricities. The values represent the error between the predicted thresholds and the experimental outcome.*

contrast discrimination thresholds for such artifacts are elevated, which requires a significant increase in depth discontinuity to make the artifact visible (Figure 6.5 (b)).

The B-1—B-3 types in Figure 6.6 show the depth discontinuity where the back-plane luminance values are higher than their front counterpart. This time the artifact pattern is embedded into the edge luminance profile, which might result in a more blurry edge appearance. Nevertheless, the HVS sensitivity for such artifact patterns is similar to the F-1—F-3 types (Figure 6.5 (b)) with remarkably close depth thresholds for the same luminance contrast (the F-1 and B-1, and F-3 and B-3 types). This observation does not hold for the F-2 and B-2 types, and it can possibly be attributed to the imperfect luminance profiles for LFS due to intensity saturation caused by constrained optimization. Interestingly, the variance in the participant responses is higher for the B-1—B-3 types than their F-1—F-3 counterparts.

While the analyses in this chapter avoid conducting a detailed analysis of the eccentricity cases (Figure 6.5 (c-d)), overall, similar observations can be made.

6.5.3 Calibrating SSIM

Similar to the analysis in Section 6.4.2, the SSIM metric is calibrated to predict the outcome of perceptual experiments (Section 6.5.1). Instead of using the previous detection threshold, the optimal SSIM is derived to detect artifacts at occlusions independently. This strategy follows the observations made in Swafford et al. [2016] and Adhikarla et al. [2017], where specific training for each artifact type led to the improvement of SSIM metric predictions.

For each combination of the luminance contrast and depth difference, the front focal images for both LFS and LB are generated and the minimum value in the SSIM map between the two algorithms is measured. In order to simulate the perceived image in the peripheral vision, a Gaussian blur with the cutoff frequency according to the quantitative HVS model by Watson and Ahumada [2011] is applied. The RMS error between the predicted and actual depth differences is the smallest, around 0.7–0.83 (Figure 6.7 (a)), and the largest value is select as the detection threshold conservatively. The depth difference thresholds as predicted by the SSIM are shown in Figure 6.7 (b–d) for various eccentricities. The errors are typically acceptable when compared to the variance in the user experiment in Figure 6.5 (b–d). However, the SSIM prediction produces depth thresholds that are consistently too large for the F-type distortions and too small for the B type (Figure 6.6). This discrepancy in the SSIM sensitivity might be attributed to differences in the distortion profiles as discussed in Section 6.5.2. Further consideration would rely on a more conservative prediction for the B type.

Using the calibrated SSIM, the depth thresholds can be predicted for larger eccentricities in display configurations with a wider field of view and extended dioptric range. Based on these predictions, combined with the experiment outcome in the fovea and near eccentricity (Section 6.5.1), The next section investigates the selection rule for finding regions to apply LFS.

6.6 Unified Optimization

The conducted analyses reveal that LB can be applied for all textured surfaces (Section 6.4). For occlusion boundaries, LFS is applied depending on the luminance contrast, depth difference, and eccentricity (Section 6.5). This section establishes the selection rule for LFS based on the occlusion analysis and proposes a unified optimization to integrate LB and LFS.

6.6.1 Selection Rule

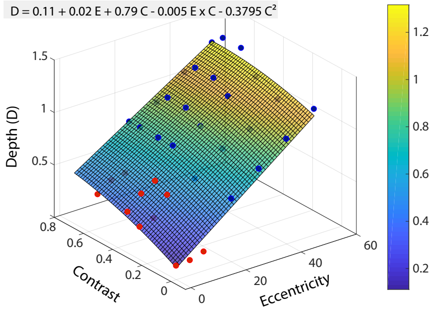


FIGURE 6.8: *The predicted depth difference thresholds from the SSIM. The goodness of fit:* $R^2 = 0.9603$*, RMSE=* 0.068*. For the measurement data only,* $R^2 = 0.8078$*, RMSE=* 0.071.

The selection rule is designed for LFS as a function of the Michelson contrast, eccentricity, and depth. First, each combination of background and foreground luminance is expressed as Michelson contrast. In this case, the F-1 and B-1, F-2 and B-2, F-3 and B-3, and E-1–E-3 types have the same contrast. Among two or three different depth thresholds for a given contrast, the smallest depth threshold is selected to be on the conservative side. In the SSIM prediction, the perception of artifacts at large eccentricities is also analyzed, which is expected to lead to larger depth thresholds. In order to check the depth separation beyond 0.6 D, a four-plane display with a 0.6 D gap between successive layers is simulated. The experimental outcome still holds for this display configuration since LB assigns the values to two nearby planes only; therefore, the behavior of LB in the two-plane display and the four-plane display is the same for edges with less than 0.6 D separation. In Figure 6.8, the depth threshold is extrapolated to 50° eccentricity. Then, a 3D surface is fitted to the predicted depth thresholds. The depth thresholds obtained from the perceptual experiments are marked with red points, and the predicted thresholds from SSIM are indicated with blue points. Although further confirmation is required with the perceptual experiments, a huge computational gain could possibly be obtained in a wide field of view multi-layered displays in the future. In the proposed method, LFS is applied to the cases where the depth difference is larger than the depth thresholds on the predicted surface.

Based on the selection rule, a mask is generated to identify the regions that require LFS. The example of a mask generation for a fish scene in Figure 6.12 is shown in Figure 6.9. From the depth map (Figure 6.9 (a)) and Michelson contrast map (Figure 6.9 (b)), a mask is generated to apply LFS for the center gaze direction (Figure 6.9 (c)). The A and B cases show the occlusion boundaries eliminated from the mask due to the decreased sensitivity at high eccentricities. The C case is an example of type E edges in Figure 6.6. Although this edge has a small depth difference, it is still masked due to its lower luminance contrast compared to nearby edges.

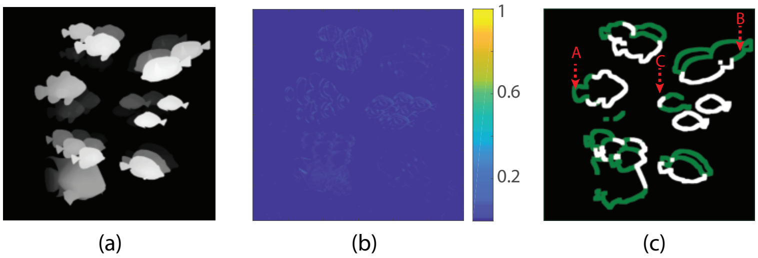


FIGURE 6.9: *Mask generation. (a) Depth map. (b) Michelson contrast map. (c) Mask. White region: masked region for the center gaze direction. Green region: masked region assuming no degradation of HVS at high eccentricities.*

6.6.2 Unified Decomposition Framework

This chapter also proposes a unified optimization scheme that solves LFS with LB as a constraint. In practice, LFS and LB can be separately calculated which can be then blended at intersection regions. However, keeping in mind that LFS requires a constrained least square optimization, using LB as the boundary condition for LFS can provide a smooth transition at intersections. The original decomposition algorithm of LFS can be written in the following form:

$$\underbrace{\begin{bmatrix} \mathbf{L}(v,u_1) \\ \mathbf{L}(v,u_2) \\ \vdots \\ \mathbf{L}(v,u_K) \end{bmatrix}}_{(KN)\times 1} = \underbrace{\begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \dots & \mathbf{P}_{1D} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}_{K1} & \mathbf{P}_{K2} & \dots & \mathbf{P}_{KD} \end{bmatrix}}_{(KN)\times(DN)} \underbrace{\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_D \end{bmatrix}}_{(DN)\times 1}. \tag{6.1}$$

Here, a two-plane parametrization of the light field is employed. v denotes the spatial coordinate on the light field plane and u denotes the spatial position of the pupil. $\mathbf{L}(v,u_k)$ is a vectorized 2D image given a viewpoint k and $\mathbf{x_d}$ is a vectorized 2D pixel value on the display layer d. K is the number of viewpoints and D is the number of layers. Without loss of generality, the number of pixels in each layer and target light field is both equal to N. In practice, the target light fields can have different resolutions. The submatrix $\mathbf{P}_{kd}$ of projection matrix $\mathbf{P}$ is defined as follows [Lee et al., 2016]: $(\mathbf{P}_{kd})_{i,j} = 1$ if $L(i,u_k)$ intersects with $(\mathbf{x}_d)_j$, and 0 otherwise.

Each component is divided into two regions: the masked and unmasked regions. then the full decomposition is applied to the masked region and the linear blending rule to the unmasked region. The subscript M denotes "masked" and U denotes "unmasked".

$$\mathbf{L}(v,u_k) = \begin{bmatrix} \mathbf{L}(v,u_k)_M \\ \mathbf{L}(v,u_k)_U \end{bmatrix}, P_{kd} = \begin{bmatrix} P_{kd,M} \\ P_{kd,U} \end{bmatrix}, \mathbf{x}_d = \begin{bmatrix} \mathbf{x}_{d,M} \\ \mathbf{x}_{d,U} \end{bmatrix} \tag{6.2}$$

Then the original equation can be rewritten as follows:

$$\underbrace{\begin{bmatrix} \mathbf{L}(v,u_1)_M \\ \mathbf{L}(v,u_1)_U \\ \mathbf{L}(v,u_2)_M \\ \mathbf{L}(v,u_2)_U \\ \vdots \\ \mathbf{L}(v,u_K)_M \\ \mathbf{L}(v,u_K)_U \end{bmatrix}}_{(KN)\times 1} = \underbrace{\begin{bmatrix} P_{11,M} & P_{12,M} & \dots & P_{1D,M} \\ P_{11,U} & P_{12,U} & \dots & P_{1D,U} \\ P_{21,M} & P_{22,M} & \dots & P_{2D,M} \\ P_{21,U} & P_{22,U} & \dots & P_{2D,U} \\ \vdots & \vdots & \vdots & \vdots \\ P_{K1,M} & P_{K2,M} & \dots & P_{KD,M} \\ P_{K1,U} & P_{K2,U} & \dots & P_{KD,U} \end{bmatrix}}_{(KN)\times(DN)} \underbrace{\begin{bmatrix} \mathbf{x_{1,M}} \\ \mathbf{x_{1,U}} \\ \mathbf{x_{2,M}} \\ \mathbf{x_{2,U}} \\ \vdots \\ \mathbf{x_{D,M}} \\ \mathbf{x_{D,U}} \end{bmatrix}}_{(DN)\times 1} \tag{6.3}$$

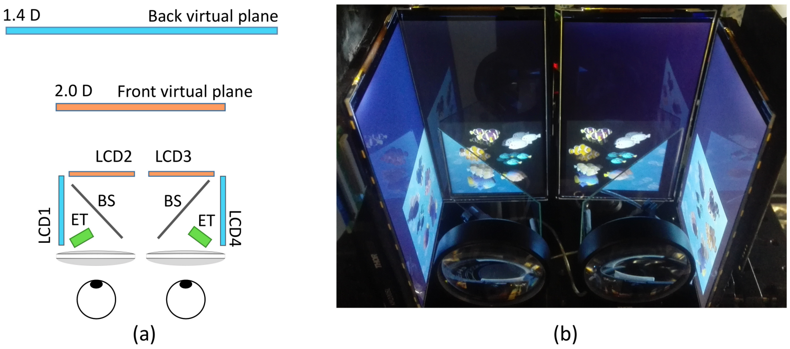


FIGURE 6.10: *The display prototype. (a) The schematic, and (b) a photograph of the display system. BS: Beam splitter, ET: Eye tracker.*

Since the linear blending rule is applied for a single image, the term $\mathbf{L}(v, u_k)_U, P_{kd,U}$ and $\mathbf{x}_{d,U}$ for $k > 1$, can be eliminated assuming that u_1 is the center viewpoint. Then, the unmasked region is handled only with the center viewpoint, $\mathbf{L}(v, u_1)_U$.

$$\underset{(N+(K-1)N_M)\times 1}{\begin{bmatrix} \mathbf{L}(v,u_1)_M \\ \mathbf{L}(v,u_1)_U \\ \mathbf{L}(v,u_2)_M \\ \vdots \\ \mathbf{L}(v,u_K)_M \end{bmatrix}} = \underset{(N+(K-1)N_M)\times(DN)}{\begin{bmatrix} P_{11,M} & P_{12,M} & \dots & P_{1D,M} \\ P_{11,U} & P_{12,U} & \dots & P_{1D,U} \\ P_{21,M} & P_{22,M} & \dots & P_{2D,M} \\ \vdots & \vdots & \vdots & \vdots \\ P_{K1,M} & P_{K2,M} & \dots & P_{KD,M} \end{bmatrix}} \underset{(DN)\times 1}{\begin{bmatrix} \mathbf{x_{1,M}} \\ \mathbf{x_{1,U}} \\ \mathbf{x_{2,M}} \\ \mathbf{x_{2,U}} \\ \vdots \\ \mathbf{x_{D,M}} \\ \mathbf{x_{D,U}} \end{bmatrix}} \tag{6.4}$$

By applying the linear blending, the dimension of the projection matrix can be reduced from $(KN) \times (DN)$ to $(N + (K-1)N_M) \times (DN)$, where N_M denotes the number of pixels in the masked region. For example, if $K = 9, D = 3, N_M = N/4$, then the dimension changes from $(9N) \times (3N)$ to $(3N) \times (3N)$.

Solving the reduced decomposition problem, however, does not provide the correct answer because $\mathbf{x}_{d,U}$ do not have enough constraints. The multi-viewpoint images impose constraints on each pixel value, but the single viewpoint cannot. Therefore, the pixel values should be calculated separately according to the linear blending rule and replaced in each iteration step.

