

UNIVERSITÄT
DES
SAARLANDES

Deep High Dynamic Range Imaging: Reconstruction, Generation and Display

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

Vorgelegt von
Chao Wang

Saarbrücken, 2025

Dean:	Prof. Dr. Roland Speicher
Date:	July 4th, 2025
Examination board:	
Chair:	Prof. Dr. Philipp Slusallek
Reviewers:	Prof. Dr. Karol Myszkowski
	Prof. Dr. Hans-Peter Seidel
	Prof. Dr. Tobias Ritschel
Academic Assistant:	Dr.Thomas Leimkühler

Abstract

High Dynamic Range (HDR) images offer significant advantages over Low Dynamic Range (LDR) images, including greater bit depth, a wider color gamut, and a higher dynamic range. These features not only provide users with an enhanced visual experience but also facilitate post-production processes in photography and filmmaking. Despite the considerable advancements in HDR technology over the years, significant challenges persist in the acquisition and display of HDR content. This thesis systematically explores the potential of leveraging deep learning techniques combined with physical prior knowledge to address these challenges. First, it investigates how implicit neural representations can be utilized to reconstruct all-in-focus HDR images from sparse, defocused LDR inputs, enabling flexible refocusing and re-exposure. Additionally, it extends the scope to the 3D domain by employing 3D Gaussian Splatting to reconstruct HDR all-in-focus fields from multi-view LDR defocused images, supporting novel view synthesis with refocusing and re-exposure capabilities. Expanding further, the thesis investigates strategies for generating HDR content from the in-the-wild LDR data or limited HDR datasets, and subsequently utilizes the resulting HDR generative models as priors to enable the transformation of LDR images into HDR. Finally, it proposes a feature contrast masking loss inspired by visual masking theory, enabling a self-supervised learning tone mapper to display the HDR content on LDR devices.

Zusammenfassung

High Dynamic Range (HDR)-Bilder bieten im Vergleich zu Low Dynamic Range (LDR)-Bildern erhebliche Vorteile, darunter eine höhere Bittiefe, ein breiteres Farbspektrum und einen größeren Dynamikbereich. Diese Eigenschaften bieten den Nutzern nicht nur eine verbesserte visuelle Erfahrung, sondern erleichtern auch die Nachbearbeitungsprozesse in der Fotografie und Filmproduktion. Trotz der erheblichen Fortschritte in der HDR-Technologie in den letzten Jahren bestehen weiterhin große Herausforderungen bei der Erfassung und Darstellung von HDR-Inhalten. Diese Dissertation untersucht systematisch das Potenzial des Einsatzes von Deep-Learning-Techniken in Kombination mit physikalischem Vorwissen, um diese Herausforderungen zu bewältigen. Zunächst wird der Einsatz impliziter neuronaler Netzwerke untersucht, um HDR-Bilder mit vollständigem Fokus aus spärlichen LDR-Bildern mit Defokus zu rekonstruieren und dabei flexible Nachfokussierungs- und Belichtungsanpassungen zu ermöglichen. Darüber hinaus wird der Anwendungsbereich auf den 3D-Bereich erweitert, indem 3D-Gaussian Splatting verwendet wird, um HDR-Felder mit vollständigem Fokus aus Multi-View-LDR-Bildern mit Defokus zu rekonstruieren, wodurch die Synthese neuer Ansichten mit Nachfokussierungs- und Belichtungsanpassungsfunktionen unterstützt wird. Weiterführend untersucht die Dissertation Strategien zur Generierung von HDR-Inhalten aus unstrukturierten LDR-Daten oder begrenzten HDR-Datensätzen und nutzt die resultierenden HDR-Generativmodelle als Prior, um die Umwandlung von LDR- in HDR-Bilder zu ermöglichen. Schließlich wird ein Feature Contrast Masking Loss vorgeschlagen, der von der Theorie der visuellen Maskierung inspiriert ist und einen selbstüberwachten Tone-Mapping-Ansatz ermöglicht, um HDR-Inhalte auf LDR-Geräten darzustellen.

