UNIVERSITY OF
CAMBRIDGE

Path from Photorealism to Perceptual Realism

Fangcheng Zhong

Wolfson College

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Fangcheng Zhong

Abstract

Path from Photorealism to Perceptual Realism

Fangcheng Zhong

Photorealism in computer graphics — rendering images that appear as realistic as photographs — has matured to the point that it is now widely used in industry. With emerging 3D display technologies, the next big challenge in graphics is to achieve *Perceptual Realism* — producing virtual imagery that is perceptually indistinguishable from real-world 3D scenes. Such a significant upgrade in the level of realism offers highly immersive and engaging experiences that have the potential to revolutionise numerous aspects of life and society, including entertainment, social connections, education, business, scientific research, engineering, and design.

While perceptual realism puts strict requirements on the quality of reproduction, the virtual scene does not have to be identical in light distributions to its physical counterpart to be perceptually realistic, providing that it is visually indistinguishable to human eyes. Due to the limitations of human vision, a significant improvement in perceptual realism can, in principle, be achieved by fulfilling the essential visual requirements with sufficient qualities and without having to reconstruct the physically accurate distribution of light. In this dissertation, we start by discussing the capabilities and limits of the human visual system, which serves as a basis for the analysis of the essential visual requirements for perceptual realism. Next, we introduce a *Perceptually Realistic Graphics (PRG) pipeline* consisting of the acquisition, representation, and reproduction of the plenoptic function of a 3D scene. Finally, we demonstrate that taking advantage of the limits and mechanisms of the human visual system can significantly improve this pipeline.

Specifically, we present three approaches to push the quality of virtual imagery towards perceptual realism. First, we introduce *DiCE*, a real-time rendering algorithm that exploits the binocular fusion mechanism of the human visual system to boost the perceived local contrast of stereoscopic displays. The method was inspired by an established model of binocular contrast fusion. To optimise the experience of binocular fusion, we proposed and empirically validated a rivalry-prediction model that better controls rivalry. Next, we introduce *Dark Stereo*, another real-time rendering algorithm that facilitates depth

perception from binocular depth cues for stereoscopic displays, especially those under low luminance. The algorithm was designed based on a proposed model of stereo constancy that predicts the precision of binocular depth cues for a given contrast and luminance. Both DiCE and Dark Stereo have been experimentally demonstrated to be effective in improving realism. Their real-time performance also makes them readily integrable into any existing VR rendering pipeline. Nonetheless, only improving rendering is not sufficient to meet all the visual requirements for perceptual realism. The overall fidelity of a typical stereoscopic VR display is still confined by its limited dynamic range, low spatial resolution, optical aberrations, and vergence-accommodation conflicts. To push the limits of the overall fidelity, we present a *High-Dynamic-Range Multi-Focal Stereo display (HDR-MF-S display)* with an end-to-end imaging and rendering system. The system can visually reproduce real-world 3D objects with high resolution, accurate colour, a wide dynamic range and contrast, and most depth cues, including binocular disparity and focal depth cues, and permits a direct comparison between real and virtual scenes. It is the first work that achieves a close perceptual match between a physical 3D object and its virtual counterpart. The fidelity of reproduction has been confirmed by a *Visual Turing Test (VTT)* where naive participants failed to discern any difference between the real and virtual objects in more than half of the trials. The test provides insights to better understand the conditions necessary to achieve perceptual realism. In the long term, we foresee this system as a crucial step in the development of perceptually realistic graphics, for not only a quality unprecedentedly achieved but also a fundamental approach that can effectively identify bottlenecks and direct future studies for perceptually realistic graphics.

Acknowledgements

As I started my PhD, I never truly expected that we would actually pass a visual Turing test with an unprecedented quality in less than three years, as presented in this dissertation. I feel extremely grateful and fortunate that hard work pays off, which oftentimes does not in research. A big thanks to my PhD advisor, Prof. Rafał K. Mantiuk, for the help and support throughout this journey, the meticulous attention to detail, and the firm belief in the prospect of perceptually realistic graphics with me. This dissertation would not have been possible without the valuable input and support from my colleagues and collaborators, both within and outside of Cambridge. They are[1] [2]: Prof. Martin S. Banks, Prof. George Drettakis, Param Hanji, Akshay Jindal, Prof. George Alex Koulieris, Prof. Karol Myszkowski, Prof. Simon J. Watt, Krzysztof Wolski, and Dr Özgür Yöntem.

[1] names in alphabetical order and titles at the time of the completion of this dissertation.
[2] co-authors of the main publications during the course of this dissertation.

Contents

Glossary

3IFC Three-interval forced choice.

ADC Analogue-to-digital converter.

BRDF Bidirectional reflectance distribution function.

CFA Colour filter array.

CFF Critical flicker frequency.

CGH Computer-generated holography.

CIE International Commission on Illumination.

cpd Cycles per degree.

CSF Contrast sensitivity function.

DiCE Dichoptic contrast enhancement.

DoF Depth of field.

DR Differentiable rendering.

DSLR camera Digital single-lens reflex camera.

FoV Field of view.

HDR High dynamic range.

HDR-MF-S display High-dynamic-range multi-focal stereo display.

HMD Head-mounted display.

HVS Human visual system.

IPD Inter-pupillary distance.

MAP Maximum a posteriori.

OLED Organic light-emitting diode.

ppd Pixels per degree.

ppi Pixels per inch.

PRG Perceptually realistic graphics.

SLM Spatial light modulator.

SPD Spectral power distribution.

ToF camera Time-of-flight camera.

VA conflict Vergence-accommodation conflict.

VR Virtual reality.

VTT Visual Turing test.

Chapter 1

Introduction

1.1 Overview

Realism is an everlasting and primary pursuit in the field of computer graphics. Well-established physically-based rendering techniques can generate images that are as realistic as photographs. Nowadays, photorealistic rendering has matured to the point that it is widely applied in the industry. From natural substances of multitudinous forms to intricate human facial expressions, artists and engineers can synthesise images of virtual scenes with complex geometry, materials, and illumination, especially in cinematography where most cutting-edge graphics algorithms in nearly all sub-fields are practised. Yet, photorealistic rendering places an upper limit on the realism achieved by a photograph. Emerging display technologies can deliver high dynamic range (HDR) and contrast, accurate colour reproduction, and a close approximation to a full set of real-world cues of 3D structure. Together, such displays can potentially exceed the realism of photographs and bring us closer to what we define as *perceptual realism* — displaying virtual scenes that are perceptually indistinguishable from real-world 3D scenes.

The increasing level of realism has the potential to significantly impact numerous aspects of life and society. For instance, in the entertainment sector such as gaming and filmmaking, *perceptually realistic graphics* (PRG) enhances the overall experience by creating a more believable and engaging environment for players and viewers. In live streaming, PRG transcends the experience of traditional media by enabling the audience to freely immerse themselves in every detail of the event. In other domains such as education, business, science, engineering, and design, incorporating PRG and 3D displays offers a valuable tool for complex concepts and ideas to be demonstrated through vivid visual aids and realistic simulation. Furthermore, PRG provides an opportunity for individuals to better connect

with friends, family, and colleagues as if they were physically present. People can also explore places from afar, such as outer space, natural wonders, museums, and historical sites, without the need for long-distance travel.

From the physics perspective, the ultimate objective of perceptually realistic graphics is to reproduce a virtual scene that faithfully approximates the true light field of the real world. Unfortunately, this entails an unreasonable requirement for storage, computing power, and physical control of light, which is currently unrealisable for any display system. However, the capability of the human visual system is limited in perceiving minor inaccuracies in the light field. The virtual reproduction of light does not have to be identical in distribution to its physical counterpart to be perceptually realistic, provided that it is visually indistinguishable to human eyes.

Limits of human vision have been widely exploited in photorealistic graphics such as level of detail, tone mapping, and colour coding. We continue this endeavour. In this dissertation, we investigate the essential visual requirements for perceptual realism and propose practical solutions that exploit the limits and mechanisms of human vision to push the quality of computer-generated 3D imagery towards perceptual realism. Throughout this dissertation, we argue that both the physical and perceptual perspectives are equally paramount in the evaluation and advancement of perceptually realistic graphics.

With this approach, we start this dissertation with a background on the *human visual system* (HVS), providing a theoretical basis for the analysis of essential visual requirements for perceptual realism from the geometric, spectral, and temporal aspects. The most relevant visual requirements that we identify as essential and fundamental for perceptual realism are retinal image, spatial resolution, depth perception, dynamic range, contrast, colour (gamut and accuracy), and temporal resolution. Such requirements provide concrete objectives for the aimed displayed qualities of perceptually realistic graphics. Next, we introduce a *perceptually realistic graphics* (PRG) *pipeline* consisting of the acquisition, representation, and reproduction of the plenoptic function of a 3D scene. We examine both the physical and perceptual perspectives in the evaluation and advancement of this pipeline. As many integral parts across the pipeline share the same techniques with photorealistic graphics, we focus on aspects that are unique or substantial to perceptual realism, such as computational 3D displays and depth reproduction, high-dynamic-range imaging, and scene representations for view synthesis with megapixel images. Finally, we present three approaches to push forward the quality of perceptually realistic graphics by exploiting the limits and mechanisms of human vision. First, we introduce *DiCE*, a dichoptic contrast enhancement method that exploits the binocular fusion mechanism of the human visual system to boost the perceived local contrast for stereoscopic displays. Next, we

introduce *Dark Stereo*, an algorithm manipulating contrast to facilitate depth perception for stereoscopic displays under low luminance. Finally, we introduce a *High-Dynamic-Range Multi-Focal Stereo display* (HDR-MF-S display) with an end-to-end imaging and rendering system that can reproduce virtual 3D objects with high fidelity to the point that they can be confused with physical ones. Overall, we position our work throughout this dissertation within a general framework such that each sub-work is dedicated to advancing certain aspects of the PRG pipeline to improve the quality for certain visual requirements for perceptual realism, as shown in Figure 1.1.

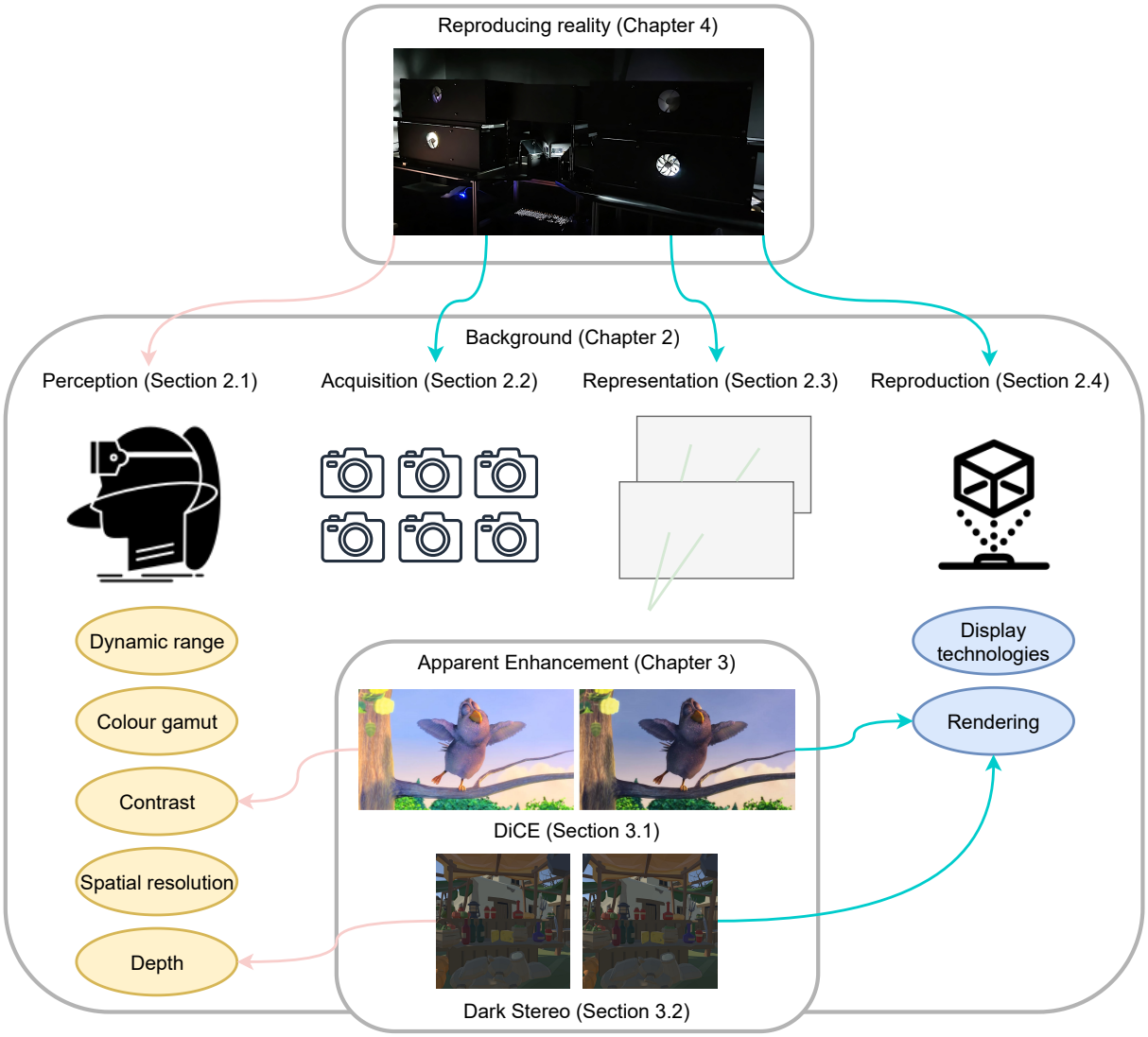


Figure 1.1: Illustration of the dissertation structure. We start with a background (Chapter 2) discussing the perception (Section 2.1), acquisition (Section 2.2), representation (Section 2.3), and reproduction (Section 2.4) of the plenoptic function. In particular, Section 2.1 identifies the essential visual requirements for perceptual realism. Sections 2.2 - 2.4 constitute a *perceptually realistic graphics* (PRG) *pipeline*. Our main work (Chapters 3 and 4) is positioned such that each sub-work focuses on improving specific aspects of the PRG pipeline to meet specific visual requirements for achieving perceptual realism.

Figure 1.2: Comparison of standard stereo images and the images with enhanced perceived contrast using DiCE [202].

1.2 Apparent enhancement rendering

The quality of visual content depends not only on the physical light distribution of the content, but also on the latent processing of it by the human visual system. As such, we can leverage particular characteristics of the HVS to improve the perceived quality of 3D scenes, transcending the limits of the display device. These approaches are referred to as *apparent enhancement* techniques.

In Chapter 3, we propose two apparent enhancement rendering algorithms designed to boost the perceived quality of contrast and depth for stereoscopic displays. In Section 3.1, we present *DiCE* [202], a dichoptic contrast enhancing method that exploits the HVS binocular fusion mechanisms to boost the perceived local contrast and visual quality of images (Figure 1.2). While this method was inspired by an established model of binocular contrast fusion, we proposed and empirically validated a rivalry-prediction model to better explain the main factors contributing to binocular rivalry when two images of different contrasts are displayed. This way we can effectively control the contrast enhancement while maintaining rivalry at a moderate level. Since the method is based on fixed tone curves, it has a negligible computational cost, and therefore, is well suited for real-time applications such as VR rendering. In Section 3.2, we present *Dark Stereo* [182], a depth-enhancing method that compensates for the deteriorated depth perception from stereo cues under

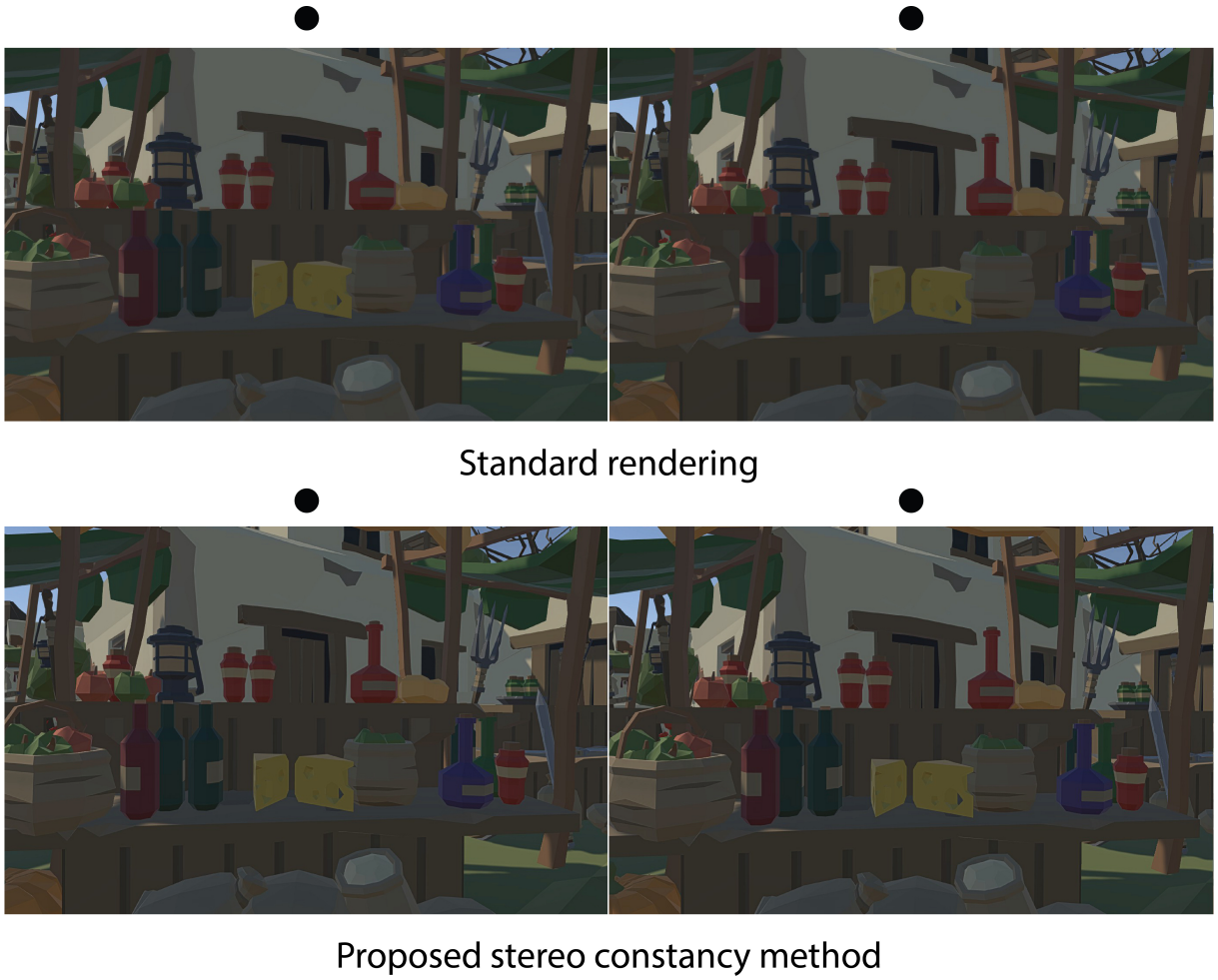


Figure 1.3: Comparison of standard stereo images and the images after stereo constancy processing using Dark Stereo [182].

low luminance (Figure 1.3). The algorithm was designed upon a proposed model of stereo constancy that predicts the precision of binocular depth cues for a given contrast and luminance. We applied the model of stereo constancy to develop a multi-scale contrast compensation method to preserve the precision of binocular depth cues at various display luminance levels. The method has been implemented in GPU shaders and thus is also well-suited for real-time applications. Both DiCE and Dark Stereo have been experimentally demonstrated to be effective in improving realism and overall visual quality.

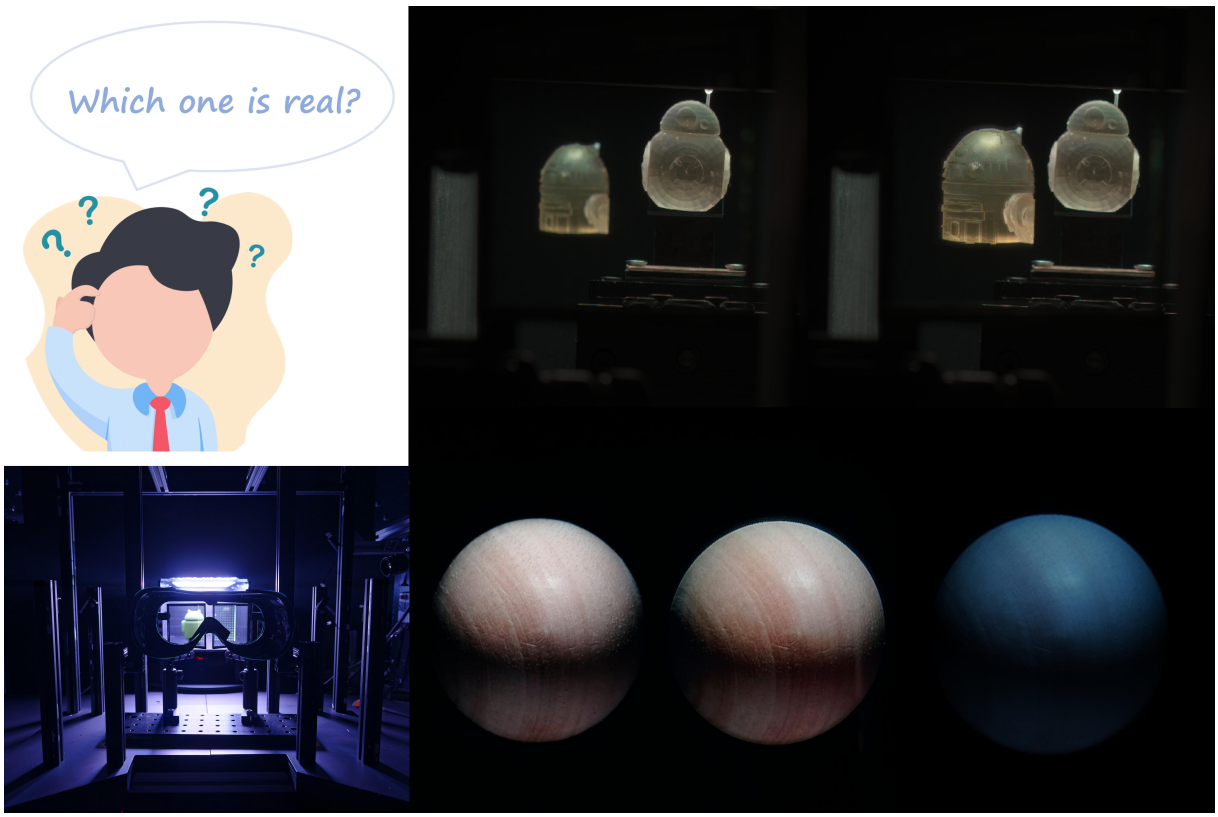


Figure 1.4: Photographs of virtual objects rendered by our High-Dynamic-Range Multi-Focal Stereo (HDR-MF-S) display [203] in comparison with real objects.

1.3 Reproducing reality

Imagine a black box containing either a physical 3D object or one virtually rendered by a 3D display. If an observer, without prior knowledge, is unable to discern the difference between these two scenarios, the display system can be said to have passed a *Visual Turing Test* (VTT) [5]. Passing a visual Turing test for arbitrarily complex scenes is the holy grail of perceptually realistic graphics. Only improving rendering as introduced in Chapter 3 is insufficient to fulfil all the visual requirements to pass a visual Turing test. The overall fidelity of a typical stereoscopic VR display is confined by limited dynamic range, low spatial resolution, lens distortions, and vergence-accommodation conflicts. Volumetric displays such as light-field or holographic displays also cannot achieve the resolution, colour accuracy, gamut, and dynamic range required for perceptual realism.

To push the limits of overall fidelity and maximise the quality of all the essential visual cues for perceptual realism, in Chapter 4, we introduce a *High-Dynamic-Range Multi-Focal Stereo display* (HDR-MF-S display) [203] with an end-to-end imaging and rendering system that can reproduce virtual 3D objects with high fidelity so that they can be confused with physical ones (Figure 1.4). By combining four custom-built HDR displays into a single-

viewer two-focal plane stereoscopic display and integrating differentiable rendering with lumigraph view synthesis and linear depth filtering, the system can acquire a real-world 3D object and reproduce it with high resolution, accurate colour, a wide dynamic range and contrast, and most depth cues, including binocular disparity and focal depth cues. Moreover, the system supports a direct comparison between the real and virtual scenes. This allows us to perform a visual Turing test to evaluate the quality of the display. We propose a strict *three-interval-forced-choice* (3IFC) visual Turing test to ensure that the virtual scene must not be visually different in any respect from the real scene. The results indicate that naive observers can only detect a discrepancy between real and displayed 3D objects with a probability of 0.44. With such a level of realism, our system can function as a testbed to facilitate a variety of studies in perceptually realistic graphics where both faithful reproductions of all visual cues and comparison to reality are paramount.

1.4 Contributions

In this dissertation, we provide a unified background for the study of perceptually realistic graphics from the perspectives of the acquisition, representation, reproduction, and perception of the plenoptic function. Next, we identify the essential visual requirements for perceptual realism and propose several approaches to push the quality of computer-generated imagery towards such requirements. Finally, we demonstrate that taking advantage of the HVS can significantly improve the perceptually realistic graphics pipeline. Specifically, we propose two apparent enhancement rendering algorithms to boost the perceived quality of contrast and depth for stereoscopic displays, without having to expand the display contrast ratio or manipulate disparity. Both algorithms have been experimentally demonstrated to improve realism and can be readily integrated with any existing VR rendering pipeline with their real-time performance. We also introduce an HDR-MF-S display apparatus with an end-to-end imaging and rendering system. The system can achieve a close perceptual match between the real and virtual objects by maximising the quality of essential visual cues, and without having to reconstruct the physically accurate light fields. This is the first work that passed a visual Turing test with a strict 3IFC criterion. We believe that this is a significant step in computer graphics that combines all different aspects towards the holy-grail goal of digitising and visually reproducing a physical 3D object. We also believe that our proposed 3IFC visual Turing test on a display apparatus allowing for a direct comparison between real and displayed scenes is a fundamental approach for the future studies of perceptually realistic graphics. Such studies not only provide insights to better understand the conditions necessary to achieve perceptual realism, but also help identify the most salient artefacts and bottlenecks of existing display technologies, which is crucial in directing the future designs of the PRG pipeline and 3D displays towards where the HVS is most sensitive.

In summary, this dissertation made the following contributions:

- Identification of the essential visual requirements for perceptual realism analysing the visual perception of the plenoptic function.
- A unified overview of the perceptually realistic graphics pipeline.
- A real-time dichoptic contrast enhancement method that improves the perceived contrast based on the binocular fusion mechanism of the human visual system, while controlling the level of rivalry based on a proposed model explaining the main factor causing the rivalry.
- A stereo constancy method that improves depth perception on dimmed displays

based on a proposed model of stereoscopic constancy on various luminance and contrast.

- A novel display apparatus with an end-to-end system capable of capturing and reproducing all the essential visual cues for a static scene of moderate size to reach a close perceptual match between the real and virtual scenes.
- A fundamental approach to study the necessary conditions for perceptual realism and evaluate the qualities of 3D displays, including a 3IFC visual Turing test and display architecture that permits a direct comparison between the real and virtual scenes.
- The first work that passed a strict 3IFC visual Turing test with a near-eye and binocular presentation of a 3D object and without any degradation of the real scene.

1.5 Publications

The following works were produced and incorporated into the main chapters during the course of this dissertation:

- Fangcheng Zhong, George Alex Koulieris, George Drettakis, Martin S. Banks, Mathieu Chambe, Frédo Durand, and Rafał K. Mantiuk. Dice: Dichoptic contrast enhancement for vr and stereo displays. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH Asia 2019, Journal Track)*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356552. URL https://doi.org/10.1145/3355089.3356552

- Fangcheng Zhong, Akshay Jindal, Ali Özgür Yöntem, Param Hanji, Simon J. Watt, and Rafał K. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH Asia 2021, Journal Track)*, 40(6), dec 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480513. URL https://doi.org/10.1145/3478513.3480513

- Krzysztof Wolski, Fangcheng Zhong, Karol Myszkowski, and Rafał K. Mantiuk. Dark stereo: Improving depth perception under low luminance. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH 2022, Journal Track)*, 41(4), jul 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530136. URL https://doi.org/10.1145/3528223.3530136

The following works were produced and partially incorporated into the background chapter during the course of this dissertation:

- Param Hanji, Fangcheng Zhong, and Rafał K. Mantiuk. Noise-aware merging of high dynamic range image stacks without camera calibration. In *Advances in Image Manipulation (ECCV workshop)*, pages 376–391. Springer, 2020. URL http://www.cl.cam.ac.uk/research/rainbow/projects/noise-aware-merging/

- Jingyu Liu*, Fangcheng Zhong*, Claire Mantel, Søren Forchhammer, and Rafał K. Mantiuk. Chapter 17 - computational 3d displays. In Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar, editors, *Immersive Video Technologies*, pages 469–500. Academic Press, 2023. ISBN 978-0-323-91755-1. doi: https://doi.org/10.1016/B978-0-32-391755-1.00023-7. URL https://www.sciencedirect.com/science/article/pii/B9780323917551000237

Chapter 2

Background

Restricting our considerations to geometric optics[1], the light distribution of a 3D scene can be fully described by a *light field*, expressed as a 7D *plenoptic function*:

$$\Phi = F(x, y, z, \theta, \varphi, \lambda, t), \tag{2.1}$$

which indicates the spectral radiance Φ (W sr^{-1} m^{-3}) in wavelength λ of a ray traversing the spatial coordinates (x, y, z) along the direction (θ, φ) at time t. The plenoptic function plays a significant role in the study of photorealistic graphics, as it can be used to synthesise photorealistic images of a 3D scene at an arbitrary viewing position, orientation, and time. The same criticality of the plenoptic function, if not greater, applies to the study of perceptually realistic graphics, as the objective is to synthesise an entire virtual light field. Therefore, in this chapter, we provide a unified background for the study of perceptually realistic graphics from the perspectives of the acquisition, representation, reproduction, and perception of the plenoptic function.

We start by formulating visual perception as the visual sampling of the plenoptic function (Section 2.1), relating the capabilities of the human vision system (HVS) to the required precision of the virtual light fields. Next, we introduce a *perceptually realistic graphics pipeline* consisting of the acquisition (Section 2.2), representation (Section 2.3), and reproduction (Section 2.4) of the light fields. We discuss the advancement and challenges of this pipeline from both the physical and perceptual perspectives. As this background reviews the entire pipeline, it is impossible to cover all the details. We focus on aspects that are unique or substantial to perceptual realism and refer to other references for further details.

[1] incoherent light and objects larger than the wavelength of light.

2.1 Visual requirements for perceptual realism

Although the plenoptic function (Equation 2.1) is a seven-dimensional continuous function, the *human vision system* (HVS) is limited in perceiving minor inaccuracies in the light fields. The virtual reproduction of light does not have to be identical to its physical counterpart, restoring the redundant information that exceeds the limits of human vision, to be perceptually realistic. For example, we do not directly perceive the spectral radiance of individual rays but the irradiance (projection) of rays coming from all directions on the retina. Spectral irradiance is further integrated over various ranges of wavelengths by the photoreceptors leading to colour vision. The spatial and temporal resolution that the HVS can resolve is also limited. These significantly simplify the visual requirements for perceptual realism. By leveraging the limitations of the HVS, it is possible to reduce the precision of the virtual light fields rendered by a 3D display while maintaining identical visual perceptions.

In this section, we explain the basics of the HVS, identifying the relevant visual requirements for perceptual realism by analysing the capabilities and limitations of the HVS in terms of its perception of the plenoptic function. We discuss such requirements from the geometric, spectral, and temporal aspects, each pertaining to the parameters $(x, y, z, \theta, \varphi)$, (λ), and (t) of the plenoptic function. We argue that, from the geometric aspect, the most relevant *visual cues* for perceptual realism are retinal image, spatial resolution, and depth perception; from the spectral aspect, the most relevant visual cues are dynamic range, contrast, and colour (gamut and accuracy). We do not prioritise considerations on the temporal aspect in this dissertation. Qualitative and quantitative requirements on such visual cues direct the designs of the perceptually realistic graphics pipeline with concrete objectives, optimising the distribution of limited resources in computation, data transmission, and display hardware to where the HVS is most sensitive. In general, it is a great challenge for a display system and its associated imaging and rendering algorithms to reproduce a virtual 3D scene that collectively meets all the visual requirements without artefacts and trade-offs.

2.1.1 Geometric considerations

We first consider the geometric parameters $(x, y, z, \theta, \varphi)$ of the plenoptic function (Equation 2.1), which specifies the origin and direction of rays. The HVS does not directly perceive the radiance of individual rays specified by these geometric parameters but their irradiance (projection) on the retina with a finite resolution. This greatly reduces the

number of rays needed to be controlled in a virtual reproduction. Such reduction is three-folded. First, due to the limited pupil size and field of view (FoV) of human eyes, only a fraction of light from the scene can reach the retina through the pupil, which has a diameter varying from 2 to 4 mm in daylight and 4 to 8 mm in the dark [166]. As shown in Figure 2.1, the maximum FoV for both eyes combined is approximately 100° vertically and 200° horizontally, with a binocular FoV (i.e. overlapped FoV seen by both eyes) of 120° [29]. An individual eye has a horizontal FoV of approximately 160°. The FoV contributes significantly to the sense of immersion but is not an essential visual cue for realism, as reducing the FoV does not necessarily degrade fidelity. Second, the visual system does not directly perceive the radiance of individual rays but a *retinal image*, the irradiance (projection) of rays onto the retina. Therefore, accurate control of individual rays is not a necessary precondition for correct image formation on the retina. Finally, human vision has a limited *acuity*, ability to distinguish small details on the retina. This is mainly determined by the density of photoreceptors on the retina and diffraction and aberration throughout the lens [155], with other factors including luminance, contrast, and colour, as can be explained by a contrast sensitivity function (Section 2.1.2). Visual acuity reaches its peak at the *fovea*, an area on the retina with the highest density of *cone* photoreceptors. It has a resolving power of approximately 120 cycles per degree (cpd) of visual angle [120], corresponding to a spatial resolution of 240 pixels per degree (ppd) for a display. To match this level of acuity, the required spatial resolution R of a display (measured by pixels per unit length (e.g. m, cm, mm)) at viewing distance d (measured by the same unit length) can be calculated as:

$$R = \frac{240}{2d \tan \frac{\pi}{360}} . \tag{2.2}$$

Although the retina only perceives the projected images of incoming rays, the HVS can acquire additional depth information of a 3D scene which is not preserved in a retinal image. *Depth perception* refers to the visual ability to perceive objects in three dimensions and infer their relative or absolute distances. It arises from a variety of *depth cues* that can be classified into *pictorial cues*, where retinal images provide the depth information, and *oculomotor cues*, where depth judgment is based on eye movements. Depth perception can also be categorised as *binocular cues* or *monocular cues*, depending on whether sensory information is observed by both eyes or a single eye.

Conventional 2D displays can provide a variety of depth cues such as shading, relative size, occlusion, and perspectives, but there are other cues unique to 3D displays. For example, *binocular disparity*, or *stereopsis*, is a binocular pictorial cue where two retinal images of the same scene are formed from disparate viewpoints of two eyes. When an object

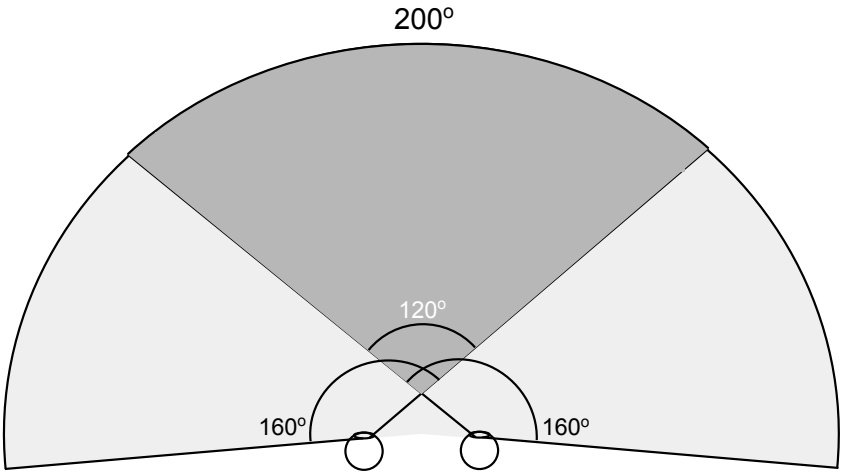


Figure 2.1: Illustration of the bonocular FoV. Each eye sees a FoV of 160°, resulting in a 200° combined horizontal FoV, 120° of which are overlapping.

is closer, the disparity between its left and right retinal images is larger, and vice versa. Inaccurate disparity causes distortions in perceived depth [63, 184]. Binocular disparity is one of the most important depth cues [28] that is commonly employed in VR and cinematographic applications to evoke stereo 3D scene appearance. *Vergence* is a binocular oculomotor cue where the optical axes of the two eyes rotate and converge towards the location of the object in focus. Kinaesthetic sensations from extraocular muscles provide information for depth perception as the depth of an object is inversely related to the angle of vergence [136]. Disparity and vergence together are referred to as *stereo cues*. *Defocus blur* is a monocular pictorial cue where objects outside the depth of field of the eyes appear blurry on the retina. Evidence has shown that focus cues affect both 3D shape perception and the apparent scale of the scene [11, 50, 173]. *Accommodation* is the mechanism that modulates the ciliary muscles to stretch or relax the lens and change the curvature of the cornea to focus on objects close or distant. Such muscle movement provides feedback to the HVS as a monocular oculomotor depth cue. As a cue weaker than defocus blur, accommodation is mainly effective within two metres [42]. Accommodation and defocus blur together are referred to as *focus cues*. A regular stereo display where the disparity is provided by presenting two separate planar images to the left and right eyes does not drive the accommodation to the correct depth. Both eyes accommodate to a fixed but incorrect distance, since all the rays are originated from the screen rather than the actual depth of the virtual object. Such incorrect accommodation cues lead to *vergence-accommodation conflict* (VA conflict) since accommodation and vergence are coupled mechanisms [66]. Their decoupling may cause an unnatural visual experience that results in discomfort [185]. Finally, *motion parallax* is a monocular pictorial cue in which the viewers consider closer objects to be moving faster than further objects.

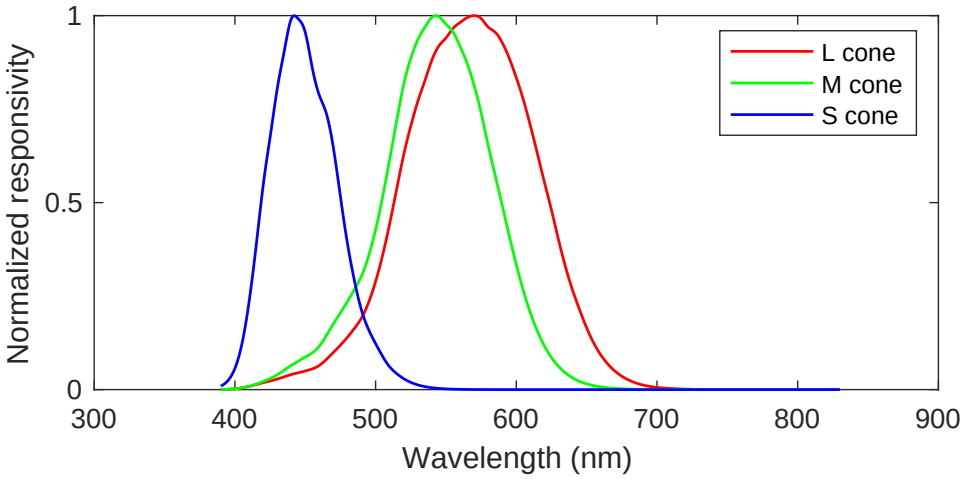


Figure 2.2: The normalised spectral responsivity of each type of cone cells, as reported by the International Commission on Illumination (CIE) in 2006 [25].

While various 3D display architectures have employed distinct approaches to support the aforementioned depth cues, it remains challenging to reproduce all the depth cues correctly and collectively without introducing spatial or temporal artefacts.

2.1.2 Spectral considerations

Now we consider the spectral parameter (λ) of the plenoptic function (Equation 2.1). Although Equation 2.1 is a function of wavelength (λ), the HVS does not perceive the spectral radiance (or irradiance) per wavelength. Instead, it integrates the spectral irradiance over the visible light spectrum, approximately ranging from 380 to 750 nanometres in wavelength, via multiple types of *photoreceptors* on the retina. *Cones* and *rods* are the photoreceptors responsible for the perception of spectral integration of light waves, with cones predominantly sensitive to *photopic* (daylight, typically over $1\,\mathrm{cd/m^2}$) vision and rods to *scotopic* (nocturnal light, typically below $10^{-3}\,\mathrm{cd/m^2}$) vision. There are three types of cones — L, M, and S cones, each peaking at a distinct wavelength in responsivity. A multi-stage neural-circuity process of the signals received by each type of photoreceptor contributes to the perception of colour.

For daylight conditions, *colour* is the result of LMS cone responses. The LMS cones integrate the light spectrum weighted by the responsivity function of the L, M, and S

cones:

$$\begin{aligned} \mathbf{L} &= \int_{\lambda} \Phi(\lambda) L(\lambda)\, d\lambda\,, \\ \mathbf{M} &= \int_{\lambda} \Phi(\lambda) M(\lambda)\, d\lambda\,, \\ \mathbf{S} &= \int_{\lambda} \Phi(\lambda) S(\lambda)\, d\lambda\,, \end{aligned} \tag{2.3}$$

where $\mathbf{L}$, $\mathbf{M}$, and $\mathbf{S}$ are the cone responses of a light spectrum with spectral radiance $\Phi(\lambda)$ per wavelength λ; $L(\lambda)$, $M(\lambda)$, and $S(\lambda)$ are the spectral responsivity of cone cells of long, medium, and short wavelength, as shown in Figure 2.2. If two nonidentical light spectra, Φ and $\overline{\Phi}$, result in the same LMS responses:

$$\begin{aligned} \int_{\lambda} \Phi(\lambda) L(\lambda)\, d\lambda &= \int_{\lambda} \overline{\Phi}\lambda) L(\lambda)\, d\lambda\,, \\ \int_{\lambda} \Phi(\lambda) M(\lambda)\, d\lambda &= \int_{\lambda} \overline{\Phi}(\lambda) M(\lambda)\, d\lambda\,, \\ \int_{\lambda} \Phi(\lambda) S(\lambda)\, d\lambda &= \int_{\lambda} \overline{\Phi}(\lambda) S(\lambda)\, d\lambda\,, \end{aligned} \tag{2.4}$$

the HVS perceive them as equivalent despite non-matching spectral power distributions (SPD). Such colours are referred to as *metamers*. Due to metamerism, displays do not have to generate light waves with physically correct spectra but a metameric match to reproduce a target colour. Therefore, it suffices to simplify Equation 2.5 from a function of wavelengths λ to a function of tristimulus colour channels c (such as LMS) for a perceptual match, reducing Φ from spectral radiance to tristimulus colour values:

$$\Phi = F(x, y, z, \theta, \varphi, c, t). \tag{2.5}$$

LMS can also be transformed into other tristimulus colour spaces such as XYZ and RGB for specific applications. The set of all possible colours up to a metameric match can be represented by a three-dimensional *gamut* of natural colours, such as one shown in Figure 2.3.

The quality of colour can be further depicted by its *chromaticity* — the relative SPD of the light waves regardless of its absolute intensities, and *luminance* — a photometric measure of the intensity. For daylight vision, luminance can be calculated by integrating the light spectrum weighted by a *photopic luminous efficiency function*:

$$Y = 683.002\ lm/W \int_{\lambda} \Phi(\lambda) \bar{y}(\lambda)\, d\lambda\,. \tag{2.6}$$

where Y is the absolute luminance ($\mathrm{cd/m^2}$) of a light spectrum with spectral radiance $\Phi(\lambda)$ per wavelength λ; $\bar{y}(\lambda)$ is the photopic luminous efficiency function, as shown in Figure 2.4,

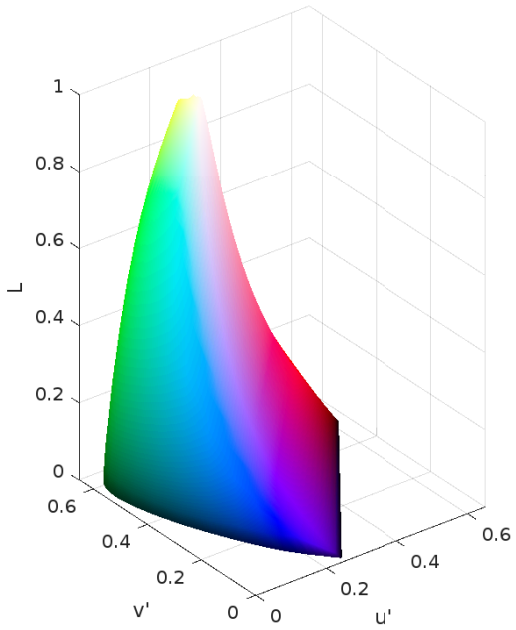


Figure 2.3: The gamut of natural colour in the CIELUV [24] colour space, where (u', v') is the chromaticity coordinate and L is a perceptual measurement of luminance.

corresponding to a weighted sum of the three cone responsivity functions according to their relative population on the retina.

The *dynamic range* of a scene, natural or displayed, refers to the ratio of its largest and smallest luminance value: $Y_{\max}/Y_{\min}$. The largest dynamic range that a display device can reproduce is also known as its *contrast ratio*. A more perceptually uniform measure of dynamic range is given by the difference of log luminance, $\log_{10}(Y_{\max}) - \log_{10}(Y_{\min})$. In natural scenes, the dynamic range spans approximately 12 to 14 orders of magnitude. While human eyes do not perceive such a large dynamic range simultaneously [186], they can adapt dynamically to shift the effective range in response to varying lighting conditions [114].