6.7 Implementation

6.7.1 Rendering and Decomposition

The proposed rendering pipeline breaks down into four steps: (1) rendering the central viewpoint image and depth map, (2) computing the mask for LFS, (3) rendering additional viewpoint images on the masked region, and (4) performing LB and the iterative decomposition using SART. In the case of full light field synthesis without any mask, nine views with a 1200 × 1200 resolution are rendered using a 1 ray/pixel. For the proposed decomposition, a single 2D image and depth map are first rendered. By analyzing the luminance contrast and depth gradient map, the masked region is calculated based on the criteria in Section 6.6.1. Next, the 8 viewpoint images, except the center images falling inside the pupil, are generated only for the masked region. Compared to the generation of full light fields, The presented selective rendering can greatly reduce the computation time. From the

TABLE 6.1: *The rendering and decomposition timings of the proposed hybrid method for various scenes. The rendering and decomposition timings are given in ms. The values in the parentheses indicate timings for full LFS rendering.*

Scene	# polygons	mask(%)	rendering	decomposition
Fish	20498	7.3	9.26 (27.48)	2.57 (4.11)
Dice	569810	6.5	14.11 (47.08)	2.44 (4.12)
Forest	16924	1.8	7.29 (28.31)	2.35 (4.19)

rendered target scene, the optimal decomposed images are calculated using the unified decomposition framework (Section 6.6.2). At this stage, the optimization time is further reduced over the conventional SART implementation by developing an efficient adaption of SART in CUDA. The rendering system is implemented using the Nvidia OptiX ray tracer, which enables selective rendering for a given mask with minimal overhead. The renderer is driven by a PC with a 3.60 GHz Xeon CPU and 32.0 GB RAM equipped with a single Nvidia GTX 1080 TI graphics card.

6.7.2 Eye-Tracked Multi-Layered Accommodative Display

In order to test the rendering strategy, a two-plane VR display is designed and built. The schematic and photograph of the setup are shown in Figure 6.10. For each eye, images from two 2560 × 1440 LCD displays (Topfoison TF60010A) are combined with a beam splitter (Edmund Optics #64-408) and magnified with an achromatic lens (Thorlabs AC508-080-A). Eye trackers (Pupil Labs) are placed right behind the two lenses. The optical system for the right eye is mounted on the linear stage for adjusting the interpupillary distance. The dioptric distances to the front and back virtual planes are set to 2.0 D and 1.4 D, respectively.

The resolution of the display is 1200 × 1200, which is significantly higher than the light field displays reported so far [Huang et al., 2015b; Mercier et al., 2017; Lee et al., 2017]. FOV is 40°, and the angular resolution of the system is 15 cpd. The designed system has a high enough resolution and large enough FoV to study the effect of foveation, while the resolution of current VR and AR systems rarely exceeds 10 cpd, which is quite limited for foveated rendering.

6.8 Results

Three different scenes are rendered to evaluate the rendering strategy. First, the computational time is measured for the proposed optimization algorithms. Then, the visual quality of the method is compared with LB and LFS on the display prototype and using simulations.

6.8.1 Performance

The total rendering time is calculated for the whole pipeline during monocular viewing. For three scenes in Figure 6.12, the rendering and decomposition timings for the proposed hybrid method and full LFS are measured in Table 1. All decompositions are performed with 10 iterations. As the shader/geometry complexity becomes higher, the rendering time increases. However, the computational saving of the hybrid method is even more pronounced with respect to full LFS since, in many scene regions, only a single view was required for LB and can avoid full LF rendering. It can be seen that the decomposition time only depends on the percentage of masked regions.

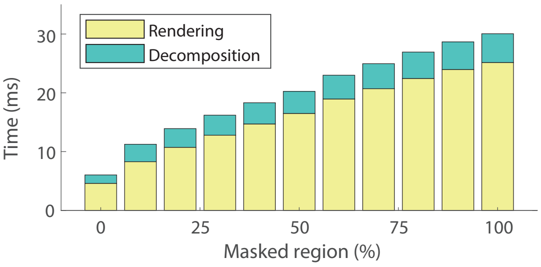


FIGURE 6.11: *The rendering and decomposition timings for various ratios of the masked region for the fish scene in Figure 6.12.*

The presented test scenes contain 5.19% of LFS region on average. The frame rates are measured as 84 Hz (×4.25), 60 Hz (×4.06), and 103 Hz (×4.50) for the fish, dice, and forest scenes. The values in the parentheses denote the speed enhancement over full LFS after subtracting a fixed cost of a single view rendering. If a scene contains many depth edges, the performance gain of the presented hybrid method reduces since most of the regions should be rendered with LFS. For binocular viewing conditions, the stereoscopic scenes are rendered sequentially. In this case, the total rendering time increases by a factor of 2. In order to test the effect of the percentage of the masked regions, the timing is also measured for various masked regions for the fish scene as shown in Figure 6.11. Here, the masks were randomly generated instead of using the mask generated by the selection rule in Section 6.6.1. The zero percentage corresponds to the LB-only rendering. The total optimization time linearly increases as the masked region grows. This trend implies that there is minimal overhead coming from selective rendering for the randomly masked region.

6.8.2 Comparison

Figure 6.12 shows, for each scene, the real photographs from the display and the simulated perceived images. Three algorithms are compared: LB, full LFS, and the proposed optimization. The masks are computed assuming the gaze is directed toward the center. For the simulated images, the SSIM is computed against ground truth, which is the focal image generated with dense light fields. The SSIM maps indicate that LB produces strong artifacts mostly along the occlusion boundaries. Furthermore, the boundaries between LFS and LB in the proposed algorithm do not produce any noticeable discontinuities, confirming the validity of the unified decomposition framework. However, halo effects are visible around edges in captured images for both LFS and the presented method. It is found that small errors in color calibration between two display layers led to such artifacts, which are not visible in the simulation.

The fish and dice scenes show various aspects of edge reconstruction analyzed in Figure 6.6. The blue fish and gray dice are examples of the E-3- and E-2-type occlusions. LB generates a sharp contrast, while LFS produces smooth transitions. On the other hand, as seen around the yellow fish and reddish-brown dice, the B-2-type edges from LB look blurry, but sharp edges are obtained in the proposed method. The edges along the white part of the fish or the orange dice present F-2-type profiles. Since the foreground objects are brighter, the differences among the three algorithms are less obvious.

The forest scene demonstrates the reconstruction quality in textured regions. The high-frequency features of slanted grass fields are preserved in LB and the presented method, but they are blurred out in LFS, which is expected from Figure 6.4(e). Although the method provides better image quality, this enhanced contrast could possibly lead to incorrect contrast curves as seen in Figure 6.2. However, far-focus images still look more blurry than the focused images on the grass field, which suggests that the failure of LB at high spatial frequencies does not affect the effectiveness of driving accommodation [MacKenzie et al., 2010]. On the other hand, low-frequency

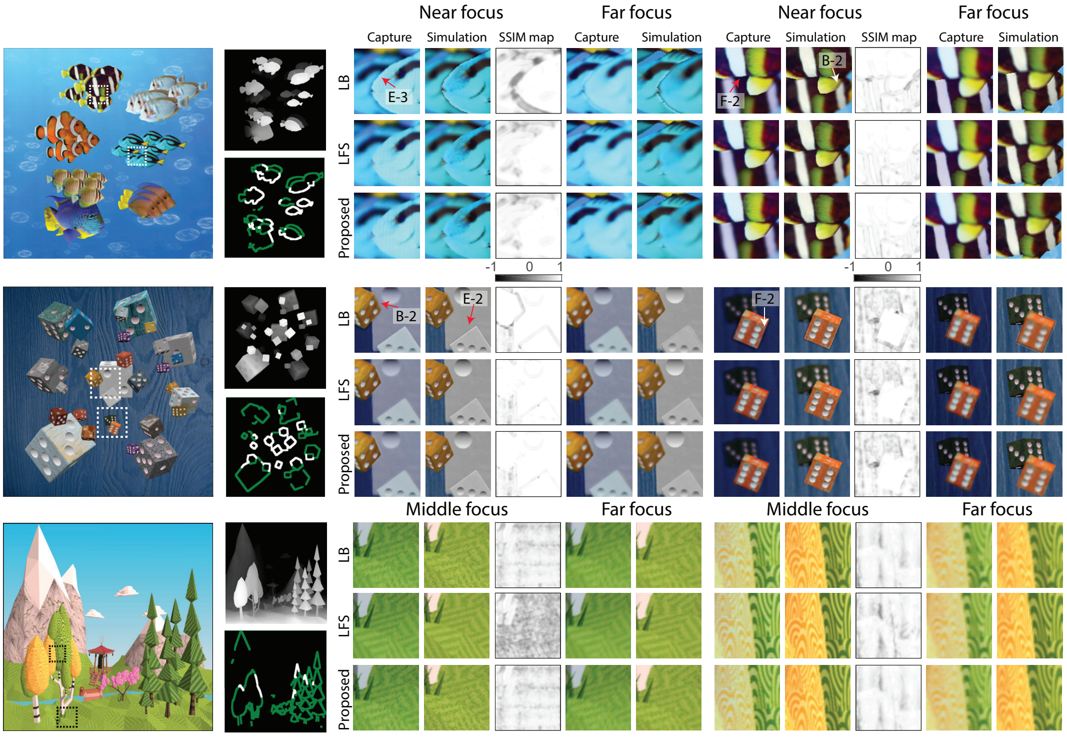


FIGURE 6.12: *Comparison of various decomposition methods for various scenes. The images (column 1) represent the target scenes. The upper and bottom images (column 2) are the corresponding depth maps and masks (white region: masked region for the center gaze direction, green region: masked region assuming no degradation of HVS at high eccentricities). For each scene, the proposed method (row 3) is compared to the LB (row 1) and full LFS (row 2). The near or middle focus images (columns 3–5, 8–10) show captured images from the display, simulated perceived image, and SSIM map between the ground truth and simulation. The captured and simulated images at far focus are also shown in columns 6–7 and 11–12.*

textures on the yellow and green trees are reconstructed with slightly higher contrast for LFS, which is expected from the low-frequency region in Figure 6.4(d). However, for those regions, the differences between the two methods are subtle, so they are not distinguishable according to the carried out perceptual experiment and SSIM predictions.

6.8.3 Temporal Coherence

As temporal coherence is a critical use case for real-time methods, the temporal coherency of the proposed method is tested on dynamic scenes. Luckily, the obtained perceptual findings allow the use of LB for textured regions, while temporal changes occur only around edges. Since the threshold functions on edges are derived based on indistinguishability between LB and LFS (Section 6.6.1), smooth transitions can be achieved when a switch between LFS and LB occurs near the threshold. First, the transition behaviors are evaluated in two dynamic scenarios. In both cases, it is assumed that the gaze is directed toward the center, which is marked with a red box. The captured and simulated videos do not show any noticeable artifacts around the edges near the gaze position. In the periphery, the transition between the two algorithms are sometimes visible, but those boundaries are not noticeable in actual viewing conditions due to the reduced sensitivity of HVS. In Scene 1, it can be observed that rendering artifacts around the high spatial frequency textures originate from the low sampling rate used (1 ray per pixel) and are unrelated to the quality of

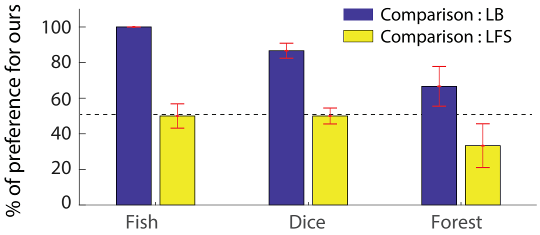


FIGURE 6.13: *The percentage of answers preferring the presented method over LB and LFS for three scenes. The error bars indicate the standard error.*

the generated mask. In order to address this issue, space-time ray-tracing methods can be employed in the future for a better rendering quality [Glassner, 1988]. For better visualization of various artifacts, the SSIM map is also computed between the images generated with the proposed method and the ground truth. Here, the ground truth is computed as focal images generated with dense light fields. The SSIM maps indicate that high spatial frequency textures show noticeable deviations due to the use of single ray per pixel and imperfect reconstruction of high spatial frequencies as discussed in Figure 6.2. The SSIM videos also show an error at occlusion boundaries in the periphery, but it is not noticeable due to the foveation. Although those artifacts are clearly seen in the SSIM maps, the rendered videos do not exhibit significant artifacts when they are observed alone without the comparison against the ground truth.

6.9 Evaluation

In order to validate the perceptual quality of the proposed design, a user experiment is conducted to compare (LB, Ours) and (gaze-contingent LFS, Ours) for four static scenes in a binocular setting. In order to simulate the gaze-contingent sampling in Section 6.3, gaze direction stimuli are shown to the user, and decomposed images are optimized for a given gaze direction. To allow the accommodation change, the gaze direction stimuli were set to rectangular boxes extending 2° of visual angles for both eyes. This small gaze change does not introduce the generation of new decomposed images in gaze-contingent LFS. The subjects are instructed to maintain the gaze direction inside the box but to judge the overall image quality. In each trial, the users are asked to choose the scene which produces better image quality. In each scene, the users compare the quality of scenes in five different gaze directions. Six subjects participated in the experiment. The outcome of the perceptual evaluation is shown in Figure 6.13. A high preference can be observed for the method over the LB, and the difference between the two methods is found statistically significant in the binomial test ($p < 0.05$ for all scenes). This can be attributed to the better reconstruction of edges in the method. In the forest scene, the difference between LB and the proposed algorithm decreases since the scene mostly consists of textured regions without occlusion. In comparison with LFS, the algorithm shows similar preferences for fish and dice scenes ($p > 0.50$), which indicates the observers are indifferent between the two methods. Considering the fact that those two scenes contain many occlusion boundaries, the outcome of the user study suggests that the masking algorithm in Section 6.6.1 successfully works. However, the subjects preferred LFS over the proposed method in the forest scene, which contains high-frequency features, and the difference is significant ($p < 0.05$). Although it was not investigated formally, subjects reported that they prefer blurred texture in LFS over ours (Figure 6.12). Since the reconstruction of high-frequency textures requires precise alignment, small pupil and head movements may lead to the perception of those features as noise patterns.