Acknowledgments

I would first like to express my deepest gratitude to my advisor, Karol Myszkowski, for bringing me to MPI and for his unwavering academic guidance throughout my doctoral journey. I still remember our first interview, which opened the door to this path. He has been a true academic mentor, whose principled research attitude, constant encouragement, and generous support have shaped both my work and mindset. His dedication to research and professionalism have deeply influenced my attitude toward academic work. I am also thankful to Hans-Peter Seidel for offering an excellent research environment. My sincere thanks go to Ana Serrano and Thomas Leimkühler for their valuable academic guidance, from which I have benefited greatly. I am especially grateful to Xingang Pan, whose encouragement and generous support helped me through difficult academic moments.

I would also like to thank my collaborators Bin Chen, Piotr Didyk, and Michal Piovarči for their insightful help on our material perception work. My appreciation extends to all my colleagues in D4 and D6. Your companionship and support have added warmth and joy to life. I am equally grateful to our administrative staff, Anne Gesellchen and Sabine Zimmer. In particular, I would like to express my sincere appreciation to Ellen Fries, whose thoughtful support during my onboarding period left a lasting impression. I will always remember her with deep respect and gratitude.

My sincere thanks also go to my colleagues at Adobe during my internship, Xuaner Zhang and Zhihao Xia. It was a great pleasure to collaborate with you, and I truly enjoyed our time working together. I also fondly recall the sunshine in California.

Finally, I wish to thank my parents and family. Your unconditional support has been an essential pillar of my academic journey, and I could not have reached this milestone without you.

Contents

Contents **ix**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Contributions . . . 5
1.3 Outline . . . 7

2 Previous Work **9**
2.1 Prerequisites for High Dynamic Range Imaging . . . 9
2.2 Acquisition and Display of HDR Images . . . 10
2.2.1 Multiple Exposure Fusion . . . 11
2.2.2 Inverse Tone Mapping . . . 12
2.2.3 Tone Mapping . . . 14
2.3 Neural Rendering . . . 16
2.3.1 Implicit Neural Representations . . . 16
2.3.2 Point Based Representations . . . 17
2.4 Generative Models . . . 18
2.4.1 Generative Adversarial Network (GAN) . . . 19
2.4.2 Diffusion Model . . . 20

3 An Implicit Neural Representation for the Image Stack **23**
3.1 Introduction . . . 23
3.2 Background . . . 25
3.2.1 Depth and AiF From Focal Stacks . . . 25
3.2.2 Implicit Neural Representations and Their Applications . . . 26
3.3 Method . . . 26
3.3.1 Depth and Image Representation with Implicit Flow . . . 27
3.3.2 Differentiable Thin Lens Model . . . 28
3.3.3 Implicit Tone Mapper . . . 29
3.3.4 Loss . . . 29
3.4 Evaluation . . . 30
3.4.1 Depth-From-Defocus and All-In-Focus Reconstruction . . . 30
3.4.2 Exploration of TAF Sampling Strategies . . . 31
3.4.3 Ablations . . . 34
3.4.4 Runtime . . . 37
3.5 Conclusions and Limitations . . . 37

4 A Real-Time HDR Radiance Fields with Depth of Field **39**
4.1 Introduction . . . 39
4.2 Background . . . 40
4.2.1 Depth-Of-Field Rendering . . . 40
4.2.2 Depth-Of-Field and High Dynamic Range Radiance Fields . . . 41
4.3 Method . . . 41

4.3.1 Prerequisite . . . 42
4.3.2 Camera Calibration . . . 42
4.3.3 HDR Radiance Field with Depth of Field . . . 43
4.3.4 Training . . . 44
4.4 Results and Evaluation . . . 45
4.4.1 Dataset . . . 45
4.4.2 Implementation . . . 47
4.4.3 Comparisons . . . 48
4.4.4 HDR Radiance Field Reconstruction and Post-Editing . . . 49
4.4.5 Ablations . . . 49
4.5 Conclusions and Limitations . . . 50

5 An Unsupervised HDR Generation Model from In-the-Wild LDR 53
5.1 Introduction . . . 53
5.2 Background . . . 55
5.2.1 High Dynamic Range Imaging . . . 55
5.2.2 Lossy Generative Adversarial Networks (GAN) . . . 55
5.3 Method . . . 56
5.3.1 GlowGAN . . . 56
5.3.2 GlowGAN-ITM . . . 58
5.4 Experiments . . . 59
5.4.1 Generation of HDR Images . . . 59
5.4.2 Ablations . . . 61
5.4.3 Application: Inverse Tone Mapping . . . 62
5.5 Conclusions and Limitations . . . 64