A closely related quantity to dynamic range and luminance is *luminance contrast*, the local difference in luminance of an object from its surroundings. Contrast can be measured in several ways subject to the spatial configuration of the stimuli. For example, the contrast of a periodic pattern such as sinusoidal gratings can be measured by *Michelson contrast*:

$$C_{\text{Michelson}} = \frac{Y_{\max} - Y_{\min}}{Y_{\max} + Y_{\min}}, \tag{2.7}$$

where $Y_{\max}$ and $Y_{\min}$ are the maximum and minimum luminances in the grating. Alternatively, *Weber contrast* can be applied to measure the contrast of patches with small

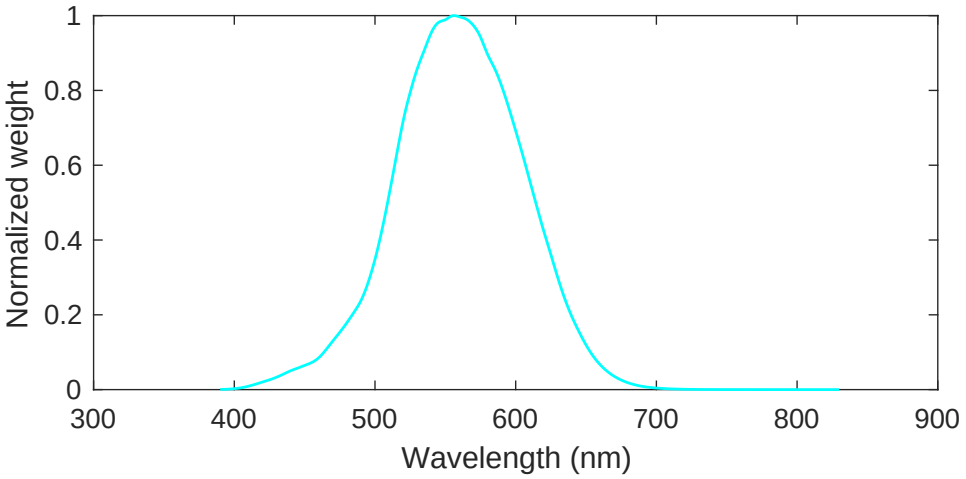


Figure 2.4: The CIE 1931 photopic luminous efficiency function [26].

foreground features superimposed on a large uniform background:

$$C_{\text{Weber}} = \frac{Y_{\text{foreground}} - Y_{\text{background}}}{Y_{\text{background}}}, \tag{2.8}$$

where $Y_{\text{foreground}}$ and $Y_{\text{background}}$ are the luminances of the foreground and background patches.

The HVS has limited sensitivity to contrast — if the contrast of a pattern is below a certain threshold, it is not detectable by human eyes. Due to such limitations, we can safely quantise the output of the plenoptic function to a certain level without resulting in a perceptual difference. *Contrast sensitivity* is defined as the inverse of the threshold detectable contrast. As contrast sensitivity varies with many factors such as the background luminance, frequency, orientation, eccentricity, and size of the stimuli, ample literature has attempted to experimentally establish a *contrast sensitivity function* (CSF) to model such variations [118]. Due to the inherent complexity of the CSF, it is typical to use *Gabor patches* — sinusoidal gratings modulated by a Gaussian envelope — as the stimuli to measure the CSF. For Gabor patches, Michelson contrast and Weber contrast are equivalent. However, although the CSF provides a reasonably accurate measurement of contrast sensitivity for Gabor patches, the detection threshold is in general higher for real images composed of numerous texture patterns that can reduce the visibility of the main feature. This phenomenon is referred to as *contrast masking* [93].

2.1.3 Temporal considerations

We also consider the temporal parameter (t) of the plenoptic function (Equation 2.1). There is a consensus that a high display refresh rate is essential to maintain a high visual quality for higher velocities of motion [87, 113]. Motion artefacts such as judder, ghosting, motion blur, and flicker can all be reduced with a higher refresh rate. However, it is difficult to determine a single threshold refresh rate above which any motion artefacts are not perceivable, as it depends on a multitude of factors such as the persistence and spatial resolution of the display, and the velocity, luminance, and contrast of the stimuli. For example, most AR/VR displays present images with *low persistence*, where an image is displayed at a higher intensity for a fraction of a frame duration and the display remains blank for the rest of the frame. Low persistence significantly reduces the motion blur caused by eye gaze moving over a discretely moving image, which is stationary on the display over the duration of a frame for a fixed refresh rate. However, while low persistence attenuates the required refresh rate to prohibit motion blur, it can introduce visible *flicker* artefacts — the perception of visual fluctuations in intensity and unsteadiness in the presence of a light stimulus — if the refresh rate is under a certain threshold [64]. *Critical flicker frequency* (CFF) measures the frequency at which an intermittent light stimulus appears to be steady without flicker artefacts. Low persistence requires a higher CFF for a steady flicker fusion. Therefore, it is difficult to determine a threshold refresh rate required for perceptually realistic motion quality for an average scenario, although studies showed that the marginal gain with a higher refresh rate significantly drops as the refresh rate rises to 300 frames per second and beyond [113].

In this dissertation, we do not prioritise temporal considerations and focus on improving realism for static scenes.

2.2 High-fidelity 3D scene acquisition

3D scene acquisition is the process of sampling the plenoptic function $\Phi = F(x, y, z, \theta, \varphi, c)$ (Equation 2.5) of a 3D scene. Since perceptual realism requires the highest quality of acquisition, we focus on scene acquisition using *digital single-lens reflex cameras* (DSLR cameras) or *mirror-less cameras* which provide better control and quality compared to other types of capture devices (such as light-field cameras, web cameras, and phone cameras), although some can be jointly employed with a mirror-less or DSLR camera to facilitate the acquisition process. For instance, *time-of-flight cameras* (ToF cameras) [75] can facilitate 3D reconstruction by applying time-of-flight techniques to resolve the distance between the camera and the scene.

Similar to Section 2.1, we consider the geometric (Section 2.2.1) and photometric (Section 2.2.2) aspects of scene acquisition, each pertaining to the parameters $(x, y, z, \theta, \varphi)$ and (c) of the plenoptic function (Equation 2.5) of the reduced form.

2.2.1 Geometric image formation

Geometric image formation establishes the geometric relationship between pixels and their sampled rays in 3D. In this subsection, we provide an overview of the geometric image formation for three camera models: pinhole, thin-lens, and realistic cameras.

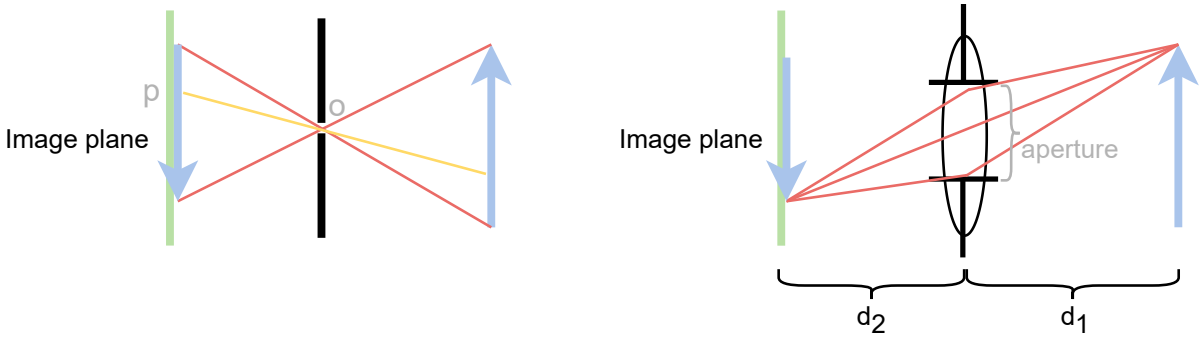


Figure 2.5: Geometric image formation for a pinhole (left) and a thin-lens (right) camera model. For a pinhole model, each infinitesimal point $\mathbf{p}$ on the image plane corresponds to the sampling of a single ray $\overrightarrow{\mathbf{op}}$ traversing the camera origin $\mathbf{o}$.

A *pinhole camera* model assumes an infinitesimal aperture that only allows for rays traversing a single point. As shown in Figure 2.5 (left), each infinitesimal pixel on the image plane corresponds to a single ray traversing the camera origin (aperture). The mapping of such a ray to its projection onto the image plane can be modelled by a *camera matrix* [161]. The pinhole model is the simplest form of geometric image formation for

an ideal situation. In the real world, an infinitesimal aperture is not physically realisable. Real cameras apply lenses to converge rays to form a sharp image with a nonzero aperture size. A *thin-lens camera* is composed of a lens with assumedly negligible thickness, as shown in Figure 2.5 (right). The lens converges rays emitting from a surface point of distance d_1 in front of the lens to a point of distance d_2 behind. According to the lens law, the relationship between d_1 and d_2 can be modelled by a *thin lens equation*:

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f}\,, \tag{2.9}$$

where f is the *focal length* of the lens. For a thin-lens model, an infinitesimal pixel on the image plane no longer corresponds to a single ray but the irradiance of multiple incoming rays from a nonzero solid angle. When the object is in focus, as shown in Figure 2.5 (right), all incoming rays are emitted from a single surface point. If the object is out of focus, an infinitesimal pixel corresponds to the irradiance of rays originating from a patch of the object's surface rather than a point, resulting in *defocus blur*. A pinhole camera can be approximated by a thin-lens camera by reducing the aperture size, which reduces blur and increases the depth of field (DoF). Yet, a small aperture may add noise (due to a deficiency of photons) and diffraction patterns to the image. A pinhole camera can also be approximated by capturing multiple images at various focal depths with thin-lens cameras and merging them to form a sharp image at all depths [88].

Both pinhole and thin-lens models characterise the major principles of geometric image formation of a camera, with assumptions on their optics in ideal cases. However, real cameras are not perfect. For example, real lenses are not infinitely thin and therefore suffer from *geometric aberrations*, including spherical aberration, coma, astigmatism, curvature of field, and distortion (radial and tangential), unless compound elements are used to correct for them. Images taken with wide-angle lenses often require proper modelling of distortion. Chromatic aberrations occur when rays of different wavelengths diverge from their point of intersection with the lens due to different refractive indices of different wavelengths. Another property of real-world cameras is *vignetting*, the tendency of darkening pixel values towards the periphery of the image. Vignetting can be caused due to natural and mechanical reasons. Natural vignetting results from the foreshortening in the object surface, projected pixel, and lens aperture, which is also present with an ideal thin lens. Mechanical vignetting is attributed to the internal blockage of rays by external objects in a lens system such as filters or secondary lenses. Finally, the pixel sensors are not infinitesimal or continuously tiled. The raw pixel value corresponds to the radiant flux received by a non-zero pixel area rather than radiance or irradiance. This requires compensation in photometric calibration, as will be explained in Section 2.2.2. Camera *geometric calibration* is the process of establishing the geometric correspondence between the 3D scene and the

2D image. This involves estimating the camera matrix and distortion, and potentially more parameters [148].

2.2.2 Photometric image formation

Section 2.2.1 explains the principles of geometric image formation which establish the geometric relationship between a pixel and its corresponding sampled rays. However, the recorded pixel value does not directly reflect the true radiance or irradiance of sampled rays. The calculation of true radiance values from raw pixel values requires compensation for exposure, aperture, noise, pixel size, and dynamic range. In this subsection, we provide a background of photometric image formation, establishing the photometric relationship between raw pixel values and radiance.

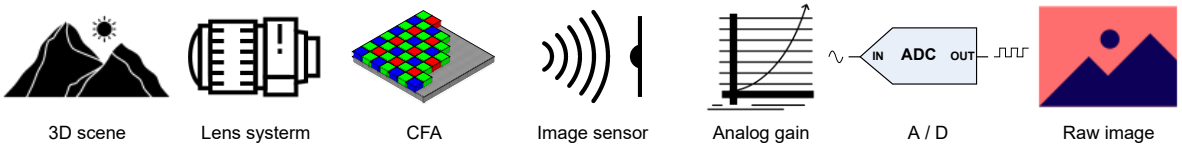


Figure 2.6: Conversion from photons to image raw pixel values in a typical photometric image formation process.

As shown in Figure 2.6, photons of a scene first traverse the camera lens system that controls the exposure time and aperture size to adjust the number of photons passing through. Before being captured by an imaging sensor, photons are filtered by a *colour filter array* (CFA) to acquire the colour information. The imaging sensor converts incident photons to electrons which is proportional to the number of registered photons. The electrons yield a voltage as an analogue signal, which can be amplified based on the settings of the *camera gain*. Finally, an *analogue-to-digital converter* (ADC) digitises the signal into discrete raw pixel intensities. Modern digital cameras provide access to this uncompressed, minimally processed data directly from the electronic imaging sensor in the form of RAW images.

As mentioned in Section 2.2.1 and above, due to a non-perfect camera lens, nonzero pixel size, and various exposure times and gain, a pixel does not directly record the radiance of a single ray but the digital signal that is linearly proportional to the total radiant energy (J) of photons carried by multiple rays received by the sensor. Since radiance measures watt per steradian per square metre (W sr^{-1} m^{-3}), the recorded raw pixel values should be compensated for that. Let $Y_i(p)$ represent the raw pixel value of the i-th colour channel of the camera native space after CFA filtering at the p-th pixel, captured with exposure

time t, gain g, aperture f-number a, and focal length f, assuming no presence of noise or vignetting, the mean radiance of the radiant energy received by the p-th pixel sensor at the i-th colour channel can be calculated as:

$$X_i(p) = k \frac{Y_i(p)}{t\, g\, \pi\, (\frac{f}{2a})^2\, s}\,, \tag{2.10}$$

where s denotes the area of the pixel sensor and k denotes the conversion factor from the digital signal to the physical radiant energy (J/1).

Equation 2.10 did not account for noise and dynamic range. The real recorded raw pixel value $Y_i(p)$ contains multiple sources of noise and thus Equation 2.10 is only an estimation of the true mean radiance. Sources of noise include *photon noise* from the inherent randomness of incoming photons which can be modelled by a Poisson distribution, *readout noise* from the voltage fluctuations while accumulating electrons, and *ADC noise* from the quantisation error in the analogue-to-digital conversion. Meanwhile, real-world scenes often span a wide dynamic range that is not possible to be captured with a single exposure, as the pixel sensor has a limited capacity for registering photons. Large exposure can result in saturated pixels while low exposure increases noise. Therefore, repeated capture is essential not only to reduce noise and the percentage error but also to account for a large dynamic range for radiance estimation. Most proposed radiance estimators from a high-dynamic-range exposure stack require an accurate calibration of noise parameters to minimise the variance of noise [31, 58, 62]. A Poisson-based estimator has been demonstrated to perform with comparable variance without the need for knowledge about sensor-specific noise parameters [61]. Finally, Equation 2.10 only models the mean radiance of registered photons after complex interactions with the lens system. Acquiring the radiance of the raw light rays described by the plenoptic function of the 3D scene requires further compensation for defocus blur, vignetting, and geometric and chromatic aberrations, as discussed in Section 2.2.1. Camera *photometric calibration* involves both radiance estimation and colourimetric calibration. Radiance estimation is usually performed per colour channel of the CFA, which can be further converted into a device-independent tristimulus colour space from the native camera RGB space via a colourimetric calibration [47]. However, camera colourimetric calibration is never 100% reliable, as the camera's RGB spectral sensitivity is different from LMS.

2.3 3D scene representation for perceptually-realistic view synthesis

In Section 2.2, we discussed the sampling of the plenoptic function using DSLR cameras. Unfortunately, the high dimensionality of the plenoptic function and the high precision required to match the limits of human vision make it unrealistic to directly sample the entire plenoptic function to the precision of perceptual realism. Insufficient input views of the captured light fields necessitate interpolation or extrapolation of the plenoptic function to unseen views, which is prone to artefacts.

A *scene representation* is a data structure that encodes the intrinsic geometric and photometric information about a 3D scene. It is a fundamental concept in graphics, upon which many algorithms and downstream applications are developed. For the purpose of this dissertation, we formulate a scene representation as a compact variant of the plenoptic function, exploiting certain known or assumed properties of the scene to reduce its dimensionality (although this is not necessarily the main consideration for the design of a representation). In photorealistic graphics, designing a scene representation to effectively and efficiently synthesise the plenoptic function at an arbitrary viewing position from a sparse sampling of the light fields has been extensively studied in *view synthesis*. Methods for photorealistic view synthesis can be extended to perceptually realistic view synthesis, where the synthesised view is rendered and evaluated on a 3D display rather than a regular 2D screen. As shown in Figure 2.7, the scene representation is an integral component of the PRG pipeline bridging acquisition and reproduction. In the pipeline, parameters of the representation are learned from images captured in acquisition. This process is known as *3D reconstruction*. The reconstructed scene is later rendered on a 3D display to reproduce a virtual light field that perceptually matches the real scene. For reconstruction, the representation is expected to be sufficiently robust to arbitrary scene complexity in geometry, topology, illumination, and materials. It should also be efficacious in retrieving the high-frequency details of the scene geometry and appearance with possibly few captures. With the emerging differentiable graphics (Section 2.3.2), the representation should also provide meaningful gradients directing the optimisation to fast and valid convergence. For reproduction, the representation is desired to be efficient in rendering and integrable with the 3D display architecture of choice.

In this section, we provide an overview of various scene representations employed for photorealistic view synthesis and discuss their extension to perceptually realistic graphics. It should be noted that for graphics in general, view synthesis is not the only application of a scene representation. There is no single representation that excels at all downstream

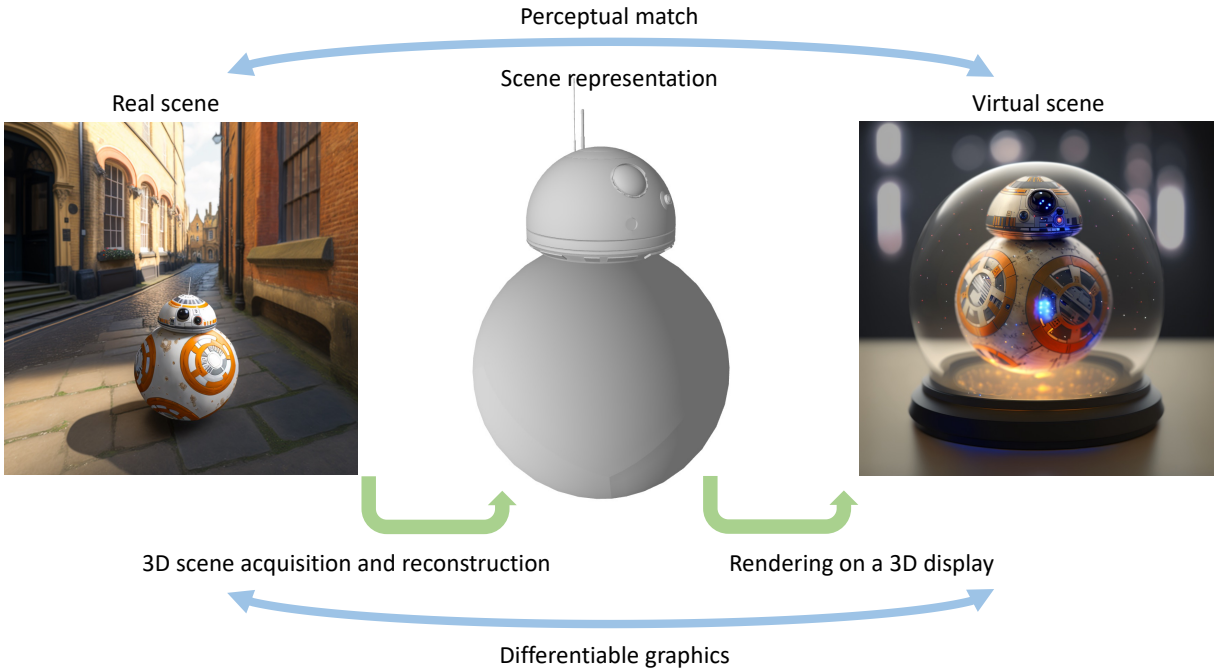


Figure 2.7: Illustration of the perceptually realistic graphics pipeline assuming a static scene. The scene representation is learned from the 3D scene acquisition and reconstruction, which can be rendered on a 3D display to reproduce a virtual light field that perceptually matches the real scene.

tasks. For example, tasks such as digital sculpture, animation, appearance editing, scene composition, and relighting require the expressiveness of a representation in geometry, materials, and lighting to perform such manipulations. However, expressiveness is not an essential ingredient for the purpose of this dissertation where realism is the foremost concern.

We structure this section with a short taxonomy of scene representations (Section 2.3.1) and their reconstruction with differentiable graphics (Section 2.3.2) for view synthesis, followed by a highlighted discussion on aspects crucial to extending photorealistic view synthesis to perceptually realistic graphics.

2.3.1 Taxonomy

A scene representation can be roughly categorised as either describing a volume or a surface, as shown in Figure 2.8.

A *volumetric representation* specifies the radiance information of every spatial location (continuous or discrete) in a 3D volume. While the plenoptic function is a proper volumetric representation, its high dimensionality makes it extremely difficult to directly reconstruct

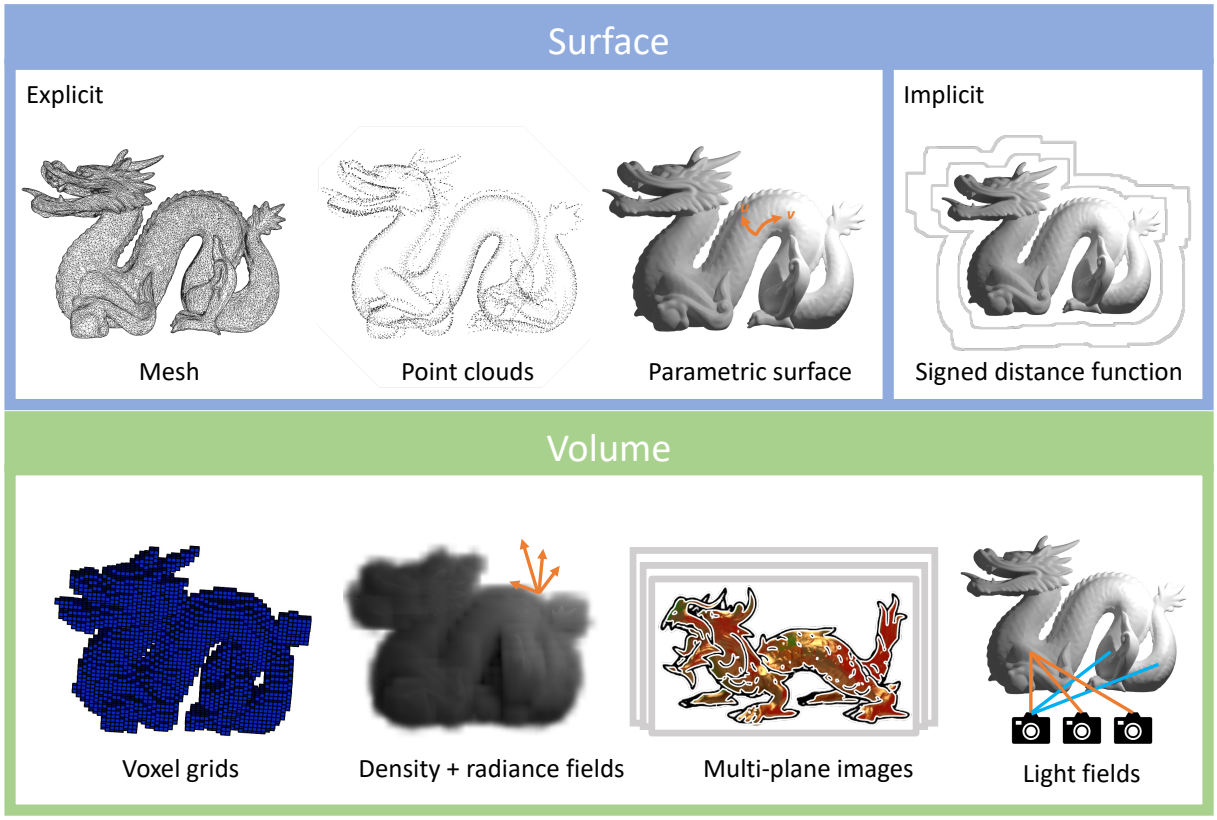


Figure 2.8: Visual illustration of common surface and volumetric representations.

it for view synthesis without dense input, or strong priors and constraints [153, 98, 4, 157]. Along with the high dimensionality is the redundancy where a large space of the scene is not occupied by any physical matter and thus does not emit or reflect light. Such redundancy can be alleviated by specifying an occupancy or density value for every discrete (*voxel grids* [101]) or continuous (*occupancy/density fields* [105, 128]) 3D position. For such representations, only space of nonzero occupancy contributes to the final accumulation of radiance, enforcing a form of multi-view consistency hardly guaranteed by an arbitrarily fitted plenoptic function. Each occupied 3D position can be additionally associated with a *radiance field* [128, 179, 147, 60, 164] to simulate a view-dependent appearance. Volumetric representation is the most general form characterising a 3D scene and is robust in modelling complex scene geometry and topology such as hair, fabric, and smoke. Therefore, it has been most widely adopted for view synthesis, especially with the recent advancement of using *neural fields* [188] to represent spatially-varying occupancy and directionally-varying radiance. Variants of volume representation with special data structures such as *multi-plane images* [204, 48, 127, 179], *octrees* [195], and *sparse voxel grids* [147] have also been proposed to reduce the computational cost of volume rendering.

In contrast to volumetric representations, a *surface representation* explicitly or implicitly specifies a 2-manifold embedded in a 3D volume as a surface. *Explicit* surface repre-

sentations specify a surface by an explicit mapping from a discrete (e.g. *mesh* [132], *point clouds* [193]) or continuous (e.g. *parametric surface* [59]) set of indices to all the surface points, while *implicit* approaches specify a surface by identifying the level set of an *implicit field* (a trivariate function such as *signed distance function* [190, 168, 191]). The implicit field and volume density can be mutually induced [133, 168, 191]. Similar to volumes, a surface point can also be additionally associated with an appearance model (e.g. *bidirectional reflectance distribution function* (BRDF) [190, 199], *lumigraph* [12, 81], *view-dependent texture map* [30], *surface light fields* [183, 20], *radiance fields* [168, 191]) to simulate view-dependent effects. Compared to volumetric representations, surface representation is less commonly adopted for pure view synthesis where surface reconstruction is not an essential intermediate step, especially for scenes containing complex geometry and topology. However, for scenes containing simple or known shapes where accurate surface reconstruction is feasible, a surface representation can potentially achieve a higher rendering speed while maintaining a high synthesis quality.

2.3.2 Differentiable graphics

In the PRG pipeline, as indicated in Figure 2.7, the parameters of a specified scene representation must be optimised to align with the input views before a novel view can be synthesised:

$$\arg\min_{\mathbf{s}} \sum_{i} \left\| \mathcal{R}(\mathbf{s}, \mathbf{c}_i) - \mathbf{I}_i \right\|, \tag{2.11}$$

where $\mathbf{s}$ indicates the unknown parameters of a specified 3D scene representation; $\mathbf{c}_i$ indicates the camera parameters of the i-th input view. Note that $\mathbf{c}_i$ can be estimated via this optimisation process as well if it is unknown. $\mathbf{I}_i$ indicates the input image captured at the i-th view; and $\mathcal{R}$ represents a rendering operator.

Equation 2.11 is an *inverse rendering* problem, i.e. inferring information about the intrinsic properties of the 3D scene from 2D images [95]. Traditionally, solving the inverse problem has been extremely difficult as the simulation of light transport by $\mathcal{R}$ is a complex process often involving non-differentiable steps, leaving efficient gradient-based optimisation methods unemployable. For example, sharp changes in visibility due to object silhouettes, occlusion, and illumination introduce discontinuities that are not differentiable.

Emerging *differentiable rendering* (DR) techniques attempt to derive effective gradients for the traditionally non-differentiable operations in rendering to facilitate solving the inverse problem. Example differentiable surface renderers include *soft rasterisation* [103, 143], *differentiable surface splatting* [193], *differentiable physically-based rendering* [97, 106, 198],

and *differentiable spherical tracing* [102, 77]. In contrast to surfaces, *volumetric ray tracing* which is suitable for rendering volumes such as voxel grids or neural radiance fields [128] is inherently differentiable as visibility is encoded into continuously-varying probabilistic density values in such representations. While this dissertation is primarily focused on applying differentiable rendering to view synthesis, differentiable rendering has profoundly wider applications in physical inference, optimal control, scene understanding, computational design, manufacturing, autonomous vehicles, and robotics.

While view synthesis has been extensively studied in the literature, it has been mostly evaluated for photorealism [46, 162] rather than perceptual realism [121, 17], i.e. evaluation on a 3D display against the view of a real-world scene. As will be discussed in Chapter 4, perceptual realism poses a lower tolerance to artefacts (such as blur, noise, and distortion) and inaccuracies in colour and appearance. Therefore, it requires a higher capacity of a scene representation to converge to i) high-dynamic-range [129] and high-resolution [130] images, and ii) scenes containing complex view-dependent appearances (such as specular reflections) [179, 164, 60]. For differentiable graphics, perceptual realism also requires the scene and rendering parameters to be optimised with respect to the real-world scene and human eyes, rather than merely images. This involves integrating accurate simulation of cameras, displays, and human eyes into the differentiable graphics pipeline [16].

2.4 3D scene reproduction with computational 3D displays

In Sections 2.2 and 2.3, we discussed the sampling and representation of the light fields, which constitute the first two stages of the perceptually realistic graphics pipeline. The last stage of the pipeline is to reproduce a virtual light field on a 3D display, aiming for a perceptual match with the real scene, as shown in Figure 2.7. While 3D scene acquisition and representation inherit techniques from photorealistic graphics, computational 3D displays are a unique constituent part of perceptually realistic graphics.

A distinguishing feature of 3D displays is that they reproduce additional depth cues compared to regular 2D screens. In the past decades, while 3D display technologies have become increasingly accessible — from stereo movies to personal head-mounted VR/AR displays, the quality and experience are still far from perceptual realism. For instance, the wide FoV of a head-mounted display (HMDs) causes an insufficient spatial resolution (measured as ppd - pixels per degree), resulting in pixelation artefacts. Optics such as Fresnel or biconvex lenses introduce noticeable lens distortions on the retinal image as well as degradation in contrast and colour. The conventional design of VR headsets by combining a stereoscope with fixed screen planes causes vergence-accommodation conflicts (VA conflict) that often lead to fatigue and sickness. On the other hand, established 2D display technologies can produce 2D imagery with a quality that fulfils many other visual requirements for perceptual realism in the absence of 3D depth cues. For example, organic light-emitting diode (OLED) displays can deliver high resolution that matches or exceeds the acuity of the human eye [119]. Dual-modulated HDR displays can achieve high dynamic range and contrast [150]. Wide, accurate colour gamut can be achieved by using more saturated primaries, or multiplexing (temporal or spatial) with more than three primaries [78, 70].

In this section, we provide an overview of 3D display systems. We categorise their architectures into *stereoscopic* displays, where the use of special headgear, glasses, or visual separators for eyes is essential to support stereo cues (disparity and vergence), and *volumetric* displays, where 3D images are created in a volume, allowing for stereo cues and true 3D viewing to naked eyes. For each display type, we explain the mechanisms they employ for depth reproduction. We also discuss the associated rendering algorithms and scene representations compatible with the display system. While certain architectures require specific representations and algorithms for rendering, many can extend existing techniques from photorealistic graphics. Finally, we analyse the performance and limitations of these display techniques by evaluating the virtual light field they produce and the visual

requirements they fulfil for perceptual realism. In general, it is a great challenge for a display system to render a virtual 3D scene that collectively meets all the essential visual requirements for perceptual realism without artefacts and trade-offs. We will discuss how various 3D display architectures take advantage of the limits of the HVS in this process. We will also discuss how well-established display technologies for 2D displays can be adopted to address the unnatural experiences created by existing 3D display techniques.

2.4.1 Volumetric displays

Volumetric displays create 3D images by emitting light approximating a true light field. They can reproduce stereo cues without the need for special headgear, glasses, or any type of visual separator for the eyes. In this section, we introduce three main types of volumetric displays, distinguished by the mechanisms they employ to reproduce the light field. Digital *holographic displays* aim to reproduce the entire distribution of light waves, including phase, amplitude, and wavelength, based on the principles of light diffraction and interference. *Light field displays* are a close alternative to holographic displays in terms of recreation of the original light distribution in a 3D volume. The distinction is that light created by a holographic display is formed by phase-conjugated rays from each hologram point, while a light field display controls the directional intensity of beams expanding from a 2D panel (usually composed of pixel cells). *Voxel-based displays* reduce the light field to a 3D volume composed of voxels.

Holographic displays

In principle, all depth cues can be automatically and simultaneously achieved by reproducing the entire light distribution of a 3D scene. This is the ultimate objective of a holographic display. Invented by Dennis Gabor [53], hologram works on the principles of light diffraction and interference. In acquisition, the object beam of a 3D scene interferes coherently with a reference beam, resulting in interference fringes going through a recording medium that records all the characteristics of light (phase, amplitude, and wavelength). In rendering, the object beam is reconstructed in a reverse manner. The reference beam is usually delivered by a single monochromatic laser. Coloured holograms can be generated by rendering three separate holograms of different wavelengths (e.g., red, green and blue) and incoherently superimposing them on one another [8]. Conventional analogue holograms are recorded and reconstructed using non-reconfigurable mediums, such as photographic emulsion. Modern *computer-generated holography* (CGH) generates holographic interference patterns using spatial light modulators (SLMs) and digital technologies [154], which

enables the rendering of holographic videos [7].

Although an ideal holographic display has been regarded as the ultimate form of 3D display, it faces challenges in both hardware and software. For example, SLMs have limited spatial resolution. Based on a 400 nm blue light, a 127 000 ppi SLM with a 200 nm pixel size is required to display a fringe width of half the wavelength of the light [21]. Current commercial SLMs can only reach 7000 ppi [22]. Even if a dense SLM is physically realisable, displaying a static 3D scene as large as a phone screen would require processing billions of pixels, placing a huge burden on computation and data transmission. For dynamic holograms, the amount of data rises to tens or hundreds of billion pixels per second. Moreover, interference of coherent wavefronts results in speckle noise that undermines the image quality in contrast and colour, despite recent progress showing that machine learning techniques can be applied to reduce the artefacts of holographic displays, including speckle noise [140, 15, 23].

Light field displays

In Equation 2.1, we use five spatial parameters to specify the starting position and direction of a ray. If we assume a constant radiance along a ray, we only need to denote where the ray hits the xy ($z = 0$) plane and can remove z from the parameterisation in Equation 2.1, reducing it to:

$$\Phi = F(x, y, \theta, \varphi, \lambda, t). \tag{2.12}$$

A generic light field display generates a four-dimensional (x, y, θ, φ) distribution of light rays from a planar light source and an optical transmission medium. This way, both positional and directional light can be recreated and modulated. The simplest method to modulate directional rays is to redistribute pixels into N horizontal views. This can be achieved by *parallax barrier* [74, 71] — an interlace of transparent and opaque stripes, or *lenticular sheet* [178] — a cylindrical micro-lens array that redirects diffused rays from pixels into specific directions. Of course, both of these approaches prohibit vertical parallax. One advantage of the lenticular sheet over the parallax barrier is that the lenticular sheet does not reduce the display luminance, since the lenticular sheet is comprised of lenses, while the parallax barrier blocks light paths. The concept of the lenticular sheet can be generalised to present both horizontal and vertical perspectives and potentially focus cues by replacing it with a 2D micro-lens array. This is the mechanism behind the *integral imaging* [99], where the light fields of a 3D object are recorded by a 2D micro-lens array and reconstructed reversely when viewed at the same distance of the object through the same lens array. One fundamental issue of these approaches, as with many other light field display architectures, is the inherent trade-off between spatial and angular resolution

(number of distinguishable views). For a display of a finite number of pixels to generate N^2 views, the spatial resolution reduces to $1/N^2$ of its full resolution. However, a too-low number of views can cause discontinuous parallax or break focus cues. Focus cues can only be achieved with a highly dense layout of angular views, as it requires at least two rays to enter the pupil. Another drawback of lenticular lenses is that they only permit a fixed number of viewing zones, limiting the view box (the range of positions from which the display can be viewed). An incorrect viewing position can cause cross-talk and other cues to be wrongly presented. To tackle these issues, such displays can be integrated with a head-tracking system [72], allowing for a dynamic adjustment of the display content to align with the viewing position. This can reduce the essential number of distinguishable views for a fixed head position to maintain a higher spatial resolution and broaden the viewing zone. The drawback is that this is usually limited to a single viewer and requires a highly accurate synchronisation between head tracking and rendering to avoid visual artefacts. Directional light rays can also be controlled by *time multiplexing*, designed to overcome the trade-off between spatial and angular resolution by superimposing each view time-sequentially [82, 91, 169]. The downside of time multiplexing is that it requires an overall refresh rate and a scanning rate of the directional device to be the product of the perceived refresh rate of each view and the number of views [21].

The fundamental issue of a light field display in trading off the spatial and angular resolution is inherently rooted in the large information bandwidth required to express a full light field. Although Equation 2.12 has one less parameter than the full expression, it still carries a redundancy of information, since 1) the change of surface colour with the viewing direction is highly correlated, and is constant for diffuse surfaces; 2) regions of uniform colours or textures exhibit small variance. *Compressive light field displays* were introduced to leverage computational methods and compressive optics to adaptively maximise the quality of the virtual light field for the displayed content [175]. They are referred to as being compressive because the number of emitted light rays can transcend the number of representing pixels, which are computationally optimised to direct the resulting rays to best approximate the target light field and minimise redundancy. *Compressive optics* usually consists of a backlight (uniform or directional) and multiplicative optical layers (e.g. LCDs). Examples of compressive light field displays include tomographic image synthesis [174], polarisation fields [92], and tensor displays [176]. One challenge of these displays is that the multiple-layer architecture introduces scattering and inter-reflections, resulting in approximation error and thus compromising the display contrast and colour. Another challenge is that compressive light field displays require a scene-based optimisation for each frame, causing a high computational cost.

Voxel-based displays

Voxel-based displays produce light originating from voxels in a 3D volume, typically by time multiplexing with image slices emitted or illuminated from switching or mechanically moving surfaces. Examples of voxel-based displays include the use of rotating display screens [45], stacks of switchable diffusers [158], the spin of a cylindrical parallax barrier and LED arrays [192], and sweeping diffusers [165]. It is possible to create strong 3D cues including stereo, parallax, and focus cues with these displays. However, the physical realisation of a voxel-based display makes it difficult to show occlusions as each voxel is semi-transparent. Those displays also cannot reproduce view-dependent surface appearance, such as specular reflections, since light rays are uniformly emitted or reflected from each voxel in all directions. Compared to a holographic or light field display, they have a confined depth range of the scene within the physical display volume, but they permit a much larger viewing zone.

2.4.2 Stereoscopic displays

The quality of a volumetric display is highly confined by its resolution, dynamic range, contrast, colour accuracy, field of view, and computational cost. After all, reproducing a full light field of sufficient size and quality requires control over billions of rays, which is currently infeasible. However, if the number of viewers is limited to a single person and the eye position can be tracked or stabilised, the subspace of a light field required to be reproduced becomes much smaller. This is one of the advantages of stereoscopic displays.

In contrast to volumetric displays, stereoscopic displays stabilise the eye position relative to the display screens[2] with special headgear or glasses, making it possible to render a 3D scene through a significantly smaller number of required rays. Moreover, existing imaging and rendering techniques (or with slight variations) for photorealistic graphics can be seamlessly integrated with stereoscopic displays. Therefore, stereoscopic 3D display products have been commercially available long before volumetric or autostereoscopic ones. The most basic design of a stereo display works by showing a separate planar image to each eye to create a stereo vision. This is for instance the case of many commercial head-mounted displays such as Oculus Quest 2 [125] and HTC Vive Flow [67]. However, such a design lacks proper focus cues. As detailed in Section 2.1.1, rays traversing each eye originate from a single planar screen at which the depth is fixed and may differ from that of the virtual object in focus. This forces the observer's accommodation mechanism

[2]This is with exceptions, such as shutter glasses and polarisation glasses. However, they cannot achieve correct accommodation cues.

to be decoupled from vergence. In this section, we introduce two variants of stereoscopic displays that support proper focus cues (i.e. accommodation and defocus blur).

Vari-focal displays

Vari-focal displays are a variant of standard stereo displays with active adjustment of the focal distance of the image plane seen by each eye via active optics such as liquid lenses [36, 1, 134]. The adjustment of the focal distance is in accordance with the observer's gaze to show a varying depth-of-field (DoF) effect. Albeit in support of focus cues, these types of displays introduce undesirable lens distortions caused by the active optics (e.g. deformable membrane mirror [36]). They also require an accurate synchronisation of the lens optics and depth-of-field rendering with the tracking of the gaze location. Inaccuracy in the optics, rendering, and the gaze of the observers leads to errors on the reproduced focal plane. The mechanisms of a vari-focal display also require its defocus blur to be synthesised in rendering [187] rather than optically reproduced since they only allow for a uniform focal depth throughout the scene for a fixed gaze.

Multi-focal displays

Multi-focal displays can be regarded as a variant of volumetric displays with a fixed viewing position. For each eye, a stack of images is rendered at a fixed number of focal planes at various distances, each plane adding a certain amount of light. Thus, a viewer can accommodate appropriately at the desired depth. These focal image planes can consist of superimposed virtual images on beam-splitters [2] or time-multiplexed image slices that sweep a 3D volume with high-speed switchable lenses [107, 18, 197]. In contrast with vari-focal displays, multifocal displays do not require a strict synchronisation of the optics and rendering with the gaze location, but maintain a high resolution and contrast as they can adopt well-established 2D display techniques [68, 203]. Architectures with fixed focal planes also prohibit optical aberrations. However, the quality of a multifocal display is especially sensitive to the accuracy of the alignment of the focal planes with the eye position, as misalignment immediately breaks sharp edges and realism. Differences in eye positions of individual observers can be compensated for with a homography correction [124] or a physical calibration [203].

Despite the aforementioned improvements, edges near depth occlusions are particularly difficult for a multi-focal display to reproduce. This is due to the additive nature of focal planes, which cannot subtract light transmission to simulate physically-correct occlusion

cues for finite pupil size. They also cannot create physically-correct accommodation cues in between the focal planes. Therefore, compensation in rendering at each focal plane is crucial for multi-focal displays to ensure a smooth perception of depth and texture. The algorithm that drives the rendering for multi-focal displays is referred to as *multi-focal decomposition*. It approximates the true light field of a 3D scene by distributing its content on a discrete number of focal planes. The simplest form of multi-focal decomposition is *nearest neighbour*, assigning the rendered object to its nearest focal plane. However, the nearest neighbour can result in sharp discontinuities for surfaces spanning the depth of multiple focal planes and artefacts at occlusion boundaries. It also drives the eye to accommodate at inaccurate depth. Alternatively, light can be distributed across focal planes via a dioptre-based linear depth filtering [2, 107]. Linear depth filtering can drive accommodation to correct depth with focal plane separations up to one dioptre [111], but may also produce visible artefacts at occlusion boundaries and for non-Lambertian surfaces. Another approach is retinal optimisation [131, 124], which approximates the retinal image of the displayed scene to be close to its real-world counterpart, especially in terms of defocus blur. It performs better at occlusion boundaries at the expense of a higher computational cost and less accurate accommodation cues [124]. Additionally, a perception-driven hybrid decomposition strategy selects the best existing decomposition method contingent on the scene content [196]. They show that in regions without occlusion boundaries, linear depth filtering typically achieves the best result among all the multi-focal decomposition algorithms.

Chapter 3

Apparent Enhancement Rendering

In Chapter 2, we established the visual cues that are most relevant for perceptual realism, including retinal image, spatial resolution, depth perception, dynamic range, contrast, colour, and temporal resolution. Although these visual cues can be physically measured, the perceived quality of such cues may vary under various viewing conditions while the physical measurement of such cues remains unchanged. This is because the perception of visual cues is a combined result of the physical light distribution of the scene and the latent processing of it by the human visual system. Therefore, we may exploit particular characteristics of the HVS to enhance the salience of visual cues that transcend the limits of display devices. Such approaches are referred to as *apparent enhancement techniques*. For example, apparent super-resolution that exceeds the display resolution can be achieved by rapid temporal pixel variations [32]. Apparent image contrast can be altered with Cornsweet illusion [142].

In this chapter, we present two rendering algorithms that can be applied to binocular stereoscopic displays to boost the perceived quality of contrast and depth. Specifically, we propose *DiCE*, a dichoptic contrast enhancing algorithm that exploits the binocular fusion mechanism of the HVS to improve the perceived contrast without having to expand the display contrast ratio (Section 3.1). We also introduce *Dark Stereo*, a depth-enhancing algorithm employing a proposed model of stereo constancy to improve the precision of depth perception from stereo cues under low luminance without having to manipulate depth or disparity (Section 3.2). Both algorithms have been experimentally demonstrated to be effective in improving realism and overall visual quality, and can be readily integrated with any existing VR rendering pipeline with their real-time performance.

3.1 Improving perceived contrast with binocular fusion

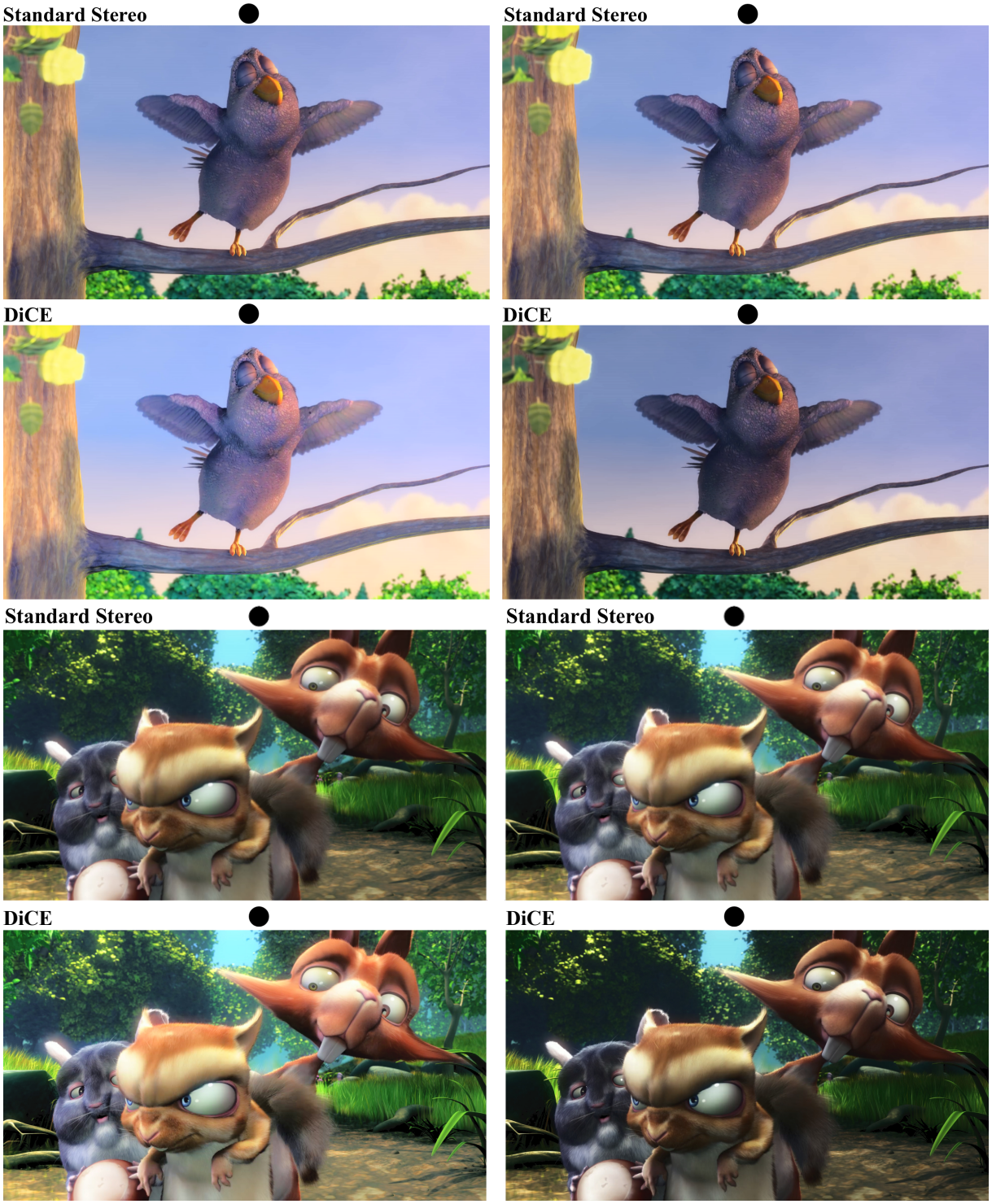


Figure 3.1: Comparison of standard stereo images and the images with enhanced perceived contrast using DiCE. The images can be cross-fused with the assistance of the dots above the images. Notice the enhanced contrast in the shadows and highlights of the scene. The stereo images are from *Big Buck Bunny* by Blender Foundation.