6.10 Limitations

Display The optical system in the display prototype is based on magnifier lenses; therefore, the system suffers from aberrations around the outer regions. The multi-plane displays with holographic optical elements [Lee et al., 2017] could be a good candidate for reducing image degradation. Although the considered display provides a relatively wide FOV of 40°, it is still smaller than current VR displays. The dioptric range of the display is also limited to two-plane displays. Therefore, further validating the occlusion analysis in wide FoV multi-plane displays with a larger dioptric range is required. The development of the displays with the extended dioptric range also enables the study of foveation rules on the occlusion boundaries in the context of defocus states, while only eccentricity has been considered in the conducted study. Similar to blurred artifacts at high eccentricities, large depth differences between the eye focus and edge can also reduce sensitivity to edge artifacts. The current prototype lacks devices measuring the accommodation states [Mercier et al., 2017; Koulieris et al., 2017]. The evaluation of the effectiveness of driving accommodation with different optimizations would be important for future work.

Perception The presented method in this chapter relied on a specific image quality metric (SSIM). As image quality evaluation is still an open problem, the suggested detailed masking aggregation could be affected by SSIM inaccuracies. As observed in the validation experiment, the prediction of SSIM does not account for artifacts induced by viewing conditions such as pupil movements and misalignment. Since the perceived images depend greatly on the focal state, a quality metric that can meaningfully compare light fields would be required. Such a metric should be applied after considering display-specific limitations in reproducing light fields. Also, a metric capable of predicting the ability to induce eye accommodation by such reproduced light fields would be desirable in deriving possibly new foveation rules in the proposed approach. All these interesting and difficult problems can be relegated to future work.

Rendering In this chapter, Lambertian scenes are only considered, whereas handling glossy objects would require the extension of the proposed masking algorithm to consider such objects as a function of the visibility of view-dependent effects. Since the boundaries between LFS and LB show smooth transitions as seen in Figure 6.12, extending the mask region should handle non-Lambertian scenes as well. Regrettably, no direct comparison is performed to the RO method. Although RO performs moderately better at 9 and 12 cpd according to the analysis in Figure 6.2, this quality improvement comes at a significant computational speed loss. It is noteworthy that the rendering speed of RO is reported as 5 FPS for a display resolution of 512×512 [Mercier et al., 2017], while the performance time of the method is faster than 60 FPS for a display resolution of 1200×1200. Furthermore, none of the methods can correctly trigger accommodation at 12 cpd; therefore, it implies that the gaze-contingent LFS is a suitable method providing similar quality offered by RO yet with much faster computational speed. Although a performance gain is expected without significant quality degradation, this comparison would be relegated to future work. The proposed strategy can be potentially used to combine RO and LB techniques, but this requires further investigation.

6.11 Conclusion

This chapter presents a hybrid decomposition framework of linear blending and light field synthesis that enables real-time rendering and high-fidelity reconstruction in multi-layered light field displays. The perceptual experiments and the SSIM analysis conducted in this chapter provide a deeper insight into visual quality produced by different decomposition algorithms. In particular, it is shown that for textured surfaces, LB and LFS are indistinguishable for low to mid-spatial frequencies, and LB is closer to the ground truth for high spatial frequencies. For occlusion boundaries, LB fails at low luminance contrast edges rather than high contrast edges, which seems counterintuitive but is a consequence of the additive combining of focused and defocused patterns at both edge sides. Moreover, those conditions for occlusions can be further relaxed for surfaces at sufficiently large eccentricities when the sensitivity of the HVS drops significantly. In order to realize the proposed selective optimization strategy, a unified optimization framework is developed to combine LB and LFS efficiently. The proposed rendering strategy is tested with a two-layer multi-plane display and validated the 60 Hz rendering time for 1200×1200 resolution with nine viewpoints.

While this chapter focuses on the additive light-field display, the presented hybrid strategy can possibly be extended to the multiplicative light-field displays since it can be formulated with the additive light-field synthesis under logarithm [Lanman et al., 2011]. Therefore, investigating the simple decomposition rules in a multiplicative architecture and integrating them with the light field synthesis algorithms would be an interesting topic. Since the major artifacts of LB around the occlusion originate from the additive nature of the light-field display, studying the edge artifact in the multiplicative display would lead to interesting perceptual insights in accommodative light-field displays.

Chapter 7

Conclusion

This thesis presented several innovative approaches to improve the quality and practicality of image-based rendering (IBR). One approach, described in Chapter 3, combines primitive image formation models with a learned scene representation to enable real-time interpolation of a scene captured at sparse views, lighting positions, and times. Although this approach is not intended to generalize across different scenes, it effectively handles the task of generalizing across changes in geometry, motion, and illumination, producing state-of-the-art visual quality at an interactive speed. Chapter 4 presents another approach that utilizes a physically-based formulation to optimize the 3D spatially-varying index of refraction based on 2D observations, allowing for the reconstruction of high-quality novel views of scenes with refractive objects in an unconstrained capturing setup. Chapter 5 presents a no-reference visibility metric designed for detecting artifacts in IBR images, along with a training strategy to minimize false positives and negatives. The proposed metric is demonstrated through two applications: accelerating automated adaptive light-field capture and providing feedback in an interactive depth manipulation system. Finally, Chapter 6 proposed a perception-driven rendering technique for multilayer accommodative displays, which achieves both real-time performance and high-fidelity results.

In addition to the specific contributions made by the methods presented in this thesis, subsequent sections delve into some insights and potential future directions that can be gained from each chapter, respectively.

7.1 Implicit Scene Representation

Unlike traditional methods of representing signals as discrete grids of pixels or parameterizing 3D shapes as grids of voxels, point clouds, or meshes, the implicit representation is a novel approach that represents a signal as a continuous function mapping a domain (such as 3D coordinates or pixel positions) to a corresponding output (such as an image). This continuous representation encodes the scene into the weights of the neural network, and the resulting inductive bias offers several potential benefits compared to traditional discrete representations, including increased flexibility and capability for more efficient processing [Tewari et al., 2022]. The X-Field representation in Chapter 3 builds upon this approach by introducing two modifications to the implicit representation. Firstly, similar to previous work by Zhou et al. [2016] and Sun et al. [2018b], it predicts texture coordinates rather than appearance. These texture coordinates are used to drive a spatial transformer [Jaderberg et al., 2015], allowing for copying details from the input images without requiring their explicit representation and enabling fast processing (20 fps). Secondly, the approach utilizes a 2D CNN instead of a 3D multilayer perception (MLP) to directly return a complete 2D per-pixel depth and correspondence map for a given input coordinate.

This is more efficient for solving X-Field problems than ray-marching and evaluating a complex MLP at each step. In general, utilizing flow maps to warp input images instead of directly generating their appearance can bring several advantages. One is that the interpolated images will retain the quality of the original input as they are constructed by directly combining the input images. Second, as the changes in flow are smoother compared to colors [Zhang et al., 2021c], the interpolation becomes easier. For instance, in a stereo vision problem where there is a planar surface with complex texture positioned at a constant depth, the motion flow (disparity) can be described by a constant value for all pixels if the surface is shifted by a small amount (e.g., ten pixels) when seen from another view. This smoothness in flow can make it easier to interpolate between different views than directly regressing the color, which may require more complex modeling. Methods that rely on directly reproducing the appearance [Barron et al., 2021; Yu et al., 2021a; Sun et al., 2022] face an even more significant challenge in preserving the intricate details of textures when working with high-resolution input images. Expanding the capacity of these models through larger MLPs or voxel grids leads to a trade-off between slower inference time and excessive storage costs. In contrast, the proposed architecture can effectively scale to high-resolution outputs by adding minimal layers without incurring a substantial computational burden. Drawing inspiration from Sun et al. [2018a], this architecture can be trained on lower-resolution images and then applied to the original high-resolution photos by simply upsampling the learned flows. Additionally, the proposed approach incorporates flow consistency into the image composition process rather than the loss functions during training. This concept could be applied to other image-based interpolation methods [Wang et al., 2021a; Li et al., 2022d]. Unlike existing methods [Wang et al., 2021a; Li et al., 2022d], which demand meticulous design of latent codes to enable semantic control of the scene, the method proposed in Chapter 3 does not require any additional effort to construct latent codes, as they are already provided in the form of well-organized space-time-light coordinates. Chapter 3 assumes that changes in the scene can be explained by flow, and scenes that do not adhere to this assumption might fail regardless of the density of input data or the capacity of the model. For instance, the strict separation between shading and albedo may be inadequate in explaining changes in global image brightness caused by casual photography with automatic exposure. Despite this, the concept may be beneficial in separating the reflective and diffuse layers for synthesizing novel views of scenes with reflective surfaces [Kopf et al., 2013; Xu et al., 2021]. On the other hand, if all necessary data is available and the model assumptions are met, a neural network (NN) must have sufficient capacity to effectively process and represent the input signals. Many current approaches to scene representation [Mildenhall et al., 2020; Sitzmann et al., 2019b], including those presented in this thesis, use a fixed-size NN to approximate a scene, regardless of its complexity. However, this can result in over- or under-representation of the scene. One potential avenue for improving the effectiveness of these representations is to design an NN whose capacity is adaptive with respect to the complexity and resolution of the input scene, similar to the approach used by Takikawa et al. [2021], Yu et al. [2021b], and Liu et al. [2020] where a sparse octree is used to fit volume grids to objects adaptively.

The X-Field representation requires structured input in the form of relative view and light positions for cameras and light sources that lie in a plane. Although the extension to unstructured view-light capture can be made possible by replacing the flow map with a depth map, provided that the transformation between the cameras is known, recent implicit volumetric representation approaches [Du et al., 2021; Zhang et al., 2021b], inspired by the seminal work of Mildenhall et al. [2020], seem to serve better in this task as they are able to produce more "consistent" 3D reconstruction through the combination of volume rendering and coordinated-based MLPs. However, these representations come with a high computational cost and typically require a large number of input views to be effective, and their performance can degrade significantly when the number of available views is very limited. Although voxel-based optimization approaches [Sun et al., 2022; Yu et al., 2021a] can help speed up the rendering process; still, a special treatment [Niemeyer et al., 2022] is required to achieve accurate results when dealing with sparse observations. Investigating the feasibility of interpolating from even sparser observations, such as using only one observation per each X-Field dimension, could be an exciting direction for future research.

Integrating the coord-conv layer is essential for the successful interpolation in X-Field. Without this layer, the output flow would become a constant value, and in the case of disparity estimation, this value would correspond to the mean disparity of the entire scene. While the coord-conv layer, first introduced in Liu et al. [2018], was intended to furnish features with information about their spatial position, making the convolution process translation-invariant, further research is required to fully comprehend the role of this layer. Alternative architectures, such as MLPs, can also be considered to represent the X-Field signals. Instead of using cascades of convolutional and upsampling layers to process (2D, 3D, or 5D) X-Field coordinates, similar to Attal et al. [2022], Li et al. [2021], and Sitzmann et al. [2021] where an MLP is trained to map 4D light field coordinates (consisting of 2D pixel positions and 2D ray directions) to corresponding color values, an MLP could take the X-Field coordinates along with 2D pixel positions as input and output flow Jacobians. It is worth noting that this approach can still be computationally efficient, as each pixel would only require a single MLP evaluation, in contrast to the hundreds of evaluations per ray that are needed for volume rendering in NeRF.

When the variations in scene content are moderate, the X-Field representation can yield output flows that are sharp and well-defined. However, in the case of sparse input images with more substantial changes, it may encounter difficulty in producing detailed flows, as evidenced in Figure 3.18. A coarse-to-fine matching strategy similar to those employed in multi-view stereo [Dabała et al., 2016], optical flow [Sun et al., 2018a], and the progressive grid in Chapter 4 can be adopted to facilitate finding the correspondence in larger changes in light, view, and time. Lastly, it would also be interesting to include editing capabilities, such as changing the style and appearance of the objects [Zhang et al., 2022; Huang et al., 2022] along with X-Field interpolation in a consistent way.

7.2 Handling Complex Lighting Effects

Transparent objects reflect and refract light, causing it to bend and scatter in various directions, and when they come into interact with light and their surroundings, they produce intricate visual effects. While a few works deal with novel view synthesis of transparent objects, most of them focus on accurately reconstructing the shape of these materials [Xu et al., 2022; Lyu et al., 2020; Li et al., 2020]. These methods often require specialized capturing equipment or large synthetic training datasets and may need knowledge of the environment and the initial shape of the transparent object with an input mask. Additionally, they are unable to handle objects with varying indices of refraction (IoR). In contrast, Chapter 4 aims to overcome these limitations by synthesizing plausible views of transparent objects in a general and unconstrained setup, using an adapted eikonal formulation [Ihrke et al., 2007] that allows for modeling of objects with varying indices of refraction. Although it is trained per scene, the optimization is done from scratch without requiring a shape prior.

Chapter 4 adopts the volumetric representation approach introduced in NeRF, which is not computationally efficient. Regrettably, it is even more demanding than NeRF as it requires a heavy gradient calculation in each eikonal step. Additionally, the volume rendering is done using stratified sampling with a high sample count. Devising a sampling strategy similar to hierarchical sampling in NeRF could help to save computation and increase the quality through allocating more samples on the surface regions and the boundary of the IoR field [Pan et al., 2022]. To enjoy even faster optimization, it is worth considering grid-based representation for modeling the IoR and absorption-emission fields [Yu et al., 2021a; Sun et al., 2022]. Further improvement in test time execution can also be achieved by pre-computing a second volume containing the gradient of IoR volume. However, the grid-based optimization needs to be regularized to ensure a smooth solution that an MLP-based representation naturally provides.

The optimization in Chapter 4 is done sequentially in which the diffuse world is learned first, and then the IoR field is learned in a second pass that requires a user to mark the bounding box of the transparent object. While it would be ideal to perform the optimization jointly and automatically with minimal user intervention, the joint optimization process is highly under-constrained. It becomes especially challenging when rays bend a lot, as it becomes harder to find correspondences between the input images and the background. Additionally, the optimization process deals with the spatial gradient of the index of refraction rather than the IoR itself, which can be numerically demanding and unstable. However, coarse-to-fine optimization through frequency annealing as suggested [Lin et al., 2021; Park et al., 2021] could be a promising approach to make joint optimization possible.

The approach described in Chapter 4 only accounts for refraction and does not consider other light phenomena such as reflection or dispersion. This is because the eikonal equations used to model the propagation of light only describe refraction, and it is not currently clear how to include reflection in these equations in a unified form. Handling other phenomena, such as dispersion (the process of white light splitting into its constituent colors), would require introducing wavelength-dependent parameters into the volume-rendering formulation. This could also enable the rendering of colored glasses, in which a specific wavelength is reflected or refracted while all other wavelengths are absorbed. However, these ideas would require further investigation to determine how they could be implemented in practice.