6 A Latent Exposure Fusion Diffusion Model for HDR Generation 67
6.1 Introduction . . . 67
6.2 Method . . . 69
6.2.1 Preliminaries . . . 69
6.2.2 Merging Latent Exposure Brackets . . . 70
6.2.3 Generating Latent Exposure Brackets . . . 71
6.3 Dataset and Training . . . 72
6.4 Applications and Evaluations . . . 72
6.4.1 HDR Content Generation . . . 72
6.4.2 LDR to HDR Image Reconstruction . . . 73
6.4.3 Ablations . . . 76
6.5 Conclusions and Limitations . . . 77

7 A Self-supervised Tone Mapping Operator 79
7.1 Introduction . . . 79
7.2 Background . . . 80
7.2.1 Contrast Perception Modeling . . . 80
7.2.2 Tone Mapping Techniques . . . 81
7.3 Method . . . 82
7.3.1 Multiple Exposure Selection . . . 82
7.3.2 Tone Mapping Network . . . 83
7.3.3 Adaptive μ-Law Compression . . . 84
7.4 Feature Contrast Masking Loss . . . 86
7.5 Results and Comparisons . . . 89
7.5.1 Competitors and Dataset . . . 89

7.5.2 Objective Evaluation 89
7.5.3 Subjective Evaluation 91
7.5.4 Ablations 93
7.6 Conclusions and Limitations 94

8 Conclusion **95**
8.1 HDR Content Reconstruction From Photographs 95
8.2 Advanced HDR Content Generation Models 98
8.3 Unsupervised and Controllable Tone Mapper 99

Bibliography - Own Work **101**

Bibliography **103**

Chapter 1

Introduction

This thesis investigates HDR (High Dynamic Range) imaging through deep learning from various perspectives. This chapter discusses the core motivation (Section 1.1), outlines the key contributions (Section 1.2), and provides a detailed overview of the thesis structure (Section 1.3).

1.1 Motivation

The accurate reproduction of real-world scenes has long been a fundamental challenge in imaging research, driven by the aspiration to replicate physical scene information on digital devices. Conventional Low Dynamic Range (LDR) formats, widely used in media content, impose significant constraints on faithfully representing scene details due to their limited color gamut, narrow luminance range, and low contrast. Commonly used formats like JPEG and H.265/AVC, relying on 8-bit encoding, inevitably discard substantial visual information, rendering them incapable of capturing critical phenomena such as self-luminous surfaces, glare, and dazzling effects.

High Dynamic Range (HDR) imaging, as an advanced imaging technology, addresses these limitations by capturing and displaying a broader luminance range and richer color information. Unlike LDR, HDR imaging accurately preserves details across the spectrum, from the darkest shadows to the brightest highlights, aligning closely with the human visual system's perception of light and color. By employing wider color gamut and supporting a wider luminance range, HDR imaging enables the faithful reproduction of scotopic, mesopic, and photopic vision, along with perceptual effects such as Hunt's effect, where brightly illuminated objects appear more colorful. This capability allows HDR images to preserve critical luminance and color details, delivering a more realistic, immersive, and high-fidelity digital experience. Figure 1.1 summarizes luminance ranges relevant for the human visual system (HVS) and various display technologies, further demonstrating that HDR representation is better aligned with the HVS.

HDR imaging brings clear benefits. Nonetheless, key challenges persist and are explored in this thesis. The first concerns the acquisition of HDR content: due to limitations in conventional camera sensors, it is still difficult to directly capture high dynamic range images. This thesis tackles the problem from two perspectives: reconstructing HDR images from multi-exposure LDR inputs and generating HDR content from latent codes. Another key issue lies in compatibility with LDR displays, which are still widely used. Effectively displaying HDR images on such devices remains a critical problem.

HDR Image Reconstruction from 2D LDR Images The acquisition of HDR images commonly relies on advanced cameras and multi-exposure fusion techniques. Due