As discussed in Chapter 2, contrast is a crucial factor that influences realism. Images with higher contrast have been demonstrated to be perceived as more realistic and three-dimensional [163]. Bright, high-dynamic-range displays can achieve high contrast, but may cause flicker in low-persistence VR/AR systems, as discussed in Section 2.1.3, and consume more power. Local tone-mapping operators can be effective at enhancing local contrast, but may lead to unnatural-looking images and artefacts in videos such as temporal incoherence [40]. They are also computationally expensive, making their use prohibitive, especially in time-critical VR/AR applications, in which every GPU cycle matters and dropping frames is not an option. As opposed to these approaches, we capitalise on the human visual system's binocular fusion mechanisms to enhance contrast and improve realism.

We exploit an inherent property of the binocular fusion mechanism to enhance perceived contrast. We introduce *DiCE*, a dichoptic contrast enhancement method that selectively applies lower or higher tone curve slopes to improve image contrast. However, a naive implementation of this approach may cause *binocular rivalry*: an unstable percept that switches between the image of one or the other eye. We empirically established the main factors causing rivalry and tune the parameters of our method accordingly. We found that the ratio of contrasts presented to both eyes is the main factor that can explain and quantify rivalry. This allows us to tune our method to maximise contrast enhancement while maintaining low rivalry. Since the dominant cause of rivalry is mostly independent of image content, our method can be implemented as a fixed set of tone curves, which have negligible computational cost and can be directly used in real-time VR rendering for any stereo displays. We evaluated our method by comparing it with previous work, showing that our solution is more successful at enhancing contrast and at the same time much more efficient. We also evaluated our method in a VR setup where users indicated that our approach improves contrast and depth compared to the baseline. Our methodology and results suggest that rendering for the binocular domain is both a computationally cheap and effective means to increase contrast in binocular displays.

We start this section with a review of preliminary concepts and related work in contrast-enhancing tone mapping (Section 3.1.1) and binocular vision (Section 3.1.2). Next, we explain our proposed binocular contrast enhancement method (Section 3.1.3) and experimentally establish the main factor that causes rivalry in enhanced images (Section 3.1.4). This lets us find the best parameters for our tone curve generation method (Section 3.1.6). Finally, we demonstrate the strengths and shortcomings of our method compared with existing dichoptic presentation techniques (Section 3.1.7).

The work presented in Section 3.1 produced the following publication:

- Fangcheng Zhong, George Alex Koulieris, George Drettakis, Martin S. Banks, Mathieu Chambe, Frédo Durand, and Rafał K. Mantiuk. Dice: Dichoptic contrast enhancement for vr and stereo displays. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH Asia 2019, Journal Track)*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356552. URL `https://doi.org/10.1145/3355089.3356552`

3.1.1 Tone mapping and contrast enhancement

Tone mapping is an image-processing operation performed to convert an image from a scene-referred colour space into a display-referred colour space. Tone mapping spans a range of techniques that can vary in their goals. Some techniques simulate specific phenomena of the visual system (glare, night vision). Others attempt to achieve the best subjective quality (colour grading, enhancement) or possibly faithful reproduction of image appearance [39]. Because scene-referred colours often exceed the dynamic range of the target display, a common goal of all tone-mapping methods is the reduction of dynamic range.

One of the most common techniques used in tone mapping is a global tone curve: a monotonic function that maps input colour/luminance values to the displayed colour/luminance values. Such a curve can be fixed and, for example, can mimic the response of a photographic film [144], or can adapt to image content [171] and a display [115]. A tone curve is typically designed to enhance contrast in visually relevant parts of the scene and compress or clip contrast in less relevant parts, which are dark or noisy [41], or contain bright highlights or light sources that cannot be easily reproduced on a display.

To revert the loss of small contrast details caused by compressive tone curves, many tone-mapping techniques involve local contrast enhancement. Such enhancement could be achieved by unsharp masking combined with edge-stopping filters [38, 41], which can avoid ringing or halo artefacts. Stronger enhancement could be achieved by operating in the gradient domain [44]. This, however, requires computationally expensive optimization. Contrast at multiple scales can be more efficiently edited using local Laplacian pyramids [137]. The main drawback of all these enhancement techniques is that they introduce a substantial computational overhead, which is unacceptable in real-time applications. Our technique replaces computationally expensive local contrast enhancement with fixed tone curves, which have negligible computational cost.

3.1.2 Binocular fusion

3.1.2.1 Tone mapping exploiting the binocular domain

Binocular fusion was exploited before in a number of tone-mapping methods for binocular displays [189, 200, 201]. We will refer to these methods as *binocular tone mapping operators* (BTMO). The goal of these techniques is to produce two tone-mapped images that are maximally different, yet comfortable to fuse. This is achieved by adjusting the parameters

of an existing [189] or newly proposed tone-mapping operator [200] in an optimization loop. The loss function is designed to maximise the difference between left- and right-eye images, leading to "richer" fused images. To ensure acceptable levels of rivalry, a *binocular viewing comfort predictor* is used to reject image pairs that are deemed too rivalrous. Neural networks can also be leveraged to generate tone-mapped images without assumptions about monocular tone operators. In concurrent work, *deep binocular tone mapping* [201] employs CNNs to generate an end-to-end binocular tone mapping operator that outputs the desired LDR pair from an HDR image. Similar to previous BTMO techniques, the loss function is designed to optimise the visual content distribution to maximise the perception of local detail and global contrast, while maintaining visual comfort. Real-time computation can be achieved with a GPU acceleration.

In contrast to BTMO techniques, our method explicitly enhances contrast based on psychophysical models and findings, rather than making images different. In Section 3.1.7, we demonstrate that this leads to much more consistent and predictable enhancement. Instead of a complex viewing comfort predictor, which combines multiple heuristics, we find a simple yet effective rivalry indicator based on new experimental findings. Our technique does not restrict the choice of tone-mapping operator and can be used with stereoscopic content. Most importantly, our technique has negligible computational cost compared to the BTMO methods, and thus can process an image pair in milliseconds rather than seconds without relying on GPUs.

3.1.2.2 Perception in dichoptic presentation

In a binocular display, *dichoptic* presentation is the presentation of different images to the two eyes and *dioptic* is the presentation of identical images to the two eyes. If the dichoptically viewed images are synthesised or photographed from two offset viewpoints at a distance approximately equal to the human interpupillary distance, they contain image disparities that elicit the illusion of depth by exploiting binocular vision. This is a *stereoscopic* image pair and always requires dichoptic presentation. Dioptic presentation cannot elicit the illusion of depth from disparities — as images for left and right eyes are identical — and thus can only show *monoscopic* images. To avoid confusion, we will refer to images without dichoptic enhancement as *standard*, regardless of whether these are monoscopic or stereoscopic images.

When the dichoptic stimuli are too dissimilar to be fused into one stable percept, the viewer experiences *binocular rivalry*. Binocular rivalry refers to a state of competition between the eyes, with one eye inhibiting the perception of the image in the other eye

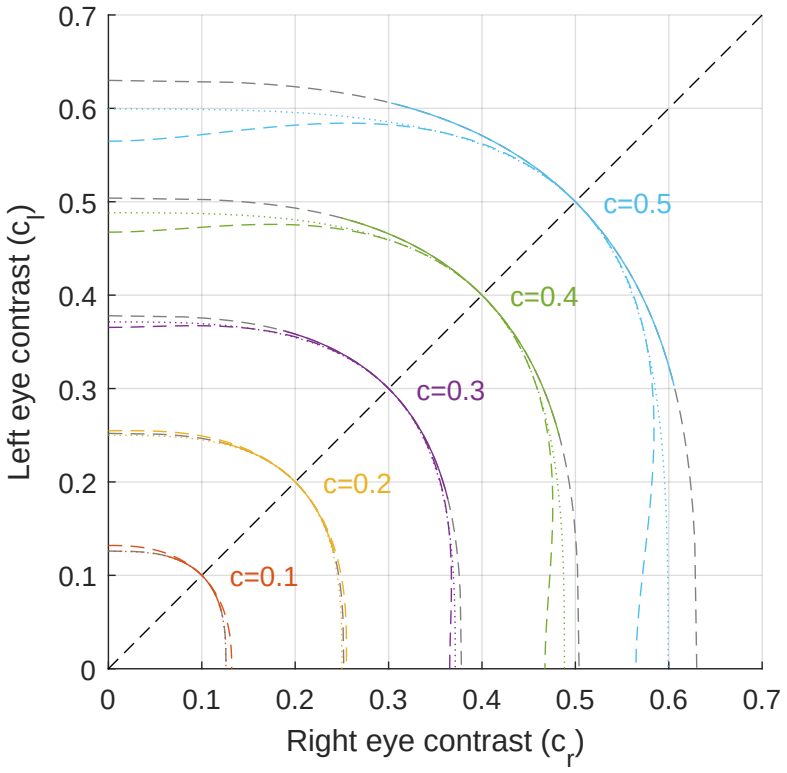


Figure 3.2: For each level of standard (dioptic) contrast ($c = 0.1..0.5$), the colour lines show the combination of the left and right eye contrast (dichoptic contrast) that produces the match. The lines are plotted according to the contrast matching model from Equation 3.2 and assuming $\beta = 3$. The black-dashed line represents standard contrast. The grey-dashed lines illustrate the range of contrast combinations that result in an unstable percept and rivalry. The colour dashed lines illustrate the same relation but according to the late summation model (Equation 3.3), and the dotted colour lines show the relation in terms of logarithmic contrast (Equation 3.4).

causing alternation between perceived images [9]. Rivalry is caused primarily by geometric differences in the two eyes' images. A special case is *lustre*, which occurs when luminance or contrast differences exist in corresponding image areas. It creates a shiny appearance in such areas.

Fusion of luminance. When a uniform patch of luminance L_l is shown to the left eye, and a patch of luminance L_r to the right eye, the fused patch can be matched to the luminance that is the (weighted) average of those:

$$L_{\text{fused}} = a\,L_l + (1-a)\,L_r\ , \tag{3.1}$$

where L_{fused} is the matching luminance (presented to both eyes) and a compensates for the dominant eye [96] ($a{\approx}0.5$).

Fusion of contrast. Legge and Rubin [94] investigated perceived contrast when two stimuli of the same spatial configuration but different contrasts are presented to the two eyes. Two stimuli were presented: The *standard* in which the same contrast is presented to the two eyes and the *test* in which a different contrast is presented to each eye. The subject adjusted the contrast of the test in one eye to create the same perceived contrast for the standard and test. They found that a generalised mean best describes their data. If we present contrast c_L to the left eye and contrast c_R to the right eye, the magnitude of the perceived, matched standard/dioptic contrast c_m is:

$$c_m = \left(\frac{c_L^\beta + c_R^\beta}{2} \right)^{\frac{1}{\beta}} . \tag{3.2}$$

β tends to be close to 3. It is the same across spatial frequencies and increases slightly with contrast. The matching contrast obtained by the above formula is illustrated as colour curves in Figure 3.2. The curves show that the fused contrast is dominated by the eye with the stronger contrast, in a manner that is close to the winner-take-all strategy.

Kingdom and Libenson [84] further show that the contrast fusion can be explained by the late summation model in which the signals from both eyes contribute to the response, R, of a contrast transducer function:

$$R(c_L, c_R) = \frac{c_L^p + c_R^p}{z + c_L^q + c_R^q} , \tag{3.3}$$

where z, p, and q are the parameters controlling the shape of the contrast transducer [93]. Curves of matching contrast resulting from the late summation model are shown as dashed colour curves in Figure 3.2. Because both models are comparable in the range where inter-ocular contrast differences are small (and the rivalry is low), we will rely on the simpler form in Equation 3.2 in further analysis.

3.1.3 Dichoptic contrast enhancement

In this section, we explain how the contrast of images seen binocularly can be enhanced beyond what can be reproduced on a typical display significantly improving image quality and realism in VR headsets and stereo displays. Our method was inspired by the observation of Legge and Rubin that the fused contrast is dominated by the image of higher contrast (Equation 3.2). We take advantage of stereoscopic displays, which can present a different image to each eye and therefore offer a separate dynamic range budget for the left and right eye. This lets us selectively use lower or higher tone curve slopes to improve image

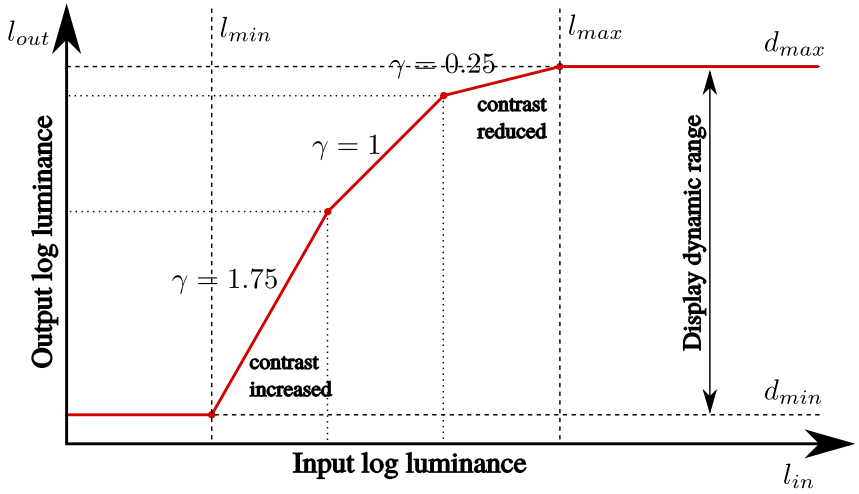


Figure 3.3: An example tone curve mapping input image log luminance to output image log luminance. The slope of the tone curve corresponds to the reduction or increase in contrast in the given tonal range of an image.

contrast. When binocularly fused, the images convey more fine detail in the shadows and highlights compared to standard tone-mapped images.

3.1.3.1 Tone curves and contrast enhancement

We define a tone curve as a function mapping the logarithmic luminance (base-10 logarithm) of the input image to the physical logarithmic luminance of the display device, as shown in Figure 3.3. Representing luminance in the logarithmic domain makes it more perceptually uniform (see Sec 2.4 in [117]) but also has the property that the slope of the tone curve in the log-log domain modulates the contrast of the corresponding tonal range. Altering the slope corresponds to multiplying log-luminance values: i.e., raising linear luminance values to a power (commonly known as gamma).

A well-selected tone curve can achieve high contrast in any relevant tonal range while mapping all pixel values to the available dynamic range. Assigning a steeper slope in one part of the tone curve boosts contrast in that range, however, a larger proportion of the output dynamic range budget is spent, necessitating contrast compression in another part of the input range. The output log-luminance is restricted by the peak luminance of the display (d_{max}) and its black level (d_{min}).

To ensure that we can rely on the contrast fusion rule when manipulating tone curves, we need to address the discrepancy in contrast units. The contrast fusion rule in Equation 3.2 is defined in terms of Michelson contrast, which we denote as c. The slope of the tone curve directly alters logarithmic contrast, which we denote as g. Logarithmic contrast is defined

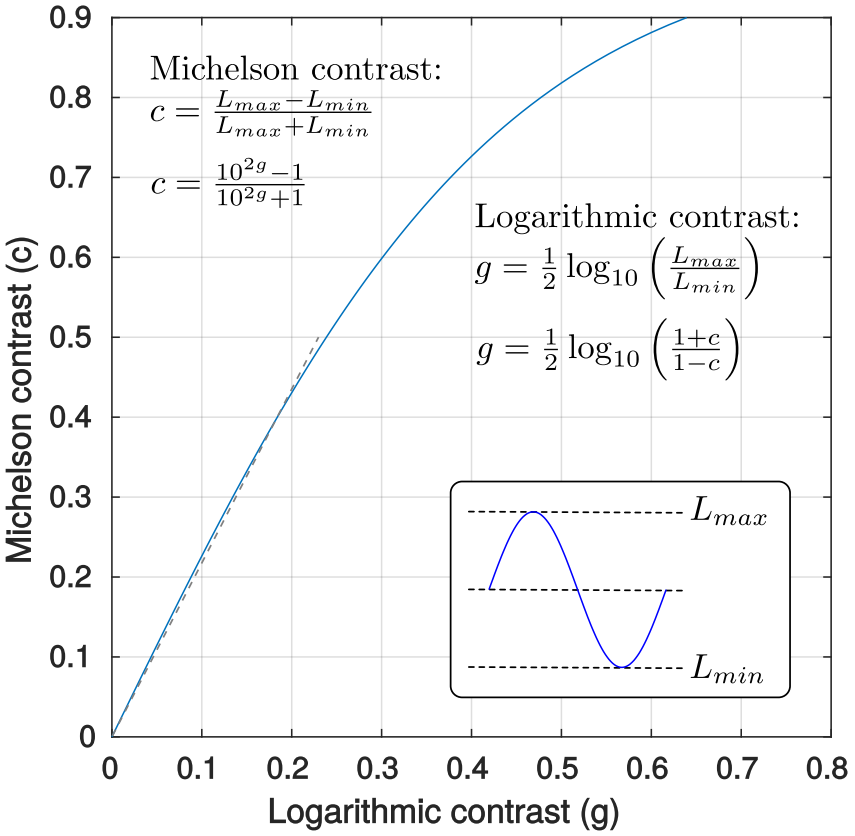


Figure 3.4: The relation between logarithmic and Michelson contrast.

as half of the logarithm of the luminance ratio, as illustrated in Figure 3.4. Logarithmic contrast is not equivalent to Michelson contrast. However, for small and medium contrasts ($c < 0.5$) that dominate natural or computer-generated imagery, both contrast measures are linearly related, as shown in Figure 3.4. Thus, the contrast fusion can be expressed in terms of logarithmic contrast:

$$g_m = \left(\frac{g_L^{\beta} + g_R^{\beta}}{2}\right)^{\frac{1}{\beta}} . \tag{3.4}$$

The new contrast matching formula, plotted as dotted lines in Figure 3.2, predicts contrast match that lies between the predictions of Equations 3.2 and 3.3.

3.1.3.2 Interleaved dichoptic tone curves

Let us consider how we can design a tone curve that would maximise contrast enhancement within the given budget of the dynamic range. A simple approach would be to create two tone curves, like those in Figure 3.5, consisting of two piece-wise linear segments. For a given tone curve segment, the slope in one eye can be increased while reduced in the other

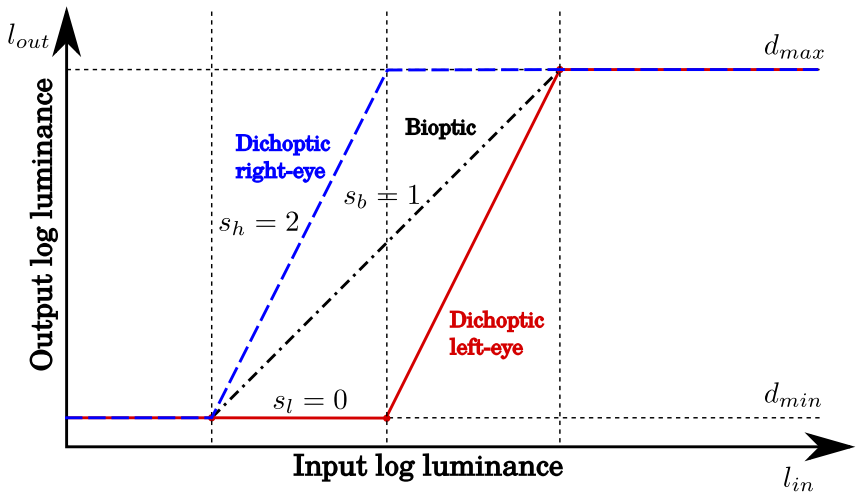


Figure 3.5: When a separate tone curve is used for each eye (dichoptic presentation) the slope of one of the curves can be up to twice as high as that for a standard dioptic presentation. The perceived contrast for the dichopic images will be 10%–50% higher (see Figure 3.6). However, such a strong separation of the tone curves will result in an image that is very uncomfortable to view.

without exceeding the dynamic range budget. If the base tone curve (black dashed line in Figure 3.5) has the slope s_b, we set the slope for one eye to s_l and the slope for the other eye to $s_h = 2\,s_b - s_l$ so that $s_l + s_h = 2\,s_b$. We will use indices l and h to denote low and high slopes (rather than left and right eyes) as the slopes will be assigned interchangeably to each eye for each segment of the tone curve. From Equation 3.2, we can find that the gain in fused contrast for the original contrast g is:

$$\Gamma = \frac{1}{g\,s_b}\left(\frac{(g\,s_l)^\beta + (g\,s_h)^\beta}{2}\right)^{\frac{1}{\beta}} = \frac{1}{s_b}\left(\frac{s_l^\beta + s_h^\beta}{2}\right)^{\frac{1}{\beta}} . \tag{3.5}$$

The gain as the function of the slope on the left and right eye is plotted in Figure 3.6. The curves clearly show that the gain in perceived contrast is greatest when the slope is maximised in one eye and minimised in another. However, such a large luminance and contrast difference could result in strong binocular rivalry.

To reduce the luminance difference and thus the potential cause of rivalry, we want the left- and right-eye tone curves to be more similar to each other. This can be achieved with an interleaved tone curve with a higher number of piece-wise linear segments, such as the one in Figure 3.7. It should be noted that increasing the number of segments does not affect the slopes of the curves in the left and right eyes and therefore does not affect contrast enhancement. However, the number of segments restricts the highest contrast that can be manipulated by the tone curve: if the contrast between two pixels is large enough to span two segments of the tone curve (i.e. be larger than Δ_{in}), it is not going to

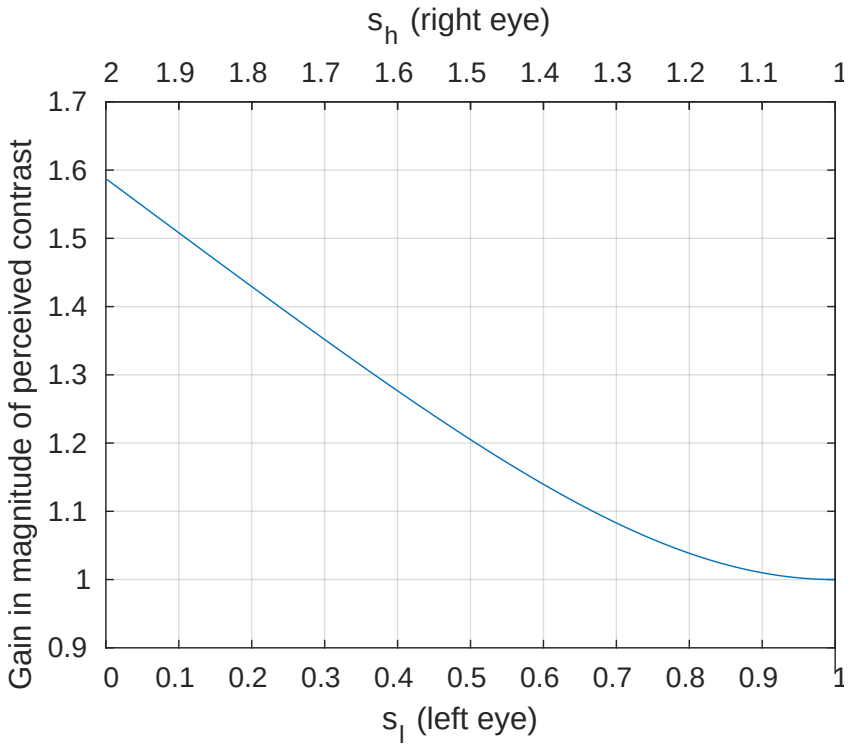


Figure 3.6: The gain in contrast due to fusing left and right eye images which are processed by the tone curves with the slopes s_l and s_h (x-axis). As the tone curve slope is reduced on the left eye (s_l), it is increased on the right eye (s_h). Such a change in slope does not reduce the dynamic range budget allocated to both eyes, but it boosts fused contrast.

be enhanced (or reduced) as intended. Finding the right number of segments and their slopes is a challenging problem and we address this problem in a series of experiments in Section 3.1.4. But first, we explain why we need to ensure the smoothness of the tone curves.

3.1.3.3 Smooth tone curves

In preliminary experiments, we observed that the piece-wise linear interleaved tone curves may result in banding artefacts when an image contains large areas with smooth gradients. These are caused by the C^1 discontinuities in our tone curves, which translate to similar discontinuities in the resulting image. The visual system is very sensitive to such discontinuities, which are interpreted as spurious contours [83]. This problem can be easily addressed by replacing the small intervals containing discontinuities in the piece-wise linear curve with a cubic Bezier curve. We set the size of the interval to be 0.1 log_{10} units. The three control points of this Bezier curve are the two endpoints on the interval and the slope-transition point, as shown in 3.7-(b). This ensures that our tone curves are C^1 continuous in the entire domain.

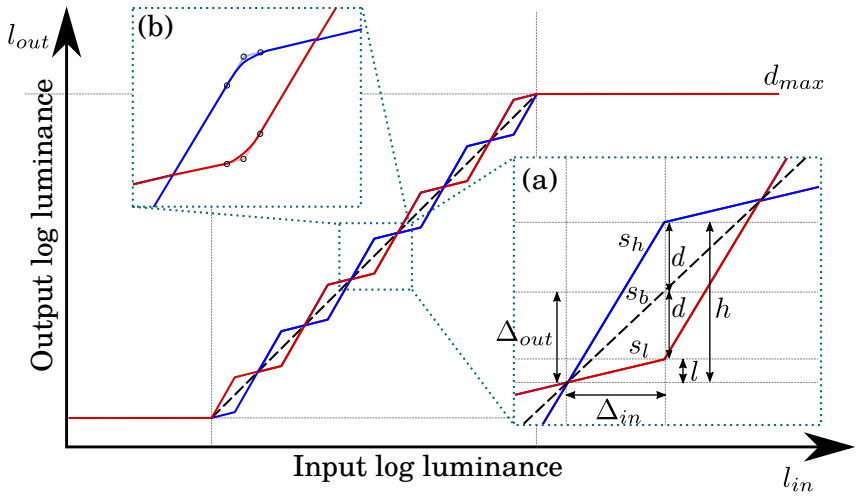


Figure 3.7: Binocular tone curves may introduce less rivalry if they are constructed so that resulting luminance values in each eye are possibly similar. The interleaved low- and high-slope segments could be used to produce such curves. Inset (a) shows the notation we use. We denote the lower slope as $s_l = l/\Delta_{in}$ and the higher slope as $s_h = h/\Delta_{in}$. We also denote the number of linear segments in the tone curve as N. For example, the segment that spans Δ_{in} is what we mean by one segment. Inset (b) shows smoothing using Bezier curves. The black circles denote the control points.

3.1.4 The predictor of rivalry

The interleaved dichoptic tone curves are controlled by two parameters: the number of segments and the slope of the interleaved tone curves. To determine the optimal choice of these parameters that would produce the strongest enhancement and acceptable level of rivalry, we conducted a perceptual experiment. The experiment was intended to test two hypotheses, each proposing a different indicator of binocular rivalry:

Hypothesis 1 If rivalry is induced by the luminance difference between the left and right eyes, a good predictor would be the maximum log-luminance difference, or $h - l$ using the notation from Figure 3.7. Note that $h - l = (s_h - s_l)\,\Delta_{in}$.

Hypothesis 2 Rivalry may also be caused by the contrast difference between the left and right eyes. A good predictor in this case would be the ratio of contrasts presented to the two eyes $s_l/s_h = l/h$.

Apparatus and participants The experiment was performed on a 24-inch NEC PA241W colourimetrically calibrated display with an attached stereoscope in a dark room (Figure 3.8). The optical path to the display was 36 cm (2.77 D). Eight volunteers

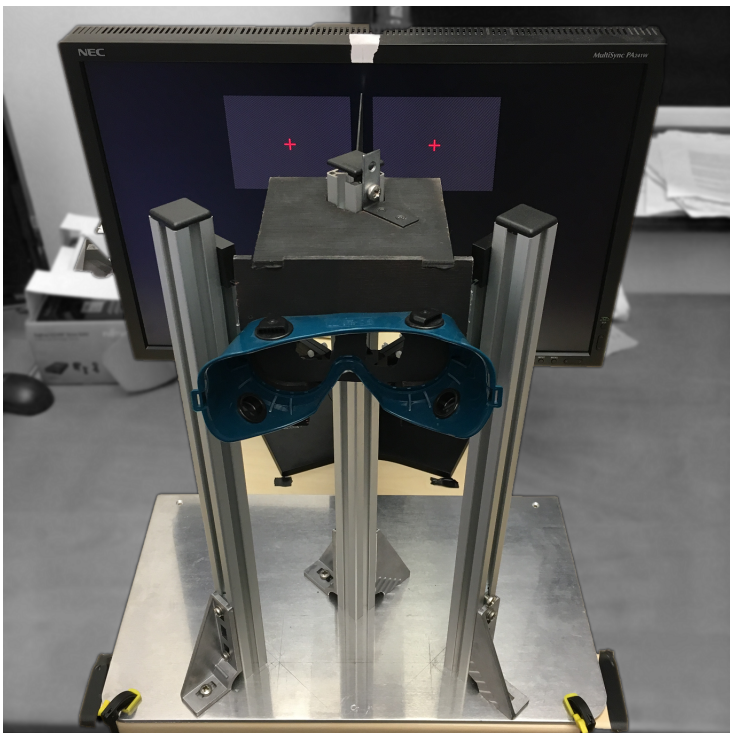


Figure 3.8: LCD display with a stereoscope used in the experiments.

participated (8 males, mean age 27.3, SD 4.2 years). Before the experiment, each participant read and signed the consent form. We also demonstrated to each participant what rivalrous and non-rivalrous stimuli look like.

Stimuli and procedure We selected 16 HDR images, which were tone mapped based on the smooth inter-leaved tone curves with N equal segments as explained in Section 3.1.3. The end-points of the tone curve were set to be at the 1st and 99th percentiles of image luminance. The dynamic range of the target display was 2.7 log-10 units (500:1 contrast).

The participants were asked to adjust the deviation d (shown in Figure 3.7) from the straight tone curve so that "the image looks sharp and comfortable to view" (exact wording on the briefing form). The critical values of d were measured using the method-of-adjustment procedure with three repetitions per image. Then, the two proposed predictors were computed accordingly as:

$$h - l = 2d \tag{3.6}$$

$$\frac{l}{h} = \frac{\Delta_{out} - d}{\Delta_{out} + d} \tag{3.7}$$

The experiment consisted of six sessions. The same HDR images were used in all of them. Four of the sessions had $N = 2, 4, 10$, and 20 interleaved segments spanning the entire

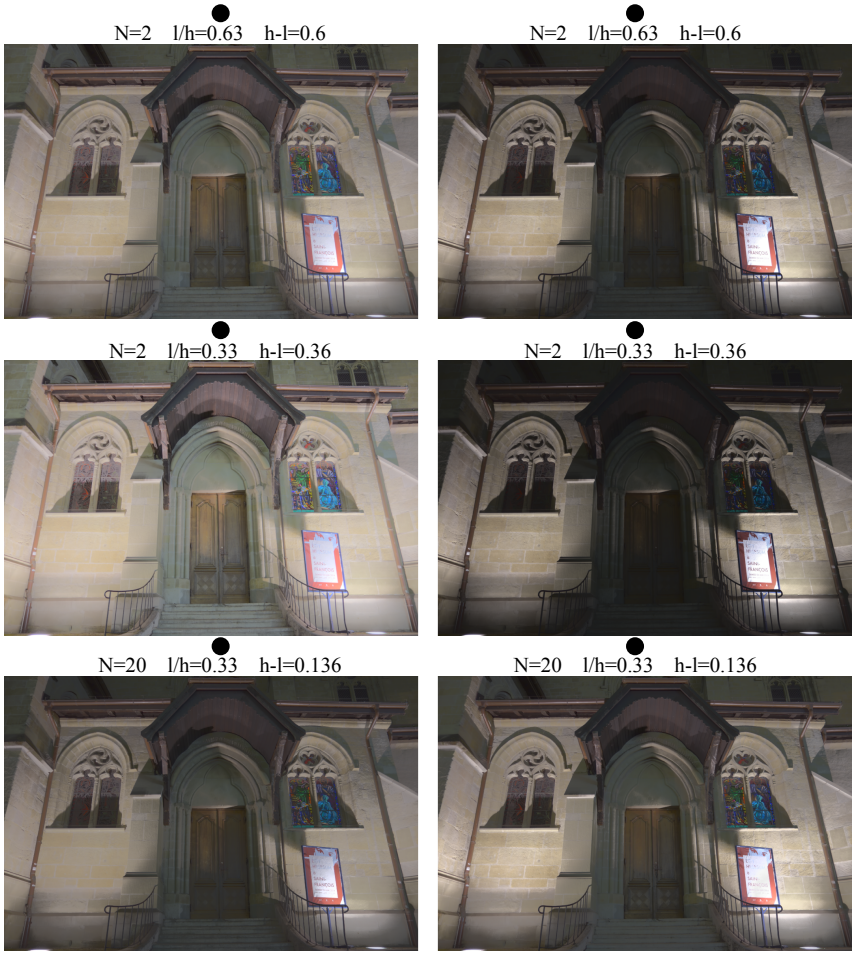


Figure 3.9: Examples of DiCE-enhanced monoscopic images from Experiment 3.1.4, with different strengths of enhancement (the enhancement is stronger at a low l/h ratio), and a different number of segments of interleaved tone curves (N). The images are suitable for cross-fusion.

dynamic range of the display, 2.7 log-10 units. The two remaining sessions had $N = 10$ segments spanning half of the display's dynamic range, 1.35 log-10 units, so that one session spanned the darker half and one the brighter half of the dynamic range. Examples of images rendered with a different number of segments and slopes are shown in Figure 3.9. The order of sessions and images was randomised.

Results The plots for the two proposed predictors and for eight participants are shown in Figure 3.10. It is evident that the ratio of contrast l/h is a much more consistent predictor than the log-luminance difference across different test conditions (number of segments, output display dynamic range). This was further tested in a leave-one-out cross-validation, where we used 7 out of 8 of the measured images to calculate a fixed value of the predictor, which was then used to predict the s_l values of the remaining images. The procedure was repeated eight times. The prediction error was computed as RMSE between the true and

Table 3.1: Each row represents the prediction errors (RMSE) for each participant using the corresponding predictors.

Participant	Log-luminance difference	Ratio of contrast
1	0.4182	0.0870
2	0.3793	0.0971
3	0.2880	0.0795
4	0.2062	0.0576
5	0.3811	0.0895
6	0.5217	0.1541
7	0.2824	0.0892
8	0.3148	0.0981

predicted s_l and is shown in Table 3.1. The results suggest that the ratio of contrast l/h is indeed the better predictor for s_l.

Discussion The results demonstrate that the magnitude of rivalry is determined by the contrast difference between the eyes (Hypothesis 2) rather than by the luminance difference (Hypothesis 1). This finding confirms the importance of contrast in visual processing [84]. There is ample evidence suggesting that low-level visual mechanisms attempt to preserve contrast but they do not encode information about absolute luminance. For example, Weber's law states that we are sensitive to ratios (contrast) rather than absolute levels. Contrast constancy preserves the appearance of supra-threshold contrast across spatial frequency and to some extent across luminance range [55, 86]. Furthermore, light-/dark-adaptation is attributed to a large extent to the retina (photoreceptors and bipolar cells) [37] and can be controlled individually per eye. This means that a per-eye luminance difference can be partially compensated by the adaptation mechanism. Therefore, it is not surprising that conflicting contrast signals evoke more rivalry than conflicting luminance signals. This finding also shows that some degree of rivalry is unavoidable as we need to introduce contrast differences for contrast enhancement. However, many observers reported that they can adapt to a moderate level of rivalry a few seconds after switching from standard to dichoptic presentation.

It should be also noted that the ratio of contrast l/h as a predictor of rivalry is independent of image content. As shown in Figure 3.10, we cannot observe a pattern for images that would be consistent across the participants. The differences in the means between observers are also small given the within-observer variance. Therefore, the high variance is likely to be due to the measurement noise, rather than systematic effects.

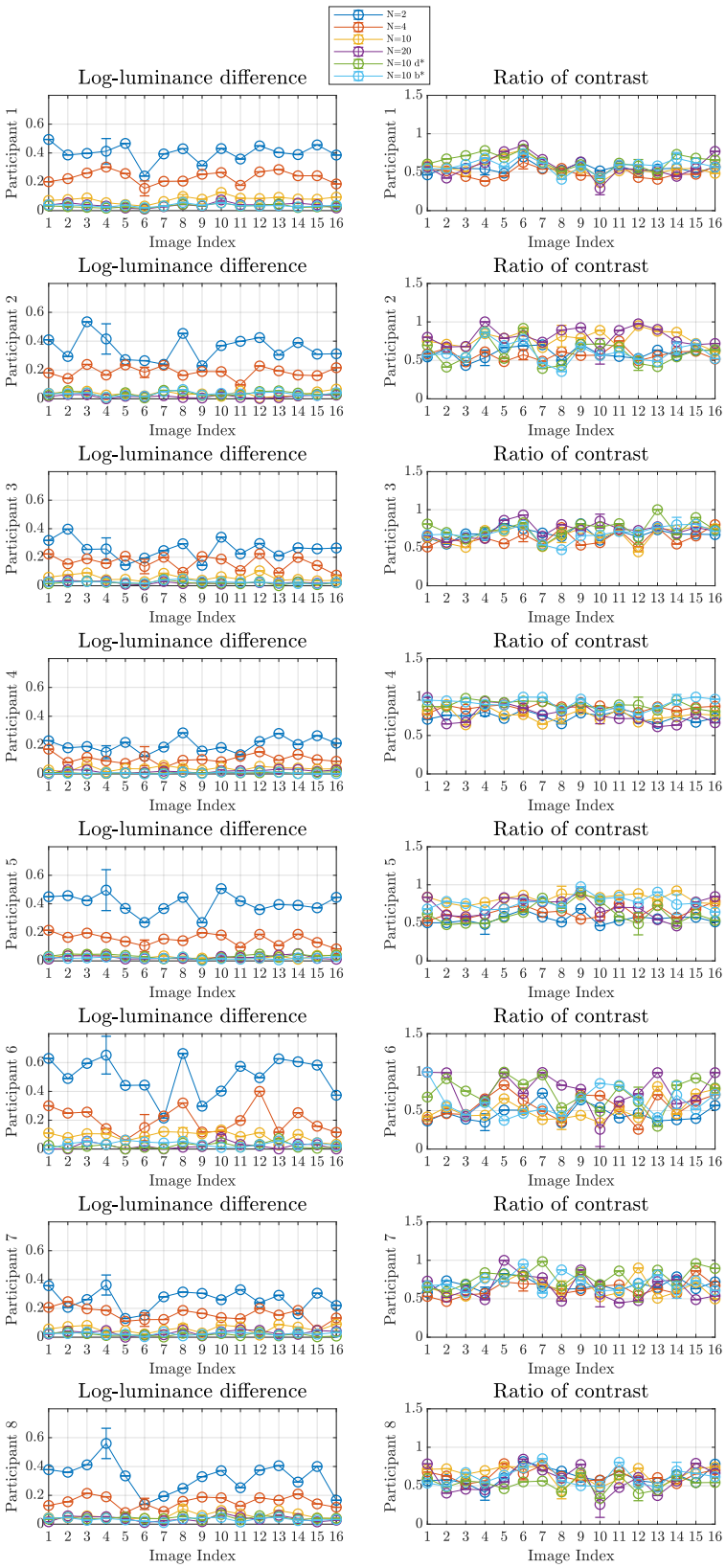


Figure 3.10: The two proposed predictors of binocular rivalry (columns) collected from Experiment 3.1.4, for eight participants (rows). The colours denote different numbers of segments N and different output display dynamic ranges (d* and b* indicate half of the display's dynamic range, with d* representing the darker half and b* representing the brighter half). The error bars represent the expected value of the standard deviation for the given set of conditions. It is evident that the ratio of contrast l/h is distributed more uniformly than the log-luminance difference.

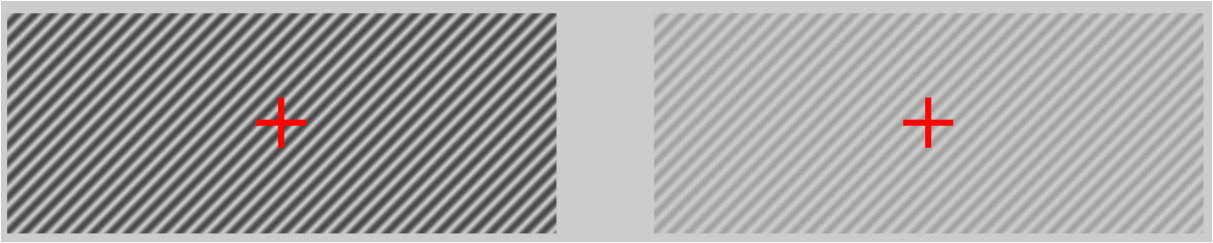


Figure 3.11: An example binocularly-fused sinusoid stimulus used in Experiment 3.1.5. The left and right stimuli were shown with different luminance.

3.1.5 Rivalry due to luminance difference

Although contrast seems to be the dominant factor in dichoptic rivalry, we cannot fully discount the effect of luminance. If we did so, we would need to assume that two images of the same contrast but very different luminance are always comfortable to fuse. To determine the maximum luminance difference that can be regarded as acceptable, we conducted one additional experiment using the same protocol as in Experiment 3.1.4.

Apparatus and Participants This experiment shares the same setup as Experiment 3.1.4. Five volunteers participated (5 males, mean age 25.2, SD 2.2 years). Before the actual experiment, they read the consent and briefing forms. In a short demo, they were shown examples of rivalrous and non-rivalrous stimuli.

Stimuli and Procedure We used sinusoid gratings as the stimuli, as shown in Figure 3.11. The gratings shown to each eye had the same contrast and frequency, but differed in luminance. Participants were asked to adjust the difference of luminance given the same criteria as in Experiment 3.1.4. Six sinusoidal gratings were generated: a factorial combination of 2 contrasts (0.2 or 0.4) and 3 frequencies (1, 3 or 5 cpd). Each condition was measured three times and the order of all trials was randomised.

Results The results indicated that most observers can tolerate the luminance difference $(h - l)$ up to 0.66 log-10 units (50th percentile). The 25th, 50th, and 75th percentiles of the data for the threshold of luminance difference are 0.51, 0.66 and 0.80 in log-10 units. We use these results to determine the best number of segments in Section 3.1.6.1.

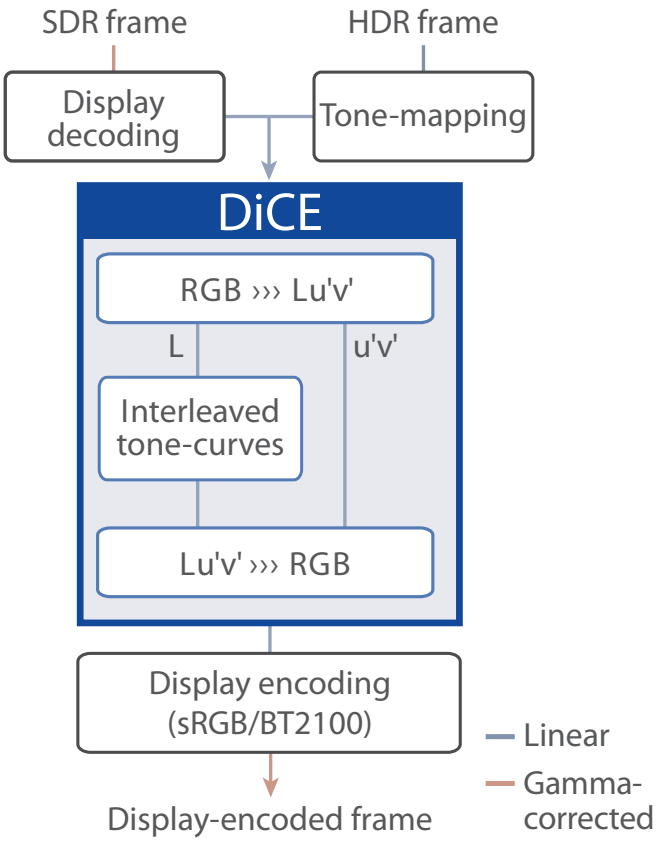


Figure 3.12: DiCE as part of a tone-mapping pipeline. The dynamic range of HDR input frames (in linear RGB colour space) can be reduced with any tone mapping operator. Alternatively, a standard SDR frame can be used. The luminance is separated from two colour-opponent channels. The per-eye interleaved tone curves are applied to the luminance channel, separately for each eye and then colour is added back. Finally, the pixel values are display-encoded into SDR (sRGB) or HDR (rec.2100) display-referred space.