In Chapter 4, the volume rendering technique [Max, 1995] utilized in NeRF is reformulated as a set of ordinary differential equations (ODEs) that are specifically tailored to the eikonal equation. The backpropagation is performed using the method introduced in Neural ODE [Chen et al., 2018b], which employs automatic differentiation to compute the gradient of the solution of the ODEs with respect to their parameters. This approach treats the underlying ODEs as a black box and employs the adjoint sensitivity method [Pontryagin, 1987] to compute gradients efficiently, thus avoiding the storage of intermediate steps. This allows volume rendering to be conducted with memory independent of the step count, enabling the efficient processing of a substantial number of rays (up to 32k) in each iteration. An alternative approach [Teh et al., 2022] involves the derivation of a backpropagation formulation specifically for the refractive ray tracing task. Similarly, this approach employs the adjoint state method to derive the derivatives with respect to the IoR field. This results in a constant memory complexity and requires significantly less memory when compared to previous differentiable rendering methods that utilize reverse-mode automatic differentiation.

One of the assumptions in Chapter 4 is that the background of the transparent object is sufficiently visible in the captured images. However, this can be a problem if the background is only presented through transparent objects, which is common in sparse-capturing setups. Although joint optimization may address this problem by simultaneously optimizing transparent objects and backgrounds to handle sparse images, a more advanced treatment is required.

The key to optimizing the IoR field in Chapter 4 is to utilize progressively coarser and finer versions of the emission and absorption models. However, how to smooth a continuous MLP-based radiance field is not immediately clear. The frequency annealing approach [Park et al., 2021] does not yield the desired band-limiting; instead, a uniform grid is fitted to the learned radiance field, and a coarse-to-fine radiance field is provided through low-pass filtering of the grid using a simple Gaussian blur kernel. However, recent research on band-limiting the coordinate-based networks [Lindell et al., 2022] can prove helpful in making these networks scale-aware. It mainly enables the control of the frequency bandwidth of the network at intermediate MLP layers, allowing for multi-scale image/volume representation.

Last but not least, similar to NeRF, the method outlined in Chapter 4 demonstrates effectiveness in novel view synthesis, although it is not specifically tailored for 3D reconstruction. The generated images are visually convincing using a physically-based rendering technique, i.e., eikonal rendering. However, the resulting IoR volume may not always be physically accurate. As depicted in Figure 4.7, the rays within a transparent object with constant IoR bend gradually rather than experiencing sharp bends at the object's boundary. This is because the direction of the light path is determined by the spatial gradient of the IoR field, and sudden bends require high spatial gradients to occur, which optimization naturally avoids. Instead, optimization strives to match the rendered image to the reference by gradually bending the rays so that they end up in the same direction as if there had been two strong, sudden bends.

7.3 Design of a No-Reference Quality Metric

A no-reference (NR) quality metric is a method of evaluating the quality of an image or video without relying on a reference signal for comparison. The assessment typically is not limited to detecting distortion but, importantly, includes judging magnitude and spatial locality. In Chapter 5, an NR metric is introduced to identify distorted regions in images generated by the IBR method. As the IBR artifacts are generally localized (e.g., appearing around the edges of occlusions), the proposed metric goes beyond the typical mean opinion scores (MOS) [Yang et al., 2022; Ke et al., 2021; Talebi and Milanfar, 2018] used to assess uniform distortions such as noise or JPEG compression and deals with a more challenging task of generating a per-pixel error map without accessing the reference. Such an error map provides analysis of the image quality at a local level and opens the potential for many novel applications, such as interactive or automatic view correction of rendering errors, as demonstrated in Section 5.4.2.

While an NR metric does not require a reference pair for evaluation, developing an NR metric relies on a data-driven approach that uses distorted-clean pairs or subjective human scores for training. Unfortunately, existing image quality assessment datasets, either full reference (FR) [Zhang et al., 2018; Lin et al., 2019; Jinjin et al., 2020] or NR datasets [Ying et al., 2019; Fang et al., 2020], do not include all types of distortions and often rely on MOS ratings. This can be a limiting factor when developing NR metrics for specialized tasks or applications. In the case of Chapter 5, the challenge is compounded by the limited availability of training data for IBR. Synthetic distortions such as noise, blur, or compression can be generated in large quantities for training purposes, but in the context of IBR, the captured images are typically sparse (e.g., 3×3, 5×5 light field); thus, only a small number of ground truth images are available. Chapter 5 addresses this challenge by presenting a training strategy that aims to minimize false positives or false negatives in the metric prediction. One aspect of this strategy involves augmenting the training data with natural images that are free from artifacts, as the number of rendered images containing artifacts is typically limited. It is essential to consider the right balance between natural and distorted training data to avoid false positives. The second aspect of the strategy entails carefully calibrating the learning process so that all reconstruction errors (in terms of their magnitude) are given equal consideration, thus avoiding false negatives. By implementing these measures, the proposed metric in Chapter 5 improved the accuracy of the NR metric in the context of IBR. As a next step, it would be interesting to examine the robustness of the metric through the use of adversarial examples [Carlini and Wagner, 2017]. This could involve adding an imperceivable noise or subtle blur to the distorted input image to test the reliability of the metric prediction.

The methodology presented in Chapter 5 is independent of the chosen metric for estimating the dissimilarity between images. It can mimic the response of well-established FR metrics such as PSNR, SSIM [Wang et al., 2004b], and LPIPS [Zhang et al., 2018], thereby enabling it to adapt to a wide range of FR metrics. This feature renders the proposed method a highly versatile tool that can potentially be used in various image quality assessment tasks.

While the considered distortions in this thesis are always IBR artifacts resulting from a specific IBR method [Dabała et al., 2016], the training approach is irrespective of the underlying IBR method, and it can be trained for other existing IBR methods [Wang et al., 2021a; Barron et al., 2021]. The proposed metric can be useful as a loss component or regularization term when training an IBR method with sparse input images [Niemeyer et al., 2022]. In that case, during the training, the novel views can be generated at arbitrary positions where no ground truth is available, and their reconstruction quality will be accessed using an NR metric.

One of the keys to successful IBR reconstruction is capturing a large number of input views; however, the main observation is that depending on the scene content, not all regions require equally dense camera placement. Diffuse, planar surfaces can be accurately predicted from other views showing the same surface, so capturing many images from these views is not necessary. On the other hand, occlusions and specularities can be more difficult to recreate because each element in the scene must be visible in at least two views for depth calculations. Capturing fewer images from these views can negatively impact reconstruction quality. The proposed metric helps to identify these challenging regions and can be used in both structured capturing setups (e.g., using a robotic arm as shown in Section 5.4.1) and potentially unstructured scenarios [Müller et al., 2022]. By providing real-time feedback to the IBR process, the capturing setup can be directed to take more photos in areas where the rendered images has a poor quality. Compared to the FR metric that requires the pair of images to be aligned, an NR metric is oblivious to the pristine reference and, thus, is not subject to misalignment. This property is beneficial in many evaluations where it is difficult or impossible to accurately align the images being compared.

Ultimately, it would be desirable to create a generic metric that is capable of detecting a wide range of artifacts, including traditional ones like compression and noise, as well as those appearing in neural network techniques like in-painting [Li et al., 2022c], face and scene generation [Karras et al., 2020; Rombach et al., 2022], and super-resolution [Wang et al., 2018b; Li et al., 2022a], which primarily involve content hallucination. However, developing such a metric would necessitate a vast dataset that encompasses a diversity of distortions, along with corresponding reference images, which presents a significant challenge. Furthermore, obtaining label data for certain scenarios, such as GANs-generated images, may prove difficult or even impossible. This raises the question of whether it is possible to train a metric using solely clean, natural images, similar to how the human visual system (HVS) can identify a distortion in an image, even if it has never encountered that specific type of distortion before. The HVS can extract a wide range of features from an image to understand and interpret the scene and uses these features to form a mental representation of the scene, which is then compared to stored memories to make judgments about it [Walinga and Stangor, 2014]. This process is not done by comparing an image to a reference but instead through experience and intuition developed from recognizing what natural photos look like and how images with artifacts differ.

7.4 Perception-Driven Rendering and Display

In recent years, head-mounted displays (HMDs) have burst onto the scene as a powerful tool for delivering virtual and augmented reality experiences. Researchers have been working to take these HMDs to the next level by introducing support for other types of cues, such as accommodation cues, which allow for a more natural and lifelike depth perception by triggering changes in the focal distance of the eye [Lambooij et al., 2009; Banks et al., 2016]. However, developing HMDs with correct accommodation cues requires generating 3D content, which demands the processing of a larger amount of data compared to 2D images. Luckily, the human visual system does have certain limitations that can be exploited to achieve an optimal balance between visual quality and computational efficiency [Weier et al., 2017]. This is where perception-driven rendering comes in, which aims to refine the visual experience by leveraging a deep understanding of the human visual system to minimize the rendering cost while still ensuring that users can perceive the full range of visual information.

Chapter 6 presents a perception-driven rendering technique that combines two decomposition methods, linear blending (LB) and light field synthesis (LFS), to achieve real-time rendering and high-fidelity reconstruction in multi-layered light field displays. The critical aspect of rendering in a multi-layered system is the efficient decomposition of an input scene into layers for proper 3D perception. The perceptual experiments conducted in this chapter offered a deeper understanding of the visual quality produced by different decomposition algorithms. The perceptual analysis results reveal that LB and LFS are indistinguishable for textured surfaces for low to mid-spatial frequencies, with LB being closer to the ground truth for high spatial frequencies. LB performs poorly at low-luminance contrast edges for occlusion boundaries but better at high-contrast edges. This may seem counterintuitive, but it is due to the additive combining of focused and defocused patterns on either side of the edge. Additionally, occlusion conditions can be further relaxed for surfaces at sufficiently large eccentricities, where the sensitivity of the human visual system decreases significantly.

The hybrid rendering strategy presented in this chapter can potentially be extended to multiplicative light field displays, which can be formulated with additive light field synthesis [Lanman et al., 2011]. Future work could include investigating the simple decomposition rules in a multiplicative architecture and integrating them with LFS algorithms. Studying the edge artifact in the multiplicative display could lead to interesting insights into accommodative light-field displays. Additionally, the perceptual evaluation of optimization algorithms for dynamic scenes could be an interesting topic of future research. Even though the incorrect boundaries of LB are clearly visible in static scenes, it is unclear whether artifacts are noticeable under motion blur. Therefore, studying the perception of artifacts in interactive and dynamic scenes could provide additional computational benefits.

The perceptual experiments outlined in Chapter 6 are conducted within a limited parametric space. To thoroughly explore the entire space, a visual quality metric, specifically the structural similarity index (SSIM), is employed to predict experimental outcomes and distinguishability in general cases. However, the SSIM measure does not take into account the viewer's gaze direction. To address this, similar to the approaches introduced in Wang and Li [2010], Zhang et al. [2014], and Sim et al. [2020] where the SSIM map is fused using a weighted mean pooling based on the content information or visual saliency in the image, one can devise a gaze-induced pooling strategy that assigns greater weight to errors in the fovea regions (the parts of the image that receive the most visual attention) and less weight to errors in the periphery regions [Mantiuk et al., 2021; Tursun et al., 2019]. Moreover, current perceptual metrics such as LPIPS and DISTS [Ding et al., 2020] can also be considered for this purpose. These metrics calculate visual differences between the extracted features from a pre-trained classification network [Ding et al., 2020] and appear to correlate well with human visual perception.

Finally, as the process of rendering light fields can be costly, a light-weight convolutional neural network can be trained to synthesize a dense LF in real-time from a sparse set of rendered RGB-D images [Xiao et al., 2018] or even directly produce the decomposed images that are utilized in multi-layer light field displays.

Bibliography - Own Work

Bemana, Mojtaba, Joachim Keinert, Karol Myszkowski, Michel Bätz, Matthias Ziegler, H-P Seidel, and Tobias Ritschel [2019]. "Learning to Predict Image-based Rendering Artifacts with Respect to a Hidden Reference Image". In: *Computer Graphics Forum*. Vol. 38. 7. Wiley Online Library, pp. 579–589.

Bemana, Mojtaba, Karol Myszkowski, Jeppe Revall Frisvad, Hans-Peter Seidel, and Tobias Ritschel [2022]. "Eikonal Fields for Refractive Novel-View Synthesis". In: *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–9.

Bemana, Mojtaba, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel [2020]. "X-fields: Implicit neural view-, light-and time-image interpolation". In: *ACM Transactions on Graphics (TOG)* 39.6, pp. 1–15.

Cogalan, Ugur, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel [2022]. "Learning HDR video reconstruction for dual-exposure sensors with temporally-alternating exposures". In: *Computers & Graphics* 105, pp. 57–72.

Çoğalan, Uğur, Mojtaba Bemana, Hans-Peter Seidel, and Karol Myszkowski [2023]. "Video frame interpolation for high dynamic range sequences captured with dual-exposure sensors". In: *Computer Graphics Forum*. Vol. 42. 2. Wiley Online Library, pp. 119–131.

Yu, Hyeonseung, Mojtaba Bemana, Marek Wernikowski, Michał Chwesiuk, Okan Tarhan Tursun, Gurprit Singh, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, and Piotr Didyk [2019]. "A perception-driven hybrid decomposition for multi-layer accommodative displays". In: *IEEE transactions on visualization and computer graphics* 25.5, pp. 1940–1950.

Ziegler, Matthias, Mojtaba Bemana, Joachim Keinert, and Karol Myszkowski [2019]. "Near Real-time Light Field Reconstruction and Rendering for On-set Capture Quality Evaluation". In: *European Light Field Imaging Workshop*. EURASIP.

Bibliography

Pediredla, A., Y. K. Chalmiani, M. G. Scopelliti, M. Chamanzar, S. Narasimhan, and I. Gkioulekas [2020]. “Path tracing estimators for refractive radiative transfer”. In: *ACM Trans. Graph.* 39.6.