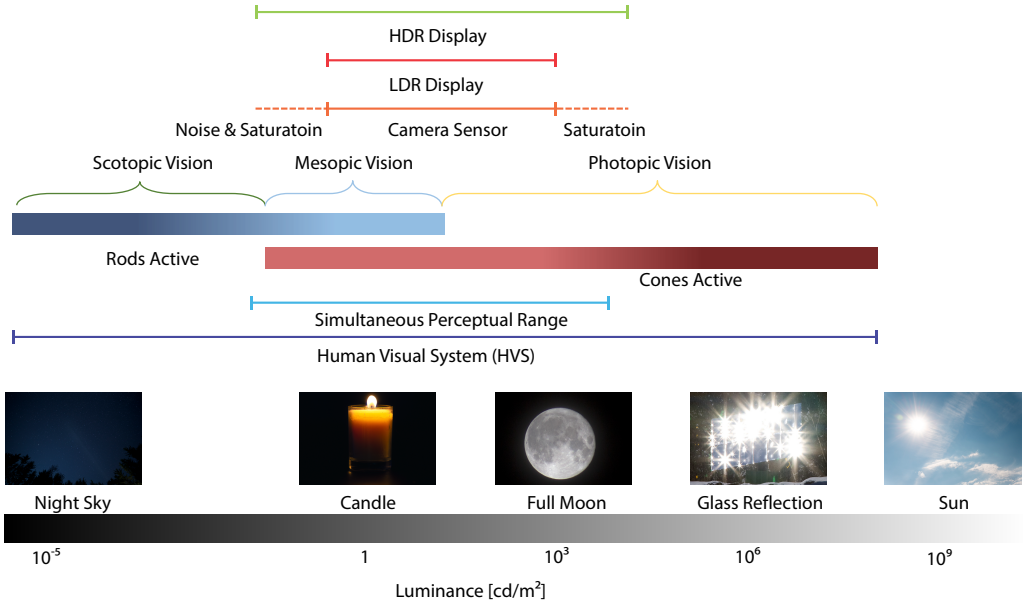


FIGURE 1.1: *Dynamic range comparison. The real world exhibits an extremely high dynamic range, spanning from very dark scenes (e.g., the night sky) to extremely bright light sources (e.g., direct sunlight). The Human Visual System (HVS) is capable of perceiving this wide range through sophisticated adaptation mechanisms. However, conventional camera sensors are inherently limited in their ability to capture this range, often resulting in noise in low-light regions and saturation in bright areas. Similarly, display technologies impose additional constraints: traditional LDR displays are restricted to a narrow dynamic range. Conversely, HDR displays can render a significantly broader range of luminance, providing a closer match to the perceptual capabilities of the HVS.*

to the high cost of advanced cameras, multi-exposure fusion has become a more popular approach for obtaining HDR images. While significant progress has been made in multi-exposure fusion techniques, several challenges remain. For instance, most fusion methods primarily address the limitations of camera sensors, such as overexposure or underexposure, while overlooking the optical imperfections of camera lenses. These imperfections can result in defocused blur and alignment issues caused by lens breathing effects, which are often ignored in existing approaches. A practical challenge in HDR images acquisition lies in recovering an all-in-focus HDR image from multi-exposure LDR images and depth of field effects. Traditional methods address this by first capturing a series of multi-exposure images at a fixed focal distance, then repeating this process for different focal distances. The resulting images are then processed through multi-exposure fusion for each focal distance, followed by multi-focus fusion to generate the final all-in-focus HDR image.

This method poses several challenges. First, the acquisition process is tedious and redundant, while the subsequent processing is highly complex and time-consuming. Furthermore, the traditional approach only produces a single all-in-focus HDR image. However, the substantial similarity between images with different exposures and defocus levels offers an opportunity for optimization. By leveraging well-established physical principles, such as the thin-lens model, which simulates defocus blur caused by variations in focal distance and aperture size, it is possible to reduce the number of captured images that are required to a sparse set while still achieving an all-in-focus HDR reconstruction. Furthermore, based on the theory of depth from defocus, depth information can be recovered from images with different defocus levels. Inspired by NeRF (Neural Radiance Fields) [Mildenhall et al., 2020] and SIREN [Sitzmann et al., 2020], we develop a unified implicit representation framework for

sparse LDR images captured with random parameters. Through a forward modeling process, the representation is optimized to match the input images, enabling the joint reconstruction of an all-in-focus HDR image and its depth map. This reconstruction supports flexible post-processing such as re-focusing and re-exposure, broadening the creative and practical potential of HDR workflows.

As with traditional HDR methods, this formulation remains confined to the 2D spatial domain. A natural and more impactful extension is to lift the problem into the 3D domain, aiming to recover a full 3D all-in-focus HDR radiance field from multi-view defocused LDR inputs. This formulation allows for simultaneous re-focusing, re-exposure, and novel view synthesis within a unified 3D scene. To support practical deployment, especially under real-time constraints, we adopt 3D Gaussian Splatting (GS) [Kerbl et al., 2023] as a compact and efficient scene representation, well-suited for high-quality and interactive rendering.