3.1.6 Implementation

Experiment 3.1.4 demonstrated that binocular rivalry is mostly induced by the contrast difference between the eyes. The variance in the perceived rivalry between the images is relatively small, therefore, we can make our enhancement method independent of image content. Our interleaved tone curves can be precomputed, and applied to an image after tone mapping (but before display coding). This is a significant advantage of our DiCE method, letting us use it with any existing tone-mapping operator, or directly with SDR images.

Figure 3.12 shows the diagram of a tone-mapping pipeline with DiCE. First, any existing tone-mapping operator can be used to reduce the dynamic range of an HDR frame and generate a display-referred frame. Alternatively, an SDR frame, decoded into a linear RGB

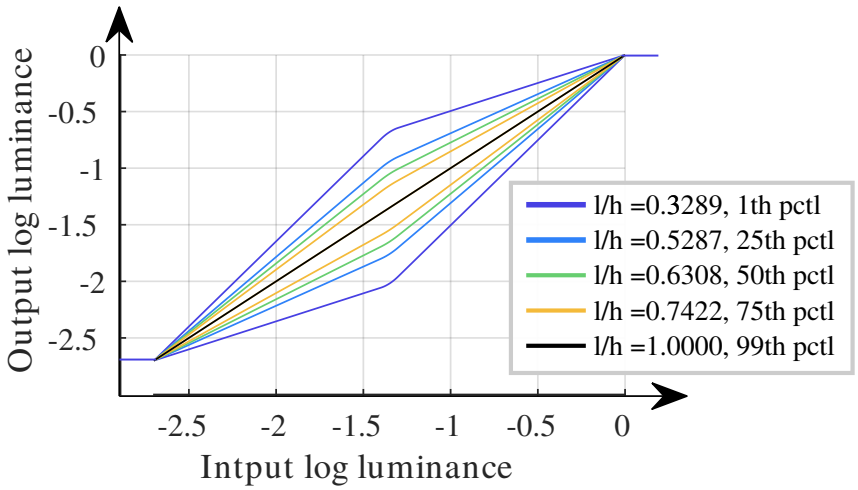


Figure 3.13: The shape of the dichoptic tone-curves at different l/h ratios. The ratios were selected to represent the 1st, 25th, 50th, 75th and 99th percentile of the data (across all images and observers) from Experiment 3.1.4.

space, can be used as input to our method. We then separate a luminance channel from CIE $u'v'$ chromaticities and apply the interleaved tone curves to the luminance channel alone. The colour is added back using an inverse colour transformation. Finally, the colours are displayed encoded and stored in a raster buffer. Depending on the target display, they can be encoded into the sRGB space for SDR displays, or one of the colour spaces from the ITU BT.2100 recommendation for HDR displays.

3.1.6.1 Selecting interleaved tone-curve parameters

Our experimental results indicate that l/h determines both contrast enhancement and the magnitude of rivalry. The l/h is also independent of the number of segments. Given that, we opt for the smallest number of segments for two reasons: a) wider segments let us enhance a broader range of spatial frequencies (as discussed in Section 3.1.3.2); and b) there is a smaller chance for banding artefacts in the region where the tone curve switches from low to the high slope (as discussed in Section 3.1.3.3). However, a small number of segments increases the maximum luminance difference, which could be another cause of rivalry, as discussed in Section 3.1.5. Therefore, in Figure 3.14 we plot the maximum luminance difference $(h - l)$ as a function of the display dynamic range and the number of segments. The plots show that $N = 2$ is the right choice for most SDR displays up to 2.8 log-10 units of the dynamic range, including OLED displays used in HTC Vive and Oculus Rift. The number of segments, however, may need to be increased to 4 for high-contrast HDR displays.

Slope selection for the interleaved tone curves creates a trade-off between contrast enhance-

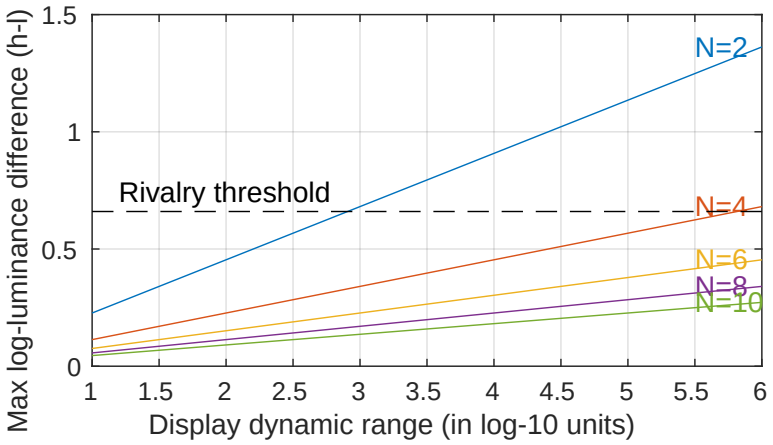


Figure 3.14: The maximum log-10-luminance difference ($h-l$) for a given display dynamic range (x-axis) and the number of segments (N, colours). The plots are drawn assuming $l/h = 0.63$ (50th percentile). The dashed line represents the rivalry threshold (50th percentile) for log-luminance difference (Experiment 3.1.5). The plot shows that for most SDR displays (dynamic range less than 2.8 log-10 units), we do not need more than 2 segments.

ment and binocular rivalry. Figure 3.4 shows that contrast enhancement is maximised for small ratios l/h, but, as found in the rivalry experiment, such small ratios increase binocular rivalry. Therefore, the ratio l/h should be set as a parameter, adjusted per user, ranging from about 0.5–0.75. The family of interleaved tone curves for the range of l/h ratios and two segments is shown in Figure 3.13.

3.1.6.2 DiCE for partial overlap HMDs

As discussed in Section 2.1.1, binocular human vision is achieved via two monocular visual fields of around 160° of horizontal visual angle each; their total horizontal field of view is approximately 200°. The combined FoV consists of three regions: an overlapping 120° central binocular region where stereopsis is achieved and two flanking monocular regions of approx. 40° each [136].

Older HMDs employed a full overlap design, in which both eyes saw the same part of the scene. This resulted in a smaller FoV as the optical design was limited by the human binocular region. Modern commercial HMDs have a partial overlap design, mimicking the human visual system. This allows for physically smaller displays while both increasing the FoV and thus immersion, and supporting wider aspect ratios [52]. In such HMDs, binocular overlap refers to the visible overlapping portion between the two eyes (see Fig. 3.15) in

the headset and describes how much of the virtual scene can be seen by both eyes, which is crucial for depth perception. In partial overlap binocular displays, only a central region of the scene is shown to both eyes, and areas to either side are seen by only one eye. This often creates interocular differences in the monocular regions [138] which often induce a perceptual effect known as *luning* which is the subjective darkening in the monocular regions or, for other users, it is experienced as a visual fragmentation of the field-of-view into three distinct regions (left, middle, right) [85].

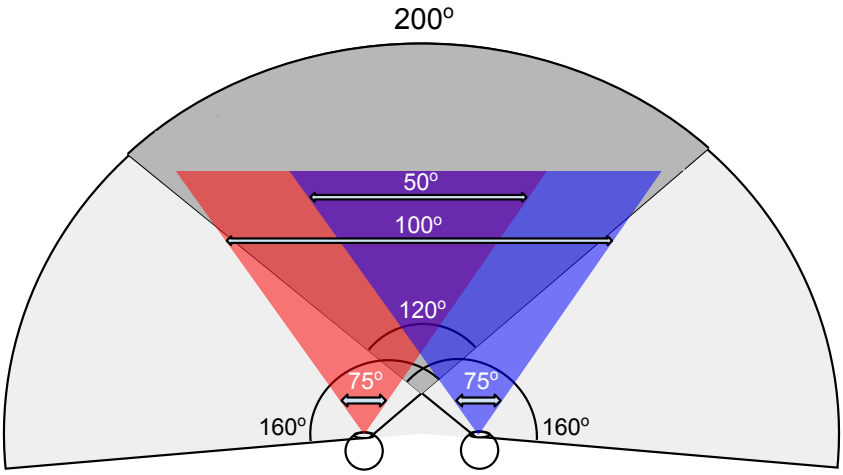


Figure 3.15: Arcs denote angles for viewing in the real world: each eye sees a field of view of about 160°. This results in a 200° combined horizontal field of view, 120° of which are overlapping and thus binocular processing or stereopsis is possible. Two-headed arrows denote angles in a modern VR headset: each eye sees a horizontal FoV of about 75°, leading to a 100° combined FoV, only 50° of which are overlapping and available for binocular processing or stereopsis.

For partial overlap HMDs (most commercial headsets) we cannot apply the interleaved tone curves to the entire FoV. If the monocular flanking regions (magenta and blue lobes in Fig. 3.15) are processed by the interleaved DiCE curves, they remain unfused and show spurious contrast modulation. This is magnified with head motion in VR, which causes the contrast appearance to change in the flanked regions. To avoid this problem, we employ a piece-wise linear blending function that ensures a gradual transition between the dichoptically tone-mapped area of the image that is viewed binocularly, to the monocularly tone-mapped flanking lobes. The binocular overlap area depends on the fixed headset optical setup and the eye relief, i.e., the distance of the eye from the lens, which itself depends both on how deep-set the eyes are in the face and how pronounced the brow is.

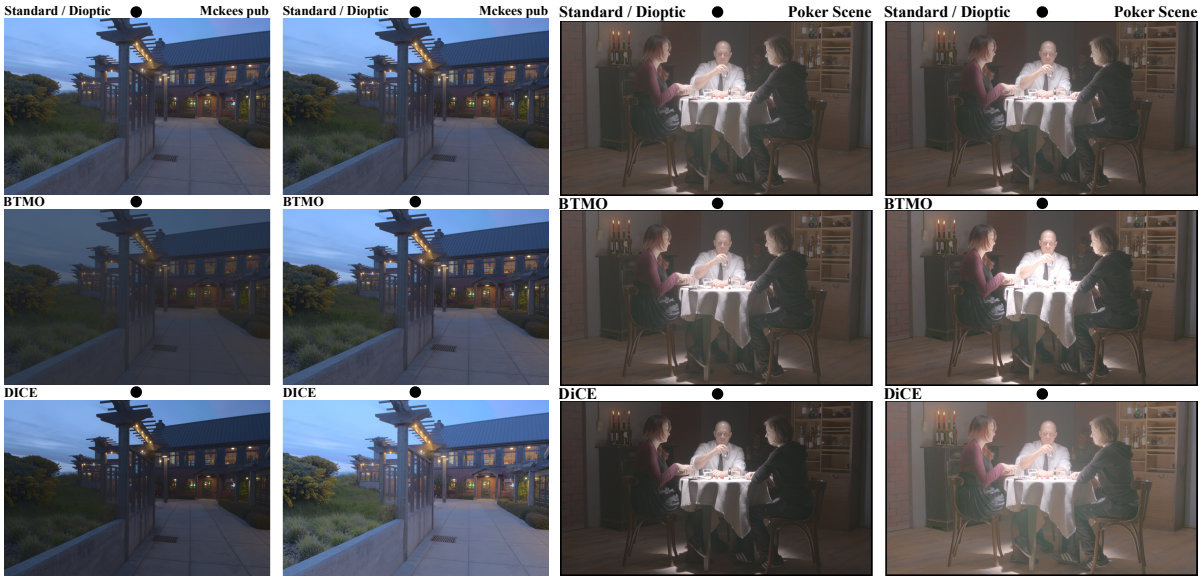


Figure 3.16: Example of two images used in Experiment 3.1.7.1 in a format suitable for cross-fusion.

3.1.7 Evaluation

We compare our method with the standard presentation and BTMO technique on the stereoscope in Experiment 3.1.7.1, and then evaluate in VR rendering in Experiment 3.1.7.2.

3.1.7.1 Validation with a stereo display

In this experiment, we compare our technique with the standard presentation (no dichoptic enhancement) and previous work (BTMO [200]) on a stereo display.

Apparatus and participants We used the same display and stereoscope as in the first experiment. 16 volunteers participated (5 females, mean age 26.8, SD 4.3 years).

Stimuli 17 monoscopic images and 2 stereoscopic images were processed with our DiCE technique and BTMO [200]. The images were kindly processed by the authors of the BTMO paper. It should be noted that both techniques serve a different purpose: BTMO is a tone-mapping operator that requires an HDR image as input. Our DiCE technique expects as input an image that has already been tone-mapped. Therefore, to reduce differences between the methods due to different tone-mapping operators, we used one of the images generated by BTMO as the standard/dioptic condition (no enhancement) and also as the input to our technique. When selecting an image, we chose the one from

the pair (left- and right-eye image) which contained fewer under- or over-exposed pixels. We used a l/h ratio of 0.63 for all DiCE-enhanced images, which was the median from Experiment 3.1.4. We selected a median rather than a higher percentile as we noted that the participants are more conservative when they are asked to self-report the rivalry threshold and can tolerate higher rivalry over time. Two images used in the experiment are shown in Figure 3.16 for cross-fusion. All other images are shown in Figure 3.17.

Procedure We used a full-design pairwise comparison experiment in which all unique combinations of conditions are compared: DiCE vs. standard/dioptic, BTMO vs. standard/dioptic, and DiCE vs. BTMO. The participants were asked two questions regarding each image pair that they saw: *which image has a higher contrast?* and *which image looks better?* The participants could switch between one and the other image in the pair using the arrow keys and they confirmed the image of higher contrast with the space key and the image they preferred with the return key. Each pair was compared three times by each observer. The order of image pairs was randomised.

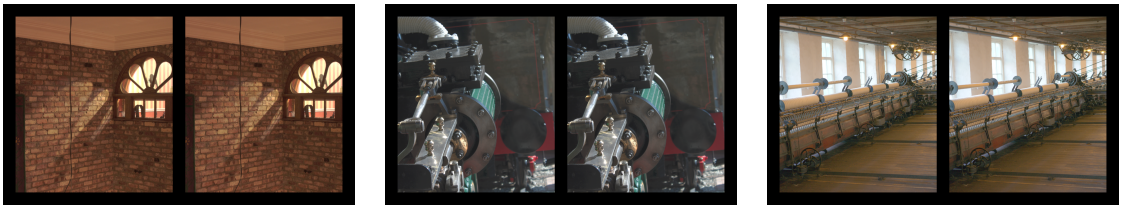
Data analysis The results of the pairwise comparison experiments were scaled using publicly available software[1] under Thurstone Model V assumptions in just-objectionable differences (JODs), which quantify the relative quality differences between the techniques. A difference of 1 JOD means that 75% of the population can spot a difference between two conditions. The details of the scaling procedure can be found in [141]. Since JOD values are relative, the bioptic (baseline) condition was fixed at 0 JOD for easier interpretation.

Results Results in Figure 3.18-a show that our DiCE method produces images of higher perceived contrast compared to their standard/dioptic counterparts, demonstrating that the contrast fusion model is effective in complex images. The BTMO results are mixed, sometimes producing images of higher, but sometimes also of lower contrast compared to the standard/dioptic condition and DiCE. It is difficult to compare DiCE and BTMO techniques in terms of contrast enhancement, as each technique can produce images of even higher contrast if the binocular rivalry metric is relaxed. This, however, will result in images that are uncomfortable to view. The main strength of DiCE over BTMO is that the enhancement is consistent across the images, demonstrating that the direct manipulation of contrast in DiCE offers better control over resulting images than the optimization used in the BTMO method.

The preference results, shown in Figure 3.18-b, are less conclusive as large subjective

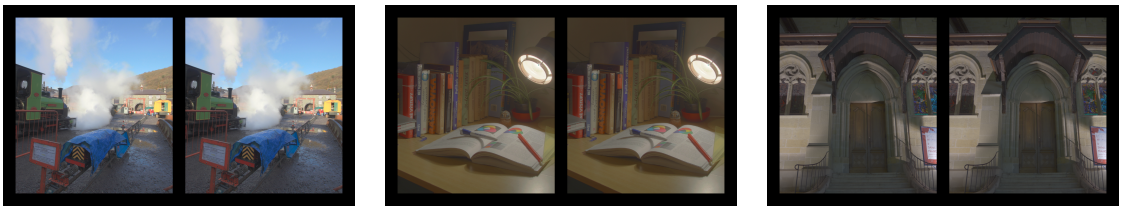
[1] pwcmp — `https://github.com/mantiuk/pwcmp`

(a) Lone rock (b) Mountain view (c) Mountain view



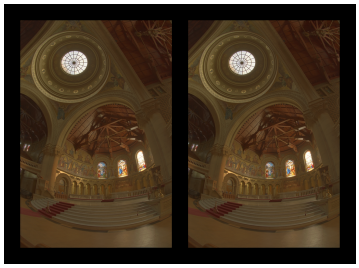
(d) Brick wall (e) Locomotive (f) Weaving machine

(g) Menai bridge (h) Moelfre (i) Snowdonia torrent



(j) Steam engine (k) Desk and lamp (l) Church door

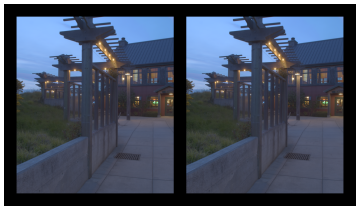
(m) Illuminated statue



(n) Memorial church

(o) Oxford church

(p) Pooh and Tiger



(q) McKees pub

(r) Landscape

(s) Poker scene

Figure 3.17: Test images used in Experiment 3.1.7.1

variations made most differences statistically insignificant. For the DiCE method, we could measure the preference difference only for the 2 (Poker scene and Moelfre) out of 19 images. These differences could still be accidental as the test does not correct for multiple comparisons. For 8 out of 10 comparisons that are statistically significant, the BTMO method produced less preferred results than standard (dioptic) presentation and only in two cases the preference was higher. This is in contrast to findings from [200], where the authors showed a strong preference for BTMO over standard presentation. We can only speculate that the effect could be due to the training of the participants; in our experiments, the participants with more exposure to dichoptic images also indicated a stronger preference for them. This could be compared to the experience of wearing new glasses, when it takes some time to get fully comfortable and used to the new correction. This result could be also explained by the broad meaning of the "preference" criterion, which could combine many factors, such as comfort, familiarity, visual quality, wow-effect, etc. The results suggest that single-dimensional "preference" may not be the best measure for the dichoptic contrast enhancement techniques.

Figure 3.16 shows an example of two images produced by each method: the one for which BTMO produces a higher contrast image (Poker scene) and the one for which DiCE produces a higher contrast image (McKees Pub).

3.1.7.2 Validation in VR

Experiment 3.1.7.1 was performed in a stereoscope, which provides high resolution and image quality, but it is less suitable for testing real-time rendering. Therefore, in the final experiment, we compare DiCE with standard presentation in VR environments. This experiment is also more relevant for the application of our method in real-time rendering. Note that we could not include BTMO in this experiment as that method is unsuitable for real-time rendering of 3D environments with 6DoF free viewing.

Apparatus and participants The VR environments were presented on an HTC Vive VR headset. Ten volunteers participated (2 females, mean age 25.8, SD 3.2 years).

Stimuli and procedure The stimuli consisted of three VR scenes shown in Figure 3.20, each seen from three different viewpoints. The participants could freely look around the scenes while seated on a swivel chair. To switch between DiCE and standard presentation, the participants pressed the trigger on the Vive controller. We used a l/h ratio of 0.55 for the DiCE method for similar reasons as in Experiment 3.1.7.1: to avoid overly

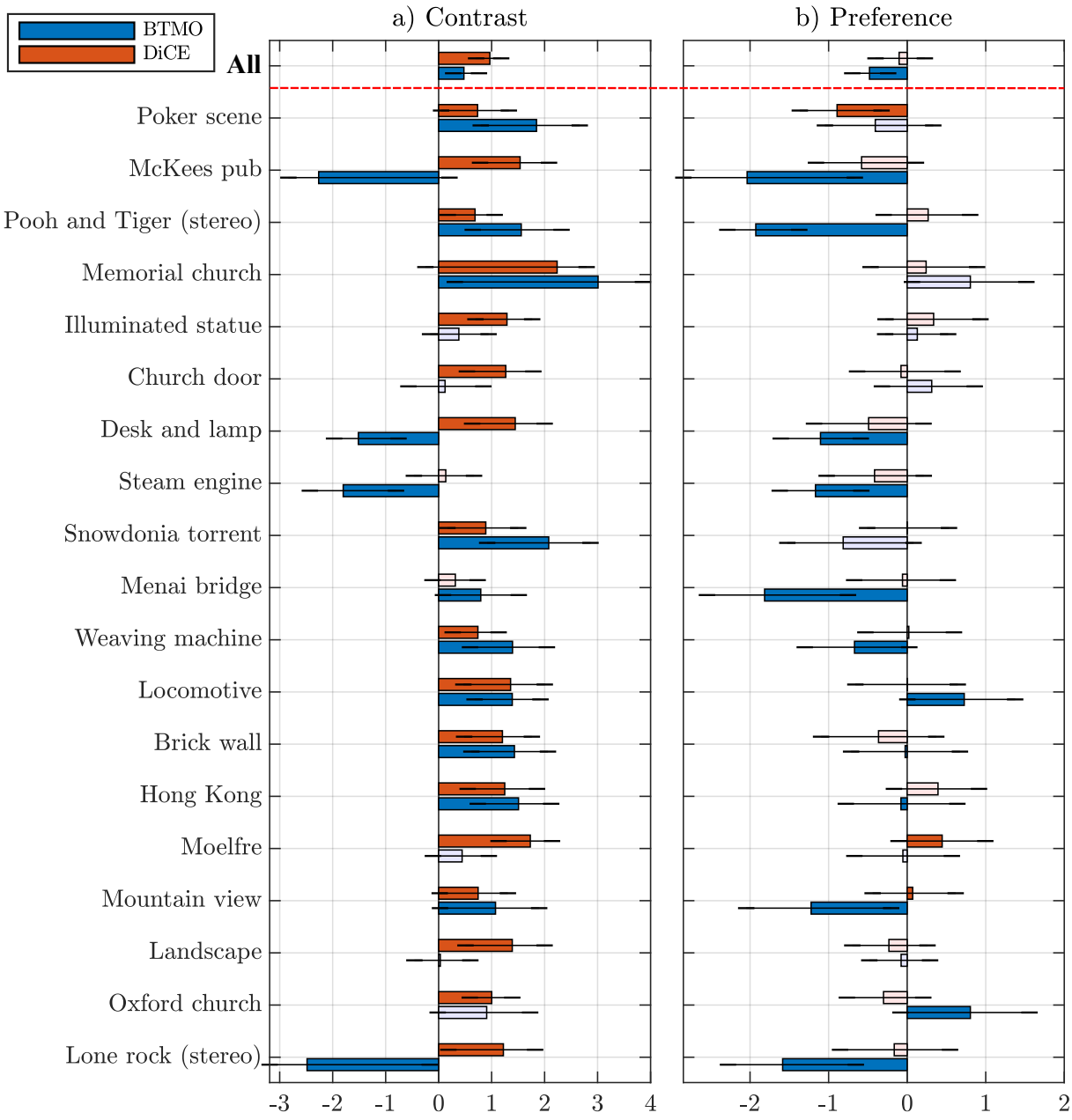


Figure 3.18: The results of the validation experiment, comparing perceived contrast (a) and preference (b). The results are reported for each scene and for the aggregated results across all the scenes. The bars indicate the quality improvement relative to the standard presentation (no dichoptic enhancement) in JOD units (the higher the better). +1 JOD in that scale means that 75% of observers select the given condition over the standard presentation. The negative values mean that the standard condition is selected more often. The grayed bars indicate that we have no statistical evidence that a given condition is different (with respect to contrast or preference) from the standard presentation. The statistical test does not include the correction for multiple comparisons.

conservative threshold adjustment and to make the method more different from the standard presentation. For each stimulus, they were asked three questions: 1) *which scene appears to be higher in contrast*?, 2) *which scene appears to have more depth*?, and 3) *in which scene the materials and textures look more realistic*?. The questions were motivated by our own observation that DiCE-enhanced images have different quality and appear more three-dimensional. We did not ask about their preference as the question did not give conclusive answers in Experiment 3.1.7.1. Before the experiment, each participant read and signed the briefing and consent forms. As part of a training session, each participant was presented with three pairs of images with examples of low/high contrast, three-dimensional/flat shading, and natural/unnatural looking textures (Figure 3.19). None of the participants reported symptoms of VR sickness after a 10-15 minute session (no formal questionnaire was used).

Results The results of Experiment 3.1.7.2 are presented in Figure 3.21 as percentages of participants who voted for DiCE when asked each of the three questions. It shows that our DiCE method produces higher contrast perception than standard presentation for all VR environments. The results also confirmed that the observers could perceive more depth with the DiCE enhancement. The effect can have a number of explanations. Ichihara et al. [73] showed that increased contrast can give an impression of depth. Binocular lustre may be causing lustrous features to pop out [181], giving the impression of false depth. Another possible explanation is that artificial disparity stemming from the different monocular images (luminance dichopticities) could give rise to a depth sensation [180]. The results for realistic-looking textures were less conclusive with only one environment, with the simplest textures and lowest complexity, showing a moderate preference for DiCE.

3.1.8 Discussion

The results of Experiment 3.1.7.1 and 3.1.7.2 confirmed that our DiCE technique can effectively enhance contrast not only for simplified stimuli, used in psychophysical models, but also for complex images. Experiment 3.1.7.2 indicated that our technique can also improve the impression of depth in images. This question emerged when we were inspecting the results of our method and noticed that they look different from typical monoscopic images because of an apparent impression of depth, even if such depth is false. We also noticed that materials change their appearance when processed with our technique. Glossy objects appear shinier, giving them a more realistic appearance. Full understanding of appearance changes caused by dichoptic presentation would require further research.

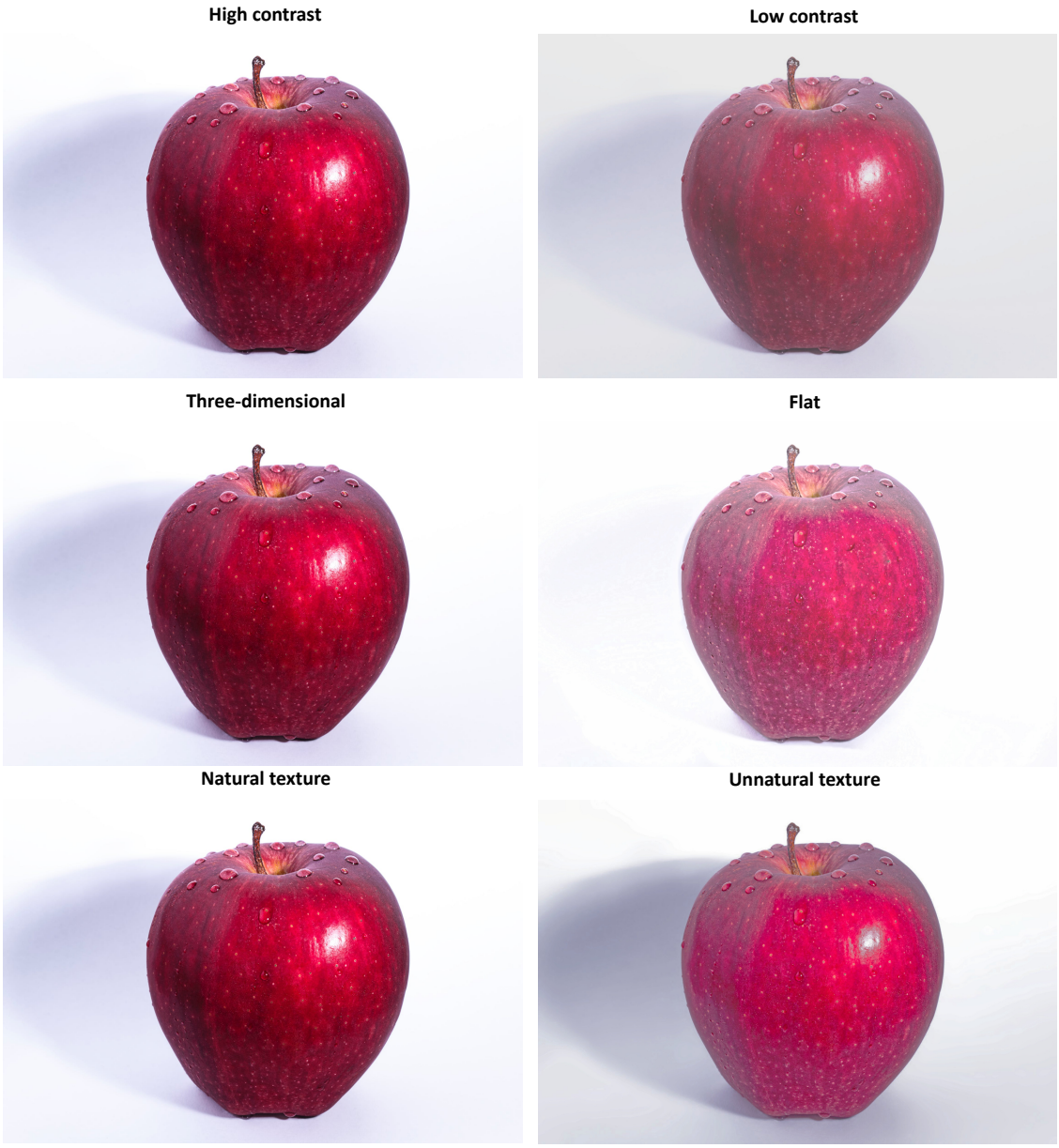


Figure 3.19: Training images for Experiment 3.1.7.2

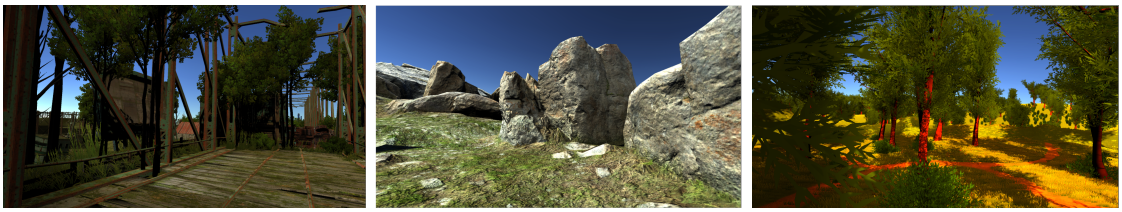


Figure 3.20: Three VR scenes in Experiment 3.1.7.2: Road, Rock, Woods (from left to right).

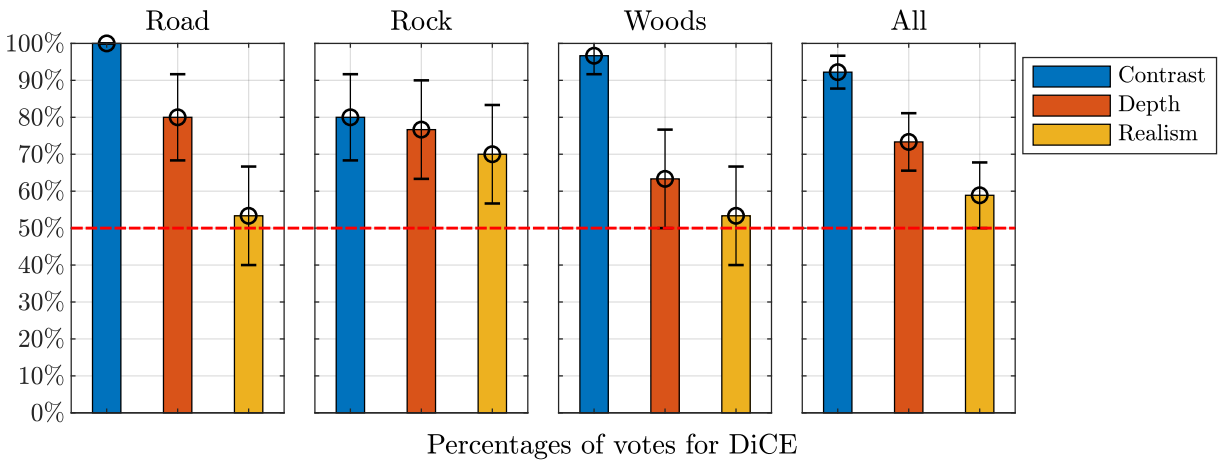


Figure 3.21: The percentages of votes for DiCE when compared to the standard presentation. The results are reported for each VR environment and for the aggregated results across all the environments. The error bars denote the confidence intervals.

Contrary to previous BTMO techniques [189, 200] that venture to make images look different to the eyes but may or may not result in contrast enhancement, our method always enhances contrast in a principled manner. BTMO methods require very expensive optimization; for the 800×600 image the authors report 22.24 seconds per single iteration for the 2012 technique and 2.36 seconds for the 2018 technique. As most rendering pipelines include a tone curve, they can be customised per eye using our interleaved tone curves at no additional cost. The simple lookup table implementation of DiCE tone curves causes zero drops in frame rates. Binocular rivalry can be controlled without the need for complex predictors, by simply changing a single parameter. We test our technique on both monoscopic and stereoscopic images, the latter being more relevant to the intended application.

One aspect of binocular fusion that our method does not directly address is the ocular dominance of the user. The visual system has a preference for one of the eyes, especially in the presence of strong rivalry. Since our method attempts to reduce rivalry, ocular dominance is less relevant than image contrast. Legge & Rubin [94] and Kingdom & Libenson [84] showed that the eye receiving the higher contrast image dictates whether it contributes more to the fused image and the effect of eye dominance is not clearly visible in their data.

The main limitation of our technique is the inherent trade-off between contrast enhancement and binocular rivalry. Stronger levels of enhancement result in more rivalry, which is perfectly acceptable for some observers (two in eight) but not for others. This was evidenced in the preference results of Experiment 3.1.7.1, where the answers were mixed even though average observers reported seeing higher contrast (Figure 3.18). Clearly, more

factors than perceived contrast contributed to the preference judgments. We suspect that the nature of the dichoptic enhancement requires some period of 'wearing-in', similar to getting used to a new pair of glasses. We did not observe any symptoms of VR sickness but such symptoms can be only revealed in a longer, purpose-designed experiment that has a control condition.

3.1.9 Summary

We propose a contrast enhancement technique for stereoscopic presentation, which is derived in a principled manner from a contrast fusion model. The main challenge of our approach is striking the right balance between contrast enhancement and visual discomfort caused by binocular rivalry. To address this challenge, we conducted a psychophysical experiment to test how content, observer, and tone curve parameters can influence binocular rivalry stemming from the dichoptic presentation. We found that the ratio of tone curve slopes can predict binocular rivalry letting us easily control the shape of the dichoptic tone curves. We validate the effectiveness of our technique in the evaluation study, in which we compare our technique with standard/dioptic presentation and previous techniques, for both monoscopic and stereoscopic images. We observed marked visual improvement in both perceived contrast and depth. In addition, glossy objects show increased shininess and are thus perceived as more realistic. The technique has a negligible computational cost (a lookup table) and only requires applying a separate tone curve for each eye. The single parameter of the curve generation may be needed to be adjusted per observer but it is content-independent so it does not require any analysis of the input images content, which is a costly operation in real-time rendering. As tone mapping is usually a part of the rendering pipeline, our technique can be easily combined with existing VR/AR rendering at no additional cost.

3.2 Improving depth perception under low luminance

Standard rendering

Proposed stereo constancy method

Figure 3.22: Stereoscopic image pairs that can be crossed-fused, demonstrating the stereo preservation under low luminance using Dark Stereo. The images should be viewed on a dimmed display (below $5\,\mathrm{cd/m^2}$).

As discussed in Chapter 2, depth perception from various depth cues, especially stereo cues, is a significant visual requirement for perceptual realism and a distinguishing feature that separates 3D displays from conventional 2D ones. However, binocular depth cues are

less reliable at low luminance because the global stereopsis mechanism requires a certain level of suprathreshold contrast to detect a binocular disparity signal [49], and the contrast detection thresholds are much higher at low luminance. On the other hand, it is often preferable to display VR content at low brightness as it brings a number of benefits in other aspects. For example, low luminance significantly reduces power consumption, as the display itself can be responsible for half of the power usage in a stand-alone, battery-powered headset. For perceptual realism, the most important benefit of keeping a low level of luminance is improved motion quality. As discussed in Section 2.1.3, low persistence is essential to reduce motion blur, but may cause flicker artefacts for a limited refresh rate. Lowering display luminance can reduce the visibility of such flickering without the need to increase the refresh rate [19].

In this section, we resolve the deterioration of depth perception under low luminance, while keeping the aforementioned other benefits by compensation for contrast to maintain a stereoscopic constancy and without manipulation of depth or disparity. In general, the undesirable effects introduced by a dimmed display include reduction of perceived contrast [6, 167], less colourful images [14, 151, 89], and undermined depth judgements based on stereoscopic depth cues. While the former two effects have been well studied and addressed in the literature, the effect of absolute luminance on depth judgements has received relatively less attention.

We measure, demonstrate, and quantify the effect of display luminance and contrast on depth judgements from binocular disparity cues, and propose an image contrast enhancement technique that can enhance depth perception on dimmed stereoscopic displays. We start this section with a review of the effects of display dimming (Section 3.2.1) and methods for depth enhancement (Section 3.2.2). Based on a series of psychophysical measurements on a prototype stereoscopic high-dynamic-range display (Chapter 4), we propose a model of *stereoscopic constancy* (Sections 3.2.3 and 3.2.4), which predicts the amount of physical contrast needed to maintain the same precision of binocular disparity depth cues across the luminance range of $0.1\,\mathrm{cd/m^2}$ to $1000\,\mathrm{cd/m^2}$. The model is then used to develop a multi-scale contrast compensation method (Section 3.2.5) that attempts to preserve the precision of binocular depth cues at different display luminance levels. The method has been implemented in GPU shaders and it can be used in real-time applications. Finally, we test our algorithm in a low-brightness VR rendering application, in which our method is both preferred and gives a better impression of depth than non-processed rendering and existing methods (Section 3.2.6).

The work presented in Section 3.2 produced the following publication:

- Krzysztof Wolski, Fangcheng Zhong, Karol Myszkowski, and Rafał K. Mantiuk. Dark stereo: Improving depth perception under low luminance. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH 2022, Journal Track)*, 41(4), jul 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530136. URL `https://doi.org/10.1145/3528223.3530136`

Author's note in collaborative work Section 3.2 contains collaborative work with other parties for the completeness of the presentation. The author contributed to the display calibration and data analysis for the 3D shape perception experiment (Section 3.2.3); the modelling of stereo consistency (Section 3.2.4); and partially the data analysis of the validation experiment (Section 3.2.6). Others were included for the completeness of the presentation.

3.2.1 Display dimming

In this subsection, we discuss the advantages and disadvantages of dimming a display and the studies on the functioning of the human visual system and colour appearance in dark and bright conditions.

Effect of display dimming on user experience Several works have studied the impact of display brightness on user experience and power consumption. Schuchhardt et al. [149] proposed an optimal dimming scheme to reduce mobile display brightness while ensuring good legibility on the screen. Erickson et al. [43] investigated the effect of colour mode on visual acuity and fatigue with VR head-mounted displays. They found that a dark background used in dark mode can reduce visual fatigue and increase visual acuity in a dim VR environment. Mantiuk et al. [116] argued for using amber and red colours on dark displays as they induce the least amount of disability glare or photophobia. They also found that the preferred display brightness was between 20 and $40\,\mathrm{cd/m^2}$ in a dark environment. Chapiro et al. [19] reported that, in low luminance conditions, judder is less visible, leading to better-perceived motion quality. These works provide solid ground for the merit of dimming VR displays. However, it is well recognised that the visual performance, including contrast and depth perception, is substantially degraded at low luminance. Although the visual system can preserve the appearance of contrast through a range of conditions [54], the contrast appears weaker and eventually disappears as the luminance is reduced, particularly the contrast that is close to the threshold [86, 139]. Lower luminance levels cause the pupil to dilate. This could result in a larger defocus blur in fixed-focus displays. Singh et al. [152] found that matching the brightness of the displayed and real object on an AR display results in more accurate depth estimation when focus cues are consistent (no vergence-accommodation conflict) or when focused on the mid-point of the tested depth range. No absolute luminance levels were reported so we cannot compare their finding with ours. According to Frisby et al. [49], the global stereopsis mechanism requires a certain level of suprathreshold contrast to detect a binocular disparity signal. Since the contrast detection thresholds are much higher at low luminance, our ability to see depth in low contrast content is greatly reduced [104]. Our work focuses on solving this issue.

Colour appearance on dimmed displays Despite the ability of our visual system to maintain colour perception across a very wide range of illumination (colour constancy), some changes in appearance are inevitable when light levels are low, in particular when the visual system transitions from cone-mediated vision (photopic) to cone- and rod-mediated

vision (mesopic) [6]. Indeed, colour appearance in mesopic vision (0.01 - 3 cd/m^2) can be influenced by the change in rod activity [156]. As a consequence, luminance levels alter the perception attributes such as hue, chroma, and lightness [89, 151]. Brightness and colourfulness also reduce with decreasing luminance [51]. Several models explaining the changes in colour appearance at mesopic light levels have been proposed [151, 14], including an extension of CIECAM02 colour appearance model [110].

Simulation and compensation of night vision Some works tried to simulate and compensate for changes between day and night visions. Wanat et al. [167] proposed a luminance re-targeting method to match the appearance of different luminance levels by altering perceived contrast and modelling hue and saturation shifts of an image. Kellnhofer et al. [79] argued that for stereoscopic displays, such changes are not sufficient to fully simulate dark conditions. Their proposed solution involves the manipulation of binocular disparity so that a scotopic stereo content displayed on a photopic monitor is perceived as the scene was scotopic. Contrary to the mentioned studies, instead of improving or simulating the appearance of a dark screen, our work focuses on improving stereo vision in low-luminance conditions.

3.2.2 Depth enhancement

In this subsection, we outline the works that manipulate image content to improve stereo vision.

Disparity manipulation Several works have proposed techniques for altering image disparity, mostly intending to reduce vergence-accommodation conflict and make images more comfortable to view. Oskam et al. [135] described a method that controls the camera convergence and interaxial separation over time to optimally map a dynamically changing scene to the desired depth range, which improves comfort. Lang et al. [90] proposed a method that controls and re-targets the depth of a stereoscopic scene in a nonlinear and locally adaptive fashion. The solution employs computed disparity and saliency estimates to compute a deformation of the input views so that they meet the desired disparities. To avoid undesirable distortions from disparity manipulation, Didyk et al. [33] introduced a perceptual model of disparity which provides a metric to evaluate perceived disparity change for stereo images. The follow-up work [35] studies the interplay of contrast and disparity on the depth discrimination. Based on their disparity-perception model, they jointly manipulate luminance contrast and disparity to reduce depth in stereoscopic images.

However, these models ignore the impact of display luminance, which is central to our work. Didyk et al. [34] also proposed a depth-enhancement technique that relies on the Cornsweet illusion in the disparity domain. While all these methods show the potential of manipulating disparity to improve the perception of depth in a displayed image, we argue that manipulating disparity in VR content might affect visual feedback to egomotion and contribute to an intensified VR sickness [76].

Depth enhancement by image content manipulation It has been shown that certain image manipulations can enhance the apparent depth. Luft et al. [109] proposed a technique that enhances contrast and colour near depth discontinuities to improve the perceptual quality of monoscopic images. Our intention is to preserve the perception of depth in stereoscopic content across different luminance levels rather than to enhance it.

3.2.3 3D shape perception

As discussed in Section 2.1.1, binocular disparity is one of the most important depth cues [28] and is commonly employed in stereoscopic displays to evoke stereo 3D scene appearance. In this section, we develop a computational model of the precision of binocular disparity cues as a function of image contrast and luminance. To this end, we designed an experiment in which observers were asked to judge the angle of a 3D hinge-like shape (Figure 3.23, left), reproduced on the display using only disparity depth cues (Figure 3.23, right). Our 3D shape perception experiment was inspired by the study of Watt et al. [172], where a similar hinge-like shape was used to examine whether focus cues have an indirect effect on depth interpretation.

Apparatus The experiment was conducted on a custom-built HDR stereo display, which shares the same architecture as the HDR multi-focal display apparatus presented in Chapter 4, except that we only used the near focal plane for this experiment. The apparatus allows a single observer to view a pair of HDR images (from $0.01\,\mathrm{cd/m^2}$ to $3000\,\mathrm{cd/m^2}$) through an optical arrangement similar to the Wheatstone mirror stereoscope, as illustrated in Figure 4.3. Further details on the display design, control software and its colourimetric and geometric calibration are explained in Chapter 4. We used HDR rather than a standard display as it can reproduce both very low and very high luminance while maintaining sufficient colour accuracy (bit-depth). The virtual images of the HDR content were placed 45 cm in front of the observers, with a resolution of 82 pixels per visual degree.

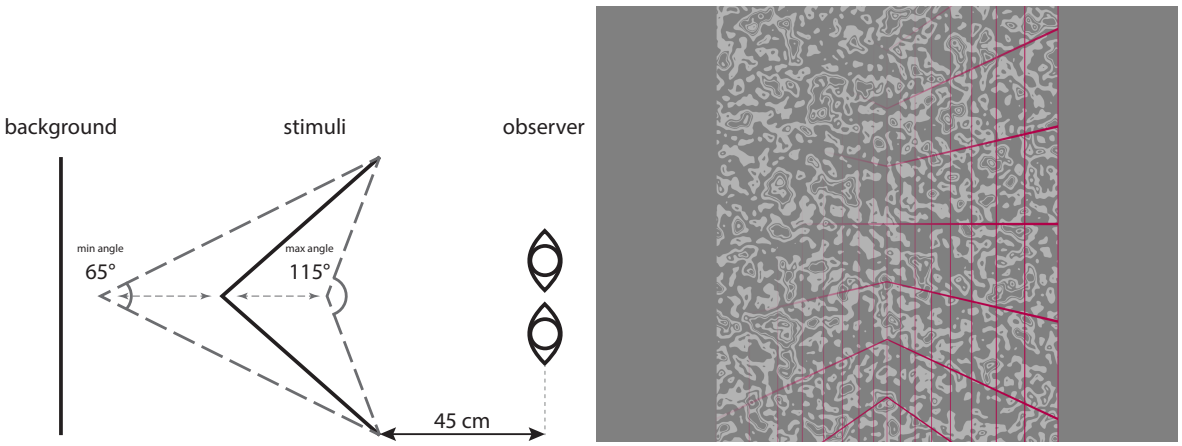


Figure 3.23: Left: The stimulus used in Experiment 3.2.3. The observer is presented with a hinge-like concave shape. The angle is changed by moving the hinge part towards or away from the observer (depicted by the arrows). Right: Procedural organic pattern on a uniform background. The superimposed grid depicts the three-dimensional shape of the stimuli. Note that the superimposed grid was only added to this figure to facilitate its 3D interpretation, while originally the hinge shape was reproduced only by the disparity cue.