Adhikarla, Vamsi Kiran, Marek Vinkler, Denis Sumin, Rafal Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk [2017]. “Towards a Quality Metric for Dense Light Fields”. In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Akeley, Kurt, Simon J. Watt, Ahna Reza Girshick, and Martin S. Banks [2004]. “A stereo display prototype with multiple focal distances”. In: *ACM Transactions on Graphics* 23.3, p. 804. ISSN: 07300301.

Aksit, Kaan, Ward Lopes, Jonghyun Kim, Peter Shirley, and David Luebke [2017]. “Near-eye varifocal augmented reality display using see-through screens”. In: *ACM Transactions on Graphics* 36.6, pp. 1–13. ISSN: 07300301.

Ament, Marco, Christoph Bergmann, and Daniel Weiskopf [2014]. “Refractive radiative transfer equation”. In: *ACM Trans. Graph.* 33.2.

Amirshahi, Seyed Ali, Marius Pedersen, and Stella X Yu [2016]. “Image Quality Assessment by Comparing CNN Features between Images”. In: *J Imag. Sci. and Technology* 60.6, pp. 60410–1.

Andersen, A [1984]. “Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm”. In: *Ultrasonic Imaging* 6.1, pp. 81–94. ISSN: 01617346.

Andersson, Pontus, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D Fairchild [2020]. “FLIP: A Difference Evaluator for Alternating Images.” In: *Proc. ACM Comput. Graph. Interact. Tech.* 3.2, pp. 15–1.

Attal, Benjamin, Jia-Bin Huang, Michael Zollhoefer, Johannes Kopf, and Changil Kim [2021]. *Learning Neural Light Fields with Ray-Space Embedding Networks*. arXiv: 2112.01523 [cs.CV].

Attal, Benjamin, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim [2022]. “Learning Neural Light Fields With Ray-Space Embedding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19819–19829.

Banks, Martin S., David M. Hoffman, Joohwan Kim, and Gordon Wetzstein [2016]. “3D Displays”. In: *Annual Review of Vision Science* 2.1, pp. 397–435. ISSN: 2374-4642.

Bao, Wenbo, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang [2019a]. “Depth-aware Video Frame Interpolation”. In: *Proc. CVPR*.

Bao, Wenbo, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang [2019b]. “MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Barron, Jonathan T., Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan [2021]. “Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields”. In: *ICCV*.

Battisti, Federica, Emilie Bosc, Marco Carli, Patrick Le Callet, and Simone Perugia [2015]. "Objective Image Quality Assessment of 3D Synthesized Views". In: *Image Commun.* 30.C, pp. 78–88.

Berger, K., C. Lipski, C. Linz, A. Sellent, and M. Magnor [2010]. "A ghosting artifact detector for interpolated image quality assessment". In: *IEEE Int. Symp. on Consumer Electronics*, pp. 1–6.

Berger, M., T. Trout, and N. Levit [1990a]. "Ray Tracing Mirages". In: *IEEE Computer Graphics and Applications* 10.3, pp. 36–41.

Berger, Marc, Nancy Levit, and Terry Trout [1990b]. "Rendering mirages and other atmospheric phenomena". In: *Eurographics*, pp. 459–468.

Bianco, Simone, Luigi Celona, Paolo Napoletano, and Raimondo Schettini [2016]. "On the use of deep learning for blind image quality assessment". In: *arXiv:1602.05531*.

Bosc, E., R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin [2011]. "Towards a New Quality Metric for 3-D Synthesized View Assessment". In: *IEEE J of Selected Topics in Signal Processing* 5.7, pp. 1332–1343.

Boss, Mark, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch [2021]. "NeRD: Neural Reflectance Decomposition from Image Collections". In: *ICCV*, pp. 12684–12694.

Bosse, S., D. Maniry, K. R. Müller, T. Wiegand, and W. Samek [2018]. "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment". In: *IEEE TIP* 27.1, pp. 206–219.

Buehler, Chris, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen [2001]. "Unstructured Lumigraph Rendering". In: *Proc. SIGGRAPH*.

Butler, D. J., J. Wulff, G. B. Stanley, and M. J. Black [Oct. 2012]. "A naturalistic open source movie for optical flow evaluation". In: *European Conf. on Computer Vision (ECCV)*. Ed. by A. Fitzgibbon et al. (Eds.) Part IV, LNCS 7577. Springer-Verlag, pp. 611–625.

Čadík, Martin, Robert Herzog, Rafał Mantiuk, Radosław Mantiuk, Karol Myszkowski, and Hans-Peter Seidel [2013]. "Learning to predict localized distortions in rendered images". In: *Comp. Graph. Forum*. Vol. 32. 7, pp. 401–10.

Cao, Xun, Zheng Li, and Qionghai Dai [2011]. "Semi-automatic 2D-to-3D conversion using disparity propagation". In: *IEEE Trans. Broad.* 57.2, pp. 491–9.

Carlini, Nicholas and David Wagner [2017]. "Towards evaluating the robustness of neural networks". In: *2017 ieee symposium on security and privacy (sp)*. Ieee, pp. 39–57.

Chandler, Damon M [2013]. "Seven Challenges in Image Quality Assessment: Past, Present, and Future Research". In: *ISRN Signal Proc.*

Chandrasekhar, Subrahmanyan [1950]. *Radiative Transfer*. Oxford University Press.

Chaurasia, Gaurav, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis [2013]. "Depth Synthesis and Local Warps for Plausible Image-based Navigation". In: *ACM Trans. Graph.* 32.3.

Chen, Anpei, Minye Wu, Yingliang Zhang, Nianyi Li, Jie Lu, Shenghua Gao, and Jingyi Yu [2018a]. "Deep Surface Light Fields". In: *Proc. i3D* 1.1, p. 14.

Chen, Anpei, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su [2022]. "TensoRF: Tensorial Radiance Fields". In: *arXiv preprint arXiv:2203.09517*.

Chen, Billy and Hendrik PA Lensch [2005]. "Light Source Interpolation for Sparsely Sampled Reflectance Fields". In: *Proc. Vision, Modeling and Visualization*, pp. 461–469.

Chen, Guanying, Kai Han, and Kwan-Yee K. Wong [2019a]. "Learning transparent object matting". In: *Int J Computer Vision* 127.10, pp. 1527–1544.

Chen, Ricky T. Q., Yulia Rubanova, Jesse Bettencourt, and David Duvenaud [2018b]. "Neural ordinary differential equations". In: *NIPS*, pp. 6572–6583.

Chen, Shenchang Eric and Lance Williams [1993]. "View Interpolation for Image Synthesis". In: *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '93. Anaheim, CA: Association for Computing Machinery, 279–288. ISBN: 0897916018.

Chen, Xu, Jie Song, and Otmar Hilliges [2019b]. "Monocular Neural Image Based Rendering with Continuous View Control". In: *ICCV*.

Chen, Zhaoxi and Ziwei Liu [2022]. "Relighting4D: Neural Relightable Human from Videos". In: *ECCV*.

Chen, Zhiqin and Hao Zhang [2019]. "Learning Implicit Fields for Generative Shape Modeling". In: *CVPR*.

Cholewiak, Steven A., Gordon D. Love, Pratul P. Srinivasan, Ren Ng, and Martin S. Banks [Nov. 2017]. "Chromablur: Rendering Chromatic Eye Aberration Improves Accommodation and Realism". In: *ACM Trans. Graph.* 36.6, 210:1–210:12. ISSN: 0730-0301.

Chuang, Yung-Yu, Douglas E. Zongker, Joel Hindorff, Brian Curless, David H. Salesin, and Richard Szeliski [2000]. "Environment matting extensions: Towards higher accuracy and real-time capture". In: *SIGGRAPH*, pp. 121–130.

Coleman, Sarah [2012]. *www.theliteratelens.com: Magnum and the Dying Art of Darkroom Printing*.

Coleman, Thomas F. and Yuying Li [1996]. "A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on Some of the Variables". In: *SIAM Journal on Optimization* 6.4, pp. 1040–1058. ISSN: 1052-6234.

Conze, Pierre-Henri, Philippe Robert, and Luce Morin [2012]. "Objective view synthesis quality assessment". In: *Proc. SPIE*.

Dabała, Łukasz, Matthias Ziegler, Piotr Didyk, Frederik Zilly, Joachim Keinert, Karol Myszkowski, Hans-Peter Seidel, Przemysław Rokita, and Tobias Ritschel [2016]. "Efficient Multi-image Correspondences for On-line Light Field Video Processing". In: *Comp. Graph. Forum (Proc. Pacific Graphics)*.

Daly, Scott J [1992]. "Visible differences predictor: an algorithm for the assessment of image fidelity". In: *Human Vision, Visual Processing, and Digital Display III*. Vol. 1666, pp. 2–16.

Debevec, Paul, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar [2000]. "Acquiring the Reflectance Field of a Human Face". In: *Proc. SIGGRAPH*, pp. 145–156.

Ding, Keyan, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma [2021]. "Locally Adaptive Structure and Texture Similarity for Image Quality Assessment". In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2483–2491.

Ding, Keyan, Kede Ma, Shiqi Wang, and Eero P. Simoncelli [2020]. "Image Quality Assessment: Unifying Structure and Texture Similarity". In: *CoRR* abs/2004.07728.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. [2020]. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, Alexey, Jost Tobias Springenberg, and Thomas Brox [2015]. "Learning to Generate Chairs with Convolutional Neural Networks". In: *CVPR*.

Du, Song-Pei, Piotr Didyk, Frédo Durand, Shi-Min Hu, and Wojciech Matusik [2014]. "Improving Visual Quality of View Transitions in Automultiscopic Displays". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 33.6.

Du, Yilun, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu [2021]. “Neural radiance flow for 4d view synthesis and video processing”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, pp. 14304–14314.

Duchowski, Andrew T., Donald H. House, Jordan Gestring, Rui I. Wang, Krzysztof Krejtz, Izabela Krejtz, Radosław Mantiuk, and Bartosz Bazyluk [2014]. “Reducing visual discomfort of 3D stereoscopic sisplays with gaze-contingent depth-of-field”. In: *Proc. ACM Symp. on Appl. Perc. (SAP)*, pp. 39–46.

Dunn, David, Cary Tippets, Kent Torell, Petr Kellnhofer, Kaan Aksit, Piotr Didyk, Karol Myszkowski, David Luebke, and Henry Fuchs [2017]. “Wide Field Of View Varifocal Near-Eye Display Using See-Through Deformable Membrane Mirrors”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.4, pp. 1322–1331. ISSN: 1077-2626.

Engel, Jesse, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan [2017]. “Neural Audio Synthesis of Musical Notes with Wavenet Autoencoders”. In: *JMLR*.

Fang, Yuming, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang [2020]. “Perceptual quality assessment of smartphone photography”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3677–3686.

Flynn, John, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker [2019]. “DeepView: View Synthesis With Learned Gradient Descent”. In: *CVPR*.

Flynn, John, Ivan Neulander, James Philbin, and Noah Snavely [2016]. “DeepStereo: Learning to Predict New Views From the World’s Imagery”. In: *CVPR*.

Forsyth, David A and Jean Ponce [2002]. *Computer Vision: a Modern Approach*. Prentice Hall Professional Technical Reference.

Fried, Ohad and Maneesh Agrawala [2019]. “Puppet Dubbing”. In: *Proc. EGSR*.

Fuchs, Martin, Volker Blanz, Hendrik P.A. Lensch, and Hans-Peter Seidel [2007]. “Adaptive Sampling of Reflectance Fields”. In: *ACM Trans. Graph.* 26.2.

Gao, Chen, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang [2021]. “Dynamic view synthesis from dynamic monocular video”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5712–5721.

Glassner, Andrew S [1988]. “Spacetime ray tracing for animation”. In: *IEEE Computer Graphics and Applications* 2, pp. 60–61.

Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow [2017]. “Unsupervised Monocular Depth Estimation with Left-Right Consistency”. In: *CVPR*.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio [2014]. “Generative adversarial nets”. In: *NIPS*, pp. 2672–80.

Gortler, Steven J, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen [1996]. “The Lumigraph”. In: *SIGGRAPH*.

Gu, Ke, Vinit Jakhetiya, Jun-Fei Qiao, Xiaoli Li, Weisi Lin, and Daniel Thalmann [2017]. “Model-based referenceless quality metric of 3D synthesized images using local image description”. In: *IEEE TIP* 27.1, pp. 394–405.

Guenter, Brian, Mark Finch, Steven Drucker, Desney Tan, and John Snyder [2012]. “Foveated 3D graphics”. In: *ACM Transactions on Graphics* 31.6, p. 1. ISSN: 07300301.

Guo, Kaiwen et al. [2019]. “The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting”. In: *ACM Trans. Graph (Proc SIGGRAPH Asia)* 38.5.

Guo, Yuan-Chen, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang [2021]. *NeRFReN: Neural Radiance Fields with Reflections*. arXiv: 2111.15234 [cs.CV].

Guthe, S., P. Schardt, M. Goesele, and D. Cunningham [2016]. "Ghosting and popping detection for image-based rendering". In: *Proc. 3DTV*, pp. 1–4.

Gutierrez, Diego, Adolfo Munoz, Oscar Anson, and Francisco J. Seron [2005]. "Non-linear Volume Photon Mapping". In: *EGSR*, pp. 291–300.

Hairer, Ernst and Gerhard Wanner [1996]. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. second revised. Springer.

Hedman, P., T. Ritschel, G. Drettakis, and G.J. Brostow [2016a]. "Scalable Inside-Out Image-Based Rendering". In: *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 35.6.

Hedman, Peter, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel J. Brostow [2018]. "Deep Blending for Free-Viewpoint Image-Based Rendering". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 37.6.

Hedman, Peter, Tobias Ritschel, George Drettakis, and Gabriel Brostow [2016b]. "Scalable Inside-Out Image-Based Rendering". In: *ACM Trans. Graph. (Proc. SIGRAPH Asia)* 35.6.

Heide, Felix, Gordon Wetzstein, Ramesh Raskar, and Wolfgang Heidrich [2013]. "Adaptive image synthesis for compressive displays". In: *ACM Transactions on Graphics* 32.4, p. 1. ISSN: 07300301.