HDR Image Generation from Latent Codes With the rapid advancement of generative models, particularly the application of large-scale vision models, Jensen Huang, the CEO of Nvidia, has even stated that "every single pixel in a video game is going to be generated, not rendered." [1]. However, due to the limitations of training data, existing generative models predominantly focus on the generation of LDR content. An important avenue of exploration, therefore, is the generation of HDR content. Unfortunately, the scarcity of HDR data makes it challenging to train a model from scratch using HDR datasets. To address this limitation, we first explore methods for generating HDR content only using LDR data, leveraging the abundance of LDR samples as a starting point. Compared to HDR data, LDR data is ubiquitous and can be easily obtained from the internet, encompassing a wide variety of categories such as landscapes, lightning, and more. LDR data can be considered as a lower-dimensional projection of HDR data along the exposure dimension. Consequently, the problem becomes one of exploring whether it is possible to generate high-dimensional HDR signals from such low-dimensional observations. Inspired by AmbientGAN [Bora et al., 2018] , we leverage a GAN [Sauer et al., 2022] framework as the backbone for our generator. By simulating the camera imaging process, we design a degradation model that projects the generated HDR content into the LDR space. This approach enables the use of abundant LDR data to guide the generation of HDR content, effectively bridging the gap between low-dimensional observations and high-dimensional reconstructions. Furthermore, the trained GAN model can serve as a prior for leveraging GAN inversion techniques [Roich et al., 2022] to transform LDR images into HDR formats, addressing the problem commonly referred to as inverse tone mapping.

Although GAN-based approaches have made progress in HDR ccontent generation, they are inherently limited by issues associated with GANs, such as the need for class-specific training data and challenges in achieving high-quality generation with robust generalization. Recently, the rapid advancement of Diffusion Models, particularly the emergence of Stable Diffusion, has demonstrated significant breakthroughs in generative capabilities. Stable Diffusion's [Rombach et al., 2022] scalability enables training on massive datasets such as LAION-5B [Schuhmann et al., 2022], achieving impressive performance in LDR content generation. However, it remains confined to LDR content, and the scarcity of HDR data poses additional challenges. To address this, we propose exploring how a diffusion model pre-trained on large-scale LDR

[1] https://www.forwardfuture.ai/p/future-gaming-every-pixel-generated-realtime: The Future of Gaming: Every Pixel Generated in Real-Time

datasets can be fine-tuned using a small amount of HDR data to enable HDR generation. As a first step, we analyze the two key components of Stable Diffusion: the VAE (Variational Autoencoder) [Kingma and Welling, 2014] and the Denoiser. The VAE primarily maps images into a latent space and back to the image space, serving as a dimensionality reduction mechanism. The Denoiser, on the other hand, operates by adding noise to and denoising latent codes, facilitating the generative process.

As outlined in Banterle et al. [2017], HDR content generation fundamentally involves hallucinating missing details in overexposed and underexposed regions, alongside addressing pixel linearization and dynamic range expansion. By analyzing the latent codes corresponding to multi-exposure images of the same scene, we observe a consistent correlation between the latent space and the image space (e.g., Latent codes encoded from overexposed images exhibit saturation as well). Inspired by multi-exposure fusion techniques in the image space, we propose a latent space multi-exposure fusion module to fine-tune the VAE (Variational Autoencoder), thereby equipping it with the capability to decode HDR images. Specifically, the module fuses multiple LDR latent codes to construct a lossless latent representation. Subsequently, the decoder is fine-tuned to map this fused latent representation into the HDR image space, effectively addressing the challenges of linearization correlated with the scene luminance and dynamic range expansion. In addition, the denoiser is fine-tuned to enable hallucination for missing details in overexposed and underexposed regions. For overexposed regions, the high-exposure latent code is concatenated as a condition with the low-exposure latent code, and noise is applied to the low-exposure latent code. The denoiser is trained to reconstruct the low-exposure latent code from the noisy input, enabling the generation of missing information in overexposed areas. Finally, the fine-tuned VAE is used to fuse and decode the resulting latent codes into HDR images. A similar process is applied symmetrically for underexposed regions. This fine-tuning approach is universally compatible with all Stable Diffusion-based architectures, enabling seamless integration for tasks such as text-to-HDR images, and text-to-HDR panorama generation. Moreover, the fine-tuned model can serve as a prior for LDR-to-HDR conversion, achieving higher quality, faster inference, and eliminating class-specific constraints compared to GAN-based methods.