Stimuli To study the influence of low luminance on binocular depth perception, we designed a stimulus that contained a controlled binocular disparity cue while minimizing the effect of other depth cues. The observers were presented with a concave, hinge-like shape on a uniform grey background. The stimuli were textured with a procedurally generated pattern (Figure 3.23, right) with only two shades of grey. To isolate only the disparity cue, the texture was projected on a surface from a position of a cyclopean eye, thus eliminating the perspective projection cue (the texture density did not change with the distance). It was also rendered without a reflection model to remove shading cues. During the experiment, observers were asked to use a chinrest to prevent head movements. The stimuli and setup are presented in Figure 3.23.

The stimuli were presented at five luminance levels: 0.1, 1, 10, 100, and 1 000 $\mathrm{cd/m^2}$. For each luminance level, four contrast levels, measured as Weber contrast (Equation 2.8), of the texture were measured: 0.05, 0.1, 0.2, or 0.4. As in the pilot experiment observers were not able to see the stimuli at 0.1 $\mathrm{cd/m^2}$ and Weber contrast of 0.05, we removed this condition from the main experiment. The order of conditions was randomised for each observer. If the luminance decreased between two conditions, we displayed a uniform field with the target luminance for a minute to ensure the observer was adapted to the new luminance level.

Experimental procedure The task was to assess whether the angle was greater or smaller than 90 degrees and confirm the decision by pressing a corresponding key on

the keyboard. For each of the 19 conditions, we used the method of constant stimuli to estimate the probability of judging the angle as acute or obtuse. The tested angles were: 65°, 75°, 85°, 95°, 105°, and 115°. Six trials were collected for each angle and each observer. We changed the angle by moving the hinge part towards and away from the observer while keeping the side edges stationary (Figure 3.23). This was done to avoid additional depth cues. The tested angles were randomised between the trials.

Each observer was asked to complete a training session at $10\,\mathrm{cd/m^2}$, in which they were given feedback on whether their answer was correct. Such feedback was not given in the main experiment. The training session helped the observers to familiarise themselves with the task and become accustomed to disparity-only stimuli. The session was also used to screen observers. We excluded three observers who were unable to properly complete the training session[2] from further experiments. The entire experiment took each participant around two hours and was split into 3–5 short sessions.

Observers Eleven volunteers (four females and seven males, mean age 31, SD 4.5 years, including three authors[3]), who passed the training session, participated in the experiment. Nine of them completed trials for all luminance levels, while two completed only the trials for luminance levels from $0.1\,\mathrm{cd/m^2}$ to $10\,\mathrm{cd/m^2}$. All observers had normal or corrected-to-normal visual acuity. All passed the Titmus stereoacuity test. All observers except the authors were naive to the purpose of the experiment. Before the experiment, each observer read and signed the consent form. The observers were rewarded for their participation. The experiment was approved by the departmental ethics board.

Results The experiment explained how well the observers could see a geometric angle at several luminance and contrast levels. The data averaged over all observers, plotted as the probability that an observer reports an obtuse angle, is shown in Figure 3.24 as red stars.

The first important observation is that the psychometric curves formed by the data points cross the 50% probability point at around a 90-degree angle regardless of luminance and contrast. This means that the perceived angles were not distorted by lower luminance and contrast. However, the slopes of the psychometric functions differ substantially between the conditions. The shallower slopes indicate that the observers more often mistook the angle at low luminance and low contrast. This means that lower luminance does not reduce the accuracy of the shape assessment task, but it reduces the precision of that task.

[2]giving purely random results as they cannot properly perceive stereo cues.

[3]The authors participated in this experiment as they did not have a preferred outcome, but rather aimed to accurately measure the precision of stereo cues.

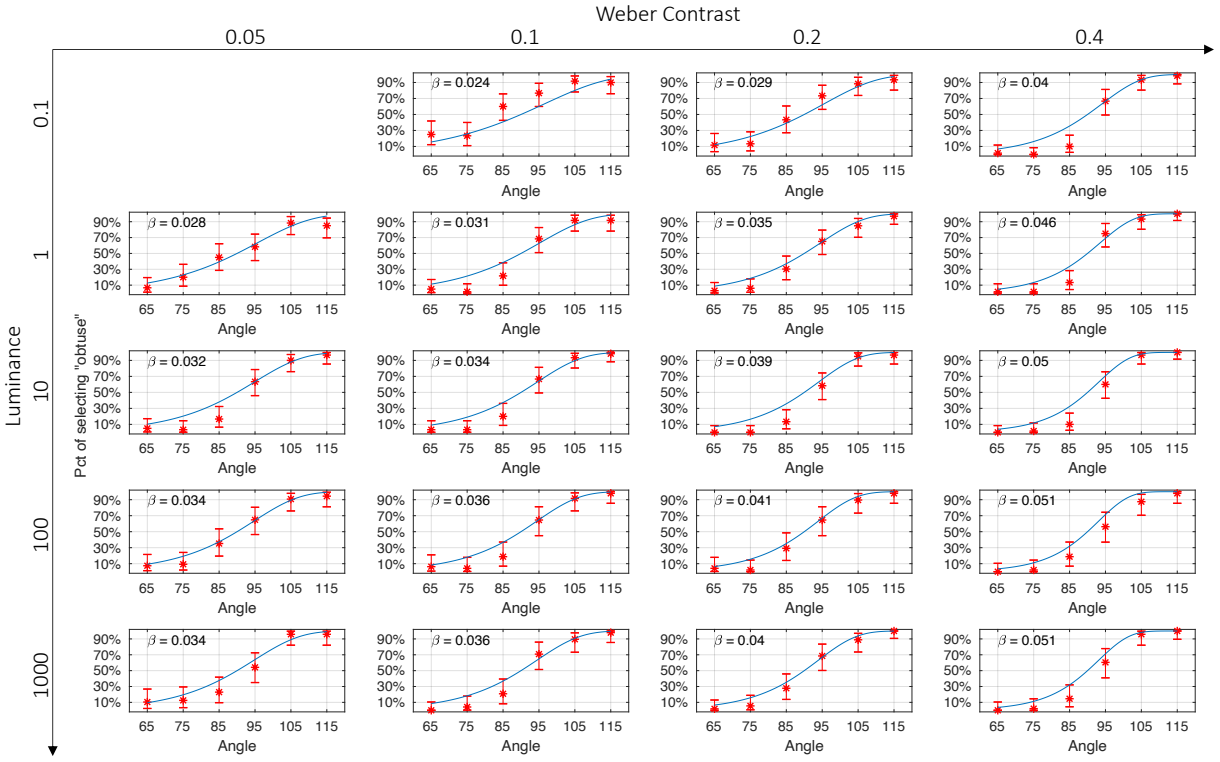


Figure 3.24: The red stars are the original data points collected from the 3D shape perception experiment (Section 3.2.3). They represent the frequency at which the participants assessed the angle as obtuse under various luminance and contrast conditions. The error bars denote the 99% confidence intervals. The blue curves represent our fitted psychometric model (Section 3.2.4). The top-left plot has no data points as it was impossible to see the stimuli in this condition.

3.2.4 Stereo constancy model

In this section, we propose a model that can predict how contrast needs to be altered to preserve the same precision of the stereo task across different luminance levels. As the first step, we assume the collected data can be explained by a psychometric function that follows the Weibull cumulative distribution function [177]. We used this psychometric function to describe the probability p of an observer perceiving the hinge-like shape (Figure 3.23) as an obtuse angle given the actual angle α, represented in degrees:

$$p(\alpha, \beta) = 1 - \exp\left(\log(0.5)10^{\beta(\alpha - \alpha_{thr})}\right) \tag{3.8}$$

where α_{thr} is the angle at which the probability of detection p is 0.5, which we assumed to be 90 degrees. The value of β controls the steepness of the function, which reflects the precision of the user performance in this task. A higher value of β means that the observer is more sensitive to the variations in the perceived angle and also that the task is easier.

Our goal is to find a model of β as a function of contrast and luminance, such that the

likelihood of the data observed in the 3D shape perception experiment is maximised. We found that β can be explained by a quadratic function of contrast and log-luminance:

$$\beta(c, L; \mathbf{w}) = w_1 L + w_2 c + w_3 L^2 + w_4 c^2 + w_5 \tag{3.9}$$

where L is the logarithm of luminance ($L = \log_{10}(Y)$), c is logarithmic contrast, and $\mathbf{w} = [w_1, ..., w_5]$ denote unknown free parameters. Note that contrast was recorded as Weber contrast C_w (Equation 2.8) in our experiment. However, our contrast enhancement method (Section 3.2.5) can be implemented more efficiently if it operates on logarithmic contrast. The Weber contrast C_w can be converted into logarithmic contrast c with the formula:

$$c = \log_{10}(C_w + 1). \tag{3.10}$$

To have better control over free parameters, and to ensure that the function is monotonic, we used the maximum a posteriori (MAP) estimation to find the values of $\mathbf{w}$. We assume $\mathbf{w} \sim N(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma})^2)$ for some $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, where μ_i and σ_i^2 are the mean and the variance of w_i respectively.

The likelihood of observing k out of n trials (of selecting an obtuse angle) can be explained by a binomial distribution, with a latent probability of p of perceiving the angle as obtuse. As indicated by Equation 3.8, the value of p is dependent on the presented angle α and detection sensitivity β, which is then parameterised by contrast c, luminance L, and $\mathbf{w}$ in Equation 3.9. Under the MAP framework, free parameters $\mathbf{w}$ can be found by minimizing the negated log-likelihood of the binomial distribution:

$$\arg\min_{\mathbf{w}} - \sum_{s} \sum_{\mathbf{d}} \log \left(\binom{n_{s,\mathbf{d}}}{k_{s,\mathbf{d}}} p_{\mathbf{d}}^{k_{s,\mathbf{d}}} \left(1 - p_{\mathbf{d}}^{(n_{s,\mathbf{d}} - k_{s,\mathbf{d}})}\right)\right) + \sum_{i=3,4} \frac{1}{2\sigma_i^2}(w_i - \mu_i)^2 \tag{3.11}$$

where s is the index of the observer, $\mathbf{d} = [\alpha, c, L]$ are the parameters of each condition, and $p_{\mathbf{d}} = p(\alpha, \beta(c, L; \mathbf{w}))$ is given by Equations 3.8 and 3.9. $n_{s,\mathbf{d}}$ in the binomial coefficient is the total number of measurements collected for observer s and condition $\mathbf{d}$ and $k_{s,\mathbf{d}}$ is the number of measurements in which obtuse angle was selected. Table 3.2 shows the final estimated parameters we found for $\mathbf{w}$, and our choices for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Under the MAP framework, $\frac{1}{2\sigma_i^2}$ becomes the weights on the regularization terms. Note that we only regularize 2^{nd}-order terms w_3 and w_4 to ensure monotonicity. With these parameters, we plot the corresponding fitted psychometric functions parameterised by β under various luminance and contrast conditions on top of the original data points in Figure 3.24 as blue curves. The plots demonstrate that the model explains well most conditions. The worse fit for some conditions (e.g. $0.1\,\text{cd/m}^2$, $C_W = 0.4$) is due to the regularization, which was necessary to make the model monotonic and thus invertible.

Table 3.2: Estimated values of free parameters of Equation 3.9 and the priors for the Maximum a Posteriori (MAP) estimation. Symbol / indicates that no prior was used.

	1	2	3	4	5
w	0.0050	0.1849	-0.0010	0.3994	0.0263
μ	/	/	-0.4	0.4	/
σ^2	/	/	0.001	0.001	/

Next, we use the fitted model to find the lines of equal precision of the task (constant β). Such lines are plotted as continuous lines in Figure 3.25. The figure shows that to maintain the same precision of the task (the same β), we need to increase the contrast at low luminance and that such an increase should be smaller for higher contrast. We will refer to this model as a stereo constancy model and use it in the next section to derive our contrast enhancement technique for dark stereo displays. To demonstrate that the effect cannot be predicted by a contrast constancy model, we plot in the same figure the contrast constancy model of Kulikowski [86] (used in [167]). The comparison shows that stereo constancy requires stronger contrast enhancement between 0.1 and 10 $\mathrm{cd/m^2}$ (mesopic and photopic range) than contrast constancy. The difference between both models is further corroborated in our validation experiments in Section 3.2.6.

3.2.5 Stereo-preserving contrast enhancement method

We use our model to design a local contrast enhancement method that preserves the precision and difficulty of stereo perception at low luminance. We follow a similar contrast retargeting algorithm as used by Wanat et al. [167] to manipulate the local contrast according to the stereo constancy model. We also improve a few processing steps for better real-time performance and temporal stability. The following subsections explain each step of the algorithm.

3.2.5.1 Colour space transformation

Since our stereo constancy model is defined in terms of physical (linear) luminance units, we first convert the rendered frame from a gamma-encoded to a linear colour space. Assuming ITU-R BT.709-6 RGB primaries and the standard gamma ($\gamma = 2.2$), the relative

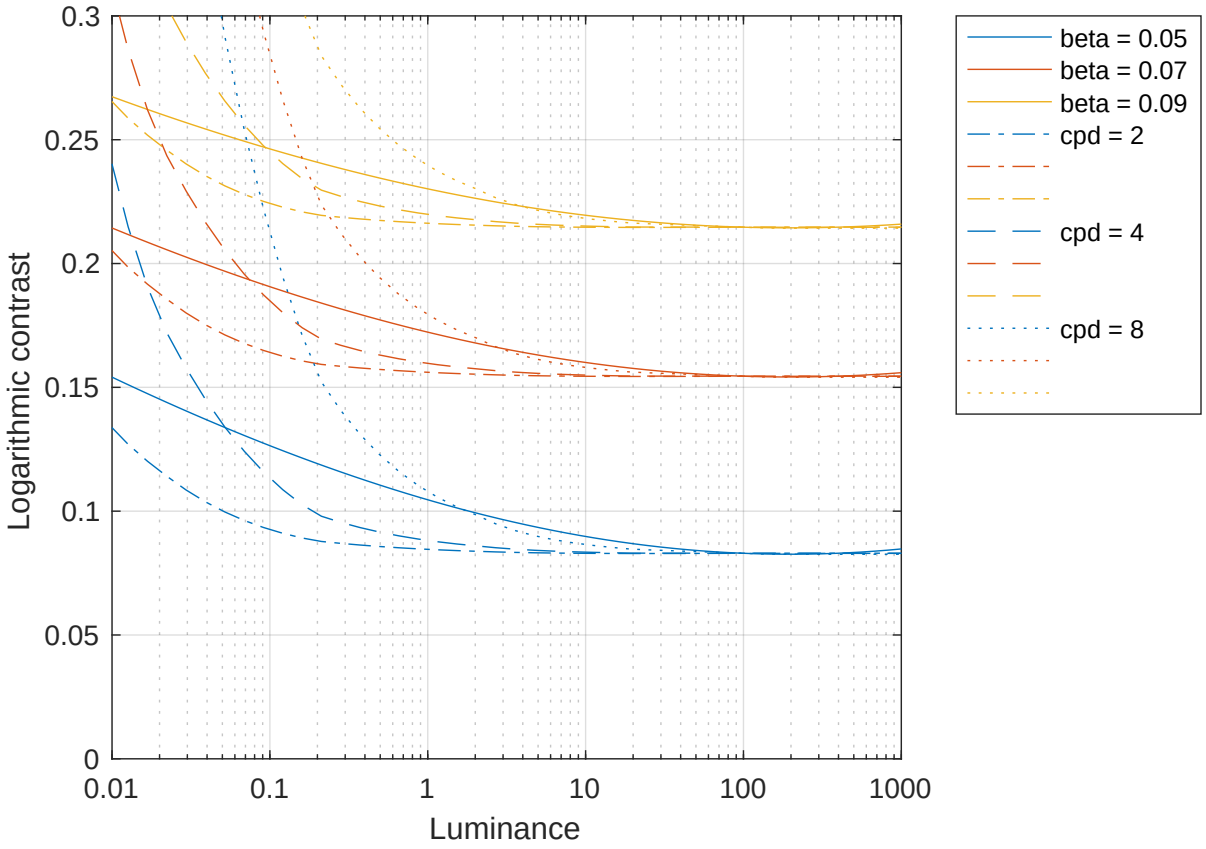


Figure 3.25: The solid lines, or equivalent-β lines, connect the contrast values that result in the same precision of perceiving depth (the same β of the psychophysical function) as a function of different display luminance levels. The lines are derived from our model of stereo task difficulty. The dashed lines represent the equivalent perceived contrast for three different spatial frequencies (2, 4, and 8 cpd) according to Kulikowski's model.

luminance, y_{input}, is computed as:

$$y_{\text{input}}(\mathbf{x}) = \sum_{k=1}^{3} v_k \, I'^{\gamma}_{input}(\mathbf{x}, k)\,, \tag{3.12}$$

where $I'(\mathbf{x}, k)$ is the gamma-encoded input value at pixel $\mathbf{x}$ and in colour channel k (in the range 0–1), while $v_k = [0.212656, 0.715158, 0.072186]$. Note that we use lower-case y for relative luminance to make it distinct from absolute luminance, Y.

3.2.5.2 Multi-scale decomposition

The proposed method compensates for the deteriorated depth perception by enhancing local image contrast. In order to operate on local image contrast, we decompose an image into frequency bands using the Laplacian pyramid. We use the classical Burt and Adelson method [13] with the coefficient $a = 0.4$ used to construct the filters. In our

implementation, we construct a Laplacian pyramid consisting of 3 levels: two band-pass levels and one low-pass level (baseband). The two band-pass levels are sufficient because of the limited effective resolution of VR headsets (in terms of pixels per degree). Such a shallow decomposition also improves the performance in real-time applications.

For computational convenience, the decomposition is performed on logarithmic values of luminance $l = \log_{10}(y_{\text{input}})$. This ensures that the coefficient of the pyramid represents logarithmic contrast (they approximate the logarithm of ratios between two levels). The Laplacian pyramid coefficient at level i is then computed as:

$$P_i(\mathbf{x}) = (g_i * l)(\mathbf{x}) - (g_{i+1} * l)(\mathbf{x}) \,, \tag{3.13}$$

where g_i is the kernel of a Gaussian pyramid at the level i and $*$ is the convolution operator.

3.2.5.3 Measure of local contrast

Our stereo constancy model requires an estimate of the local contrast to find a corresponding equivalent contrast in the target image. Although the coefficients of the Laplacian pyramid can be used for this purpose, it can result in over-enhancement and artefacts at sharp contrast edges [167]. We follow the same approach as in [167] and compute a root-mean-squared (RMS) measure of local contrast.

The localised root-mean-square (RMS) contrast can be computed as:

$$c_i(\mathbf{x}) = \sqrt{(g_\sigma * l^2)(\mathbf{x}) - ((g_\sigma * l)(\mathbf{x}))^2} \,, \tag{3.14}$$

where l is the logarithm of relative luminance and g is a Gaussian kernel with standard deviation σ. We use the kernels with larger σ at the lower frequency pyramid levels. To avoid computing additional convolutions, we can instead reuse the Gaussian pyramid and estimate the local RMS contrast as:

$$c_i(\mathbf{x}) = \sqrt{H_i(\mathbf{x}) - G_i^2(\mathbf{x})} \,, \tag{3.15}$$

where G_i is a Gaussian pyramid built from log-luminance l and H_i is a Gaussian pyramid built from squared log-luminance l^2. Computing a second pyramid H is inexpensive on a GPU, as it can be done by operating on a 2-channel texture, where the first channel contains log-luminance and the second channel contains squared log-luminance.

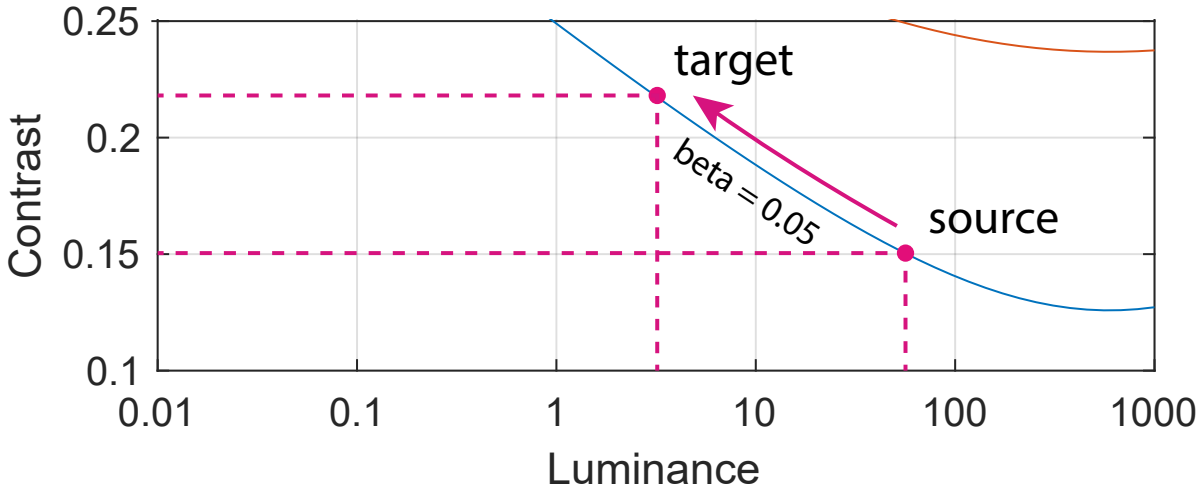


Figure 3.26: Method of finding equivalent contrast that preserves the precision of binocular disparity cues. Similar to Figure 3.25, for a given input contrast and source luminance, our stereo constancy model gives the curves of equivalent contrast (constant β, blue line). This lets us find the desired contrast for any target display luminance.

3.2.5.4 Contrast retargeting

Once the Laplacian pyramid and contrast magnitude are computed, we can map the contrast for a given source luminance (Y_{in}) to the contrast that provides the same stereoacuity when seen at target luminance (Y_{out}). This can be done by executing the following steps for every frequency band except the low-pass band (baseband), which does not encode contrast.

Finding contrast enhancement factor We need to find an equivalent contrast at another (target) luminance level, which results in the same stereo precision (β) as the original contrast. We can rearrange Equation 3.9 to compute the equivalent contrast c_{eq} for the desired logarithmic contrast c, source (Y_{in}) and target (Y_{out}) luminance:

$$c_{eq}(c, Y_{\text{in}}, Y_{\text{out}}) = \frac{-w_2 + \sqrt{w_2^2 - 4w_4 t}}{2w_4}, \tag{3.16}$$

where

$$\begin{aligned} t &= w_1 L_{\text{out}} + w_3 L_{\text{out}}^2 + w_5 - \beta(c, L_{\text{in}}) \\ L_{\text{in}} &= \log_{10} Y_{\text{in}} \qquad L_{\text{out}} = \log_{10} Y_{\text{out}} \end{aligned} \tag{3.17}$$

with the parameters $w_1, ..., w_5$ reported in Table 3.2. Function $\beta(\cdot)$ is given in Equation 3.9. The process of mapping contrast between luminance levels is further illustrated in Figure 3.26.

Instead of directly modifying contrast in the Laplacian pyramid, we compute a contrast enhancement factor:

$$m_i(\mathbf{x}) = \frac{c_{\text{eq}}\Big(c_i(\mathbf{x}), Y_{\text{in}}(\mathbf{x}), Y_{\text{out}}(\mathbf{x})\Big)}{c_i(\mathbf{x})}, \tag{3.18}$$

where c_i is an input RMS contrast computed according to Equation 3.15 and $c_{\text{eq}}()$ is the equivalent contrast function from Equation 3.16. Y_{in} and Y_{out} are source and target luminance which are computed as:

$$\begin{aligned} Y_{\text{in}}(\mathbf{x}) &= 10^{G_N(\mathbf{x})} \cdot Y_{\text{peak,src}}\,, \\ Y_{\text{out}}(\mathbf{x}) &= 10^{G_N(\mathbf{x})} \cdot Y_{\text{peak,trg}}\,, \end{aligned} \tag{3.19}$$

where G_N is the base band of the Gaussian pyramid (as explained in Section 3.2.5.2). $Y_{\text{peak,src}}$ is the peak luminance of the source display (before dimming) and $Y_{\text{peak,trg}}$ is the peak luminance of the target (dimmed) display. We use $Y_{\text{peak,src}} = 80\,\text{cd/m}^2$ in all our experiments.

Contrast enhancement Given the local contrast estimate computed in Section 3.2.5.2, we retarget it, enhancing locally the Laplacian pyramid:

$$\tilde{P}_i(\mathbf{x}) = P_i(\mathbf{x}) \cdot m_i(\mathbf{x}), \tag{3.20}$$

where P_i is the i-th level of Laplacian pyramid ($i = 1, ..., N-1$, excluding the base-band) and m_i is a corresponding enhancement factor from Equation 3.18.

We reconstruct the resulting enhanced luminance channel y_{enh} by summing all N levels of pyramid $\tilde{P}$ including the base band:

$$y_{\text{enh}}(\mathbf{x}) = 10^{\sum_{i=1}^{N} \tilde{P}_i(\mathbf{x})}. \tag{3.21}$$

3.2.5.5 Reconstructing colour image

The enhanced colour image I_{enh} is produced by multiplying input colour (RGB) image in linear space I_{input} by the ratio of enhanced and input luminances:

$$I_{\text{enh}}(\mathbf{x}, k) = I_{\text{input}}(\mathbf{x}, k)\, \frac{y_{\text{enh}}(\mathbf{x})}{y_{\text{input}}(\mathbf{x})}\,, \tag{3.22}$$

where k is the index of the colour channel ($k \in \{1, 2, 3\}$). Such an approach may, however, result in out-of-gamut colours (one of the colour channels values greater than 1) and

distorted or desaturated colours. To prevent this, we compute how much a particular pixel can be enhanced until the pixel exceeds the gamut:

$$m_{\max}(\mathbf{x}) = \frac{1}{\max_k \left\{ I_{input}(\mathbf{x}, k) \right\}} . \tag{3.23}$$

Next, we introduce this term into the previous grey to colour conversion equation:

$$I_{\text{enh}}(\mathbf{x}, k) = I_{\text{input}}(\mathbf{x}, k) \cdot \min \left\{ \frac{y_{\text{enh}}(\mathbf{x})}{y_{\text{input}}(\mathbf{x})}, m_{\max}(\mathbf{x}) \right\} . \tag{3.24}$$

As the last step, we convert the linear colour channels to display-ready gamma-encoded ones with a gamma function: $I'(\mathbf{x}, k) = I^{1/\gamma}(\mathbf{x}, k)$.

3.2.6 Validation

We evaluated the effectiveness of our method in a validation experiment in which we compared the proposed enhancement algorithm with the most closely related method of Wanat et al. [167] and standard rendering. The methods were compared in terms of the impression of three-dimensionality and the appearance of the presented scene.

VR headset The experiments were prepared for the Valve Index VR headset. We chose Valve Index because it offers a relatively high display resolution of a maximum of 16 pixels per visual degree and its drivers allow the user to dim its display. We set the peak luminance to be $5\,\text{cd/m}^2$.

Stimuli The test scene was built from stylised assets that provided a good balance between good quality content, similar to those found in most VR experiences, and performance (no complex geometry). An example screenshot from the scene is shown in Figure 3.27. Figure 3.28 shows three rendering modes used in the experiment: the proposed stereo-constancy model (top), standard rendering (middle), and Wanat's method (bottom).

Procedure During the experiment, we placed the observers in the virtual environment and teleported them to five different locations. They were allowed to look around freely and switch between two rendering methods using the trackpad on the right controller. In each trial, they compared our method with either standard rendering (no post-processing)

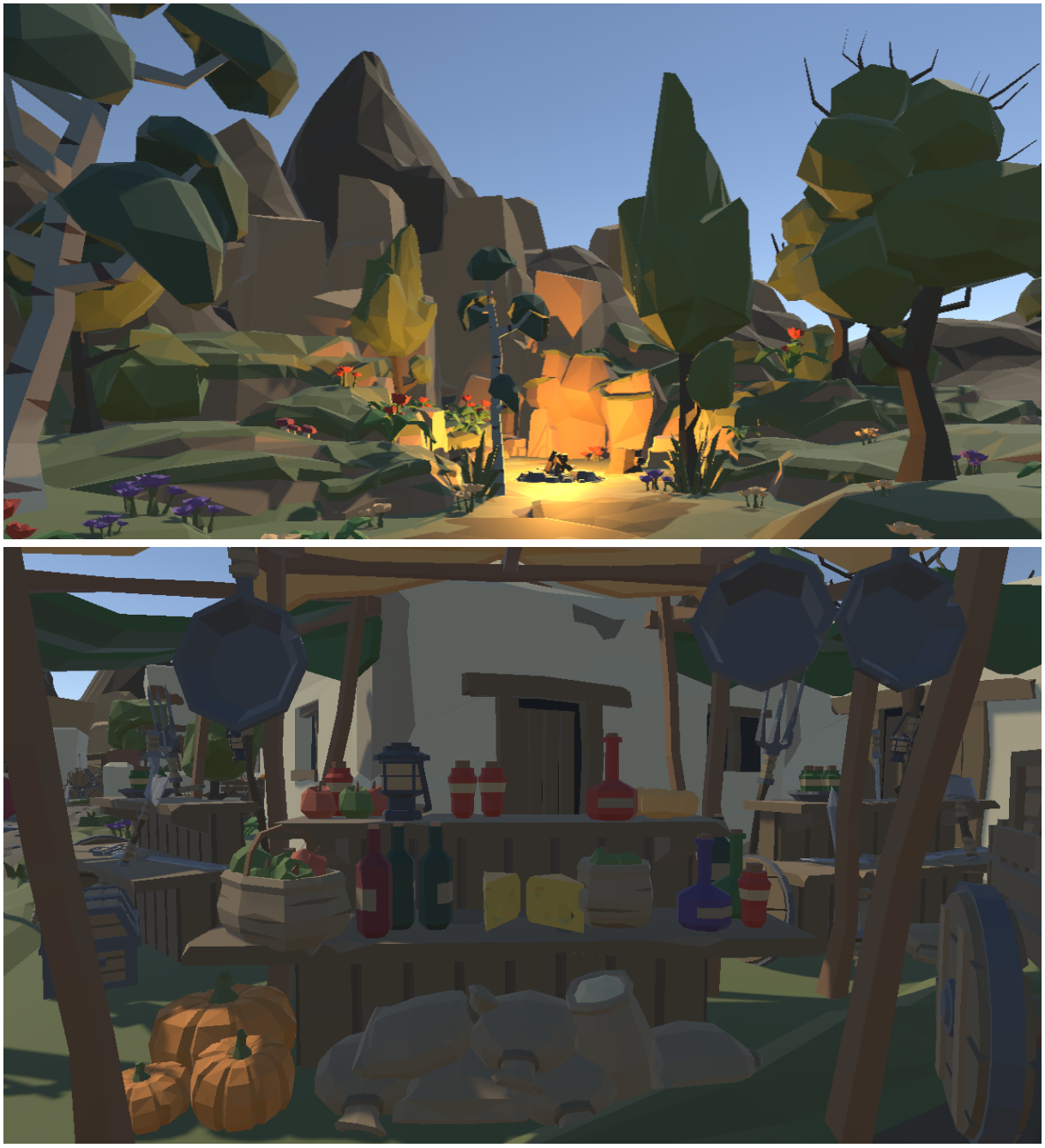


Figure 3.27: Preview of the scene presented to the observer in the preference experiment. The assets are a part of the POLYGON series prepared by Synty Store. The images show a non-enhanced (standard) rendering of the scene.

or the method of Wanat et al. [167].

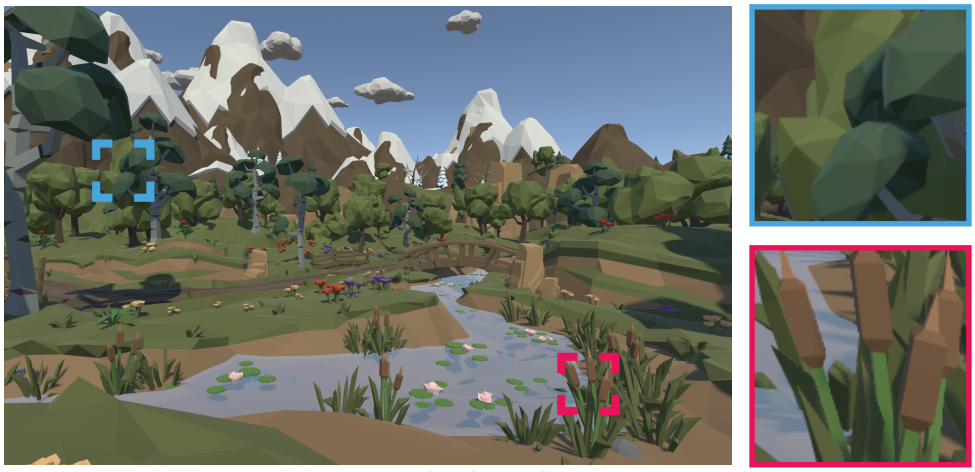
The experiment was split into two parts. In each part, the observers visited the same locations (in random order), regardless of the question being asked. In the first part, the observers were asked to select the rendering mode that looks *more three dimensional* and in the second part they were asked to select the rendering mode that *looks better*. The observers gave the answer by pressing the right trigger while the selected rendering method was active. The participants were allowed to make a selection only after viewing both rendering modes. For both questions and every condition, each location was shown to the participant five times, each time using a different direction of the camera (random rotation around the up vector). The order of trials and parts was randomised. We also displayed information about the current progress of the experiment and the assessment criterion (depth or preference) at the bottom of the viewport.

Observers Nine observers (one female and eight males, mean age 26.7, SD 3.2 years) were recruited among students and researchers. All observers had a normal or corrected-to-normal vision and were also naïve to the purpose of the experiment. Before the experiment, each participant read and signed the consent form. The participants were screened for stereoacuity in a test performed in VR, in which they had to choose a closer square from a pair (akin to the Titmus fly test).

Results The bars in Figure 3.29 show the percentage of trials in which our method was chosen over the alternative method. The yellow circles indicate per-observer results. Eight out of nine observers agreed that the image modified with our method looks more three-dimensional and also better than the image enhanced with Wanat et al.'s method and standard rendering. We further validated these results with a one-sided binomial test and a null hypothesis of random selection for both preference and impression of three-dimensionality. The tests confirmed that the results were significant and our contrast enhancement for stereo-constancy improves the perception of 3D shapes and produces more preferred images. In a post-experimental survey, we asked the observers whether the colour seen in the VR headset appeared natural. None of the observers reported an artefact in colour appearance.

3.2.7 Discussion

No distortion of depth The most important conclusion from our 3D shape perception experiment (Section 3.2.3) is that low luminance levels (0.1–10 $\mathrm{cd/m^2}$) do not distort depth.



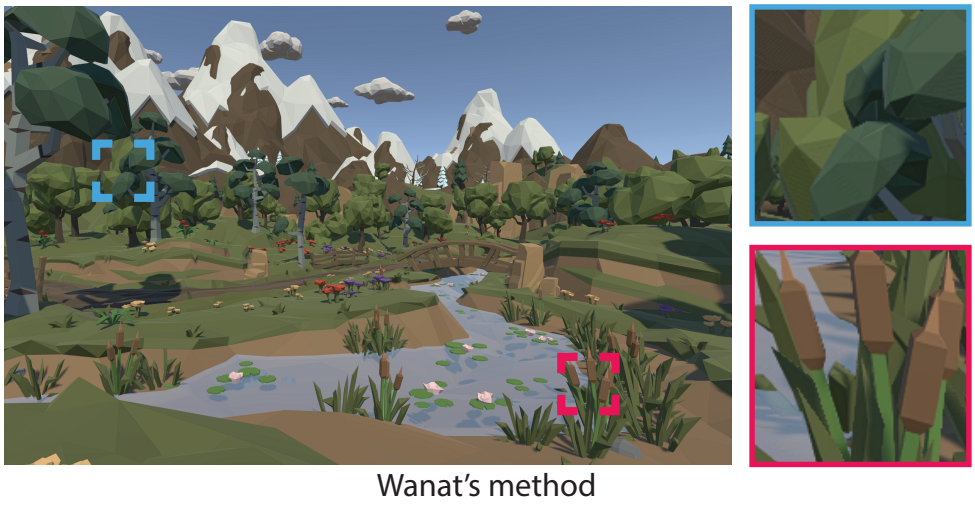


Figure 3.28: Three rendering methods used in the preference experiment: image enhanced with the proposed stereo-constancy model (top), standard rendering with no enhancement (middle), and image enhanced with Wanat's method (bottom). The insets show close-ups of the selected image areas. It can be observed that Wanat's and the proposed stereo constancy methods increase local contrast and result in a sharper image, but for different purposes.

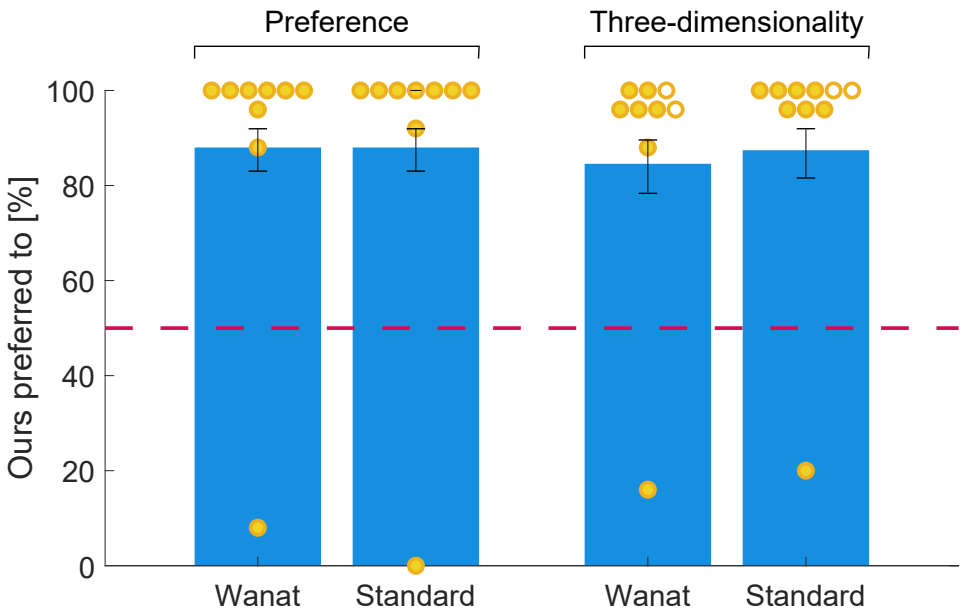


Figure 3.29: Results of Experiment 3.2.6, assessing the preference and the impression of three-dimensionality. The yellow circles represent the per-observer results and the empty circles denote observers who failed the stereoacuity test. The height of the bars presents the percentage of trials in which our method was chosen over the method given below (excluding disqualified observers). *Standard* stands for a no enhancement and *Wanat* for the method proposed by Wanat et al. The error bars present a 95% confidence interval and the red dashed horizontal line indicates the guess rate.

This is in contrast to the observations of Kellnhofer et al. [79], who reported compression of depth at low luminance. Our experiment showed that, even at $0.1\,\mathrm{cd/m^2}$, observers could correctly assess the angle without bias (high accuracy), however with larger variance in their responses (lower precision). Had the perceived depth been compressed at low luminance, the results would have been biased towards obtuse angles, as the observers compensated for the reduced disparity. However, we did not experiment with light levels below $0.1\,\mathrm{cd/m^2}$, so we cannot confirm whether the perception of 3D shapes is affected at these luminance levels.

Contrast vs. disparity manipulation Several works [33, 35, 80] manipulate disparity to improve depth perception. While such an approach is practical for 3D cinema content, it is unsuitable for VR environments, in which depth must be faithfully reproduced to give accurate visual feedback to egomotion. Disparity manipulation in VR is likely to result in conflicting visual and vestibular sensations leading to VR-sickness [76].

Colour appearance in mesopic vision Degradation in stereoacuity is not the only issue of showing VR content at low brightness. It is well-established that colour appearance

also degrades as luminance decreases to mesopic vision [6, 51]. Those models could be incorporated into our enhancement technique, as was done in [167], however, we did not find the changes in colour to be substantial enough to require additional processing. None of the experiment participants reported an unnatural colour appearance in our post-experiment questionnaire (Section 3.2.6). Our observation is also supported by the results of Kwak et al. [89, Fig. 6,7] who reported negligible changes in colourfulness and hue, measured using magnitude estimation, when the reference white was reduced to only $1\,\mathrm{cd/m^2}$. Larger changes in colour appearance can be observed when two luminance conditions are presented to the observer simultaneously in an asymmetric (haploscopic) matching experiment [151]. Such artificial presentation, however, is not representative of viewing content on a dimmed VR headset.

Limitations Our model was fitted to the data collected in the luminance range between $0.1\,\mathrm{cd/m^2}$ and $1\,000\,\mathrm{cd/m^2}$, which may limit the ability of our model to generalize to very low luminance levels. We cannot generalize our model to the scotopic levels much below $0.1\,\mathrm{cd/m^2}$, but we argue that such low luminance is less relevant for displays. We also do not consider the influence of tone mapping on depth perception. Since tone mapping often involves contrast compression, we expect increased difficulty in inferring depth from tone-mapped stereo images.

3.2.8 Summary

Dimming a display can be beneficial for VR experience as it reduces the visibility of flicker, saves power, prolongs battery life, and reduces the cost of the device. The major downside of this approach is the reduced sensitivity to stereoscopic depth cues, which are major visual cues for perceptually realistic graphics that differs from photorealistic graphics. Contrary to previous works [79], we do not find the distortion of 3D depth at low luminance (0.1-$1\ \mathrm{cd/m^2}$), but instead, we find increased difficulty and lower precision (larger variance) of assessing 3D shapes based on binocular cues. This motivates our method for enhancing contrast at low luminance levels, intended at improving the reliability of stereoscopic depth cues. We demonstrate that such contrast enhancement can be implemented in the real-time rendering of VR environments. We further show the effectiveness of such depth enhancement in a perceptual experiment asking about qualitative aspects of preference and impression of depth. The experiment demonstrates that depth perception can be effectively restored by contrast enhancement and overall visual quality can be improved. The proposed method can improve the user experience for VR headsets that need to operate at low power or those that cannot achieve high refresh rates.

Chapter 4

Reproducing Reality

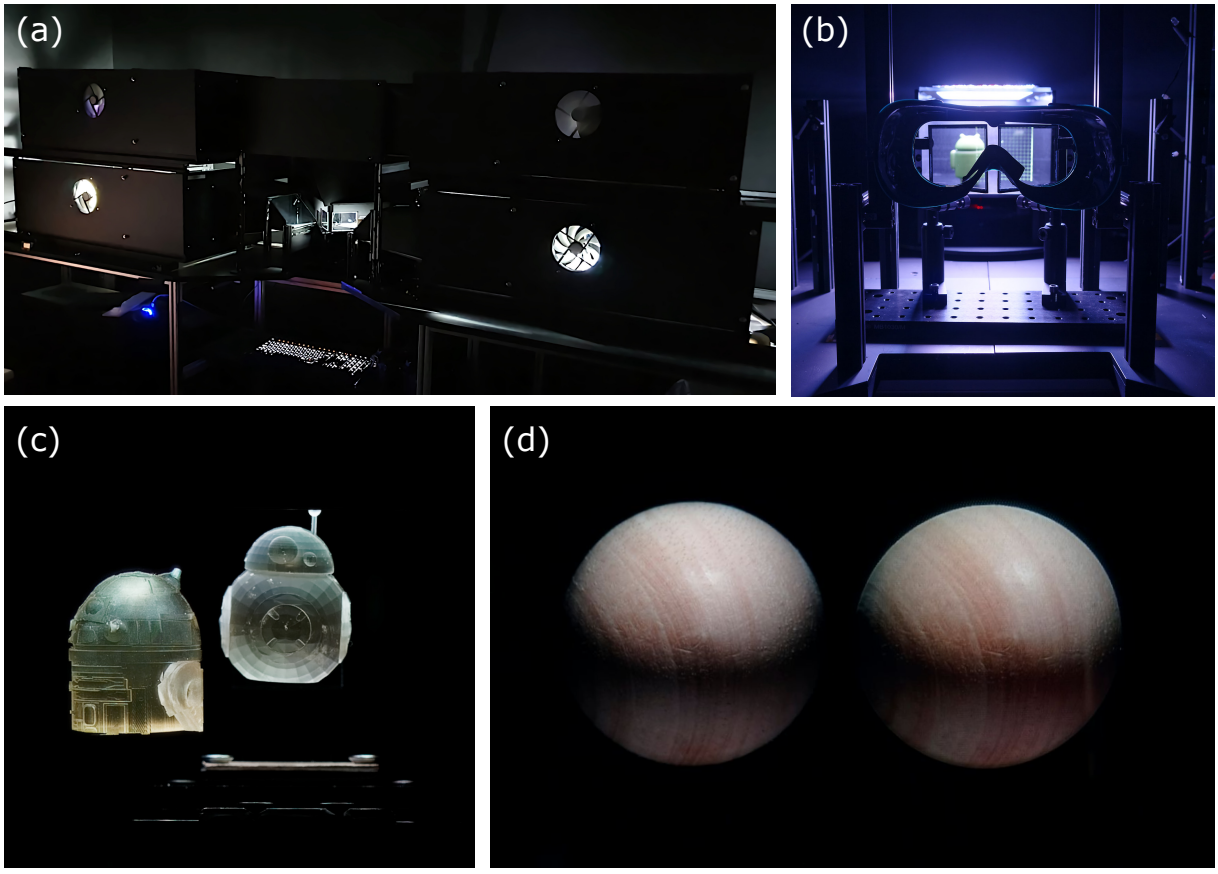


Figure 4.1: We built a High-Dynamic-Range Multi-Focal Stereo display (a) which allows for a direct comparison with a physical scene located in front of the observer (b). The display can reproduce real-world 3D objects with accurate colour, contrast, disparity, and a range of focal depth, making it hard to distinguish between real and virtual scenes (c, d).

Imagine a black box that contains either a physical 3D object or one virtually rendered by a 3D display, with a naive observer tasked to distinguish between the two scenarios. This is the notion of a *visual Turing test* [5] - an extension of the Turing test to the visual

domain to evaluate perceptual realism. Passing a visual Turing test for arbitrarily complex scenes is the holy grail of perceptually realistic graphics. Among all the possible quality metrics, the visual Turing test is also one of the most efficacious and indicative methods for evaluating perceptual realism, as physically measuring perceived quality can be nontrivial.