Henzler, Philipp, Niloy J Mitra, and Tobias Ritschel [2019a]. "Escaping Plato's Cave: 3D Shape From Adversarial Rendering". In.

Henzler, Philipp, Niloy J. Mitra, and Tobias Ritschel [2019b]. "Escaping Plato's cave: 3D shape from adversarial rendering". In: *ICCV*, pp. 9984–9993.

Herzog, Robert, Martin Čadík, Tunç O. Aydin, Kwang In Kim, Karol Myszkowski, and Hans-Peter Seidel [2012]. "NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis". In: *Comp. Graph. Forum* 31.2, pp. 545–54.

Hinton, Geoffrey E and Ruslan R Salakhutdinov [2006]. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786.

Hu, Xinda and Hong Hua [2014]. "High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics." In: *Optics express* 22.11, pp. 13896–903. ISSN: 1094-4087.

Hua, Hong [2017]. "Enabling Focus Cues in Head-Mounted Displays". In: *Proceedings of the IEEE* 105.5, pp. 805–824. ISSN: 0018-9219.

Hua, Hong and Bahram Javidi [2014]. "A 3D integral imaging optical see-through head-mounted display". In: *Optics Express* 22.11, p. 13484. ISSN: 1094-4087.

Huang, Fu-Chung, Kevin Chen, and Gordon Wetzstein [2015a]. "The light field stereoscope". In: *ACM Transactions on Graphics* 34.4, 60:1–60:12. ISSN: 07300301.

Huang, Fu-Chung, David Luebke, and Gordon Wetzstein [2015b]. "The light field stereoscope". In: *ACM SIGGRAPH 2015 Emerging Technologies on - SIGGRAPH '15*. New York, New York, USA: ACM Press, pp. 1–1. ISBN: 9781450336352.

Huang, Xin, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang [2021]. *HDR-NeRF: High Dynamic Range Neural Radiance Fields*. arXiv: `2111.14451 [cs.CV]`.

Huang, Yi-Hua, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao [2022]. "StylizedNeRF: consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18342–18352.

Hullin, Matthias B., Martin Fuchs, Ivo Ihrke, Hans-Peter Seidel, and Hendrik P. A. Lensch [2008]. "Fluorescent Immersion Range Scanning". In: *ACM Trans. Graph.* 27.3.

Ichnowski, Jeffrey, Yahav Avigal, Justin Kerr, and Ken Goldberg [2021]. *Dex-NeRF: Using a Neural Radiance Field to Grasp Transparent Objects*. arXiv: `2110.14217 [cs.RO]`.

Ihrke, Ivo, Kiriakos N Kutulakos, Hendrik PA Lensch, Marcus Magnor, and Wolfgang Heidrich [2008]. "State of the art in transparent and specular object reconstruction". In: *Eurographics STAR*.

Ihrke, Ivo, Kiriakos N Kutulakos, Hendrik PA Lensch, Marcus Magnor, and Wolfgang Heidrich [2010]. "Transparent and specular object reconstruction". In: *Comp. Graph. Forum* 29.8, pp. 2400–2426.

Ihrke, Ivo, Gernot Ziegler, Art Tevs, Christian Theobalt, Marcus Magnor, and Hans-Peter Seidel [2007]. "Eikonal rendering: Efficient light transport in refractive objects". In: *ACM Trans. Graph.* 26.3.

Iqbal, Umar, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, and Jan Kautz [2022]. "RANA: Relightable Articulated Neural Avatars". In: *arXiv preprint arXiv:2212.03237*.

Isola, Phillip, Jun Zhu, Tinghui Zhou, and Alexei A Efros [2017]. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR*.

Jacobs, David, Orazio Gallo, Emily Cooper, Kari Pulli, and Marc Levoy [2015]. "Simulating the visual experience of very bright and very dark scenes". In: *ACM Trans Graph (TOG)* 34.3, p. 25.

Jaderberg, Max, Karen Simonyan, Andrew Zisserman, et al. [2015]. "Spatial Transformer Networks". In: *Proc. NIPS*.

Jegou, Herve, Matthijs Douze, and Cordelia Schmid [2008]. "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search". In: *ECCV*, pp. 304–317.

Jiang, Huaizu, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz [2018]. "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation". In: *CVPR*.

Jinjin, Gu, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao [2020]. "Pipal: a large-scale image quality assessment dataset for perceptual image restoration". In: *European Conference on Computer Vision*. Springer, pp. 633–651.

Kalantari, Nima Khademi, Ting-Chun Wang, and Ravi Ramamoorthi [2016]. "Learning-based View Synthesis for Light Field Cameras". In: *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 35.6.

Kang, Le, Peng Ye, Yi Li, and David Doermann [2014]. "Convolutional neural networks for no-reference image quality assessment". In: *CVPR*, pp. 1733–40.

Kap-Kee, KIM [2015]. *Apparatus and method for correcting disparity map*. US Patent 9,208,541.

Karras, Tero, Samuli Laine, and Timo Aila [2019]. "A Style-based Generator Architecture for Generative Adversarial Networks". In: *CVPR*, pp. 4401–4410.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila [2020]. "Analyzing and Improving the Image Quality of StyleGAN". In: *Proc. CVPR*.

Ke, Junjie, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang [2021]. "Musiq: Multi-scale image quality transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157.

Kellnhofer, Petr, Piotr Didyk, Karol Myszkowski, Mohamed M Hefeeda, Hans-Peter Seidel, and Wojciech Matusik [2016a]. "GazeStereo3D: Seamless disparity manipulations". In: *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 35.4.

Kellnhofer, Petr, Piotr Didyk, Szu-Po Wang, Pitchaya Sitthi-Amorn, William Freeman, Fredo Durand, and Wojciech Matusik [2017]. "3DTV at Home: Eulerian-Lagrangian Stereo-to-Multiview Conversion". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 36.4.

Kellnhofer, Petr, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel [2016b]. "Transformation-aware perceptual image metric". In: *J Electronic Imaging* 25.5, p. 053014.

Khaing, May Phyo and Mukunoki Masayuki [2019]. "Transparent Object Detection Using Convolutional Neural Network". In: *Big Data Analysis and Deep Learning Applications*.

Khan, Erum Arif, Erik Reinhard, Roland W. Fleming, and Heinrich H. Bülthoff [2006]. "Image-based material editing". In: *ACM Trans. Graph.* 25.3, pp. 654–663.

Kim, J. and S. Lee [2017]. "Fully Deep Blind Image Quality Predictor". In: *IEEE J Sel. Topics in Signal Processing* 11.1, pp. 206–220.

Kim, J., H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik [2017a]. "Deep Convolutional Neural Models for Picture-Quality Prediction: Challenges and Solutions to Data-Driven Image Quality Assessment". In: *IEEE Signal Processing Magazine* 34.6, pp. 130–141.

Kim, Jaewon, Ilya Reshetouski, and Abhijeet Ghosh [2017b]. "Acquiring Axially-Symmetric Transparent Objects Using Single-View Transmission Imaging". In: *CVPR*, pp. 3559–3567.

Kingma, Diederik P and Max Welling [2013]. "Auto-encoding Variational Bayes". In: *Proc. ICLR*.

Kolos, Maria, Artem Sevastopolsky, and Victor Lempitsky [2020]. "TRANSPR: Transparency Ray-Accumulating Neural 3D Scene Point Renderer". In: *3DV*, pp. 1167–1175.

Konrad, Robert, Nitish Padmanaban, Keenan Molner, Emily A Cooper, and Gordon Wetzstein [2017]. "Accommodation-invariant computational near-eye displays". In: *ACM Transactions on Graphics* 36.4, pp. 1–12. ISSN: 07300301.

Kopf, Johannes, Fabian Langguth, Daniel Scharstein, Richard Szeliski, and Michael Goesele [2013]. "Image-Based Rendering in the Gradient Domain". In: *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 32.6.

Koulieris, George-Alex, Bee Bui, Martin S Banks, and George Drettakis [2017]. "Accommodation and Comfort in Head-Mounted Displays". In: *ACM Transactions on Graphics* 36.4, pp. 1–11.

Kramida, Gregory [2016]. "Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays". In: *IEEE Transactions on Visualization and Computer Graphics* 22.7, pp. 1912–1931. ISSN: 1077-2626.

Kutulakos, Kiriakos and Eron Steger [2008]. "A Theory of Refractive and Specular 3D Shape by Light-Path Triangulation". In: *Int J Computer Vision* 76.1, 13—29.

Lambooij, Marc, Wijnand IJsselsteijn, Marten Fortuin, and Ingrid Heynderickx [2009]. "Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review". In: *Journal of Imaging Science and Technology* 53.3, p. 030201. ISSN: 10623701.

Lanman, Douglas and David Luebke [2013]. "Near-eye light field displays". In: *ACM Transactions on Graphics* 32.6, pp. 1–10. ISSN: 07300301.

Lanman, Douglas, Gordon Wetzstein, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar [2011]. "Polarization fields". In: *ACM Transactions on Graphics* 30.6, p. 1. ISSN: 07300301.

Lee, Seungjae, Jaebum Cho, Byounghyo Lee, Youngjin Jo, Changwon Jang, Dongyeon Kim, and Byoungho Lee [2017]. "Foveated Retinal Optimization for See-through Near-Eye Multi-Layer Displays (Invited Paper)". In: *IEEE Access* 4.c, pp. 1–1. ISSN: 2169-3536.

Lee, Seungjae, Changwon Jang, Seokil Moon, Jaebum Cho, and Byoungho Lee [2016]. "Additive light field displays". In: *ACM Transactions on Graphics* 35.4, pp. 1–13. ISSN: 07300301.

Legge, G.E. and J.M. Foley [1980]. “Contrast masking in human vision”. In: *Journal of the Optical Society of America* 70.12, pp. 1458–1471.

Levin, Anat and Fredo Durand [2010]. “Linear view synthesis using a dimensionality gap light field prior”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1831–1838. ISBN: 978-1-4244-6984-0.

Levoy, Marc and Pat Hanrahan [1996]. “Light field rendering”. In: *SIGGRAPH*, pp. 31–42.

Li, Haoying, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen [2022a]. “Srdiff: Single image super-resolution with diffusion probabilistic models”. In: *Neurocomputing* 479, pp. 47–59.

Li, Tianye, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. [2022b]. “Neural 3D Video Synthesis From Multi-View Video”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5521–5531.

Li, Wenbo, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia [2022c]. “MAT: Mask-Aware Transformer for Large Hole Image Inpainting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10758–10768.

Li, Zhengqi, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely [2022d]. “DynIBaR: Neural Dynamic Image-Based Rendering”. In: *arXiv preprint arXiv:2211.11082*.

Li, Zhengqin, Yu-Ying Yeh, and Manmohan Chandraker [2020]. “Through the looking glass: neural 3D reconstruction of transparent shapes”. In: *CVPR*, pp. 1262–1271.

Li, Zhong, Liangchen Song, Celong Liu, Junsong Yuan, and Yi Xu [2021]. “Neulf: Efficient novel view synthesis with neural 4d light field”. In: *arXiv preprint arXiv:2105.07112*.

Lin, Caizhang, Chris Varekamp, Karel Hinnen, and Gerard De Haan [2012]. “Interactive disparity map post-processing”. In: *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pp. 448–455.

Lin, Chen-Hsuan, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey [2021]. “Barf: Bundle-adjusting neural radiance fields”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5741–5751.

Lin, Hanhe, Vlad Hosu, and Dietmar Saupe [2019]. “KADID-10k: A Large-scale Artificially Distorted IQA Database”. In: *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 1–3.

Lin, Kwan Yee and Guanxiang Wang [2018]. “Hallucinated-IQA: No-Reference Image Quality Assessment via Adversarial Learning”. In: *CVPR*.

Lindell, David B, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein [2022]. “Bacon: Band-limited coordinate networks for multiscale scene representation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16252–16262.

Ling, Suiyi and Patrick Le Callet [2018]. “How to Learn the Effect of Non-Uniform Distortion on Perceived Visual Quality? Case Study Using Convolutional Sparse Coding for Quality Assessment of Synthesized Views”. In: *ICIP*, pp. 286–290.

Ling, Suiyi, Jing Li, Junle Wang, and Patrick Le Callet [2019]. “GANs-NQM: A Generative Adversarial Networks based No Reference Quality Assessment Metric for RGB-D Synthesized Views”. In: *arXiv:1903.12088*.

Lipski, Christian, Christian Linz, Kai Berger, Anita Sellent, and Marcus Magnor [2010]. “Virtual Video Camera: Image-Based Viewpoint Navigation Through Space and Time”. In: *Computer Graphics Forum* 29.8.

Liu, K. and J. Y. Yang [1989]. "Reconstruction Of 3-D Refractive Index Fields From Multi-Frame Interfetometric Data". In: *New Methods in Microscopy and Low Light Imaging*. Vol. 1161. Proc. SPIE, pp. 42–46.

Liu, Lingjie, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt [2020]. "Neural sparse voxel fields". In: *Advances in Neural Information Processing Systems* 33, pp. 15651–15663.

Liu, Rosanne, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski [2018]. "An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution". In: *Proc. NIPS*.

Lombardi, Stephen, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh [2019a]. "Neural Volumes: Learning Dynamic Renderable Volumes from Images". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 38.4.

Lombardi, Stephen, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh [2019b]. "Neural volumes: learning dynamic renderable volumes from images". In: *ACM Trans. Graph.* 38.4.

Love, Gordon D, David M Hoffman, Philip J W Hands, James Gao, Andrew K Kirby, and Martin S Banks [2009]. "High-speed switchable lens enables the development of a volumetric stereoscopic display." In: *Optics express* 17.18, pp. 15716–25. ISSN: 1094-4087. arXiv: NIHMS150003.

Lubin, Jeffrey [1995]. "Vision Models for Target Detection and Recognition". In: ed. by Eli Peli. World Scientific. Chap. A Visual Discrimination Model for Imaging System Design and Evaluation, pp. 245–283.

Lyu, Jiahui, Bojian Wu, Dani Lischinski, Daniel Cohen-Or, and Hui Huang [2020]. "Differentiable refraction-tracing for mesh reconstruction of transparent objects". In: *ACM Trans. Graph.* 39.6.