HDR Image Display on LDR Devices HDR represents a more advanced image format; however, due to the widespread use of LDR display devices, it is necessary to develop an effective compression process to render HDR content on these displays. This process, referred to as tone mapping [Reinhard et al., 2010], is essential for adapting HDR content to the constraints of LDR devices while preserving visual fidelity. Through a detailed analysis of the objectives and mechanisms of tone mapping, we identified contrast reproduction as the critical component. Drawing upon visual masking theory [Daly et al., 2000], we modeled image contrast in the feature space, incorporating both local contrast and neighborhood contrast. This approach enables contrast enhancement in low-light regions and compression in high-brightness areas, aligning more closely with the human visual system's perception. Building on this model, we developed a Feature Contrast Masking loss designed to optimize a simple CNN network for tone mapping. This method operates directly on HDR images, thereby eliminating the reliance on paired datasets (HDR-LDR images) while achieving superior visual quality compared to other state-of-the-art methods.

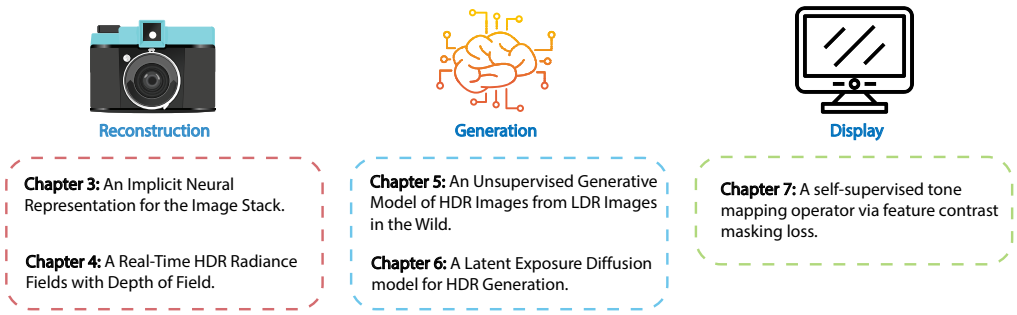


FIGURE 1.2: *This section outlines the main contributions of the thesis. In Chapters 3 [Wang et al., 2023b] and 4 [Wang et al., 2024a] present methods for reconstructing HDR images from captured, lossy LDR inputs. In Chapters 5 [Wang et al., 2023a] and 6 [Wang et al., 2025] focus on building generative models for HDR image synthesis. Finally, in Chapter 7 [Wang et al., 2022a] addresses the problem of effectively visualizing HDR content on LDR displays through tone mapping.*

1.2 Contributions

This thesis primarily investigates HDR imaging from three perspectives: reconstruction, generation, and display as shown in Figure 1.2.

HDR Images Reconstruction Traditional HDR images reconstruction techniques primarily focus on multi-exposure fusion while overlooking defocus blur caused by lens imperfections. Additionally, their reliance on 2D grid-based representations inevitably introduces challenges such as cropping during alignment operations and limitations in enabling flexible post-processing edits. Recently, implicit neural representations, such as SIREN [Sitzmann et al., 2020] and NeRF [Mildenhall et al., 2020], have demonstrated significant advantages in signal fitting. Inspired by these advancements, Chapter 3 (based on Wang et al. [2023b]) proposes **Time-Aperture-Focus Distance (TAF)**,

- A unified implicit neural representation capable of handling unstructured TAF stacks while supporting all-in-foucs (AiF) HDR image and depth reconstruction with flexible post-editing,
- Superior performance in Depth from Defocus (DfD) and All-in-focus (AiF) HDR image reconstruction, surpassing state-of-the-art methods, and
- A novel dataset featuring synthetic and captured focal stacks for comprehensive analysis and validation.