In Chapter 3, we presented two rendering algorithms for stereoscopic displays to boost the perceived quality of contrast and depth. Nonetheless, only improving rendering for contrast and depth is not sufficient to meet all the visual requirements for perceptual realism and pass a visual Turing test. The overall fidelity of a typical stereoscopic VR display is confined by limited dynamic range, low spatial resolution, lens distortions, and vergence-accommodation conflicts. To push the limits of overall fidelity and challenge the visual Turing test, we present a *High-Dynamic-Range Multi-Focal Stereo display (HDR-MF-S display)* with an end-to-end imaging and rendering system aimed to maximise the quality of all the essential visual cues for perceptual realism, as shown in Figure 4.1.

Passing a visual Turing test puts very strict requirements on the quality of reproduction. To make the task feasible, we aim for a visual reproduction of a static scene encompassing a moderate field of view (27° ×21.8°) and seen from a fixed viewing position (no motion parallax). As analysed in Chapter 2, such a scene can in principle be reproduced with perceptually-realistic fidelity if sufficient quality and accuracy can be achieved in terms of the retinal image, spatial resolution, depth cues, dynamic range, contrast, and colour. The fact that human perception integrates across different cues, creating a 'holistic' percept, raises the possibility that almost inevitable small differences to the real world in terms of individual attributes may not be noticeable provided the other visual cues are collectively presented with sufficient quality.

The first objective of this work is to build a display apparatus and a 3D scene acquisition and rendering system that combines high spatial resolution with accurate colours, luminance levels, and cues to 3D structure (including focal distance). Our display apparatus combines four custom-built HDR displays into a single-viewer two-focal plane stereoscopic display. It can deliver a brightness level up to 3000 $\mathrm{cd/m^2}$ and below 0.01 $\mathrm{cd/m^2}$, a spatial resolution of at least 85 pixels per degree[1] at a viewing distance of 462 mm, a colour gamut of BT.709, correct disparity, and variations in focal depth from 462 mm (2.16 D/dioptres) to 740 mm (1.35 D). These capabilities are sufficient to reproduce a small scene inside a box of size 200 mm × 160 mm × 300 mm (width × height × depth) with levels of realism that exceed what existing display technologies can offer. Furthermore, the display is constructed in such a way that a viewer can simultaneously, or selectively, see a physical box containing

[1] Although our display resolution does not reach the peak resolving power at the fovea (240 ppds [120], see Section 2.1.1), it is sufficient for the majority of the population who do not normally have a 20/20 vision, as tested by our VTT experiment (Section 4.5).

real objects and compare them with displayed ones in the same spatial position. This enables a set of new perceptual experiments that have not been possible before. To deliver high-quality content for such a display, we create a system for acquiring, reconstructing, and rendering 3D scenes with a surface lumigraph [12] (light field with a proxy mesh). The system involves the capture of multi-exposure image stacks from multiple viewpoints with a high-resolution mirror-less camera, camera pose estimation with photogrammetry, colour calibration with a spectrometer, proxy mesh registration with differentiable rasterization, lumigraph view synthesis with view-dependent UV maps, multi-focal rendering with linear depth filtering, and a custom-designed focal plane calibration to compensate for different viewing positions of observers.

The second objective of this work is to apply this system to visually reproduce a moderate-size stationary object at a close distance to the observer (0.5 m) with a high fidelity such that it can be confused with a physical 3D object. The fidelity of reproduction should be confirmed by a visual Turing test with a strict criterion: the virtual scene must not be visually different in any respect from the real scene. To this end, we propose and performed a visual Turing test in a *three-interval-forced-choice* (3IFC) experiment where we asked naive observers to choose a scene that appears different when presented with two real and one virtual scenes, or one real and two virtual scenes. In this way, as opposed to a regular 2IFC test, we evaluate realism objectively and eliminate subjective interpretations of realism from prior experiences. The experiment results show that naive observers can only discriminate between real and displayed 3D objects with a probability of 0.44. To our knowledge, this is the first work that achieves a close perceptual match between a real-world 3D object and its displayed counterpart in both geometry and appearance. In contrast to previous work [126, 10, 119], we achieved this with a near-eye and binocular presentation of the stimuli and a much more challenging 3IFC test, and without any optical degradation of the real scene. The attempt at this challenge provides insights to better understand the conditions necessary to achieve perceptual realism. In the long term, we foresee this approach as an important step in the study of future display technologies, including AR and VR, to determine what display capabilities are most critical in achieving perceptual realism. Our display apparatus can also be useful in further studies of essential visual cues for realism such as material perception, colour appearance, and depth perception, in which realistic objects and scenes need to be faithfully reproduced.

We start this chapter with a review of early attempts at the visual Turing test (Section 4.1). Next, we introduce the architecture and hardware setup (Section 4.2), and the imaging and rendering pipeline (Section 4.3) of our HDR-MF-S display apparatus. We include several visual demonstrations to show the characteristic capabilities of our display system and performed a qualitative evaluation of its limitations (Section 4.4). Finally, we explain

the procedure of our 3IFC visual Turing test and discuss the results (Section 4.5).

The work presented in Chapter 4 produced the following publication:

- Fangcheng Zhong, Akshay Jindal, Ali Özgür Yöntem, Param Hanji, Simon J. Watt, and Rafał K. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH Asia 2021, Journal Track)*, 40(6), dec 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480513. URL `https://doi.org/10.1145/3478513.3480513`

Author's note in collaborative work Chapter 4 contains collaborative work with other parties for the completeness of the presentation. The author contributed to the overall design and implementation of the HDR-MF-S imaging and rendering pipeline (Section 4.3), including the light-field acquisition, lumigraph reconstruction, differentiable rendering, and multi-focal calibration and rendering; the qualitative evaluation (Section 4.4); the design of the VTT experiment (Section 4.5) and its data collection and analysis; and the geometric and photometric calibration of the data camera and the HDR-MF-S display apparatus (Section 4.2).

4.1 Early attempts of visual Turing test

Obtaining realistic results has been one of the main pursuits of computer graphics and particularly rendering. Global illumination and physically based rendering allowed for accurate simulation of light [56]. When combined with tone mapping methods [39], simulation of lens glare [145], and camera response [144], these techniques can produce photorealistic images, indistinguishable from photographs of real-world scenes. However, since the focus of our work lies beyond photorealism, we review the studies that attempted to achieve perceptual realism by matching a virtual scene with a physical one.

Meyer [126] was the first to compare rendering shown on a display with a real scene in an experiment. The participants saw the real scene and a CRT screen with its reproduction side by side, via viewfinders of two cameras with telephoto lenses. Additional Fresnel lenses were added to enlarge the viewfinder images so that they could be seen from 112 cm. Despite the lack of binocular depth cues and the low resolution of the CRT screen, the authors reported that neither naive observers nor experts could tell which image was computer generated. Although this was an impressive result, it was helped by the degradation of the real-scene images, due to lens distortions, and their small size (9.2×9.2 cm seen from 112 cm, or 4.7°).

Borg et al. [10] reported a graphics Turing test experiment, in which they successfully reproduced the result of Meyer without the need to see the stimuli via a viewfinder. The participants viewed either a real object (a pyramid or a sphere), or a display seen through a small aperture in a 2 m long box. The stimuli were viewed with one eye. Also, because the authors could not achieve the required dynamic range on their display they asked the participants to view the images from 10 cm away from the box in a non-dark room (50 lux) so that the display black level was masked by glare in the eye and adaptation.

Masaoka et al. [119] measured how the impression of realism is degraded with the reduction of resolution. The authors conducted a pairwise comparison experiment, in which one of the conditions was a real scene and the other conditions were images of gradually reduced resolution. The results of comparisons were scaled using a Bradley-Terry model to give a measure of the sense of realness, proportional to JND units. The images and the real scene were seen through a synopter so there were no binocular disparities, and the distance was 480 cm to ensure sufficient angular resolution and minimise the influence of variations in focal distance. The study found that a resolution between 60 and 120 cycles per degree is required to achieve the perceived realism of a real scene.

None of the above studies attempted to reproduce binocular depth cues but instead reduced

their influence by using large viewing distances and optics. These studies also reported difficulties in reproducing the real-world dynamic range. Both of these aspects were addressed in the study of Vangorp et al. [163], in which the virtual scene was reproduced on an HDR display (SIM2 HDR47E) seen through a stereoscope, albeit at low resolution (30 ppd). The goal of the study was to measure how binocular disparity and contrast contribute to realism, in a manner similar to Masaoka et al.'s study of resolution. The task was to compare two displayed scenes, each with a certain amount of both contrast and disparity modification, and choose the one closer to the real scene. The participants could look at the real scene at their discretion, but it was not included in the compared conditions, so the experiment could not test for a perceptual match. The authors found that the participants were more sensitive to changes in contrast than in disparity, and selected as more realistic either natural or moderately enhanced contrast.

In addition to the above-mentioned visual Turing-test experiments, a comparison with a real scene has also been used to evaluate the reproduction of brightness [122] and tone mapping [194], but these studies did not attempt to achieve a perceptual match with a real scene.

Although the studies of Meyer, Borg et al., and Masaoka et al. reported a perceptual match of the display and real scenes, they were achieved only in monocular view or using optics that degraded the visual quality of the real scene. Our work aims to go beyond these efforts. We reproduce all visual cues, including depth and dynamic range, and match a real object seen at a small viewing distance, and with no optical aberrations.

4.2 HDR-MF-S display

The main objective of the design of our HDR-MF-S display is to maximise the visual quality and realism of the displayed images for all the following capability dimensions: physical luminance, dynamic range (contrast), colour gamut, binocular and focal depth cues. The goal is to deliver all these capabilities altogether with sufficient qualities rather than focusing on maximising a single one. While there are several fundamentally distinct approaches to 3D display architectures as discussed in Section 2.4, not all of them meet the requirements for our objective. For example, accurate depth cues, matching light distributions in the real world, can be potentially achieved with holographic [108] or light field [159] displays. However, the current state-of-the-art of these technologies does not allow us to achieve the field of view, colour accuracy, resolution, or dynamic range required for perceptual realism. Reproducing a four-dimensional light field of sufficient size and quality with these technologies requires control over billions of pixels, which is currently

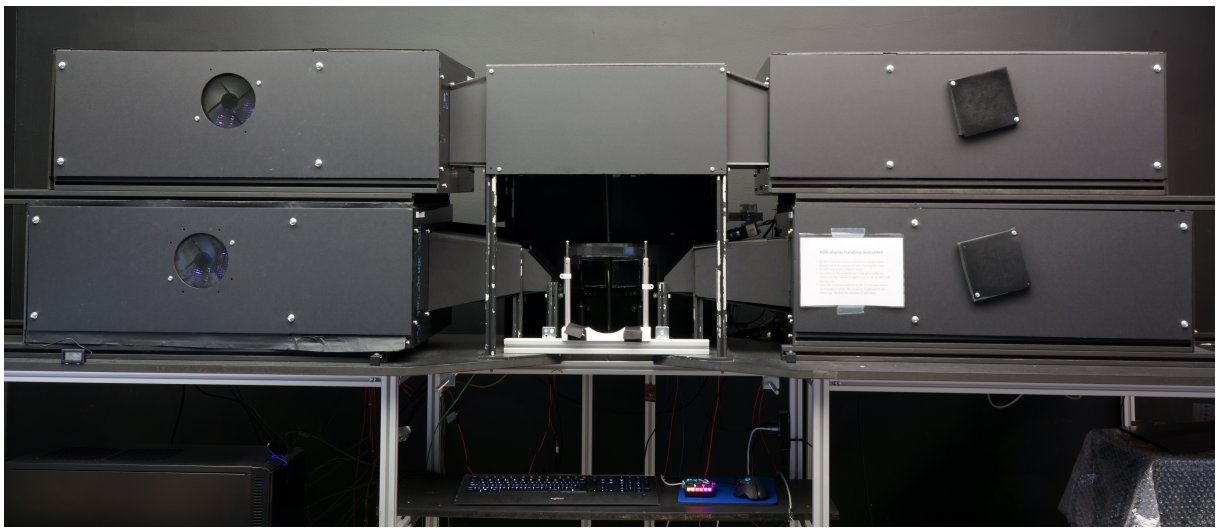


Figure 4.2: The front view of the display.

infeasible. However, if we can either stabilise or track the viewing position, the subspace of a light field that we need to reproduce is much smaller, making it possible to build a display of required capabilities.

One approach to producing the required light distribution, given either fixed or known eye position, is to use a stereoscopic multi-focal display [2]. In such displays, the eye sees the sum of light from multiple superimposed planes at different focal distances. Such displays can effectively drive accommodation to any point between the planes if the plane separation is small enough ($\sim$0.6 D to $\sim$0.9 D) [111, 112], while retaining desirable capabilities of conventional displays (resolution, colour gamut). Moreover, this uncomplicated design, without any refractive or diffractive optical components in the viewing path, generates images without additional optical distortions. This is in contrast with vari-focal displays [36] or near-eye light field displays [69], which are likely to introduce noticeable aberrations. One important limitation of a multi-focal display is that the addition of focal planes reduces dynamic range. The additive nature of the beam splitters elevates the black level, and their transmission limits the peak brightness of each plane. We address this problem by combining a multi-focal stereoscopic display design with high-dynamic-range displays, making a high-dynamic-range multi-focal stereoscopic (HDR-MF-S) display. In the following subsections, we explain the details of the design of our HDR-MF-S display and how it achieves the capability dimensions that we desire.

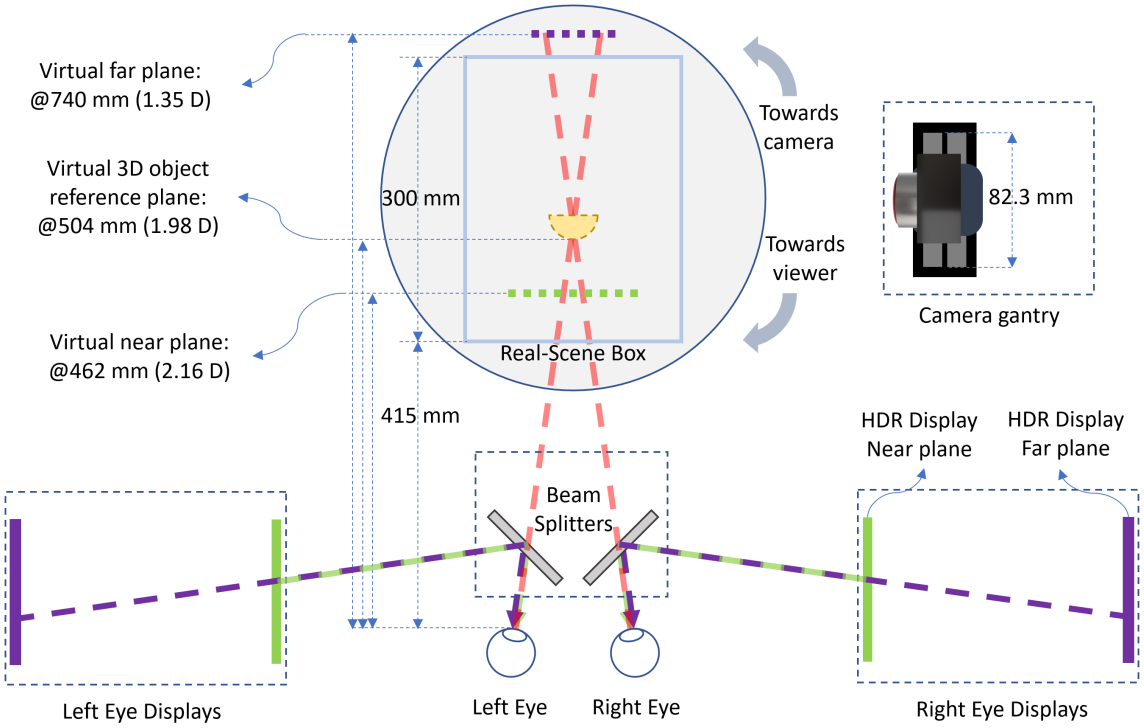


Figure 4.3: Schematic of the high-dynamic-range multi-focal stereo display apparatus. (Note that to simplify the schematic, not all the folding mirrors and beam splitters are shown.) The apparatus creates two image planes (green and blue dashed lines inside the real-scene box) per eye and the observer sees respective images via beam splitters. The real-scene box is observed through the same beam splitters. The real-scene box is on a manually rotating platform moving toward a fixed capturing position or a fixed display position. The camera gantry is on another manually movable platform (not shown in the figure) which can move towards or away from the real-scene box allowing coarse adjustment of the field of view.

4.2.1 Apparatus overview

Figure 4.2 shows a photograph of the front view of our display apparatus. The apparatus comprises three main components as shown in Figure 4.3: a Wheatstone stereoscope with four high-dynamic-range displays and two focal planes; a real-scene box in front of the observer that is seen through a pair of beam splitters; and a motorised camera slider capable of capturing dense horizontal light fields of the real-scene box. In this setup, a small physical scene is arranged in the real-scene box. This box normally faces the observer, but can be rotated to face the camera rig in order to capture its light field as shown in Figure 4.3. When facing the observer, the real scene and its rendered counterpart are spatially superimposed. We can instantly switch between the real and displayed scenes by controlling the light in the real-scene box and the display. We discuss the details of each component in the following subsections.

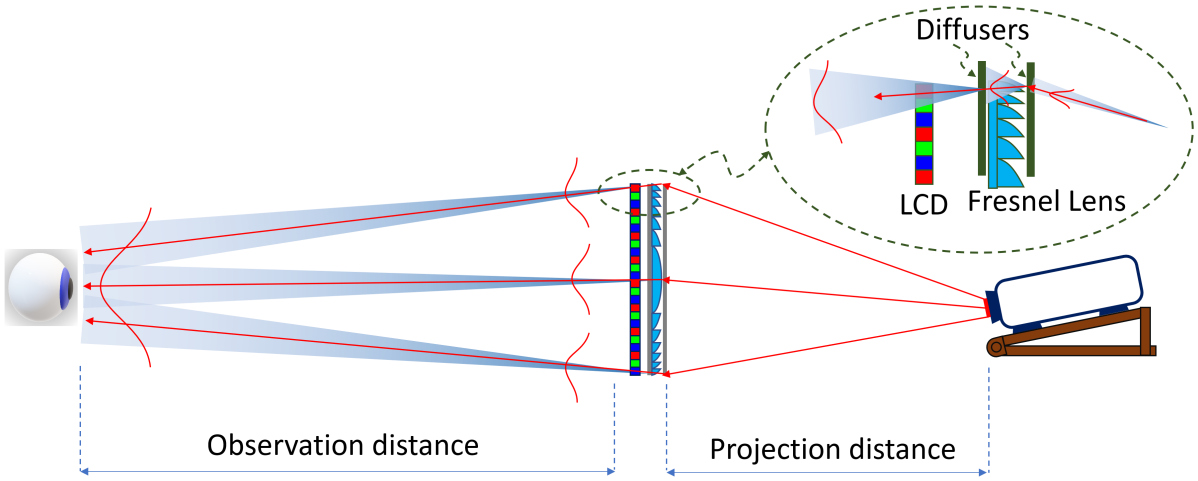


Figure 4.4: Each HDR display comprises a projector acting as a backlight for an LCD panel with factory backlighting removed. A Fresnel lens sandwiched between two narrow-angle diffusers, with scattering angles of 10 and 5 degrees. An image from the projector is formed on the first diffuser acting as the backlighting of the LCD. The Fresnel lens helps to steer the backlighting toward the eye uniformly. The second diffuser prevents reflections between LCD glass and the Fresnel lens substrate.

4.2.2 HDR displays

The key feature of our display is the capability of reproducing a high dynamic range, with a peak luminance of 3000 cd/m^2 and the black level much below 0.01 cd/m^2. Such a low black level practically eliminates any stray light in areas of an image that should remain black. The HDR reproduction is delivered by four projector-based dual-modulation displays, similar in design to those used in one of the first HDR displays [150][2].

The software for controlling each display implemented the standard two-spatial-modulator factorization algorithm [150] running on a GPU. However, we took special care to achieve accurate geometric alignment and high colour accuracy. The geometric alignment was achieved by taking images with a DSLR camera of a calibration pattern (a grid of points) displayed separately on the LCD and the DLP and then aligning them using homography and mesh-based warping. The point-spread function of the DLP was measured for the same grid of points and approximated with a Gaussian function. The colourimetric calibration was achieved by measuring the colour ramps with a spectro-radiometer (Specbos 1211) and fitting a gamma-offset-gain model to the LCD panel and using a dense look-up table for the DLP. The dense look-up table was necessary as the response of the projector was non-monotonic after removing the colour wheel. The effective bit-depth of both displays

[2]More details of the HDR display hardware and our improvements over [150] can be found in [203]

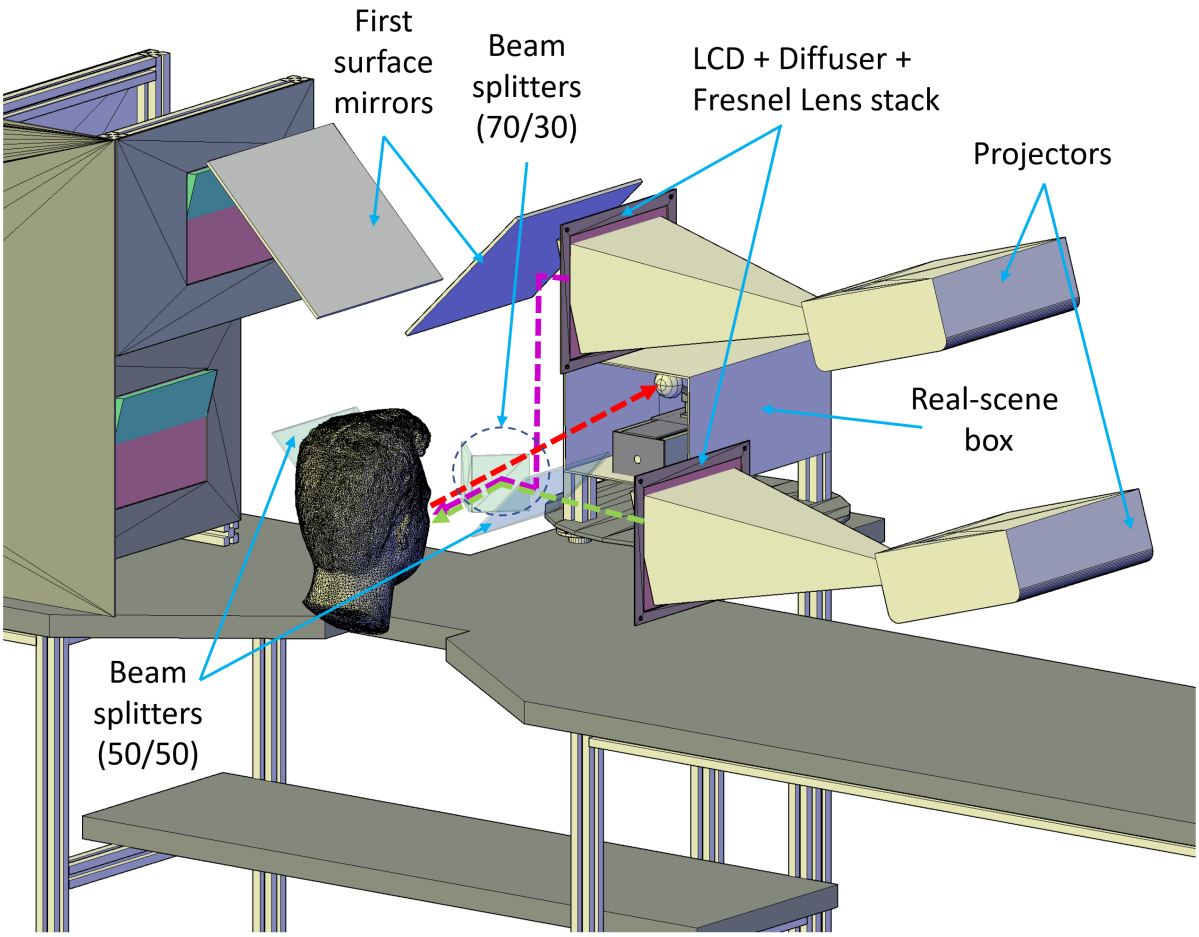


Figure 4.5: The schematic showing the light paths from two display focal planes (green and purple dashed lines) and from the real-scene box, for the right eye. The red dashed line shows the viewing direction of the observer towards the real-scene box. The line colours are consistent with Figure 4.3.

was increased to 10 bits by bit-stealing (DLP) and spatio-temporal dithering (both DLP and LCD). The uniformity of the display was compensated by taking an image with a DSLR camera and using it for compensation of the DLP image.

4.2.3 Focal planes and optics

To vary the focal distance, similar to a multi-focal display [2], our display can generate images at two focal planes, at the distances of 462 mm (2.16 D) and 740 mm (1.35 D) from the viewer, providing a 0.81 diopter separation between the planes. The separation was selected to ensure that the images shown on two planes provide cues for accommodation for any distance between the two planes [111]. Such distances also ensure a resolution of at least 85 pixels per degree for the observer. These distances are adjustable by moving the HDR displays on their mounting rails.

Figure 4.6: Frames of the glasses with an IR LED. The participants were asked to wear these frames to track their head position.

Figure 4.5 shows the optical paths for near and far virtual images on the right-hand side. The image of the far plane is formed by reflecting the real image of the top right HDR display through a mirror, and two beam splitters. The purple dashed line in the figure indicates its optical path. The near-plane image is formed by reflecting the real image of the bottom HDR display from a single beam-splitter, depicted by the green dashed lines. This is symmetrical for the left-hand side of the setup. We opted for this simple optical design without any refractive [111] or varifocal [18] optics to avoid aberrations, which would introduce detectable imperfections and also reduce the dynamic range due to scattering of the light. The real-scene box is observed through 70R/30T (reflection/transmittance, Edmund Optics, 64-409) beam splitters, located in front of the observer's eyes. The red dashed line shows the viewing direction through these beam splitters. This reflection/transmittance ratio was selected to achieve a higher brightness of the display. The second beam-splitter 50R/50T (weidner-glas.de) on the side is used to combine the images from far and near planes. Since the system has several optical paths crossing each other, we enclosed all the image-delivering paths separately to avoid cross-talk images. At the optical exit where the observer views the scene and the displays, we placed a chinrest and forehead rest to fix the viewing direction and limit head movements. We also placed blinders on either side of the chinrest posts to prevent a direct line of sight of the near-plane LCD screens.

Multi-focal plane displays are very sensitive to misalignment due to head movement and often require either bite-bars [111] to eliminate such movements or active correction in rendering through eye tracking [124]. We aimed to build a setup similar to the latter using an IR LED fixed onto a glasses frame without lenses (Figure 4.6). The observers were asked to wear the frame while viewing, and the LED was tracked using a high frame rate machine vision camera (iDS UI-3140CP), with 25 mm C-mount lens (Fujinon HF25HA-1B) and a visible light filter. This allowed us to track the observer's head position in real time. We later use the data from the head tracker in our experiment (Section 4.5) to determine the invalid trials.

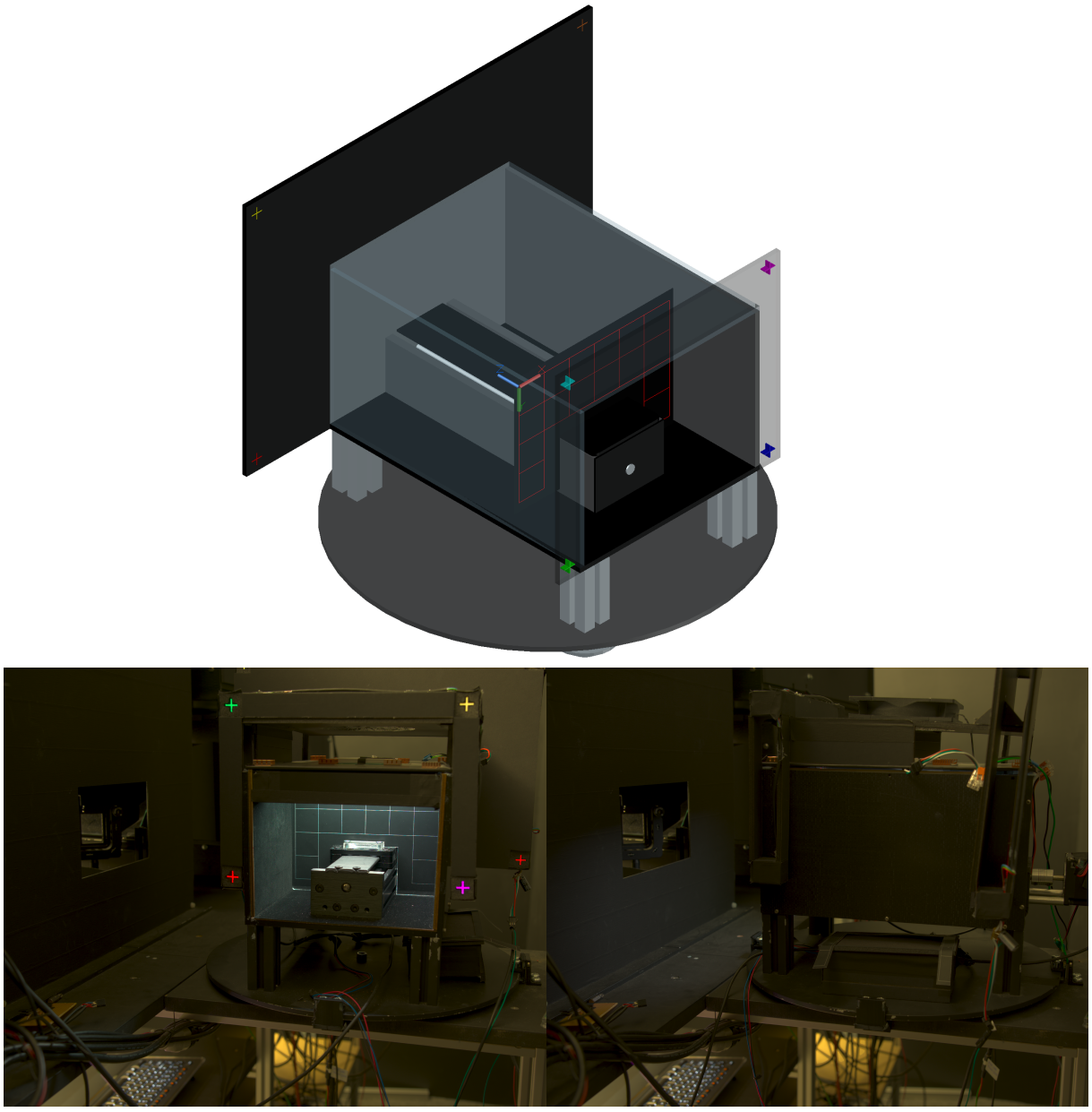


Figure 4.7: The front and side view of the real-scene box and schematic of the calibration target inside. The calibration target has a grid of four-by-six squares of the size 30 mm x 30 mm, which defines a world coordinate system. The red, green, and blue arrows in the figure represent the origin and orientations of the X, Y, and Z axes, respectively. We define the upper-left corner of the grid as the origin for the X and Y axes and the target placed at the front location as the $Z = 0$ plane.

4.2.4 Real-scene box

The real-scene box has the inner dimensions of 200 mm × 160 mm × 300 mm (width × height × depth). It was made of black acrylic, which was covered on the inside with high-absorption blackout material (Thorlabs: Black Flocked Self-Adhesive Paper). The ceiling was fitted with an LED array light source with 225 individually addressable RGB LEDs (WS2812B). The real-scene box was fixed on a platform, supported by ball transfer units, allowing it to be freely rotated towards the observer for viewing, or towards the camera for light-field capture, as shown in Figure 4.7. The real-scene box rotation was fixed in either of the two positions using custom magnetic mounts.

To facilitate several calibration procedures for our imaging system (Section 4.3), we defined world space coordinates for the real-scene box. We placed a removable calibration target on a gantry plate inside the real-scene box, as shown in Figure 4.7. The gantry (Oozenest, 250 mm C-Beam Linear Actuator) can be controlled to move the target freely from the entry of the real-scene box to its end. The calibration target had a grid of four-by-six squares of the size 30 mm x 30 mm. We used the grid to define a world coordinate space, as shown in Figure 4.7.

In addition to the calibration target, the real-scene box also included eight cross-shaped calibration markers placed outside the box, as shown in Figure 4.7. The markers were used as reference points to register the camera pose when the calibration target inside the box had to be removed. The markers were carved on the two foamboards and illuminated by an RGB LED (WS2812B) with a diffuser to improve their visibility.

4.2.5 Data camera for light field capture

To capture a horizontal light field of the real-scene box, we mounted a Sony α7R3 mirrorless camera with a Sony G OSS zoom lens (focal length 24-105 mm) on a motorised camera slider (Figure 4.8) at a distance of 415 mm from the real-scene box, similar to the distance from the viewing position to the real-scene box. The camera slider traversed a baseline of 82.3 mm with an accuracy of 5 µm.

4.3 HDR-MF-S imaging & rendering system

To achieve a perceptual match between real and virtual scenes, we need not only a display capable of reproducing all relevant cues, but also an imaging and rendering system, which

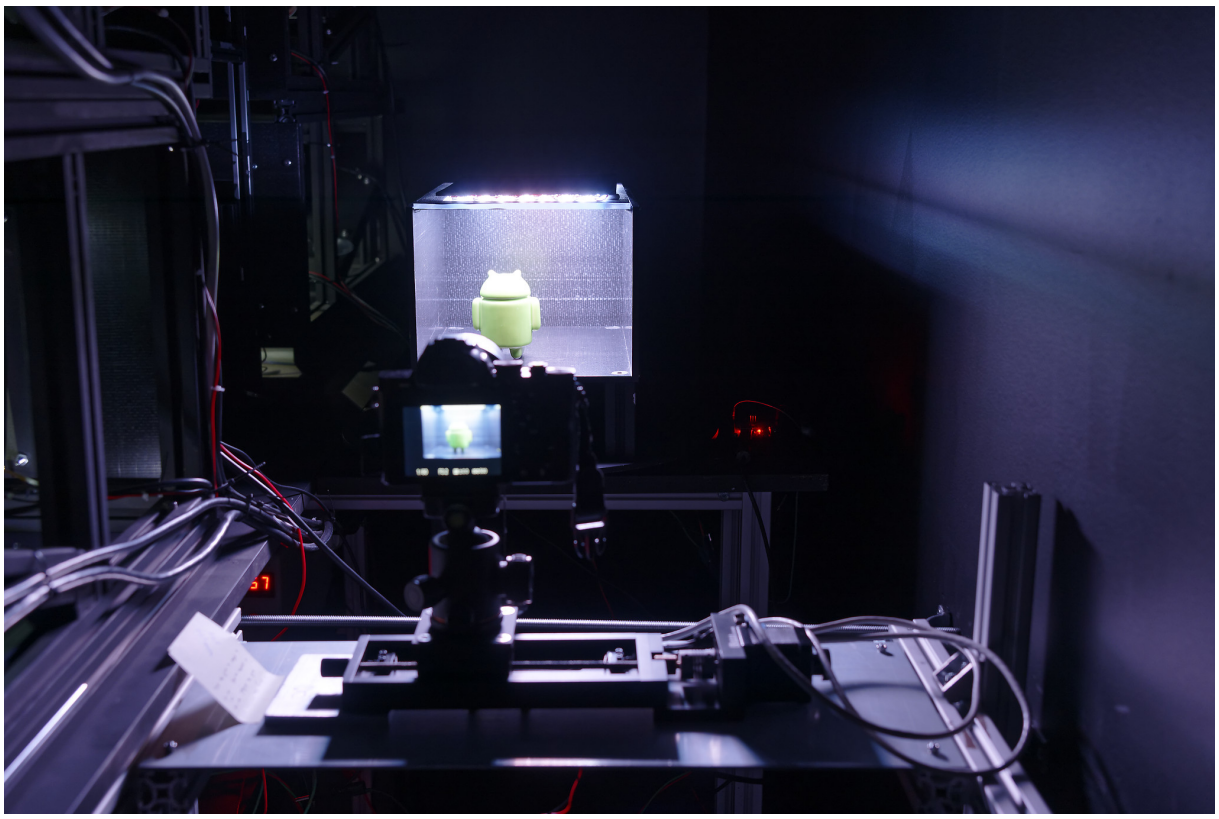


Figure 4.8: The data camera and motorised slider for light field capture.

can capture a real scene and reproduce it with sufficient quality. Most importantly, the rendered scene should match the viewpoint of the observer. Our system is currently limited to processing scenes of relatively simple or known geometry, but can handle complex non-Lambertian materials and high-dynamic-range illumination.

Figure 4.9 shows a diagram of our HDR-MF-S imaging and rendering system. We start with the capture of a horizontal HDR light field, which is colour-calibrated for the spectra of the scene illumination (Section 4.3.1). Next, we employ photogrammetry to perform a 3D reconstruction and estimate camera matrices (Section 4.3.2). After that, we apply a differentiable rasterizer to register a proxy mesh of the main object with its silhouette in each HDR light field image (Section 4.3.2), so we can project the fitted mesh to each light field image to obtain a view-dependent UV map and texture. Before rendering, we find the position of each focal plane of the display with respect to the eye position and the calibration target in the real-scene box (Section 4.3.3). Finally, we integrate lumigraph view synthesis with linear depth filtering [2] to render the final scene on our HDR-MF-S display (Section 4.3.4).

We found lumigraph to be the most suitable 3D representation for our purpose as it models non-Lambertian surfaces, is robust to processing high-resolution textures, and performs rendering in real time. We have also experimented with dense light fields, either captured

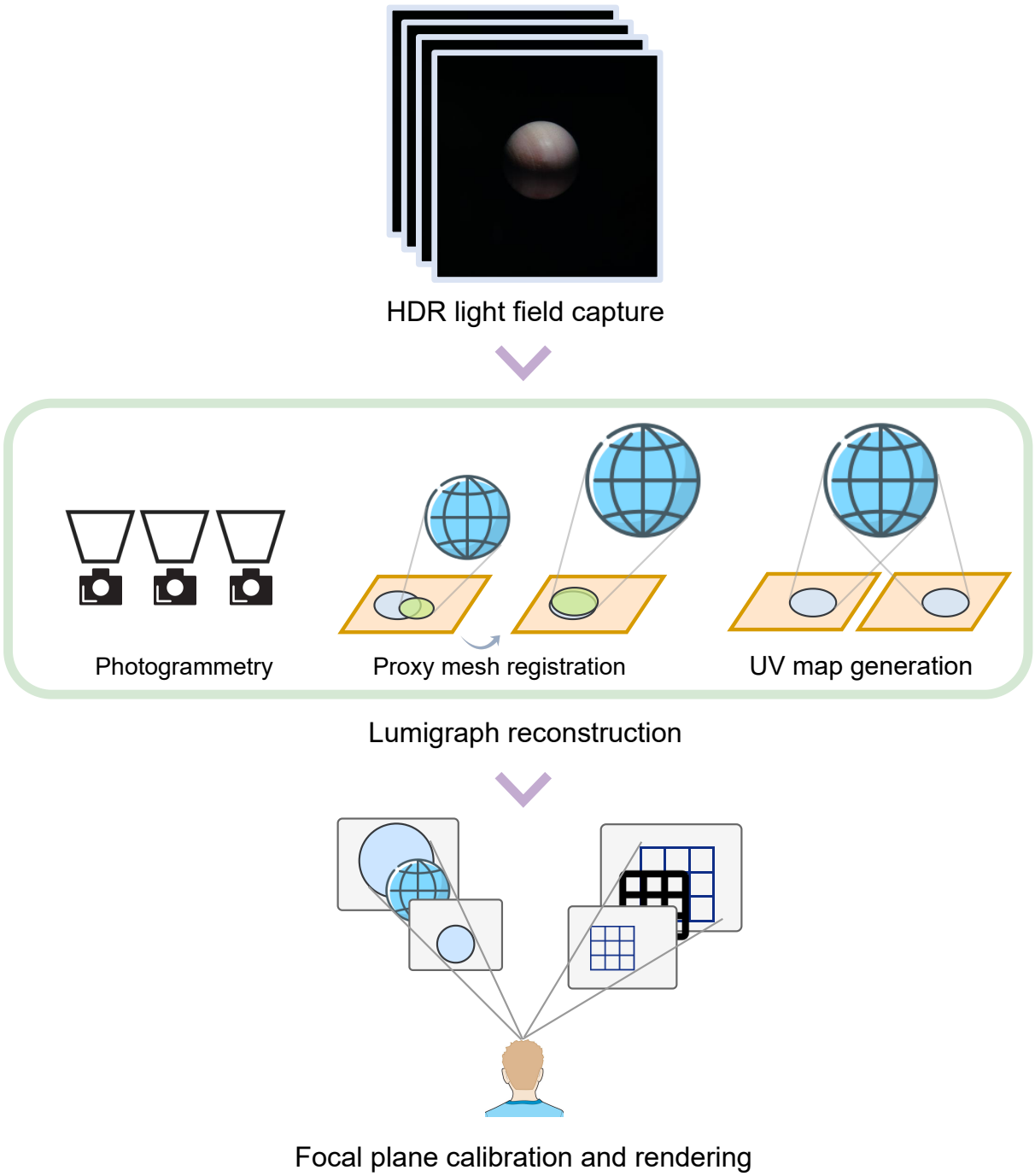


Figure 4.9: The process of capturing and rendering contents for our HDR-MF-S display. Refer to Section 4.3 for the explanation.

or reconstructed using neural radiance fields [128], but they did not match the quality required for perceptual realism. For example, according to Equation 2.2, capturing our real-scene box (at the viewing distance of 415 mm) up to the visual acuity at the fovea requires a spatial resolution of 33.13 pixels per millimetre. This approximately corresponds to an image in the size of 6629 × 5302 pixels given the size of the real-scene box (200 mm × 160 mm). Training, convergence, and rendering with such image size for view synthesis remains an actively studied problem, although significant improvements [130] have been made subsequent to our work. Therefore, in this work, we combined photogrammetry and differentiable rendering to align known geometry with the captured HDR images to reconstruct a lumigraph.

4.3.1 HDR light field capture

Using our data camera discussed in Section 4.2.5, we first capture a high-resolution (7360 x 4912 pixels) light field consisting of 16 views with a separation of 5 mm between them. For each camera view, we capture an HDR exposure stack consisting of up to five RAW images spaced two stops apart in exposure time and ISO of 100. We merge the RAW images to increase the dynamic range and reduce noise using a Poisson photon noise estimator [61]. Next, we demosaic the merged images using the *DDFAPD algorithm* [123]. To calibrate for colours, we measure the spectra of a colour checker passport (X-Rite) positioned inside the real-scene box with a spectroradiometer (Specbos 1211, Jeti). Then, we compensate for the measured spectral transmission of the 70/30 beam-splitter and recover trichromatic coordinates using the CIE XYZ 1931 colour-matching functions. The XYZ colour coordinates are used to find the matrix that transforms from native camera linear RGB space into CIE XYZ and which results in the smallest RMSE of DeltaE 2000 colour differences. The white patch in the colour checker is used for white balance. Finally, we apply the matrix to convert the merged HDR images from their native camera linear RGB space to the BT.709 space used by our display.

4.3.2 Lumigraph reconstruction

The objective of this stage is to construct a *surface lumigraph* [57, 12] (a light field projected on a proxy geometry), represented by a proxy mesh and view-dependent UV maps and textures [30], of the captured scene.

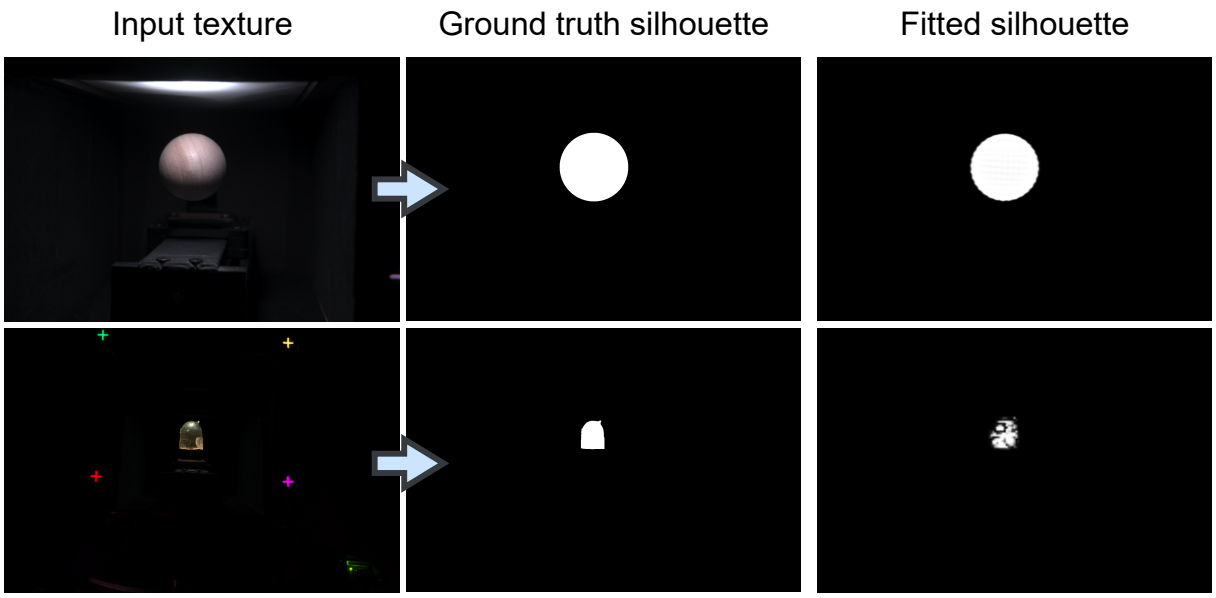


Figure 4.10: Results of the fitted silhouettes of the proxy mesh after registration optimised by differentiable rasterization.

Photogrammetry We first use *Meshroom* [3], a photogrammetry software, to perform a multi-view stereo reconstruction of the scene. We supply Meshroom with the HDR light field images captured from the gantry and additional single-exposure images captured with the camera mounted on a Magic Arm (Manfrotto) and positioned at multiple locations around the front of the real-scene box. These additional images are necessary for the 3D reconstruction but are not used for textures. After the reconstruction, Meshroom returns a noisy scene mesh (including the main object, the real-scene box, the calibration markers, etc.) with estimated camera extrinsic and intrinsic matrices. Note that at this stage, the scene mesh is in an arbitrary local camera space. The camera matrices are also calculated with respect to this space. We record the coordinates of each reconstructed calibration marker in local space, which we later use for a coordinate transform.