MacKenzie, Kevin J, David M Hoffman, and Simon J Watt [2010]. "Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control". In: *Journal of Vision* 10.8, pp. 22–22. ISSN: 1534-7362.

Mahajan, Dhruv, Fu-Chung Huang, Wojciech Matusik, Ravi Ramamoorthi, and Peter Belhumeur [2009]. "Moving Gradients: a Path-based Method for Plausible Image Interpolation". In: *ACM Trans. Graph.* Vol. 28. 3, p. 42.

Maimone, Andrew and Henry Fuchs [2013]. "Computational augmented reality eyeglasses". In: *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. October. IEEE, pp. 29–38. ISBN: 978-1-4799-2869-9.

Maimone, Andrew, Andreas Georgiou, and Joel S Kollin [2017]. "Holographic Near-Eye Displays for Virtual and Augmented Reality". In: *ACM Transactions on Graphics* 36.4, pp. 1–16.

Malzbender, Tom, Dan Gelb, and Hans Wolters [2001]. "Polynomial Texture Maps". In: *Proc. SIGGRAPH*.

Manning, Russell A. and Charles R. Dyer [1999]. "Interpolating View and Scene Motion by Dynamic View Morphing". In: *CVPR*. Vol. 1, pp. 388–394.

Mantiuk, Rafal, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich [2011a]. "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions". In: *ACM Trans. Graph. (Proc. SIGGRAPH)*.

Mantiuk, Rafał K, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney [2021]. "Fovvideovdp: A visible difference predictor for wide field-of-view video". In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–19.

Mantiuk, Rafat, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich [July 2011b]. "HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions". In: *ACM Trans. Graph.* 30.4, 40:1–40:14. ISSN: 0730-0301.

Mark, William R, Leonard McMillan, and Gary Bishop [1997]. "Post-rendering 3D Warping". In: *Proc. i3D*.

Mathews, Steven and Philip B Kruger [1994]. "Spatiotemporal transfer function of human accommodation". In: *Vision Research* 34.15, pp. 1965–1980. ISSN: 00426989.

Matsuda, Nathan, Alexander Fix, and Douglas Lanman [2017]. "Focal surface displays". In: *ACM Transactions on Graphics* 36.4, pp. 1–14. ISSN: 07300301.

Matusik, Wojciech, Hanspeter Pfister, Remo Ziegler, Addy Ngan, and Leonard Mcmillan [2002]. "Acquisition and Rendering of Transparent and Refractive Objects". In: *EGWR*, pp. 267–278.

Mauderer, Michael, Simone Conte, Miguel A. Nacenta, and Dhanraj Vishwanath [2014]. "Depth perception with gaze-contingent depth of field". In: *Proc Human Fact in Comp Sys (CHI)*, pp. 217–226.

Max, Nelson [1995]. "Optical models for direct volume rendering". In: *IEEE Trans Vis Comput Graph* 1.2, pp. 99–108.

Maximov, Maxim, Laura Leal-Taixé, Mario Fritz, and Tobias Ritschel [2019]. "Deep Appearance Maps". In: *Proc. ICCV*.

McMillan, Leonard and Gary Bishop [1995]. "Plenoptic modeling: An image-based rendering system". In: *SIGGRAPH*, pp. 39–46.

Meka, Abhimitra, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. [2020]. "Deep relightable textures: volumetric performance capture with neural rendering". In: *ACM Transactions on Graphics (TOG)* 39.6, pp. 1–21.

Meka, Abhimitra et al. [2019]. "Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination". In: *ACM Trans. Graph (Proc SIGGRAPH)* 38.4.

Mercier, Olivier, Yusufu Sulai, Kevin Mackenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman [2017]. "Fast gaze-contingent optimal decompositions for multifocal displays". In: *ACM Transactions on Graphics* 36.6, pp. 1–15. ISSN: 07300301.

Mildenhall, Ben, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar [2019]. "Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 38.4.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng [2020]. "NeRF: Representing scenes as neural radiance fields for view synthesis". In: *ECCV*, pp. 405–421.

Moon, Seokil, Chang-Kun Lee, Dukho Lee, Changwon Jang, and Byoungho Lee [2017]. "Layered Display with Accommodation Cue using Scattering Polarizers". In: *IEEE Journal of Selected Topics in Signal Processing* 4553.c, pp. 1–1. ISSN: 1932-4553.

Moorthy, A.K. and A.C. Bovik [2010]. "A Two-Step Framework for Constructing Blind Image Quality Indices". In: *IEEE Signal Proc. Letters* 17.5, pp. 513–16. ISSN: 1070-9908.

Müller, Thomas, Alex Evans, Christoph Schied, and Alexander Keller [2022]. "Instant neural graphics primitives with a multiresolution hash encoding". In: *arXiv preprint arXiv:2201.05989*.

Musgrave, F. Kenton [1990]. "A Note on Ray Tracing Mirages". In: *IEEE Computer Graphics and Applications* 10.6, pp. 10–12.

Narain, Rahul, Rachel A. Albert, Abdullah Bulbul, Gregory J. Ward, Martin S. Banks, and James F. O'Brien [2015]. "Optimal presentation of imagery with focus cues on multi-plane displays". In: *ACM Transactions on Graphics* 34.4, 59:1–59:12. ISSN: 07300301.

Narwaria, Manish and Weisi Lin [2010]. "Objective image quality assessment based on support vector regression". In: *IEEE Trans. Neural Networks* 21.3, pp. 515–9.

Nguyen Phuoc, Thu, Chuan Li, Stephen Balaban, and Yongliang Yang [2018]. "RenderNet: A deep Convolutional Network for Differentiable Rendering from 3D Shapes". In.

Nguyen Phuoc, Thu, Chuan Li, Lucas Theis, Christian Richardt, and Yongliang Yang [2019]. "HoloGAN: Unsupervised Learning of 3D Representations From Natural Images". In: *ICCV*.

Niemeyer, Michael, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan [2022]. "RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs". In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Niemeyer, Michael, Lars Mescheder, Michael Oechsle, and Andreas Geiger [2019]. "Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics". In: *Proc. ICCV*.

Oechsle, Michael, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger [2019]. "Texture Fields: Learning Texture Representations in Function Space". In: *ICCV*.

Pan, Jen-I, Jheng-Wei Su, Kai-Wen Hsiao, Ting-Yu Yen, and Hung-Kuo Chu [2022]. "Sampling Neural Radiance Fields for Refractive Objects". In: *SIGGRAPH Asia 2022 Technical Communications*, pp. 1–4.

Pandey, Rohit, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello [2021]. "Total relighting: learning to relight portraits for background replacement". In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–21.

Park, Keunhong, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla [2021]. "Nerfies: Deformable Neural Radiance Fields". In: *ICCV*, pp. 5865–5874.

Patney, Anjul and Aaron Lefohn [2018]. "Detecting Aliasing Artifacts in Image Sequences Using Deep Neural Networks". In: *Proc. HPG*.

Patney, Anjul, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn [2016]. "Towards foveated rendering for gaze-tracked virtual reality". In: *ACM Transactions on Graphics* 35.6, pp. 1–12. ISSN: 07300301.

Peers, Pieter and Philip Dutré [2003]. "Wavelet environment matting". In: *EGSR*, pp. 157–166.

Peers, Pieter, Dhruv K. Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec [2009]. "Compressive Light Transport Sensing". In: *ACM Trans. Graph.* 28.1.

Penner, Eric and Li Zhang [2017a]. "Soft 3D Reconstruction for View Synthesis". In: *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 36.6.

Penner, Eric and Li Zhang [2017b]. "Soft 3D Reconstruction for View Synthesis". In: *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 36.6.

Philip, Julien, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis [2019]. "Multi-View Relighting Using a Geometry-Aware Network". In: *ACM Trans. Graph.* 38.4.

Ponomarenko, N., V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti [2009]. "TID2008 - A database for evaluation of full-reference visual quality assessment metrics". In: *Advances of Modern Radioelectronics* 10, pp. 30–45.

Pontryagin, Lev Semenovich [1987]. *Mathematical theory of optimal processes*. CRC press.

Preisendorfer, Rudolph W. [1957]. "A mathematical foundation for radiative transfer theory". In: *Journal of Mathematics and Mechanics* 6.6, pp. 685–730.

Pumarola, Albert, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer [2020]. "D-NeRF: Neural Radiance Fields for Dynamic Scenes". In: *CVPR*.

Pumarola, Albert, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer [2021]. "D-nerf: Neural radiance fields for dynamic scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327.

Radford, Alec, Luke Metz, and Soumith Chintala [2015]. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *Arxiv 1511.06434*.

Ravikumar, Sowmya, Kurt Akeley, and Martin S. Banks [2011]. "Creating effective focus cues in multi-plane 3D displays". In: *Optics Express* 19.21, p. 20940. ISSN: 1094-4087.

Reda, Fitsum, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless [2022]. "Film: Frame interpolation for large motion". In: *European Conference on Computer Vision*. Springer, pp. 250–266.

Reddy, Dikpal, Ravi Ramamoorthi, and Brian Curless [2012]. "Frequency-Space Decomposition and Acquisition of Light Transport under Spatially Varying Illumination". In: *Proc. ECCV*, 596–610.

Reed, Scott E, Yi Zhang, Yuting Zhang, and Honglak Lee [2015]. "Deep Visual Analogy-making". In: *NIPS*.

Ren, Peiran, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo [2015]. "Image Based Relighting Using Neural Networks". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 34.4.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer [2022]. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

Ronneberger, O., P.Fischer, and T. Brox [2015]. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *MICCAI*, pp. 234–241.

Saad, M.A., A.C. Bovik, and C. Charrier [2012]. "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain". In: *IEEE TIP* 21.8, pp. 3339–3352.

Sabater, Neus, Guillaume Boisson, Benoit Vandame, Paul Kerbiriou, Frederic Babon, Matthieu Hog, Tristan Langlois, Remy Gendrot, Olivier Bureller, Arno Schubert, and Valerie Allie [2017]. "Dataset and Pipeline for Multi-View Light-Field Video". In: *CVPR Workshops*.

Saito, Shunsuke, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li [2019]. "PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization". In: *Proc. ICCV*, pp. 2304–2314.

Sajjan, Shreeyak, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song [2020]. "ClearGrasp: 3D shape estimation of transparent objects for manipulation". In: *ICRA*, pp. 3634–3642.

Schönberger, Johannes Lutz and Jan-Michael Frahm [2016]. "Structure-from-Motion Revisited". In: *CVPR*, pp. 4104–4113.

Serrano, Ana, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia [2019]. "Motion parallax for 360 RGBD video". In: *IEEE Trans. Vis. & Comp. Graph. (Proc. IEEE VR)* 25.

Sheikh, H.R., M.F. Sabir, and A.C. Bovik [2006]. “A statistical evaluation of recent full reference image quality assessment algorithms”. In: *IEEE TIP* 15.11, pp. 3440–3451.

Sim, Hyeonjun, Jihyong Oh, and Munchurl Kim [2021]. “Xvfi: Extreme video frame interpolation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14489–14498.

Sim, Kyohoon, Jiachen Yang, Wen Lu, and Xinbo Gao [2020]. “MaD-DLS: mean and deviation of deep and local similarity for image quality assessment”. In: *IEEE Transactions on Multimedia* 23, pp. 4037–4048.

Simonyan, Karen and Andrew Zisserman [2014]. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv:1409.1556*.

Sinha, Sudipta N., Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski [2012]. “Image-Based Rendering for Scenes with Reflections”. In: *ACM Trans. Graph.* 31.4.

Sitzmann, Vincent, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand [2021]. “Light field networks: Neural scene representations with single-evaluation rendering”. In: *Advances in Neural Information Processing Systems* 34, pp. 19313–19325.

Sitzmann, Vincent, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer [2019a]. “DeepVoxels: Learning Persistent 3D Feature Embeddings”. In: *CVPR*.

Sitzmann, Vincent, Michael Zollhöfer, and Gordon Wetzstein [2019b]. “Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations”. In: *NeurIPS*.

Smith, Jonathan D., Kamyar Azizzadenesheli, and Zachary E. Ross [2021]. “EikoNet: Solving the Eikonal Equation With Deep Neural Networks”. In: *IEEE Trans Geoscience and Remote Sensing* 59.12, pp. 10685–10696.

Solh, M., G. AlRegib, and J. M. Bauza [2011]. “3VQM: A vision-based quality measure for DIBR-based 3D videos”. In: *2011 IEEE Int. Conf. on Multimedia and Expo*, pp. 1–6.

Srinivasan, Pratul P, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron [2021]. “Nerv: Neural reflectance and visibility fields for relighting and view synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7495–7504.

Stam, Jos [2020]. “Computing Light Transport Gradients using the Adjoint Method”. In: *arXiv preprint arXiv:2006.15059*.

Stam, Jos and Eric Languénou [1996]. “Ray tracing in non-constant media”. In: *EGWR*, pp. 225–234.

Stets, J. D., A. Dal Corso, J. B. Nielsen, R. A. Lyngby, S. H. N. Jensen, J. Wilm, M. B. Doest, C. Gundlach, E. R. Eiriksson, K. Conradsen, A. B. Dahl, J. A. Bærentzen, J. R. Frisvad, and H. Aanæs [2017]. “Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering”. In: *Applied Optics* 56.27, pp. 7679–7690.

Stets, Jonathan, Zhengqin Li, Jeppe Revall Frisvad, and Manmohan Chandraker [2019]. “Single-shot analysis of refractive shape using convolutional neural networks”. In: *IEEE WACV*, pp. 995–1003.

Sun, Cheng, Min Sun, and Hwann-Tzong Chen [2022]. “Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5459–5469.

Sun, D., E. B. Sudderth, and M. J. Black [2012]. “Layered Segmentation and Optical Flow Estimation Over Time”. In: *CVPR*, pp. 1768–1775.

Sun, Deqing, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz [2018a]. "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume". In: *CVPR*.

Sun, Qi, Fu-Chung Huang, Joohwan Kim, Li-Yi Wei, David Luebke, and Arie Kaufman [2017]. "Perceptually-guided foveation for light field displays". In: *ACM Transactions on Graphics* 36.6, pp. 1–13. ISSN: 07300301.