Building on this foundation, we extend the problem to the recovery of All-in-Focus HDR radiance fields from multi-view defocused LDR images with varying exposures. The simulation of defocus inherently involves convolution operations, where the aperture's convolution kernel is commonly approximated by a Gaussian kernel. Given the closed-form property of Gaussian convolutions, which remain Gaussian after convolution, we adopt the 3D Gaussian as the scene representation to ensure real-time performance. Accordingly, we propose **CinematicGS**, Chapter 4 (based on Wang et al. [2024a]),

- A unified framework for reconstructing HDR radiance fields with depth of field,

- Real-time post-editing capabilities, including exposure adjustment and arbitrary refocusing, and
- State-of-the-art performance in HDR and all-in-focus reconstruction, along with new challenging benchmark datasets.

HDR Images Generation Although current generative models are capable of producing high-quality images, they are limited to LDR content due to the constraints of training data. On the one hand, as previously discussed, HDR imaging offers unparalleled advantages over LDR, making HDR content generation a highly promising direction. On the other hand, the scarcity of high-quality HDR data makes it challenging to train an HDR generative model from scratch using only HDR datasets. However, LDR and HDR images exhibit significant semantic similarities, with LDR images effectively serving as projections of HDR images along the exposure dimension. Inspired by AmbientGAN [Bora et al., 2018], we propose **GlowGAN**, Chapter 5 (based on Wang et al. [2023a]),

- A novel framework that bridges HDR and LDR spaces via a camera model, enabling unsupervised learning of HDR images from in-the-wild LDR data and synthesizing high-quality HDR images with a significantly higher dynamic range, eliminating the reliance on ground-truth HDR datasets, and
- An unsupervised inverse tone mapping method, GlowGAN-ITM, leveraging GlowGAN as a prior to achieve state-of-the-art reconstruction of large overexposed regions.

While GlowGAN has shown promising results in HDR generation, it suffers from inherent limitations—most notably, the need to train separate models for each scene type, which significantly constrains scalability and generalization. To overcome these limitations and further enhance generative capability, we turn to diffusion models. Recent advances, particularly Stable Diffusion [Rombach et al., 2022], have demonstrated impressive image synthesis quality by leveraging large-scale training datasets. In addition to high-fidelity generation, these pre-trained models encode strong visual priors that can be repurposed for tasks such as depth [Ke et al., 2024] and surface normal [Fu et al., 2025] estimation via fine-tuning. Motivated by this, we explore whether a pre-trained LDR diffusion model can be effectively adapted using only limited HDR data to achieve HDR image generation. To address this, we propose **LEDiff**, Chapter 6 (based on Wang et al. [2025]),

- A fine-tuning approach that retains the pre-trained LDR latent space while leveraging a small set of HDR data to expand dynamic range and enhance details in deep shadows and bright highlights,
- A method for enabling HDR generation in existing image and video generative models, achieving state-of-the-art performance in inverse tone mapping, including linearization and photorealistic reconstruction of clipped regions in real LDR images.

HDR Images Display While HDR technology offers significant advantages, a practical challenge remains: the majority of existing display devices are LDR. Visualizing HDR content on such devices requires a compression process known as tone mapping. Traditional learning-based tone mapping approaches often depend on manually annotated paired (HDR-LDR images) datasets, which are inherently subjective and

labor-intensive to produce. To address this limitation, we conducted a fundamental analysis of tone mapping, identifying its primary objective as reproducing HDR contrast on LDR devices. Inspired by the Visual Masking theory [Zeng et al., 2000], we developed a contrast modeling approach in the feature space. Building on this foundation, we designed a novel loss function and proposed **Self-TMO**, Chapter 7 (based on Wang et al. [2022a]),

- A self-supervised tone mapping network that processes multiple exposures derived from HDR images to produce state-of-the-art tone mapping results, eliminating the need for annotated HDR-tone mapped image pairs during training,
- A perceptually-inspired feature contrast masking loss function, derived in the feature space, that directly supervises the tone mapping process using HDR image content, minimizing perceived luminance contrast loss in the tone-mapped outputs, and
- An adaptive μ-law compression technique that adjusts HDR histograms to resemble those of tone-mapped LDR images, facilitating effective self-supervision using the input HDR image.

The author of this thesis is the principal contributor to the research presented in each chapter. Additionally, during the course of the Ph.D. studies, the author has actively contributed to other research projects in related areas, including material perception [Chen et al., 2023a; Chen et al., 2022b; Chen et al., 2021; Serrano et al., 2021] and tone mapping [Zhang et al., 2021].