Proxy mesh registration and UV map generation The mesh reconstructed from photogrammetry does not meet the accuracy of perceptual match required by our experiment. Hence, we choose to experiment with objects with simple or known geometry and pre-generate the mesh files, as mesh reconstruction is not the main focus of this work. However, we still need to register the mesh to the correct coordinates. It is crucial to ensure that the projected silhouette of the registered mesh is near-identical to the ground truth. Otherwise, the rendering would appear distorted once we project the mesh onto light field images to construct the lumigraph. We employ *SoftRas* [103, 143], a differentiable rasterizer, to find an optimal spatial transformation to align the mesh with the silhouettes in captured images. Specifically, the optimal parameters of a spatial transformation $\mathcal{T}$

including scaling, rotation, and translation can be found by

$$\arg\min_{\mathcal{T}} \sum_{i} ||\mathcal{R}(\mathcal{T}(\mathbf{M}), \mathbf{C}_i) - \mathbf{I}_i||, \tag{4.1}$$

where $\mathcal{R}$ is a differentiable renderer that rasterizes a grey-scale silhouette image, $\mathbf{M}$ is the unregistered mesh, $\mathbf{C}_i$ is the i-th camera matrix, and $\mathbf{I}_i$ is the extracted ground-truth silhouette from the i-th camera view. We apply the *GrabCut algorithm* [146] to extract the ground-truth silhouettes of the main object. Figure 4.10 shows the results of the silhouette fitting. After the registration of the proxy mesh, we generate the UV coordinates by projecting the mesh vertices onto each HDR texture using the camera matrices obtained from photogrammetry.

Local-to-world coordinate transformation To facilitate the following calibration steps, it is convenient to have the scene geometry represented in world coordinates expressed in physical units (meters). To do this, we determine the coordinates of the calibration markers in both local space (Section 4.3.2) and world space (Section 4.2.4) and apply the *orthogonal Procrustes algorithm* to find an optimal change-of-coordinates transformation from the local to the world space.

4.3.3 View-dependent focal plane calibration

Both pairs of display focal planes must be well-aligned with the positions of the observer's eyes to correctly align the two focal planes and match the scene shown in the real-scene box. To map the coordinates of each display to the world coordinates of the real-scene box, we perform a manual focal plane calibration. As different observers have different inter-pupillary distances (IPDs) and may put their heads at different positions, this calibration needs to be performed per observer.

During the calibration, the observer is asked to put their head on the chin rest and press against a rigid forehead rest. The forehead rest provides additional stability and limits head movements. As shown in Figure 4.11, each eye is presented with four crosses on one of the HDR displays. They move the four crosses to align them to the corresponding specified crossings of the calibration target in the real-scene box. The observers perform this alignment for each of the two focal planes per eye and for the calibration target positioned at two different depths. The gantry inside the box moves the target to their desired locations. After this calibration, we obtain a correspondence of eight points in world space and in image space. They are used to find the transformation from the 2D

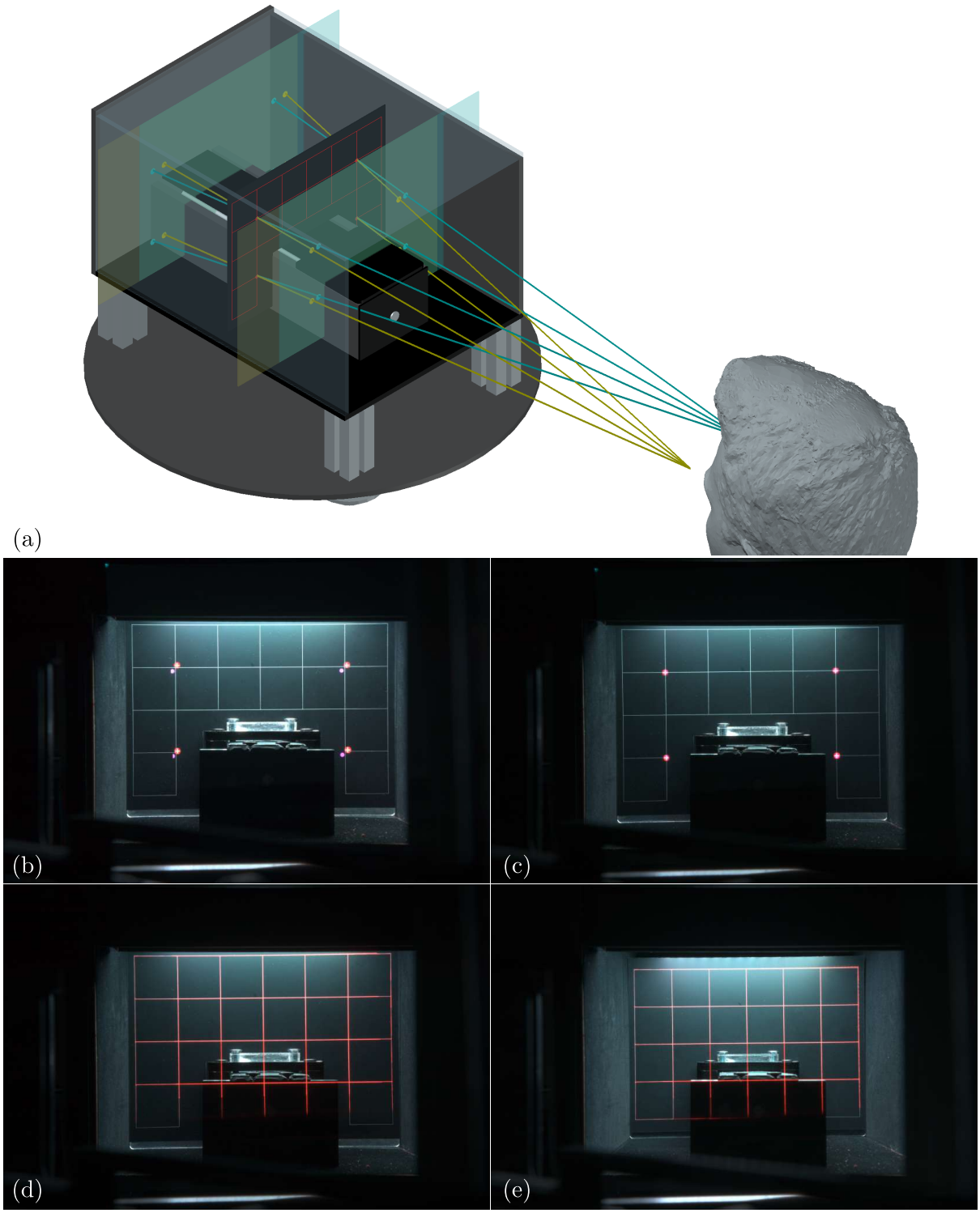


Figure 4.11: (a) Schematic of the focal plane calibration. We use yellow and cyan to indicate the view of the left and right eye. (b, c) Left-eye view of the focal plane calibration interface. Observers drag the red (near plane) and pink dots (far plane) to align with the corresponding positions on the calibration target. (d, e) Rendering of the calibration grids at different gantry positions after calibration.

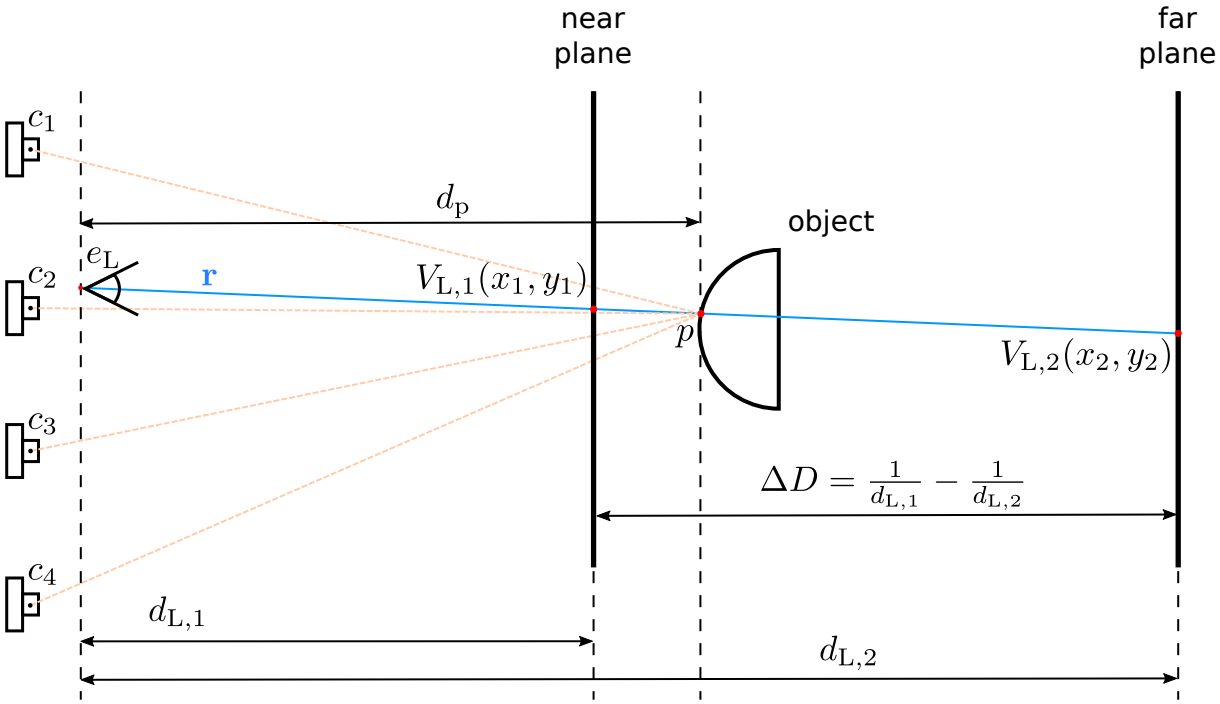


Figure 4.12: The radiance computation for the near and far focal planes for the left eye. $c_1, \ldots, c_4$ are the positions of data cameras. e_L is the viewing position of the left eye.

coordinates on each focal plane (an image shown on each HDR display) to the world coordinates. We use the *direct linear transformation algorithm* (DLT) [160] to find a rendering matrix M which maps the world coordinates to the clip space for each focal plane. Finally, we apply an RQ decomposition to decompose the rendering matrix into a view (extrinsic) matrix V and a projection (intrinsic) matrix P, i.e. $M = P\,V$. With the view matrix, we are able to compute the observer's view (eye) positions and orientations, which is required for the lumigraph view synthesis and multi-focal decomposition in the rendering stage.

4.3.4 Multi-focal lumigraph rendering

To find the value of each pixel of the near and far display focal planes, we use lumigraph rendering [57], combined with linear depth filtering in the diopter space [2]. We choose simple linear filtering as our test scene does not contain any occlusions, which would require more advanced methods [131, 124, 196], as discussed in Section 2.4. Specifically, the value of the pixel (x, y) on the j-th focal plane (1 – near, 2 – far) for the left eye (index L) is computed by filtering across the focal planes and cameras (similarly for the right

eye):

$$V_{L,j}(x,y) = \underbrace{\frac{|D_{\mathrm{p}} - D_{\mathrm{L,j}}|}{\Delta D}}_{\substack{\text{linear depth} \\ \text{filtering}}} \underbrace{\sum_{k=1}^{K} T_k(u_k, v_k)\, w_k}_{\substack{\text{view synthesis across} \\ K \text{ camera views}}}, \tag{4.2}$$

where the symbols are illustrated in Figure 4.12. We use lower case symbol d to represent distances in meters and upper case symbol D to represent distances in diopters, so that $D = 1/d$. In particular, $D_{\mathrm{L,j}}$ is the diopter of the j-th focal plane from the viewing position e_L. D_{p} is the distance (in diopters) of the intersection point p of the ray $\mathbf{r}$ with the object, where $\mathbf{r}$ originates from e_L and passes through pixel (x, y). ΔD indicates the diopter difference between the near and far focal planes. $T_k(u_k, v_k)$ represents the value of the HDR texture associated with the data camera k for the texture coordinate (u_k, v_k) at the intersection point p. We calculate T_k by rasterising the texture-mapped registered mesh (Section 4.3.2) with the rendering matrices generated during the focal plane calibration (Section 4.3.3). The texture is filtered with standard mipmap. The value of w_k is the weight associated with each data camera. As we assume a static eye position, we always select the nearest neighbour in our current implementation to avoid blur artefacts:

$$w_k = \begin{cases} 1, & \text{if } ||e_L - c_k|| = \min_j ||e_L - c_j||, \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where the values of e and c (data camera origins) are obtained from the focal plane calibration (Section 4.3.3) and lumigraph reconstruction (Section 4.3.2) respectively.

4.4 Results

Although it is difficult to convey the three-dimensionality and color appearance of the scenes shown on our display using photographs, in this section, we include a few to demonstrate some of its characteristic capabilities. We captured images of several displayed and real objects using a Sony α7R3 camera with a 55 mm lens (SEL55F18Z). We set the aperture to F9.5 so that its diameter matched the expected pupil diameter for our scene (5.8 mm). We also performed the focal plane calibration (Section 4.3.3) for the viewing position of the camera.

Figure 4.13 demonstrates a close perceptual match between the real and virtual objects achieved by our system. The accurate spatial alignment of the virtual object overlaying the physical object demonstrates the perceptual match in geometry (Figure 4.13(a)). We also achieved a close match in appearance and shading (see the overlapping shadows and

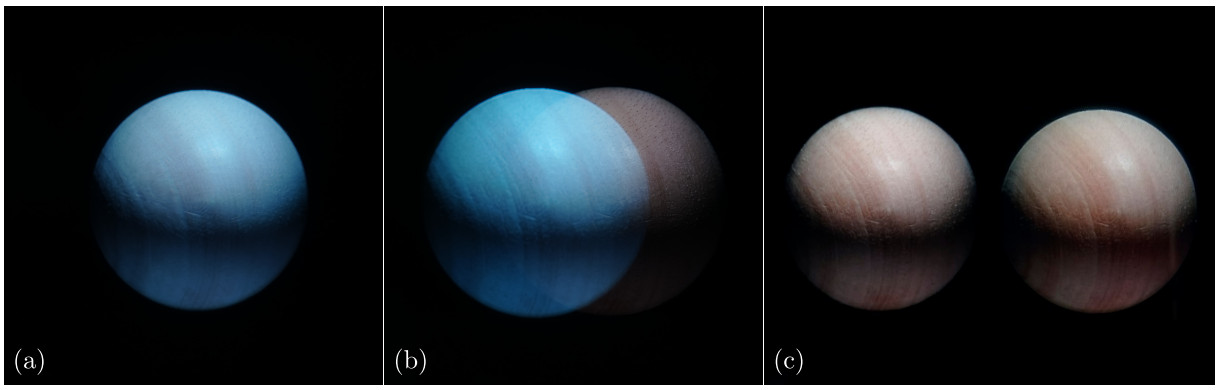


Figure 4.13: (a) Photograph of a virtual object displayed on top of the real object. We changed the hue of the texture to show a mixed-reality effect. (b) The real object can be seen more clearly with the displayed object slightly shifted away. (c) Photograph of the displayed object (right) next to the real object (left). The small white strip visible on the bottom right corner of the right object is not a display artefact but a reflection from the background.

specular reflections in Figure 4.13(a) and the side-by-side comparison in Figure 4.13(c)). With such a level of precision, we are able to show many mixed-reality effects that would not be possible otherwise such as changing the hue of the physical object without changing the shadows or textures.

Figure 4.14 shows photographs of a rendered 3D-printed robot figure, displayed at three distances while the camera was set to one of those three focal distances. As expected, the display shows a desired defocus blur when the object is shown at a different focal depth from that of the camera lens. However, since there is no display focal plane in the mid-distance, the image shown at the centre is a superimposition of the two defocused images from both focal planes, which results in a visually incorrect blur. The amount of such blur can be reduced by bringing both focal planes closer.

To evaluate the resolution limit and the aforementioned incorrect defocus blur (when the virtual object is placed between the two focal planes) of our display, we reproduced a 1951 USAF resolution test chart (ThorLabs, R3L3S1P, positive, 3"×3") and photographed it in comparison with the physical chart (Figure 4.15). We built a custom lightbox to illuminate the chart from the back, producing a high-contrast resolution pattern. We displayed either the real or rendered virtual chart at one of three distances[3]: 500 mm (near), 577 mm (middle), and 654 mm (far). The camera focus was also set to one of these distances. To reduce the Moiré pattern resulting from the interference of the LCD and camera sensor pixel grids, we reduced the aperture to F16 and processed the images using *DxO PhotoLab 4.3.0* with only Moiré filtering enabled. Note that the Moiré pattern was

[3]For this evaluation, we moved the near focal plane close to the near distance and the far focal plane close to the far distance.

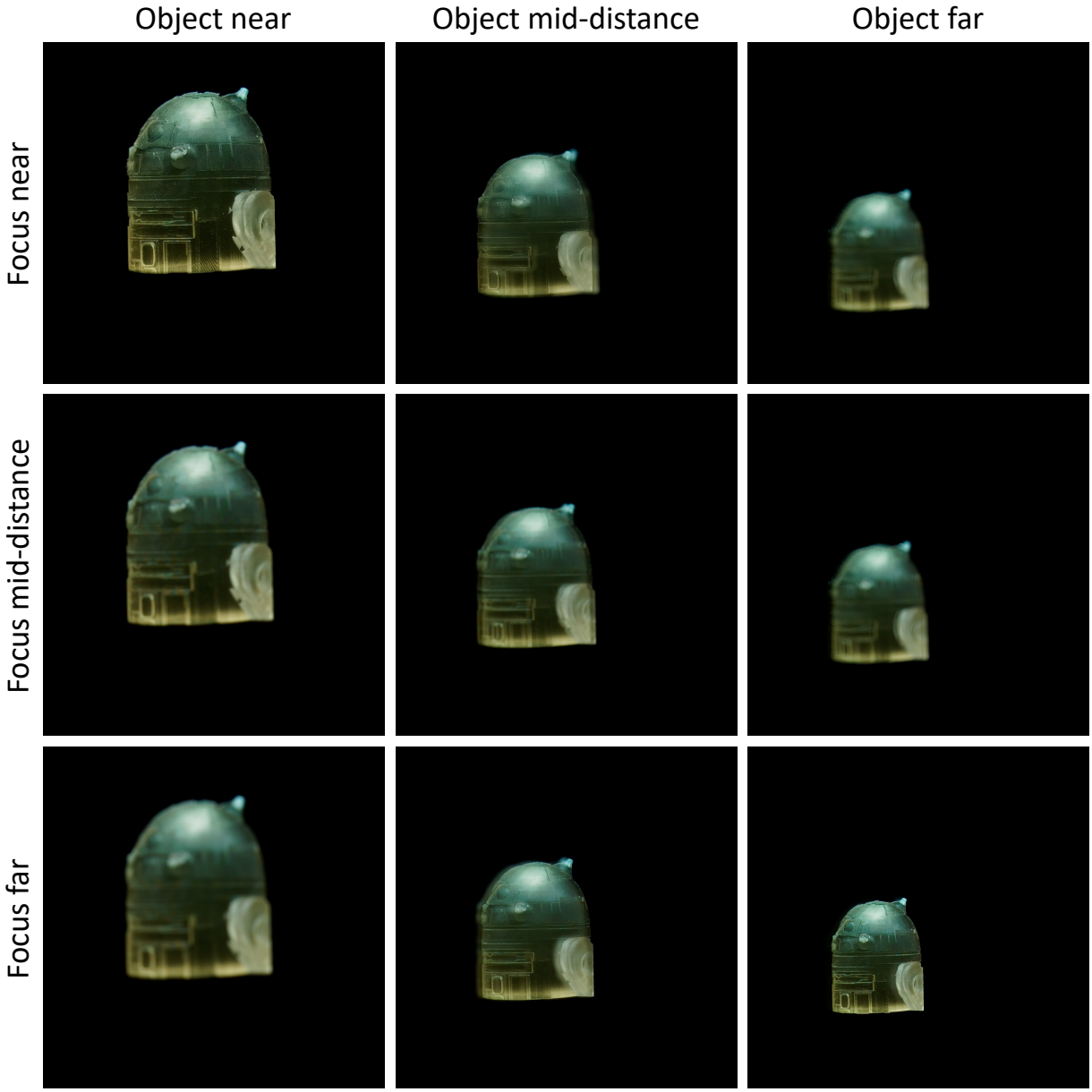


Figure 4.14: Photographs of an object rendered on our display at different depths (columns) while the camera focus was set to one of the three fixed focal distances (rows). The photographs demonstrate the performance of defocus blur due to the multi-focal plane rendering. Note that the subpixel structure, seen in magnification, is not noticeable when the object is seen by the eye. The position of the object changes in the field of view since the camera optical axis was not aligned with the object depth axis.

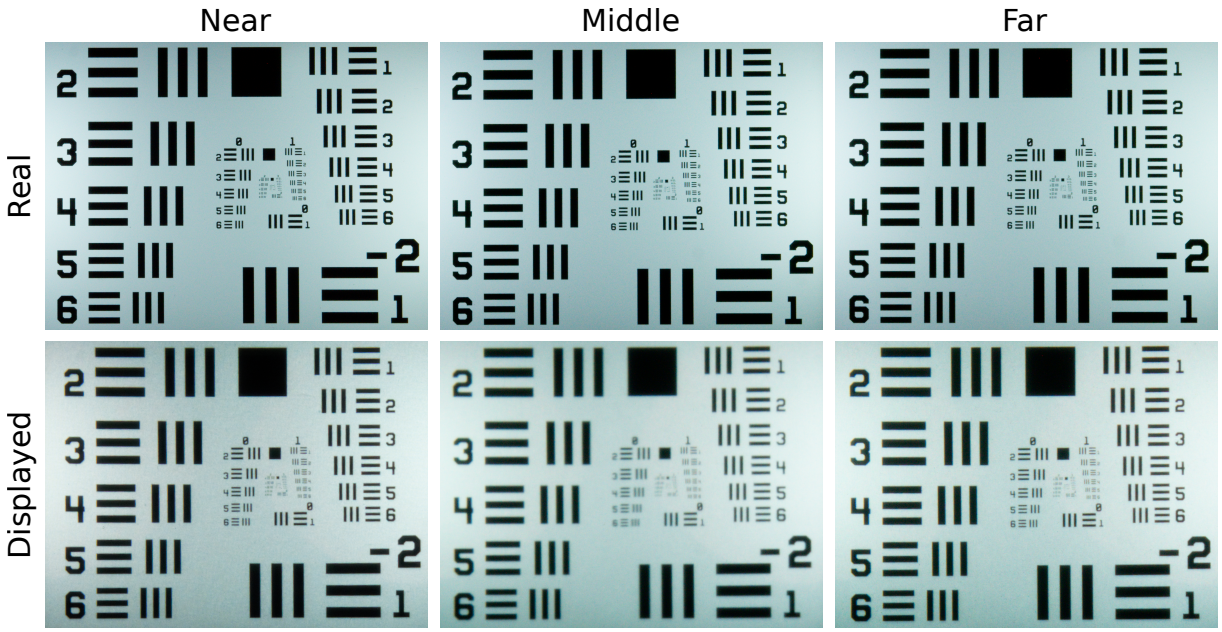


Figure 4.15: Photographs of a physical 2D resolution chart (top) in comparison with its displayed counterpart (bottom) placed at different depths (columns).

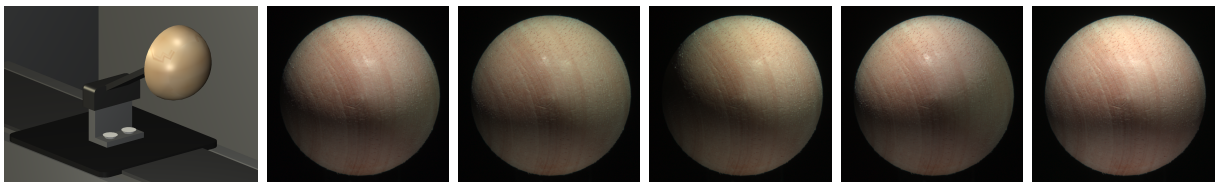


Figure 4.16: The 3D CAD model of the object (left) and its photographs under the five illumination patterns used in the experiment. The base of the wooden hemisphere had a diameter of 47 mm. The photographs have been tone-mapped with $\gamma = 2.2$ to preserve the original colours.

not visible to naked eyes. Assuming that the resolution limit is the point at which the lines blend together and cannot be regarded as separate, our display can reproduce up to 4.0 $^{\text{lp}}/_{\text{mm}}$ at 500 mm (0.58 $^{\text{lp}}/_{\text{arcmin}}$), 2.83 $^{\text{lp}}/_{\text{mm}}$ at 577 mm (0.48 $^{\text{lp}}/_{\text{arcmin}}$) and 4.0 $^{\text{lp}}/_{\text{mm}}$ at 654 mm (0.76 $^{\text{lp}}/_{\text{arcmin}}$). This shows a dip for the middle distance, at which the displayed image is a superimposition of two defocused focal planes (Figure 4.15, 2nd row, 2nd column).

4.5 Visual Turing test

We designed an experiment to test whether participants can distinguish between real and virtual objects shown by our system. The experiment is inspired by the early work of Meyer et al. [126] and many follow-up studies, which attempted to create a system that

passes a *computer graphics Turing test* or *virtual reality visual Turing test*. In contrast to these studies, which have reproduced only 2D images of limited dynamic range, we have created a capture-and-display system that can deliver all necessary visual cues. The secondary objective of our experiment is to test the sensitivity of the visual system to the degradation of different cues (contrast, in this experiment) when all other cues are present. We hope that such data will facilitate understanding of what trade-offs are acceptable in the fidelity of individual display properties, while still delivering highly realistic content — valuable information for building practical display systems.

Stimuli Our test object was a wooden hemisphere (a prop used to teach geometry) that was lightly sanded and stained, but retained the texture of wood and produced an imperfect specular reflection of moderate intensity (refer to Figure 4.16). We chose to work with a simple primitive shape as reconstructing geometry is not the main focus of this project where we treated the ground-truth mesh as given. However, we still need to perform a geometric registration using differentiable rendering and our pipeline can be easily extended to reconstruct unknown shapes (see Section 4.6). As shown on the left of Figure 4.16, the hemisphere was attached to a 3D printed holder (504 mm from the viewer) on the flat side and had its spherical side directed toward the viewer so that it appeared as a sphere to a participant. We selected this object for its simple geometry and complex material and texture properties.

The sphere was illuminated by one of five different light patterns, produced by the RGB LED array on the ceiling of the real-scene box. The patterns were created by switching on a set of 2 LEDs at different positions in the LED array so that the object was illuminated from a slightly different angle each time (while keeping overall brightness approximately the same). To indirectly illuminate the object from the bottom, a piece of white cardboard acting as a diffuse reflector was placed under the object. Different illumination patterns are an important part of our experiment design as they vary the stimulus between the trials so that the participants cannot memorise small differences in appearance across the trials.

A rectangular aperture, made of black cardboard, was placed on the front side of the real-scene box so that only the illuminated hemisphere can be seen. The illumination was reduced to the point at which only the hemisphere can be seen but not any part of the real-scene box (the peak luminance of the object was $2\,\mathrm{cd/m^2}$).

In addition to the *standard* condition, which was our best reproduction of the real object, we created a distorted condition, in which we artificially reduced contrast. The contrast

was reduced by modifying pixel values:

$$I_{\text{mod}}(x, y, c) = \left(\frac{I_{\text{org}}(x, y, c)}{I_{\text{med}}} \right)^{\gamma} I_{\text{med}} \,, \tag{4.4}$$

where I_{med} is the median luminance of the image, I_{org} and I_{mod} are the original and modified images (in linear RGB colour space), and (x, y, c) are pixel and colour channel indices. We determined in a pilot experiment that $\gamma = 0.8$ produced results that were detectable but sufficiently challenging. Not only does this condition let us evaluate the effects of reducing contrast per se, but it also plays important role in our experiment design that it allows us to exclude the possibility that the task given to the participants was too difficult to be feasible (or that they are not paying adequate attention). Consider the case where we reduce presentation time, or luminance, such that none of the participants can detect the real stimulus amongst rendered alternatives. This pattern of data would resemble passing the visual Turing test, but for an entirely trivial reason. Showing that people *can* detect small reductions in contrast with our chosen experiment parameters, however, would demonstrate that they did perform the discrimination task satisfactorily, and so a failure to discriminate in the standard condition can be interpreted at face value.

The object was rendered either on the near focal plane of our display using nearest-neighbour rendering, or on both focal planes using linear depth filtering, as explained in Section 4.3.4. We tested both conditions to understand the importance and challenges of delivering correct focal depth.

Procedure We used a three-interval-forced-choice (3IFC), or odd-one-out, procedure. In each trial, the participant was shown three intervals, for 2 seconds each, from which either two were real and one virtual, or two were virtual and one real. The participant was given the instruction: *You will see three objects, one after another. Select the object that appears different from the two others.* We intentionally avoided asking a question about realism as such a question would be open to subjective interpretations of what *real* looks like, and may lead observers to attend to some aspects of the stimulus while ignoring others. With an oddity task, the observer was instead free to use any aspect of the stimulus to make their judgement, making it a true test of the ability to discriminate real from rendered images. Indeed, the 3IFC task can be considered a very strict test of our display, given that in practical use observers will often evaluate the realism of a rendered scene without the presence of an equivalent real comparison. To avoid after-images causing identical stimuli to appear differently between intervals, we showed a plane with a noise texture of the same average luminance as the object and at the same distance. Our procedure aims to objectively measure whether observers are able to distinguish a real object from a

virtual one without being provided with any training, prior knowledge, or experience for the given task.

The experimental session consisted of 120 trials, which took on average 40 minutes to complete, split into two sessions with a short break. Each participant completed 30 repetitions of each condition. In each trial, we randomly selected either a standard stimulus or one with reduced contrast condition, and presented it using either 2-focal plane rendering, or only on the front focal plane (4 conditions in total). One of five illumination patterns was randomly selected for each trial (the same pattern was used in all three intervals). As the alignment of two focal planes is crucial for the reproduction of focal distance, we displayed an alignment grid (similar to Figure 4.11) before each trial. The participants pressed a key to continue only when a good alignment was achieved. They also had an option to repeat the trial if they were distracted or accidentally moved their heads. Finally, we asked the participants to wear glasses frames with an IR-LED (Figure 4.6), which was used to track and record their head position before and after each interval. We removed the measurements for the trials in which the movement reported by the head tracking was above a certain threshold while multi-focal rendering was used ($\approx 15\%$ of the measurements).

Participants 12 participants (3 females and 9 males, mean age 27.8, SD 4.1 years) completed the experiment. Each participant was screened for normal stereo acuity with the *Titmus fly test* and for normal colour vision with the *Ishihara test*. The participants were instructed to wear their corrective optics. They were compensated for their participation.

Results The participants' answers give us a measure of the probability of selecting the correct interval, $P(\text{correct})$. Since the participants can select the correct answer by chance, we need to correct for that by modelling:

$$\begin{aligned} P(\text{correct}) &= P(\text{chance} \cup \text{detected}) \\ &= P(\text{chance}) + P(\text{detected}) - P(\text{chance})P(\text{detected})\,, \end{aligned} \tag{4.5}$$

where $P(\text{chance}) = 1/3$ in a 3IFC experiment. $P(\text{detected})$ does not depend on the protocol (2IFC or 3IFC) and a zero $P(\text{detected})$ indicates a complete perceptual equivalence between the real and virtual objects. The resulting probability of detecting the interval that appears different, $P(\text{detected})$, is plotted in Figure 4.17 for all 12 participants. As expected, the results show that the reduced contrast increases the probability of detecting the different object, proving that the participants can perform the task. However, for most participants, multi-focal rendering on both planes made it easier to perform the task compared to

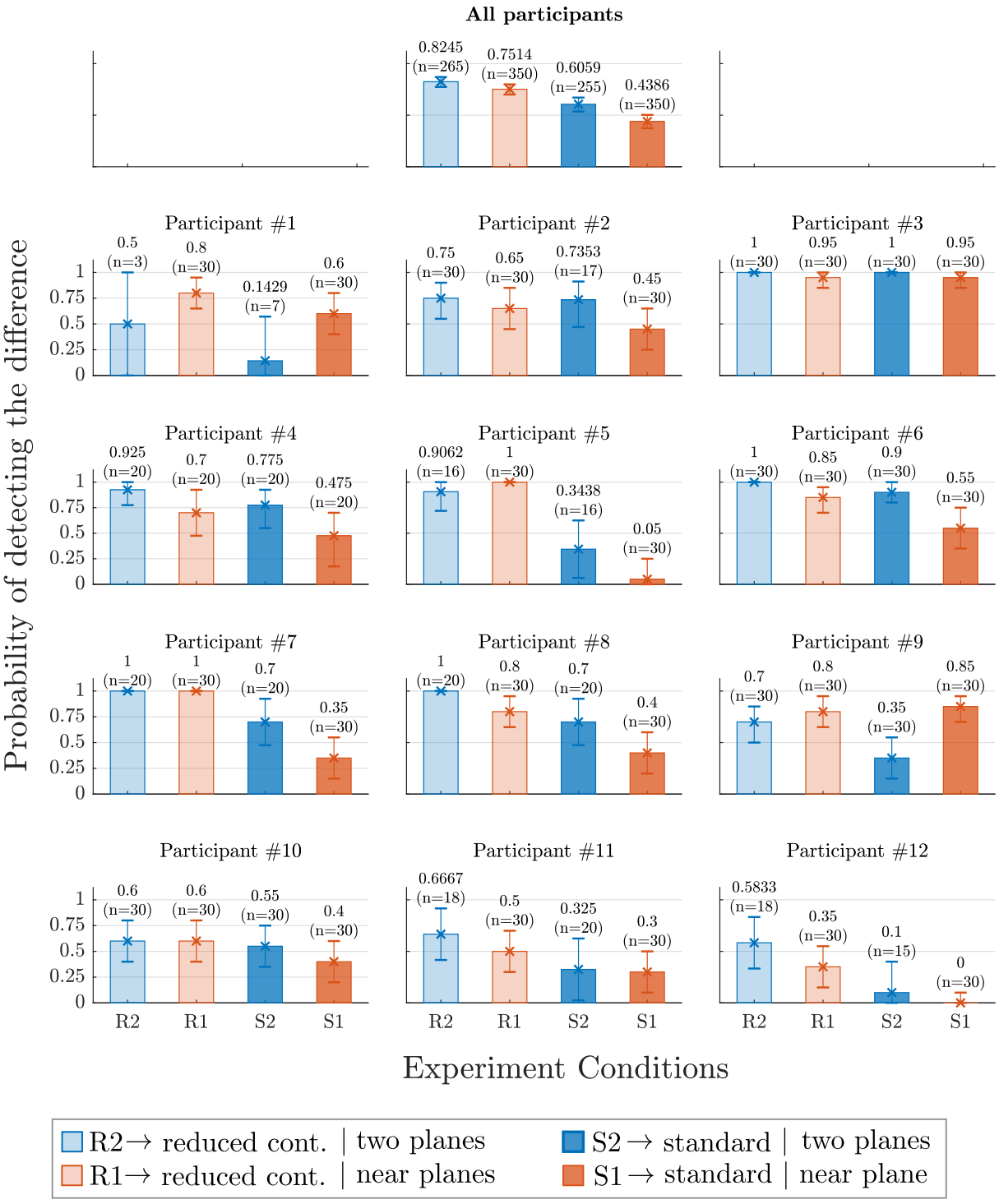


Figure 4.17: The probability of detecting the correct interval (compensated for the guess rate) for each condition and participant.

Table 4.1: The results of the post-experiment questionnaire in which the participants were asked to tick one or more differences they could see between the real and virtual objects.

Options	Votes
different colour	*2*
different sharpness	*6*
different brightness	*4*
different shape or size	*0*
different position or orientation	*0*
different illumination	*1*
one object or other objects appeared flatter	*0*
one object or other objects appeared less shiny	*6*
material appeared different	*1*

rendering only to the near plane[4]. We discuss potential factors that contribute to this outcome in Section 3.2.7.

The results also show large individual differences in detection probabilities across participants. This is most likely because different participants tend to pay attention to different aspects of the stimuli. Meanwhile, although our display resolution did not reach the peak visual acuity (see Section 2.1.1) at the fovea, only participant #3 who reported a 20/20 vision was able to detect tiny differences in resolution and detail when comparing with a physical object. We collected a post-experiment questionnaire to better understand how the participants attempted to identify the different objects. In the questionnaire, we asked: *What made the selected object stand out from the other objects?* and gave a set of possible answers listed in Table 4.1. Table 4.1 shows that among the 12 participants, six participants ticked *sharpness*, which could be a result of the incorrect defocus blur discussed in Section 4.4 or the insufficient resolution (compared to human sensitivity) of our display. The option *one object or other objects appeared less shiny* was also ticked by six participants. This is potentially due to an inaccuracy of our lumigraph synthesis approach, since the shininess of an object is attributed to specular reflections. Four participants selected *brightness* while two selected *colour*, indicating room for improvements in our photometric calibration and colour reproduction. We elaborate on the aforementioned issues in Section 3.2.7. All participants reported that none of the virtual stimuli appeared unnatural when viewed in isolation and if they had not been asked to look for differences from a physical stimulus, they would have deemed the virtual stimuli to be real.

[4]This is with the exception of participant #1 and #9. However, participant #1 only had seven valid trials for standard two-plane rendering, resulting in large confidence intervals. During a post-experiment study, participant #9 indicated that the scene rendered on both focal planes better matched the real scene in terms of colour and contrast. Participant #9 also reported not paying attention to the edges of the object where defocus artefacts can be more salient with multi-focal rendering.

We use our measurements across four conditions to further isolate the factors that contributed to the detection. Assuming that all factors are independent but multiple factors can trigger the detection, we can model the probability of detection as the probability summation:

$$P(\text{detected}) = 1 - (1 - P(\text{f1}))\,(1 - P(\text{f2}))(1 - P(\text{contrast}))\,, \tag{4.6}$$

where $P(\text{f1})$ is the probability of detecting the difference due to single focal plane rendering, $P(\text{f2})$ is the probability of detecting the artefacts due to the limitations of two-focal plane rendering (excluding all factors contributing to $P(\text{f1})$) and $P(\text{contrast})$ is the probability of detecting reduced contrast. We use maximum likelihood estimation to compute those probabilities across all participants and get:

$$P(\text{f1}) = 0.44 \quad P(\text{f2}) = 0.3 \quad P(\text{contrast}) = 0.56\,. \tag{4.7}$$

This shows the observers have 44% chance of detecting the difference between real and virtual objects shown by our display and that two-focal plane rendering increases that chance by 30%[5]. The isolated probability of detecting the contrast reduction by 20% ($\gamma = 0.8$) is 56%, which corresponds to about 1 JND unit (78% for a 2IFC protocol). The reduced contrast conditions serve as an example of a procedure that can be used to scale other relevant "distortions", such as the change of luminance, disparity or black level.

4.6 Discussion

3IFC task The outcome of our experiment, showing that observers can detect the virtual object in 44% of the cases may appear worse than the results reported in other works [126, 10]. However, we need to consider that this is the first time a direct comparison was made between a display and a 3D object seen from a short distance. We also used a much more challenging 3IFC procedure, which removed the subjective assessment of "realism" from our task, and made our test sensitive to very small differences between displayed and real objects. Such differences in certain insignificant aspects (such as viewing angle, object size, position, etc.) do not necessarily degrade the quality of realism for images viewed in isolation.

[5]Note that the probability of detecting limitations of single-focal-plane or two-focal-plane rendering (or both) is: $P(\text{f1} \cup \text{f2}) = P(\text{f1}) + P(\text{f2}) - P(\text{f1})P(\text{f2}) = 0.61$.

Distorted conditions Most visual experiments in graphics either test preference (does A look better than B) or measure similarity to a "reference", which is often obtained from costly renderings, such as path tracing. Both approaches can only be used to determine relative improvements with regard to another rendering method, which may or may not capture the desired visual qualities. Our reduced contrast condition demonstrated how a (simulated) rendering method (or a display limitation) can be directly compared against the ultimate reference of a real-world object. Such absolute measures can tell us that a certain percentage of observers across a population will not notice any observable difference to the real-world object (P(contrast)), while discounting the existing imperfections of the display (P(f1) and P(f2)). We plan to use such a methodology to quantify the importance of various display capabilities, such as the dynamic range, absolute luminance, disparity, focal distance, accommodation, and others.

Eye tracking Multi-focal rendering requires very precise alignment across the focal planes. Effective alignment without uncomfortable restraining of the head position requires active tracking and compensation for the head position. Our IR LED tracker was a first step toward this goal. Latency of the tracking, and the limited refresh rate of the display, did not let us implement active compensation for head movement yet. These are not fundamental limitations of the approach, however.

Multi-focal rendering Our experiment showed a result that rendering on two planes with linear depth filtering made it easier for most observers to detect discrepancies. One explanation could be that while linear depth filtering with the current two-plane separation distance can drive accommodation to the correct depth, it causes an increased defocus blur compared to real scenes. Any multi-focal-plane display with a practical number of focal planes necessarily samples focal depth coarsely, and so most scene points will not coincide precisely with a focal plane. Accommodation can be driven to the appropriate inter-plane distances by linear depth filtering (with plane separations up to and even exceeding that used here, [111]). Yet, at least one image plane must be defocused (because two cannot be focused on simultaneously), resulting in potentially detectable blur compared to a real scene. The results suggest defocus blur plays a more important role in perceptual realism than the accommodation response. As we are relatively insensitive to accommodation state, and it is a weak depth cue, incorrect accommodation is likely to provide weaker cues to realism than blur. Several steps can be taken, however, to reduce this defocus blur compared to the present study. Due to light scattering inside the real-scene box, we used dim illumination, which increased the pupil size, thereby increasing defocus blur. In rendered scenes this problem can be reduced by using higher luminance (including

HDR) scenes. Also, in this study, the far focal plane was far from the stimulus. Either adding an additional intermediate focal plane, or moving the planes to optimal positions with respect to the scene content, would reduce the focal depth inaccuracies that lead to increased defocus blur. Finally, more advanced multi-focal decomposition algorithms may be able to compensate for the loss of high spatial frequencies that characterizes defocus blur [131, 124]. Since a correct stimulus to accommodation is necessary to avoid vergence-accommodation conflicts [65], there is great value in attempting to optimise multi-focal displays for reproducing realism. We hope that our display can be used to explore the trade-offs involved in doing this. For example, does tolerance to incorrect focal depth increase if other aspects of the scene are delivered with very high fidelity?

Reproducible stimuli Our system has several limitations in terms of the stimuli it can reproduce. While our system can synthesize non-Lambertian materials with specular reflections, the quality may not reach the level of perceptual equivalence, as indicated by our post-experiment questionnaire. Specular highlights are sensitively dependent on viewing positions, making them difficult to be reproduced as it is unlikely that our data camera perfectly overlaps with the observer's eye position. We anticipate that such inaccuracies can be reduced by capturing more light field views or incorporating more advanced neural scene representations [179, 164, 60]. We did not explore this direction as training and convergence of scene-representation networks with large-size data (8k images in our case) remained an actively studied problem at the time of our work. In the future, we plan to evaluate various view synthesis approaches with our apparatus in terms of perceptual realism, whereas existing works only focus on photorealism. Our system is also currently limited to simple or known geometry. Nonetheless, this is not a fundamental limitation of our approach — we can modify the loss function in Equation 4.1 such that it not only optimises for a spatial transform but also for a per-vertex deformation to fit an unknown geometry. However, this approach also requires capturing many more views around the object. Since 3D reconstruction is not the main focus of this work, we chose to work with known geometry and a horizontal light field. Our rendering method is currently unable to reproduce edge occlusions of objects at different depths without introducing visible artefacts. Our intention is to test more advanced multi-plane rendering methods [131, 124, 196] in the future. Our display has an advantage over the previously built multi-focal plane displays in that it can reproduce a much higher dynamic range, which gives more flexibility in optimising for multi-plane decomposition (for example, greater headroom for compensating for the loss of high spatial frequencies). In addition, although the resolution of our display is much higher than that found in the previous work [163], it is still lower than the levels required for a perceptual match, as reported by Masaoka et al. [119]. Achieving the highest resolution reported in their paper (120 cpd)

would require tripling the resolution of our LCD panels. This is currently impossible when using off-the-shelf components. As above, it will be interesting to explore whether tolerance to lower-than-optimal resolution is increased when other aspects of the scene are delivered with high fidelity. Finally, our colour calibration currently relies on CIE XYZ 1931 colour matching functions, which are known to be inaccurate for short wavelengths [27]. It also did not account for the contribution of rods to colour perception or individual differences. Better colour matching may require capturing multispectral images and individual corrections to compensate for the differences in cone sensitivities.

4.7 Summary

The main objective of our work is to build an end-to-end system that can acquire a small 3-dimensional object and reproduce it faithfully with all the necessary visual cues on a display. Being able to do so is an important step for perceptually realistic graphics, in which the depicted imagery is indistinguishable from the real world. A direct comparison with real-world objects lets us better understand the limitations of not only the visual system but also those of display technologies, 3D representations, and rendering techniques. For example, we found that defocus blur could play a more important role than accommodation response in perceptual realism, together with the need for accurate view-point tracking, as one of the main limitations of multi-focal plane displays.

We demonstrate that the first iteration of our HDR-MF-S display can deliver virtual imagery that is in only 44% of the cases detected as different from its real-world counterpart. This result was obtained when asking the question *is it different?* rather than *is it real?*, making the task more objective but also requiring higher accuracy from a display system. This work is also the first attempt to reproduce a 3D object at a short distance, with an essential set of visual cues. Finally, our experiment design with a "control" distorted condition ensured that the participants were correctly completing the task.

The display is a platform for a wide range of experimental studies, in which both faithful reproductions of all visual cues and comparison to reality are paramount. For example, it can be used for studies on gloss and material perception, physics-based rendering, global illumination, tone mapping, view synthesis, augmented & mixed reality, and many more. All these studies can take advantage of full control over each display capability dimension, such as dynamic range or luminance. The displays can also simulate a wide range of see-through AR displays, by using a real-scene box as a real environment and offering a much higher dynamic range and peak luminance than that of most head-mounted displays.

Chapter 5

Conclusion and Future Work

In this dissertation, we provide a comprehensive unified overview of perceptually realistic graphics (PRG), an emerging field gaining increasing attention and impact within the graphics community, along with several proposed advancements in the field. We demonstrate that perceptual realism can be approached by maximising the quality of essential visual cues rather than reproducing a physically correct distribution of light. From an application perspective, we believe that PRG has the potential to revolutionise the way people entertain and interact with the digital world. Although achieving this would require further research on the interaction between digital and physical content, realism remains an essential and enduring requirement for the experience of such applications. From a research perspective, PRG opens the door to several new branches in the study of computer graphics, as traditional graphics was primarily focused on photorealism and rendering on a single image plane. As demonstrated by this dissertation, human perception must be in the loop in the design of new approaches throughout the PRG pipeline. Traditional algorithms for 3D scene acquisition, representation, and rendering must also be adapted to meet more stringent visual requirements and accommodate novel 3D display architectures.