Sun, Shao-Hua, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim [2018b]. "Multi-view to Novel View: Synthesizing Novel Views with Self-learned Confidence". In: *Proc. ECCV*.

Swafford, Nicholas T., José A. Iglesias-Guitian, Charalampos Koniaris, Bochang Moon, Darren Cosker, and Kenny Mitchell [2016]. "User, Metric, and Computational Evaluation of Foveated Rendering Methods". In: *Proceedings of the ACM Symposium on Applied Perception*. SAP '16, pp. 7–14.

Sweeney, D. W. and C. M. Vest [1973]. "Reconstruction of Three-Dimensional Refractive Index Fields from Multidirectional Interferometric Data". In: *Applied Optics* 12.11, pp. 2649–2664.

Takikawa, Towaki, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler [2021]. "Neural geometric level of detail: Real-time rendering with implicit 3D shapes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11358–11367.

Talebi, Hossein and Peyman Milanfar [2018]. "Nima: Neural image assessment". In: *IEEE TIP* 27.8, pp. 3998–4011.

Tang, Huixuan, N. Joshi, and A. Kapoor [2011]. "Learning a blind measure of perceptual image quality". In: *CVPR*, pp. 305–12.

Teh, Arjun, Matthew O'Toole, and Ioannis Gkioulekas [2022]. "Adjoint nonlinear ray tracing". In: *ACM Transactions on Graphics (TOG)* 41.4, pp. 1–13.

Tewari, A., O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer [2020]. "State of the art on neural rendering". In: *Comp. Graph. Forum* 39.2.

Tewari, Ayush, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. [2022]. "Advances in neural rendering". In: *Computer Graphics Forum*. Vol. 41. 2. Wiley Online Library, pp. 703–735.

Thies, Justus, Michael Zollhöfer, and Matthias Nießner [2019]. "Deferred Neural Rendering: Image Synthesis using Neural Textures". In: *ACM Trans. Graph. (Proc. SIGGRAPH)*.

Tian, Chao, Yongying Yang, Yongmo Zhuo, Tao Wei, and Tong Ling [2011]. "Tomographic reconstruction of three-dimensional refractive index fields by use of a regularized phase-tracking technique and a polynomial approximation method". In: *Applied Optics* 50, pp. 6495–6504.

Tian, S., L. Zhang, L. Morin, and O. Déforges [2018]. "NIQSV+: A No-Reference Synthesized View Quality Assessment Metric". In: *IEEE TIP* 27.4, pp. 1652–64.

Tretschk, Edgar, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt [2021a]. "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12959–12970.

Tretschk, Edgar, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt [2021b]. "Non-Rigid Neural Radiance Fields:

Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video". In: *ICCV*.

Trifonov, Borislav, Derek Bradley, and Wolfgang Heidrich [2006a]. "Tomographic Reconstruction of Transparent Objects". In: *EGSR*, pp. 51–60.

Trifonov, Borislav, Derek Bradley, and Wolfgang Heidrich [2006b]. "Tomographic reconstruction of transparent objects". In: *ACM SIGGRAPH Sketches*, p. 55.

Tsai, Chia-Yin, Ashok Veeraraghavan, and Aswin Sankaranarayanan [2015]. "What does a single light-ray reveal about a transparent object?" In: *Proc. IEEE ICIP*.

Tursun, Okan Tarhan, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk [2019]. "Luminance-contrast-aware foveated rendering". In: *ACM Transactions on Graphics (TOG)* 38.4, pp. 1–14.

Vangorp, Peter, Gaurav Chaurasia, P-Y Laffont, Roland W Fleming, and George Drettakis [2011]. "Perception of Visual Artifacts in Image-Based Rendering of Façades". In: *Comp. Graph. Forum*. Vol. 30. 4, pp. 1241–50.

Vogels, Thijs, Fabrice Rousselle, Brian McWilliams, Gerhard Röthlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák [2018]. "Denoising with Kernel Prediction and Asymmetric Loss Functions". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 37.4, 124:1–124:15.

Waechter, Michael, Mate Beljan, Simon Fuhrmann, Nils Moehrle, Johannes Kopf, and Michael Goesele [2017]. "Virtual Rephotography: Novel View Prediction Error for 3D Reconstruction". In: *ACM Trans. Graph.* 36.1.

Walinga, Jennifer and Charles Stangor [2014]. "Introduction to psychology-1st canadian edition". In.

Wang, Huamin and Ruigang Yang [2005]. "Towards Space: Time Light Field Rendering". In: *Proc. i3D*.

Wang, Qianqian, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser [2021a]. "Ibrnet: Learning multi-view image-based rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699.

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro [2018a]. "Video-to-Video Synthesis". In: *NeurIPS*.

Wang, Ting-Chun, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A. Efros, and Ravi Ramamoorthi [2017]. "Light Field Video Capture Using a Learning-Based Hybrid Imaging System". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 36.4.

Wang, Xintao, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy [2018b]. "Esrgan: Enhanced super-resolution generative adversarial networks". In: *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0.

Wang, Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli [2004a]. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612. ISSN: 1057-7149.

Wang, Zhou and Alan C. Bovik [2006]. *Modern Image Quality Assessment*. Morgan & Claypool Publishers.

Wang, Zhou, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli [2004b]. "Image quality assessment: from error visibility to structural similarity". In: *IEEE TIP* 13.4, pp. 600–12.

Wang, Zhou and Qiang Li [2010]. "Information content weighting for perceptual image quality assessment". In: *IEEE Transactions on image processing* 20.5, pp. 1185–1198.

Wang, Ziyu, Liao Wang, Fuqiang Zhao, Minye Wu, Lan Xu, and Jingyi Yu [2021b]. *MirrorNeRF: One-Shot Neural Portrait Radiance Field from Multi-Mirror Catadioptric Imaging*. arXiv: `2104.02607 [cs.CV]`.

Watson, Andrew B. and Albert J. Ahumada [2011]. "Blur clarified: A review and synthesis of blur discrimination". In: *Journal of Vision* 11.5, p. 10.

Watson, Andrew B. and Denis G. Pelli [1983]. "Quest: A Bayesian adaptive psychometric method". In: *Perception & Psychophysics* 33.2, pp. 113–120. ISSN: 1532-5962.

Weier, Martin, Michael Stengel, Thorsten Roth, Piotr Didyk, Elmar Eisemann, Martin Eisemann, Steve Grogorick, André Hinkenjann, Ernst Kruijff, Marcus Magnor, et al. [2017]. "Perception-driven accelerated rendering". In: *Computer Graphics Forum*. Vol. 36. 2. Wiley Online Library, pp. 611–643.

Wetzstein, Gordon, Douglas Lanman, Wolfgang Heidrich, and Ramesh Raskar [2011a]. "Layered 3D". In: *ACM Transactions on Graphics* 30.4, p. 1. ISSN: 07300301.

Wetzstein, Gordon, David Roodnick, Wolfgang Heidrich, and Ramesh Raskar [2011b]. "Refractive shape from light field distortion". In: *ICCV*, pp. 1180–1186.

Wexler, Yonatan, Andrew W. Fitzgibbon, and Andrew Zisserman [2002]. "Image-Based Environment Matting". In: *EGSR*, pp. 279–290.

White, Tom [2016]. "Sampling Generative Networks". In: *Arxiv 1609.04468*.

Wilburn, Bennett, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy [2005]. "High performance imaging using large camera arrays". In: *ACM SIGGRAPH*, pp. 765–76.

Wildeboer, Meindert Onno, Norishige Fukushima, Tomohiro Yendo, Mehrdad Panahpour Tehrani, Toshiaki Fujii, and Masayuki Tanimoto [2011]. "A Semi-Automatic Depth Estimation Method for FTV". In: *Information and Media Technologies* 6.2, pp. 501–507.

Williams, Lance [1978]. "Casting Curved Shadows on Curved Surfaces". In: *SIGGRAPH Comput. Graph.* 12.3, pp. 270–4.

Wizadwongsa, Suttisak, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn [2021]. "NeX: Real-time View Synthesis with Neural Basis Expansion". In: *CVPR*, pp. 8534–8543.

Wolski, Krzysztof, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radoslaw Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K. Mantiuk [2018a]. "Dataset and Metrics for Predicting Local Visible Differences". In: *ACM Trans. Graph.*

Wolski, Krzysztof, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K Mantiuk [2018b]. "Dataset and metrics for predicting local visible differences". In: *ACM Transactions on Graphics (TOG)* 37.5, p. 172.

Wu, Bojian, Yang Zhou, Yiming Qian, Minglun Cong, and Hui Huang [2018]. "Full 3D reconstruction of transparent objects". In: *ACM Trans. Graph.* 37.4.

Xiao, Lei, Anton Kaplanyan, Alexander Fix, Matt Chapman, and Douglas Lanman [2018]. "Deepfocus: Learned image synthesis for computational display". In: *ACM SIGGRAPH 2018 Talks*, pp. 1–2.

Xie, Enze, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo [2021]. *Segmenting Transparent Object in the Wild with Transformer*. arXiv: `2101.08461 [cs.CV]`.

Xu, Jiamin, Xiuchao Wu, Zihan Zhu, Qixing Huang, Yin Yang, Hujun Bao, and Weiwei Xu [2021]. "Scalable Image-Based Indoor Scene Rendering with Reflections". In: *ACM Trans. Graph.* 40.4.

Xu, Jiamin, Zihan Zhu, Hujun Bao, and Wewei Xu [2022]. "A Hybrid Mesh-neural Representation for 3D Transparent Object Reconstruction". In: *arXiv preprint arXiv:2203.12613*.

Xu, Zexiang, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi [2019]. "Deep View Synthesis from Sparse Photometric Images". In: *ACM Transactions on Graphics (TOG)* 38.4, p. 76.

Xu, Zexiang, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi [2018]. "Deep Image-based Relighting from Optimal Sparse Samples". In: *ACM Transactions on Graphics (TOG)* 37.4, p. 126.

Yang, Sidi, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang [2022]. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200.

Yeh, Yu-Ying, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang [2022]. "Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation". In: *ACM Transactions on Graphics (TOG)*.

Yeom, Han-Ju, Hee-Jae Kim, Seong-Bok Kim, HuiJun Zhang, BoNi Li, Yeong-Min Ji, Sang-Hoo Kim, and Jae-Hyeung Park [2015]. "3D holographic head mounted display using holographic optical elements with astigmatism aberration compensation". In: *Optics Express* 23.25, p. 32025. ISSN: 1094-4087.

Yeung, Sai-Kit, Chi-Keung Tang, Michael S. Brown, and Sing Bing Kang [2011]. "Matting and Compositing of Transparent and Refractive Objects". In: *ACM Trans. Graph.* 30.1.

Ying, Zhenqi ang, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik [2019]. "From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality". In: *arXiv preprint arXiv:1912.10088*.

You, Junyong and Jari Korhonen [2021]. "Transformer for image quality assessment". In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 1389–1393.

Yu, Alex, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa [2021a]. *Plenoxels: Radiance Fields without Neural Networks*. arXiv: `2112.05131 [cs.CV]`.

Yu, Alex, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa [2021b]. "PlenOctrees for Real-time Rendering of Neural Radiance Fields". In: *ICCV*, pp. 5752–5761.

Zamir, Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang [2022]. "Restormer: Efficient transformer for high-resolution image restoration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739.

Zannoli, Marina, Gordon D Love, Rahul Narain, and Martin S Banks [2016]. "Blur and the perception of depth at occlusions". In: *Journal of Vision* 16.6, p. 17. ISSN: 1534-7362.

Zhang, Kai, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely [2022]. "Arf: Artistic radiance fields". In: *European Conference on Computer Vision*. Springer, pp. 717–733.

Zhang, Kai, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely [2021a]. "PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting". In: *CVPR*, pp. 5453–5462.

Zhang, L., L. Zhang, X. Mou, and D. Zhang [2011]. "FSIM: A Feature Similarity Index for Image Quality Assessment". In: *IEEE TIP* 20.8, pp. 2378–2386.

Zhang, Lin, Ying Shen, and Hongyu Li [2014]. "VSI: A visual saliency-induced index for perceptual image quality assessment". In: *IEEE Transactions on Image processing* 23.10, pp. 4270–4281.

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang [2018]. "The unreasonable effectiveness of deep features as a perceptual metric". In: *CVPR*, pp. 586–95.

Zhang, X. and B. A. Wandell [1997]. "A spatial extension of CIELAB for digital color-image reproduction". In: *J ISD* 5.1, p. 61. ISSN: 10710922.

Zhang, Xiuming, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron [2021b]. "NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination". In: *ACM Trans. Graph.* 40.6.

Zhang, Zhoutong, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel [2021c]. "Consistent depth of moving objects in video". In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–12.

Zhao, Yuheng, Jinjing Jiang, Yi Chen, Richen Liu, Yalong Yang, Xiangyang Xue, and Siming Chen [2022]. "Metaverse: Perspectives from graphics, interactions and visualization". In: *Visual Informatics* 6.1, pp. 56–67. ISSN: 2468-502X.

Zhou, Tinghui, Matthew Brown, Noah Snavely, and David G. Lowe [2017]. "Unsupervised Learning of Depth and Ego-Motion from Video". In: *CVPR*.

Zhou, Tinghui, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely [2018]. "Stereo Magnification: Learning View Synthesis Using Multiplane Images". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 37.4.

Zhou, Tinghui, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros [2016]. "View Synthesis by Appearance Flow". In: *ECCV*.

Zitnick, C Lawrence, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski [2004]. "High-quality Video View Interpolation Using a Layered Representation". In: *ACM Trans. Graph.* Vol. 23. 3.

Zongker, Douglas E., Dawn M. Werner, Brian Curless, and David H. Salesin [1999]. "Environment Matting and Compositing". In: *SIGGRAPH*, pp. 205–214.

Zou, Yuliang, Zelun Luo, and Jia-Bin Huang [2018]. "Df-net: Unsupervised Joint Learning of Depth and Flow Using Cross-task Consistency". In: *ECCV*, pp. 36–53.