Given the nascent nature of perceptually realistic graphics, this dissertation represents merely the beginning of our exploration of the field. We anticipate a multitude of research questions to emerge in the future. To start, this dissertation primarily focused on maximising the perceived quality of static 3D scenes, with limited exploration of visual requirements related to temporal aspects. Acquisition, representation, and rendering of dynamic scenes must meet stricter requirements to achieve perceptual realism in motion. Meanwhile, as briefly discussed in Section 2.3, perceptual realism requires optimising scene and rendering parameters with respect to real-world scenes and human vision, rather than merely images. This requires integrating accurate simulation of cameras, displays, and

human vision into the differentiable graphics pipeline. For example, artefacts such as aberration (geometric and chromatic) and blur introduced by the camera must be modelled and corrected in the optimisation loop. Factors such as pupil size, lens, and chromatic aberration of human eyes also influence the formation of the retinal image. Furthermore, while traditional scene manipulation algorithms were designed to process scene content in digital forms, new methods are required in PRG to directly manipulate the perception of physical scenes and interactions between real and virtual objects. For example, in Chapter 4, we demonstrate that our HDR-MF-S display is able to alter the colour and brightness of the physical objects by superimposing a virtual mask. However, reducing light transmission in additive displays to darken the physical objects or simulating a darkening effect remains a longstanding challenge. More advanced effects such as relighting, recolouring, removal, and occlusion of physical objects are also underdeveloped. Finally, an effective 3D quality metric (parallel to image quality metrics such as SSIM [170] and PSNR) is needed to quantitatively evaluate the qualities of virtual 3D scenes rendered on a 3D display, as the visual Turing test can be laborious and unscalable. Such a metric may also consider the tradeoffs associated with the importance of individual visual cues. In Chapter 2, we provide a discussion on a mixture of qualitative and quantitative criteria to achieve perceptual realism for the worst-case scenario, as we consider the best capabilities of human vision in its limit. However, factors such as contrast sensitivity, visual acuity, and motion perception can be scene-dependent. A desirable 3D quality metric is expected to adapt to varying viewing conditions and scene content, enabling the identification of the minimum requirements on individual visual cues for perceptual realism.

References

[1] Kaan Akşit, Ward Lopes, Jonghyun Kim, Peter Shirley, and David Luebke. Near-eye varifocal augmented reality display using see-through screens. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130892. URL https://doi.org/10.1145/3130800.3130892.

[2] Kurt Akeley, Simon J. Watt, Ahna Reza Girshick, and Martin S. Banks. A stereo display prototype with multiple focal distances. *ACM Trans. Graph.*, 23(3):804–813, August 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015804. URL https://doi.org/10.1145/1015706.1015804.

[3] AliceVision. Meshroom: A 3D reconstruction software., 2018. URL https://github.com/alicevision/meshroom.

[4] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[5] Martin S. Banks, David M. Hoffman, Joohwan Kim, and Gordon Wetzstein. 3d displays. *Annual Review of Vision Science*, 2(1):397–435, 2016. doi: 10.1146/annurev-vision-082114-035800. URL https://doi.org/10.1146/annurev-vision-082114-035800. PMID: 28532351.

[6] J. L. Barbur and A. Stockman. Photopic, mesopic and scotopic vision and changes in visual performance. *Encyclopedia of the Eye*, 3:323–331, 2010.

[7] Stephen A. Benton and Jr. Bove, V. Michael. *Holographic Imaging*. Wiley-Interscience, USA, 2008. ISBN 047006806X.

[8] Hans I. Bjelkhagen and Evangelos Mirlis. Color holography to produce highly realistic three-dimensional images. *Appl. Opt.*, 47(4):A123–A133, Feb 2008. doi: 10.1364/AO.47.00A123. URL http://opg.optica.org/ao/abstract.cfm?URI=ao-47-4-A123.

[9] Randolph Blake. A neural theory of binocular rivalry. *Psychological review*, 96(1): 145, 1989.

[10] M. Borg, S. S. Johansen, D. L. Thomsen, and M. Kraus. Practical implementation of a graphics turing test. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Charless Fowlkes, Sen Wang, Min-Hyung Choi, Stephan Mantler, Jürgen Schulze, Daniel Acevedo, Klaus Mueller, and Michael Papka, editors, *Advances in Visual Computing*, pages 305–313, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33191-6. URL https://doi.org/10.1007/978-3-642-33191-6_30.

[11] David Buckley and John P Frisby. Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges. *Vision research*, 33(7):919–933, 1993.

[12] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. SIGGRAPH '01, page 425–432, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 158113374X. doi: 10.1145/383259.383309. URL https://doi.org/10.1145/383259.383309.

[13] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[14] Dingcai Cao, Joel Pokorny, Vivianne C Smith, and Andrew J Zele. Rod contributions to color perception: linear with rod contrast. *Vision Research*, 48(26):2586–92, nov 2008. ISSN 1878-5646. doi: 10.1016/j.visres.2008.05.001.

[15] Praneeth Chakravarthula, Ethan Tseng, Tarun Srivastava, Henry Fuchs, and Felix Heide. Learned hardware-in-the-loop phase retrieval for holographic near-eye displays. *ACM Transactions on Graphics (TOG)*, 39(6):186, 2020. URL https://doi.org/10.1145/3414685.3417846.

[16] Praneeth Chakravarthula, Florian Schiffers, Oliver Cossairt, Ethan Tseng, Seung-Hwan Baek, and Felix Heide. Differentiable cameras and displays. In *ACM SIGGRAPH 2022 Courses*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393621. doi: 10.1145/3532720.3535685. URL https://doi.org/10.1145/3532720.3535685.

[17] Alan Chalmers and Andrej Ferko. Levels of realism: From virtual reality to real virtuality. In *Proceedings of the 24th Spring Conference on Computer Graphics*, pages 19–25, 2008.

[18] Jen-Hao Rick Chang, B. V. K. Vijaya Kumar, and Aswin C. Sankaranarayanan. Towards multifocal displays with dense focal stacks. *ACM Trans. Graph.*, 37(6), December 2018. ISSN 0730-0301. doi: 10.1145/3272127.3275015. URL https://doi.org/10.1145/3272127.3275015.

[19] Alexandre Chapiro, Robin Atkins, and Scott Daly. A luminance-aware model of judder perception. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 38(5), 2019. ISSN 0730-0301. doi: 10.1145/3338696. URL https://doi.org/10.1145/3338696.

[20] Anpei Chen, Minye Wu, Yingliang Zhang, Nianyi Li, Jie Lu, Shenghua Gao, and Jingyi Yu. Deep surface light fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(1):1–17, Jul 2018. ISSN 2577-6193. doi: 10.1145/3203192. URL http://dx.doi.org/10.1145/3203192.

[21] Fuhao Chen, Chengfeng Qiu, and Zhaojun Liu. Investigation of autostereoscopic displays based on various display technologies. *Nanomaterials*, 12(3), 2022. ISSN 2079-4991. doi: 10.3390/nano12030429. URL https://www.mdpi.com/2079-4991/12/3/429.

[22] Huang-Ming Philip Chen, Jhou-Pu Yang, Hao-Ting Yen, Zheng-Ning Hsu, Yuge Huang, and Shin-Tson Wu. Pursuing high quality phase-only liquid crystal on silicon (lcos) devices. *Applied Sciences*, 8(11), 2018. ISSN 2076-3417. doi: 10.3390/app8112323. URL https://www.mdpi.com/2076-3417/8/11/2323.

[23] Suyeon Choi, Manu Gopakumar, Yifan Peng, Jonghyun Kim, and Gordon Wetzstein. Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays. *ACM Trans. Graph. (SIGGRAPH Asia)*, 2021.

[24] CIE. Colorimetry, cie publication no. 15. Technical report, 2004.

[25] CIE. Fundamental chromaticity diagram with physiological axes–part 1. *Commission Internationale de l'Eclairage*, pages 170–1, 2006.

[26] CUPC CIE. Commission internationale de l'eclairage proceedings, 1931. *Cambridge University, Cambridge*, 1932.

[27] CIE170-1:2006. Fundamental chromacity diagram with psychological axes - part 1. Technical report, Central Bureau of the Commission Internationale de l' Éclairage, 2016. URL http://www.cie.co.at/publications/fundamental-chromaticity-diagram-physiological-axes-part-1.

[28] J. Cutting and P. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein and S. Rogers, editors, *Perception of Space and Motion (Handbook Of Perception And Cognition)*, pages 69–117. Academic Press, 1995.

[29] Gislin Dagnelie. *Visual prosthetics: Physiology, bioengineering, rehabilitation*. 01 2011. doi: 10.1007/978-1-4419-0754-7.

[30] Paul Debevec, Yizhou Yu, and George Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. In *Eurographics Workshop on Rendering Techniques*, pages 105–116. Springer, 1998.

[31] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of SIGGRAPH '97*, pages 369–378, Los Angeles, CA. USA, 1997. ACM Press. ISBN 0897918967.

[32] Piotr Didyk, Elmar Eisemann, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. Apparent display resolution enhancement for moving images. In *ACM SIGGRAPH 2010 Papers*, SIGGRAPH '10, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450302104. doi: 10.1145/1833349.1778850. URL `https://doi.org/10.1145/1833349.1778850`.

[33] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. A perceptual model for disparity. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2011. doi: 10.1145/1964921.1964991. URL `https://doi.org/10.1145/1964921.1964991`.

[34] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. Apparent stereo: the Cornsweet illusion can enhance perceived depth. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder, editors, *Human Vision and Electronic Imaging XVII*, volume 8291, pages 180 – 191. International Society for Optics and Photonics, SPIE, 2012. doi: 10.1117/12.907612. URL `https://doi.org/10.1117/12.907612`.

[35] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, Hans-Peter Seidel, and Wojciech Matusik. A luminance-contrast-aware disparity model and applications. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 31(6), 2012.

[36] D. Dunn, C. Tippets, K. Torell, P. Kellnhofer, K. Akşit, P. Didyk, K. Myszkowski, D. Luebke, and H. Fuchs. Wide field of view varifocal near-eye display using see-through deformable membrane mirrors. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1322–1331, 2017. doi: 10.1109/TVCG.2017.2657058.

[37] Felice A Dunn and Fred Rieke. Single-photon absorptions evoke synaptic depression in the retina to extend the operational range of rod vision. *Neuron*, 57(6):894–904, mar 2008. ISSN 1097-4199. doi: 10.1016/j.neuron.2008.01.031.

[38] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics*, 21(3):257–266, jul 2002. ISSN 07300301. doi: 10.1145/566654.566574.

[39] G. Eilertsen, R. K. Mantiuk, and J. Unger. A comparative review of tone-mapping algorithms for high dynamic range video. *Computer Graphics Forum*, 36(2):565–592, may 2017. ISSN 01677055. doi: 10.1111/cgf.13148. URL `http://doi.wiley.com/10.1111/cgf.13148`.

[40] Gabriel Eilertsen, Robert Wanat, Rafał K Mantiuk, and Jonas Unger. Evaluation of Tone Mapping Operators for HDR-Video. *Computer Graphics Forum*, 32(7):275–284, oct 2013. ISSN 01677055. doi: 10.1111/cgf.12235. URL `http://doi.wiley.com/10.1111/cgf.12235`.

[41] Gabriel Eilertsen, Rafał K. Mantiuk, and Jonas Unger. Real-time noise-aware tone mapping. *ACM Transactions on Graphics*, 34(6):1–15, oct 2015. ISSN 07300301. doi: 10.1145/2816795.2818092.

[42] W. Epstein and S. Rogers. Handbook of perception and cognition. volume 5: Perception of space and motion., 1995.

[43] A. Erickson, K. Kim, G. Bruder, and G. F. Welch. Effects of dark mode graphics on visual acuity and fatigue with virtual reality head-mounted displays. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 434–442, 2020.

[44] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. In *ACM SIGGRAPH 2008*, pages 1–10. ACM, 2008. URL `http://portal.acm.org/citation.cfm?id=1399504.1360666`.

[45] G. E. Favalora, J. Napoli, D. M. Hall, R. K. Dorval, M. Giovinco, M. J. Richmond, and W. S. Chun. 100-million-voxel volumetric display. In D. G. Hopper, editor, *Cockpit Displays IX: Displays for Defense Applications*, volume 4712 of , pages 300–312, August 2002. doi: 10.1117/12.480930.

[46] James A Ferwerda. Three varieties of realism in computer graphics. In *Human vision and electronic imaging viii*, volume 5007, pages 290–297. SPIE, 2003.

[47] James A. Ferwerda and Chester F. Carlson. Fundamentals of color science. In *ACM SIGGRAPH 2020 Courses*, SIGGRAPH '20, New York, NY, USA, 2020. Association

for Computing Machinery. ISBN 9781450379724. doi: 10.1145/3388769.3407479. URL https://doi.org/10.1145/3388769.3407479.

[48] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent, 2019.

[49] John P Frisby and John E W Mayhew. Contrast sensitivity function for stereopsis. *Perception*, 7(4):423–429, 1978. doi: 10.1068/p070423. URL https://doi.org/10.1068/p070423. PMID: 704272.

[50] John P Frisby, David Buckley, and Janet M Horsman. Integration of stereo, texture, and outline cues during pinhole viewing of real ridge-shaped objects and stereograms of ridges. *Perception*, 24(2):181–198, 1995.

[51] Chenyang Fu, Changjun Li, Guihua Cui, M. Ronnier Luo, Robert W. G. Hunt, and Michael R. Pointer. An investigation of colour appearance for unrelated colours under photopic and mesopic vision. *Color Research & Application*, 37(4):238–254, 2012. ISSN 0361-2317. doi: 10.1002/col.20691.

[52] Philippe Fuchs. *Virtual reality headsets-a theoretical and pragmatic approach*. CRC Press, 2017.

[53] D. Gabor. A New Microscopic Principle. *Nature*, 161(4098):777–778, May 1948. ISSN 0028-0836, 1476-4687. doi: 10.1038/161777a0. URL https://www.nature.com/articles/161777a0.

[54] B Y M A Georgeson and G D Sullivan. Contrast constancy: deblurring in human vision by spatial frequency channels. *The Journal of Physiology*, 252(3):627–656, 1975.

[55] M A Georgeson and G D Sullivan. Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol.*, 252(3):627–656, nov 1975.

[56] Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. *ACM SIGGRAPH Computer Graphics*, 18(3):213–222, jul 1984. ISSN 0097-8930. doi: 10.1145/964965.808601. URL https://dl.acm.org/doi/10.1145/964965.808601.

[57] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proc. of SIGGRAPH '96*, pages 43–54. ACM Press, 1996. ISBN 0897917464. doi: 10.1145/237170.237200. URL http://portal.acm.org/citation.cfm?doid=237170.237200.

[58] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. Optimal hdr reconstruction with linear digital cameras. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 215–222. IEEE, 2010.

[59] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[60] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18409–18418, June 2022.

[61] Param Hanji, Fangcheng Zhong, and Rafał K. Mantiuk. Noise-aware merging of high dynamic range image stacks without camera calibration. In *Advances in Image Manipulation (ECCV workshop)*, pages 376–391. Springer, 2020. URL `http://www.cl.cam.ac.uk/research/rainbow/projects/noise-aware-merging/`.

[62] S.W. Hasinoff, F. Durand, and W.T. Freeman. Noise-optimal capture for high dynamic range photography. In *CVPR*, pages 553–560. IEEE, 2010.

[63] Robert T. Held and Martin S. Banks. Misperceptions in stereoscopic displays: A vision science perspective. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, APGV '08, page 23–32, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939814. doi: 10.1145/1394281.1394285. URL `https://doi.org/10.1145/1394281.1394285`.

[64] David M. Hoffman and Grace Lee. Temporal Requirements for VR Displays to Create a More Comfortable and Immersive Visual Experience. *Information Display*, 35(2):9–39, mar 2019. ISSN 0362-0972. doi: 10.1002/msid.1018. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/msid.1018`.

[65] David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8(3):33, mar 2008. ISSN 1534-7362. doi: 10.1167/8.3.33. URL `http://jov.arvojournals.org/article.aspx?doi=10.1167/8.3.33`.

[66] Ian P. Howard and Brian J. Rogers. *Binocular vision and stereopsis*. Number no. 29 in Oxford psychology series. Oxford University Press, New York, 1995. ISBN 9780195084764.

[67] HTC. Vive flow, 2021. URL https://www.vive.com/us/product/.

[68] Xinda Hu and Hong Hua. High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics. *Opt. Express*, 22(11):13896–13903, Jun 2014. doi: 10.1364/OE.22.013896. URL http://opg.optica.org/oe/abstract.cfm?URI=oe-22-11-13896.

[69] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: Immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Trans. Graph.*, 34(4), July 2015. ISSN 0730-0301. doi: 10.1145/2766922. URL https://doi.org/10.1145/2766922.

[70] Fu-Chung Huang, Dawid Pajak, Jonghyun Kim, Jan Kautz, and David Luebke. Mixed-primary factorization for dual-frame computational displays. *ACM Trans. Graph.*, 36(4), jul 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073654. URL https://doi.org/10.1145/3072959.3073654.

[71] Kuo-Chung Huang, Yi-Heng Chou, Lang-chin Lin, Hoang Yan Lin, Fu-Hao Chen, Ching-Chiu Liao, Yi-Han Chen, Kuen Lee, and Wan-Hsuan Hsu. A study of optimal viewing distance in a parallax barrier 3d display. *Journal of the Society for Information Display*, 21(6):263–270, 2013. doi: https://doi.org/10.1002/jsid.172. URL https://sid.onlinelibrary.wiley.com/doi/abs/10.1002/jsid.172.

[72] Tianqi Huang, Boxuan Han, Xinran Zhang, and Hongen Liao. High-performance autostereoscopic display based on the lenticular tracking method. *Opt. Express*, 27 (15):20421–20434, Jul 2019. doi: 10.1364/OE.27.020421. URL http://opg.optica.org/oe/abstract.cfm?URI=oe-27-15-20421.

[73] Shigeru Ichihara, Norimichi Kitagawa, and Hiromi Akutsu. Contrast and depth perception: Effects of texture contrast and area contrast. *Perception*, 36(5):686–695, 2007. doi: 10.1068/p5696. URL https://doi.org/10.1068/p5696. PMID: 17624115.

[74] Frederic E. Ives. A novel stereogram. *Journal of the Franklin Institute*, 153(1):51–52, 1902. ISSN 0016-0032. doi: https://doi.org/10.1016/S0016-0032(02)90195-X. URL https://www.sciencedirect.com/science/article/pii/S001600320290195X.

[75] Shahram Izadi, Ayush Bhandari, Achuta Kadambi, and Ramesh Raskar. 3d imaging with time of flight cameras: Theory, algorithms and applications. In *ACM SIGGRAPH 2014 Courses*, SIGGRAPH '14, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329620. doi: 10.1145/2614028.2615433. URL https://doi.org/10.1145/2614028.2615433.

[76] Jochen Jacobs, Xi Wang, and Marc Alexa. Keep it simple: Depth-based dynamic adjustment of rendering for head-mounted displays decreases visual comfort. *ACM Trans. Appl. Percept.*, 16(3), 2019.

[77] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[78] Isaac Kauvar, Samuel J. Yang, Liang Shi, Ian McDowall, and Gordon Wetzstein. Adaptive color display via perceptually-driven factored spectral projection. *ACM Trans. Graph.*, 34(6), oct 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818070. URL `https://doi.org/10.1145/2816795.2818070`.

[79] Petr Kellnhofer, Tobias Ritschel, Peter Vangorp, Karol Myszkowski, and Hans-Peter Seidel. Stereo day-for-night: Retargeting disparity for scotopic vision. *ACM Trans. Appl. Percept.*, 11(3), 2014.

[80] Petr Kellnhofer, Piotr Didyk, Karol Myszkowski, Mohamed M. Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. Gazestereo3d: Seamless disparity manipulations. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 35(4), 2016. doi: 10.1145/2897824.2925866. URL `http://dx.doi.org/10.1145/2897824.2925866`.

[81] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *CVPR*, 2021.

[82] Dae-Sik Kim, Sergey Shestak, Kyung-Hoon Cha, Sang-Moo Park, and Seon-Deok Hwang. Time-sequential autostereoscopic OLED display with segmented scanning parallax barrier. In Bahram Javidi, Jung-Young Son, Manuel Martinez-Corral, Fumio Okano, and Wolfgang Osten, editors, *Three-Dimensional Imaging, Visualization, and Display 2009*, volume 7329, pages 236 – 242. International Society for Optics and Photonics, SPIE, 2009. doi: 10.1117/12.820313. URL `https://doi.org/10.1117/12.820313`.

[83] Fred Kingdom and Bernard Moulden. Border effects on brightness: A review of findings, models and issues. *Spatial Vision*, 3(4):225–262, jan 1988. ISSN 01691015. doi: 10.1163/156856888X00140.

[84] Frederick A. A. Kingdom and Lauren Libenson. Dichoptic color saturation mixture: Binocular luminance contrast promotes perceptual averaging. *Journal of Vision*, 15 (5):2, apr 2015. ISSN 1534-7362. doi: 10.1167/15.5.2.

[85] Victor Klymenko, Robert W Verona, Howard H Beasley, and John S Martin. Convergent and divergent viewing affect luning, visual thresholds, and field-of-view

fragmentation in partial binocular overlap helmet-mounted displays. In *Helmet-and Head-Mounted Displays and Symbology Design Requirements*, volume 2218, pages 82–97. International Society for Optics and Photonics, 1994.

[86] J.J. Kulikowski. Effective contrast constancy and linearity of contrast sensation. *Vision Research*, 16(12):1419–1431, jan 1976. ISSN 00426989. doi: 10.1016/0042-6989(76)90161-9. URL http://www.sciencedirect.com/science/article/pii/0042698976901619.

[87] Yoshihiko Kuroki, Tomohiro Nishi, Seiji Kobayashi, Hideki Oyaizu, and Shinichi Yoshimura. A psychophysical study of improvements in motion-image quality by using high frame rates. *Journal of the Society for Information Display*, 15(1):61–68, 2007.

[88] Kiriakos N. Kutulakos and Samuel W. Hasinoff. Focal stack photography: High-performance photography with a conventional camera. In *In IAPR Machine Vision Appl*, 2009.

[89] Youngshin Kwak, Lindsay William MacDonald, and M. Ronnier Luo. Mesopic color appearance. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging VIII*, volume 5007, pages 161 – 169. International Society for Optics and Photonics, SPIE, 2003. doi: 10.1117/12.477371. URL https://doi.org/10.1117/12.477371.

[90] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross. Nonlinear disparity mapping for stereoscopic 3d. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 29(4), 2010. ISSN 0730-0301. doi: 10.1145/1778765.1778812. URL https://doi.org/10.1145/1778765.1778812.

[91] Douglas Lanman, Matthew Hirsch, Yunhee Kim, and Ramesh Raskar. Content-adaptive parallax barriers: Optimizing dual-layer 3d displays using low-rank light field factorization. 29(6), dec 2010. ISSN 0730-0301. doi: 10.1145/1882261.1866164. URL https://doi.org/10.1145/1882261.1866164.

[92] Douglas Lanman, Gordon Wetzstein, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar. Polarization fields: Dynamic light field display using multi-layer lcds, 2011.

[93] Gordon E. Legge and John M. Foley. Contrast masking in human vision. *J. Opt. Soc. Am.*, 70(12):1458–1471, Dec 1980. doi: 10.1364/JOSA.70.001458. URL http://opg.optica.org/abstract.cfm?URI=josa-70-12-1458.

[94] Gordon E Legge and Gary S Rubin. Binocular interactions in suprathreshold contrast perception. *Attention, Perception, & Psychophysics*, 30(1):49–61, 1981.

[95] Jed Lengyel. The convergence of graphics and vision. *Computer*, 31(7):46–53, 1998.

[96] Willem JM Levelt. Binocular brightness averaging and contour information. *British journal of psychology*, 56(1):1–13, 1965.

[97] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):222:1–222:11, 2018.

[98] Zhong Li, Liangchen Song, Celong Liu, Junsong Yuan, and Yi Xu. NeuLF: Efficient Novel View Synthesis with Neural 4D Light Field. In Abhijeet Ghosh and Li-Yi Wei, editors, *Eurographics Symposium on Rendering*. The Eurographics Association, 2022. ISBN 978-3-03868-187-8. doi: 10.2312/sr.20221156.

[99] G. Lippmann. Épreuves réversibles donnant la sensation du relief. *Journal de Physique Théorique et Appliquée*, 7(1):821–825, 1908. ISSN 0368-3893. doi: 10.1051/jphystap:019080070082100. URL `http://www.edpsciences.org/10.1051/jphystap:019080070082100`.

[100] Jingyu Liu*, Fangcheng Zhong*, Claire Mantel, Søren Forchhammer, and Rafał K. Mantiuk. Chapter 17 - computational 3d displays. In Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar, editors, *Immersive Video Technologies*, pages 469–500. Academic Press, 2023. ISBN 978-0-323-91755-1. doi: https://doi.org/10.1016/B978-0-32-391755-1.00023-7. URL `https://www.sciencedirect.com/science/article/pii/B9780323917551000237`.

[101] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020.

[102] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[103] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL `https://arxiv.org/abs/1904.01786`.

[104] Margaret S. Livingstone and David H. Hubel. Stereopsis and positional acuity under dark adaptation. *Vision Research*, 34(6):799–802, 1994.

[105] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019.

[106] Guillaume Loubet, Nicolas Holzschuch, and Wenzel Jakob. Reparameterizing discontinuous integrands for differentiable rendering. *ACM Trans. Graph.*, 38 (6), nov 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356510. URL `https://doi.org/10.1145/3355089.3356510`.

[107] Gordon D. Love, David M. Hoffman, Philip J.W. Hands, James Gao, Andrew K. Kirby, and Martin S. Banks. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Opt. Express*, 17(18):15716–15725, Aug 2009. doi: 10.1364/OE.17.015716. URL `http://www.opticsexpress.org/abstract.cfm?URI=oe-17-18-15716`.

[108] Mark E. Lucente. Electronic holographic displays: 20 years of interactive spatial imaging. In Cranton Wayne Fihn Mark Chen, Janglin, editor, *Handbook of Visual Display Technology*, pages 2721–2740. Springer-Verlag Berlin Heidelberg, Bristol, 2012.

[109] Thomas Luft, Carsten Colditz, and Oliver Deussen. Image enhancement by unsharp masking the depth buffer. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 25(3):1206–1213, 2006. ISSN 0730-0301. doi: 10.1145/1141911.1142016. URL `https://doi.org/10.1145/1141911.1142016`.

[110] Ming Ronnier Luo and Changjun Li. *CIECAM02 and Its Recent Developments*, pages 19–58. Springer New York, New York, NY, 2013. ISBN 978-1-4419-6190-7. doi: 10.1007/978-1-4419-6190-7_2. URL `https://doi.org/10.1007/978-1-4419-6190-7_2`.

[111] Kevin J MacKenzie, David M Hoffman, and Simon J Watt. Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control. *Journal of vision*, 10(8):22, jan 2010. ISSN 1534-7362. doi: 10.1167/10.8.22. URL `http://www.ncbi.nlm.nih.gov/pubmed/20884597`.

[112] Kevin J. MacKenzie, Ruth A. Dickson, and Simon J. Watt. Vergence and accommodation to multiple-image-plane stereoscopic displays: "real world" responses with practical image-plane separations? *Journal of Electronic Imaging*, 21(1):011002, feb 2012. ISSN 1017-9909. doi: 10.1117/1.JEI.21.1.011002. URL `http://electronicimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JEI.21.1.011002`.

[113] Alex Mackin, Katy C. Noland, and David R. Bull. The visibility of motion artifacts and their effect on motion quality. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2435–2439, 2016. doi: 10.1109/ICIP.2016.7532796.

[114] Aditi Majumder and Michael S. Brown. *Practical Multi-projector Display Design*. A. K. Peters, Ltd., Natick, MA, USA, 2007. ISBN 1568813104.

[115] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Transactions on Graphics*, 27(3):68, 2008. doi: 10.1145/1360612.1360667. URL `http://portal.acm.org/citation.cfm?id=1399504.1360667`.

[116] Rafał Mantiuk, Allan G. Rempel, and Wolfgang Heidrich. Display considerations for night and low-illumination viewing. In *Proc. of Symposium on Applied Perception in Graphics and Visualization - APGV '09*, pages 53–58, 2009. ISBN 9781605587431. doi: 10.1145/1620993.1621005. URL `http://portal.acm.org/citation.cfm?doid=1620993.1621005`.

[117] Rafał K. Mantiuk, Karol Myszkowski, and Hans-peter Seidel. High Dynamic Range Imaging. In *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–42. John Wiley & Sons, Inc., Hoboken, NJ, USA, jun 2015. doi: 10.1002/047134608X.W8265. URL `http://doi.wiley.com/10.1002/047134608X.W8265`.

[118] Rafał K. Mantiuk, Maliha Ashraf, and Alexandre Chapiro. Stelacsf: A unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Trans. Graph.*, 41(4), jul 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530115. URL `https://doi.org/10.1145/3528223.3530115`.

[119] K. Masaoka, Y. Nishida, M. Sugawara, E. Nakasu, and Y. Nojiri. Sensation of Realness From High-Resolution Images of Real Objects. *IEEE Transactions on Broadcasting*, 59(1):72–83, mar 2013. ISSN 0018-9316. doi: 10.1109/TBC.2012.2232491. URL `http://ieeexplore.ieee.org/document/6407850/`.

[120] Kenichiro Masaoka, Yukihiro Nishida, Masayuki Sugawara, Eisuke Nakasu, and Yuji Nojiri. Sensation of realness from high-resolution images of real objects. *IEEE transactions on broadcasting*, 59(1):72–83, 2013.

[121] Ann McNamara, Alan Chalmers, Tom Troscianko, and Iain Gilchrist. Comparing real & synthetic scenes using human judgements of lightness. In *Rendering Techniques 2000: Proceedings of the Eurographics Workshop in Brno, Czech Republic, June 26–28, 2000 11*, pages 207–218. Springer, 2000.

[122] Ann. M. McNamara. Exploring perceptual equivalence between real and simulated imagery. In *Proceedings of the 2nd symposium on Appied perception in graphics and*

visualization - APGV '05, pages 123–128, New York, New York, USA, 2005. ACM Press. ISBN 1595931392. doi: 10.1145/1080402.1080425. URL http://portal.acm.org/citation.cfm?doid=1080402.1080425.

[123] Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. Demosaicing with directional filtering and a posteriori decision. *IEEE Transactions on Image Processing*, 16(1):132–141, 2006. URL https://doi.org/10.1109/TIP.2006.884928.

[124] Olivier Mercier, Yusufu Sulai, Kevin Mackenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman. Fast gaze-contingent optimal decompositions for multifocal displays. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130846. URL https://doi.org/10.1145/3130800.3130846.

[125] Meta. Oculus quest 2, 2020. URL https://store.facebook.com/quest/products/quest-2/.

[126] G. W. Meyer. An experimental evaluation of computer graphics imagery. *ACM Transactions on Graphics*, 5(1):30–50, jan 1986. ISSN 0730-0301. doi: 10.1145/7529.7920. URL https://dl.acm.org/doi/10.1145/7529.7920.

[127] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. URL https://doi.org/10.1145/3306346.3322980.

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

[129] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. *arXiv*, 2021.

[130] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

[131] Rahul Narain, Rachel A. Albert, Abdullah Bulbul, Gregory J. Ward, Martin S. Banks, and James F. O'Brien. Optimal presentation of imagery with focus cues on multi-plane displays. 34(4), July 2015. ISSN 0730-0301. doi: 10.1145/2766909. URL https://doi.org/10.1145/2766909.

[132] Baptiste Nicolet, Alec Jacobson, and Wenzel Jakob. Large steps in inverse rendering of geometry. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 40 (6), December 2021. doi: 10.1145/3478513.3480501.

[133] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[134] Oculus. Half dome 3. `https://www.oculus.com/blog/half-dome-updates-frl-explores-more-comfortable-compact-vr-prototypes-for-worl` 2020.

[135] Thomas Oskam, Alexander Hornung, Huw Bowles, Kenny Mitchell, and Markus Gross. Oscam - optimized stereoscopic camera control for interactive 3d. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 30(6):1–8, 2011. ISSN 0730-0301. doi: 10.1145/2070781.2024223. URL `https://doi.org/10.1145/2070781.2024223`.

[136] Stephen E Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999.

[137] Sylvain Paris, Samuel W. Hasinoff, and Jan Kautz. Local Laplacian filters. *ACM Transactions on Graphics*, 30(4):1, jul 2011. ISSN 07300301. doi: 10.1145/2010324.1964963.

[138] Robert Patterson, Marc Winterbottom, Byron Pierce, and Robert Fox. Binocular rivalry and head-worn displays. *Human factors*, 49(6):1083–1096, 2007.

[139] Eli Peli, Jian Yang, Robert Goldstein, and Adam Reeves. Effect of luminance on suprathreshold contrast perception. *Journal of the Optical Society of America A*, 8(8):1352, aug 1991. ISSN 1084-7529. doi: 10.1364/JOSAA.8.001352. URL `https://www.osapublishing.org/abstract.cfm?URI=josaa-8-8-1352`.

[140] Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein. Neural Holography with Camera-in-the-loop Training. *ACM Trans. Graph. (SIGGRAPH Asia)*, 2020. URL `https://doi.org/10.1145/3414685.3417802`.

[141] Maria Perez-Ortiz and Rafal K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint*, dec 2017. URL `http://arxiv.org/abs/1712.03686`.

[142] Dale Purves, Amita Shimpi, and R. Beau Lotto. An empirical explanation of the cornsweet effect. *Journal of Neuroscience*, 19(19):8542–8551, 1999. ISSN 0270-6474.

doi: 10.1523/JNEUROSCI.19-19-08542.1999. URL https://www.jneurosci.org/content/19/19/8542.

[143] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[144] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267, jul 2002. ISSN 07300301. doi: 10.1145/566654.566575. URL http://scholar.google.ca/scholar?q=ferwerda+model+of+visual+adaptation&hl=en&btnG=Search#2http://portal.acm.org/citation.cfm?doid=566570.566575http://portal.acm.org/citation.cfm?doid=566654.566575.

[145] T Ritschel, M Ihrke, J. R. Frisvad, J. Coppens, K. Myszkowski, and H.-P. Seidel. Temporal Glare: Real-Time Dynamic Simulation of the Scattering in the Human Eye. *Computer Graphics Forum*, 28(2):183–192, apr 2009. ISSN 01677055. doi: 10.1111/j.1467-8659.2009.01357.x. URL http://doi.wiley.com/10.1111/j.1467-8659.2009.01357.x.

[146] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, August 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015720. URL https://doi.org/10.1145/1015706.1015720.

[147] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.

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

[149] M. Schuchhardt, S. Jha, R. Ayoub, M. Kishinevsky, and G. Memik. Optimizing mobile display brightness by leveraging human visual perception. In *2015 International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES)*, pages 11–20, 2015.

[150] Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. High dynamic

range display systems. *ACM Trans. Graph.*, 23(3):760–768, August 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015797. URL https://doi.org/10.1145/1015706.1015797.

[151] Jae Chul Shin, Hirohisa Yaguchi, and Satoshi Shioiri. Change of color appearance in photopic, mesopic and scotopic vision. *Optical Review*, 11(4):265–271, 2004. ISSN 1340-6000. doi: 10.1007/s10043-004-0265-2.

[152] Gurjot Singh, Stephen R. Ellis, and J. Edward Swan. The Effect of Focal Distance, Age, and Brightness on Near-Field Augmented Reality Depth Matching. *IEEE Transactions on Visualization and Computer Graphics*, 26(2):1385–1398, 2018. ISSN 19410506. doi: 10.1109/TVCG.2018.2869729.

[153] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021.

[154] Chris Slinger, Colin Cameron, and Maurice Stanley. Computer-generated holography as a generic display technology. *Computer*, 38(8):46–53, 2005.

[155] George Smith and David Atchison. *The Eye and Visual Optical Instruments*. Cambridge University Press, New York, USA, 1997. URL https://eprints.qut.edu.au/5163/.

[156] Bjørn Stabell and Ulf Stabell. Chromatic rod–cone interaction during dark adaptation. *J. Opt. Soc. Am. A*, 15(11):2809–2815, Nov 1998. doi: 10.1364/JOSAA.15.002809. URL http://josaa.osa.org/abstract.cfm?URI=josaa-15-11-2809.

[157] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering, 2021. URL https://arxiv.org/abs/2112.09687.

[158] A. Sullivan. DepthCube solid-state 3D volumetric display. In A. J. Woods, J. O. Merritt, S. A. Benton, and M. T. Bolas, editors, *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291 of , pages 279–284, May 2004. doi: 10.1117/12.527543.

[159] Phil Surman and Ian Sexton. Emerging autostereoscopic displays. In Cranton Wayne Fihn Mark Chen, Janglin, editor, *Handbook of Visual Display Technology*, pages 2652–2667. Springer-Verlag Berlin Heidelberg, Bristol, 2012.

[160] I.E. Sutherland. Three-dimensional data input by tablet. *Proceedings of the IEEE*, 62(4):453–461, 1974. doi: 10.1109/PROC.1974.9449.

[161] Richard Szeliski. *Computer vision: Algorithms and applications*. Springer, 2010.

[162] Christiane Ulbricht, Alexander Wilkie, and Werner Purgathofer. Verification of physically based rendering algorithms. In *Computer Graphics Forum*, volume 25, pages 237–255. Wiley Online Library, 2006.

[163] Peter Vangorp, Rafat K Mantiuk, Bartosz Bazyluk, Karol Myszkowski, Radosław Mantiuk, Simon J Watt, and Hans-Peter Seidel. Depth from HDR: depth induction or increased realism? In *ACM Symposium on Applied Perception - SAP '14*, pages 71–78, New York, New York, USA, 2014. ACM Press. ISBN 9781450330091. doi: 10.1145/2628257.2628258. URL `http://dl.acm.org/citation.cfm?doid=2628257.2628258`.

[164] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022.

[165] Voxon. Voxon vx1. `https://voxon.co/technology/`, 2020.

[166] H. Kenneth Walker, W. Dallas Hall, and J. Willis Hurst, editors. *Clinical methods: the history, physical, and laboratory examinations*. Butterworths, Boston, 3rd ed edition, 1990. ISBN 9780409900774.

[167] Robert Wanat and Rafał K. Mantiuk. Simulating and compensating changes in appearance between day and night vision. *ACM Trans. Graph.*, 33:147:1–147:12, 2014.

[168] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021.

[169] Xuan Wang and Hong Hua. Time-multiplexed integral imaging based light field displays. In Bernard C. Kress and Christophe Peroz, editors, *Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR) II*, volume 11765, pages 156 – 162. International Society for Optics and Photonics, SPIE, 2021. doi: 10.1117/12.2576809. URL `https://doi.org/10.1117/12.2576809`.

[170] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[171] Gregory Ward-Larson, Holly Rushmeier, and Christine Piatko. A Visibility Matching Tone Reproduction Operator for High Dynamic Range Scenes. *IEEE Transactions on Visualization and Computer Graphics*, 3(4):291–306, 1997. doi: 10.1109/2945.646233.

[172] Simon J. Watt, Kurt Akeley, Marc O. Ernst, and Martin S. Banks. Focus cues affect perceived depth. *Journal of Vision*, 5(10):7–7, 12 2005. ISSN 1534-7362. doi: 10.1167/5.10.7. URL https://doi.org/10.1167/5.10.7.

[173] Simon J Watt, Kurt Akeley, Marc O Ernst, and Martin S Banks. Focus cues affect perceived depth. *Journal of vision*, 5(10):7–7, 2005.

[174] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar. Layered 3D: Tomographic Image Synthesis for Attenuation-based Light Field and High Dynamic Range Displays. *ACM Trans. Graph. (Siggraph)*, 2011.

[175] Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar. Compressive light field displays. *IEEE Computer Graphics and Applications*, 32(5):6–11, 2012. doi: 10.1109/MCG.2012.99.

[176] Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, and Ramesh Raskar. Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. *ACM Trans. Graph.*, 31(4), July 2012. ISSN 0730-0301. doi: 10.1145/2185520.2185576. URL https://doi.org/10.1145/2185520.2185576.

[177] Felix A Wichmann and N Jeremy Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313, 2001.

[178] Oscar H. Willemsen, Siebe T. de Zwart, Wilbert L. IJzerman, Martin G. H. Hiddink, and Tim Dekker. 2D/3D switchable displays. In Ari Tervonen, Malgorzata Kujawinska, Wilbert IJzerman, and Herbert De Smet, editors, *Photonics in Multimedia*, volume 6196, pages 150 – 161. International Society for Optics and Photonics, SPIE, 2006. doi: 10.1117/12.661911. URL https://doi.org/10.1117/12.661911.

[179] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[180] Jeremy M Wolfe. Stereopsis and binocular rivalry. *Psychological review*, 93(3):269, 1986.

[181] Jeremy M. Wolfe and Susan L. Franzel. Binocularity and visual search. *Perception & Psychophysics*, 44(1):81–93, Jan 1988. ISSN 1532-5962. doi: 10.3758/BF03207480. URL https://doi.org/10.3758/BF03207480.

[182] Krzysztof Wolski, Fangcheng Zhong, Karol Myszkowski, and Rafał K. Mantiuk. Dark stereo: Improving depth perception under low luminance. *ACM Trans. Graph.*

(Proceedings of ACM SIGGRAPH 2022, Journal Track), 41(4), jul 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530136. URL https://doi.org/10.1145/3528223.3530136.

[183] Daniel N. Wood, Daniel I. Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H. Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, page 287–296, USA, 2000. ACM Press/Addison-Wesley Publishing Co. ISBN 1581132085. doi: 10.1145/344779.344925. URL https://doi.org/10.1145/344779.344925.

[184] Andrew J. Woods, Tom Docherty, and Rolf Koch. Image distortions in stereoscopic video systems. pages 36–48, San Jose, CA, September 1993. doi: 10.1117/12.157041. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1009586.

[185] Matthias Wöpking. Viewing comfort with stereoscopic pictures : an experimental study on the subjective effects of disparity magnitude and depth of focus. *Journal of The Society for Information Display*, 3:101–103, 1995.

[186] Feng Xiao, Jeffrey DiCarlo, Peter Catrysse, and Brian Wandell. High dynamic range imaging of natural scenes. volume 10, pages 337–342, 01 2002.

[187] Lei Xiao, Anton Kaplanyan, Alexander Fix, Matthew Chapman, and Douglas Lanman. DeepFocus: Learned Image Synthesis for Computational Displays. *ACM Trans. Graph.*, 37(6):200:1–200:13, December 2018. ISSN 0730-0301. doi: 10.1145/3272127.3275032. URL http://doi.acm.org/10.1145/3272127.3275032.

[188] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14505.

[189] Xuan Yang, Linling Zhang, Tien-Tsin Wong, and Pheng-Ann Heng. Binocular tone mapping. *ACM Transactions on Graphics (TOG)*, 31(4):93, 2012.

[190] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020.

[191] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[192] Tomohiro Yendo, Naoki Kawakami, and Susumu Tachi. Seelinder: the cylindrical lightfield display. 07 2005. doi: 10.1145/1187297.1187314.

[193] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (proceedings of ACM SIGGRAPH ASIA)*, 38(6), 2019.

[194] Akiko Yoshida, Rafał Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. Analysis of Reproducing Real-World Appearance on Displays of Varying Dynamic Range. *Computer Graphics Forum (Proc. of Eurographics)*, 25(3):415–426, sep 2006. ISSN 0167-7055. doi: 10.1111/j.1467-8659.2006.00961.x. URL `http://doi.wiley.com/10.1111/j.1467-8659.2006.00961.x`.

[195] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021.

[196] Hyeonseung Yu, Mojtaba Bemana, Marek Wernikowski, Michał Chwesiuk, Okan Tursun, Gurprit Singh, Karol Myszkowski, Radoslaw Mantiuk, Hans-Peter Seidel, and Piotr Didyk. A perception-driven hybrid decomposition for multi-layer accommodative displays. *IEEE transactions on visualization and computer graphics*, PP, 02 2019. doi: 10.1109/TVCG.2019.2898821.

[197] Roberts Zabels, Krišs Osmanis, Mārtiņš Narels, Uģis Gertners, Ainārs Ozols, Kārlis Rūtenbergs, and Ilmārs Osmanis. Ar displays: Next-generation technologies to solve the vergence–accommodation conflict. *Applied Sciences*, 9(15):3147, Aug 2019. ISSN 2076-3417. doi: 10.3390/app9153147. URL `http://dx.doi.org/10.3390/app9153147`.

[198] Cheng Zhang, Bailey Miller, Kai Yan, Ioannis Gkioulekas, and Shuang Zhao. Path-space differentiable rendering. *ACM Trans. Graph.*, 39(4), jul 2020. ISSN 0730-0301. doi: 10.1145/3386569.3392383. URL `https://doi.org/10.1145/3386569.3392383`.

[199] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Trans. Graph.*, 40(6), dec 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480496. URL `https://doi.org/10.1145/3478513.3480496`.

[200] Zhuming Zhang, Xinghong Hu, Xueting Liu, and Tien-Tsin Wong. Binocular Tone Mapping with Improved Overall Contrast and Local Details. *Comput. Graph. Forum*, 37(7):433–442, 2018. doi: 10.1111/cgf.13580.

[201] Zhuming Zhang, Chu Han, Shengfeng He, Xueting Liu, Haichao Zhu, Xinghong Hu, and Tien-Tsin Wong. Deep binocular tone mapping. *The Visual Computer*, 35 (6):997–1011, Jun 2019. ISSN 1432-2315. doi: 10.1007/s00371-019-01669-8. URL https://doi.org/10.1007/s00371-019-01669-8.

[202] Fangcheng Zhong, George Alex Koulieris, George Drettakis, Martin S. Banks, Mathieu Chambe, Frédo Durand, and Rafał K. Mantiuk. Dice: Dichoptic contrast enhancement for vr and stereo displays. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH Asia 2019, Journal Track)*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356552. URL https://doi.org/10.1145/3355089.3356552.

[203] Fangcheng Zhong, Akshay Jindal, Ali Özgür Yöntem, Param Hanji, Simon J. Watt, and Rafał K. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH Asia 2021, Journal Track)*, 40(6), dec 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480513. URL https://doi.org/10.1145/3478513.3480513.

[204] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. URL https://doi.org/10.1145/3197517.3201323.