1.3 Outline

The structure of this thesis is as follows: Chapter 2 provides a review and discussion of relevant literature. Chapter 3 details the proposed implicit representation for reconstructing All-in-Focus HDR images from multiple defocused LDR inputs, while enabling advanced post-editing capabilities. Chapter 4 introduces a unified 3D Gaussian representation for recovering All-in-Focus HDR radiance fields from multi-view defocused LDR images, facilitating real-time post-editing operations such as refocusing and re-exposure. Moving forward to HDR generation, Chapter 5 introduces the first HDR generator trained on unstructured, in-the-wild LDR data, while Chapter 6 investigates enabling HDR generation capabilities in diffusion models through fine-tuning with limited HDR datasets. Lastly, Chapter 7 proposes a self-supervised framework for visualizing HDR images on LDR devices, utilizing only HDR data as supervision. Chapter 8 presents the conclusion of this thesis and discusses potential directions for future research.

Chapter 2

Previous Work

This chapter reviews the related work pertinent to this thesis. Section 2.1 discusses the foundational concepts of High Dynamic Range (HDR) imaging, emphasizing its practical benefits and academic relevance. Section 2.2 outlines common HDR acquisition methods, including Multiple Exposure Fusion (MEF) (Section 2.2.1) and Inverse Tone Mapping (ITM) (Section 2.2.2), and Tone Mapping (TM) methods (Section 2.2.3) for displaying HDR images on LDR display. Section 2.3 examines advancements in neural rendering, covering both implicit neural representations (Section 2.3) and point-based representations (Section 2.3.2). In addition, Section 2.4 provides an overview of prominent generative models, including GAN-based approaches (Section 2.4.1) as well as diffusion-based models (Section 2.4.2).

2.1 Prerequisites for High Dynamic Range Imaging

Dynamic Range Dynamic range defined as the ratio between the largest and smallest measurable quantity, varies across applications. In display systems, the contrast ratio quantifies the luminance ratio of the brightest (white) to the darkest (black) colors a display can produce, normalized to avoid infinity in HDR displays (e.g., 1000:1). For scenes, the log exposure range measures dynamic range as the difference in base-10 logarithms between the brightest and darkest luminance values, providing a perceptually accurate representation of luminance variation. In film photography, exposure latitude is expressed in base-2 logarithms (stops) and represents the difference between the film's capture range and the scene's luminance range, indicating the margin of error allowed in exposure settings. In human visual perception, the perceivable dynamic range is determined by factors such as light scattering in the eye and local adaptation, allowing the perception of dynamic ranges exceeding 4 log-10 units in certain scenes. Preferences for higher contrast ratios, often above 1000:1 on HDR displays, highlight the significance of dynamic range, though limitations in display technologies, such as OLED's restricted peak brightness, can diminish perceived contrast in low-luminance regions [Mantiuk et al., 2015a].

LDR vs. HDR Images To understand the differences between LDR and HDR imaging, it is essential to first clarify the concepts of display-referred (device-referred) and scene-referred representations. Display-referred formats, such as JPEG, and PNG, are designed to align with the capabilities of specific display devices, encoding pixel values tailored to their luminance range, contrast, and color gamut. These formats, typical in LDR imaging, prioritize immediate display readiness but compromise visual information during acquisition or storage by applying non-linear transformations, such as gamma correction, which compress the dynamic range and distort the relationship between pixel values and the actual photometric properties of the scene.

LDR images are commonly stored with 8-bit precision, which introduces quantization errors and limits their ability to capture fine details. As a result, pixel values in LDR images are non-linearly related to luminance, and the term "luma" is instead used to describe perceived light intensity in these images.

In contrast, HDR imaging employs scene-referred representations that encode the actual photometric properties of a scene, such as luminance or spectral radiance, independently of specific display devices. Pixel values in HDR images are linearly related to luminance, providing a closer approximation to the perceived intensity of light per surface area, albeit with some deviation due to the spectral sensitivities of cameras used for HDR capture. HDR images are typically stored using 32-bit floating-point precision (such as, OpenEXR and Radiance HDR), significantly reducing quantization errors and enabling more detailed reproduction of subtle variations in brightness and color. These deviations are usually within a range of 10% for achromatic surfaces and up to 30% for colored objects. Although HDR pixel values are often referred to as radiance for each color channel, a more accurate term is trichromatic color values, as they reflect the spectral sensitivities of the capture system rather than the physical definition of radiance [Reinhard et al., 2010; Mantiuk et al., 2015a; Eilertsen, 2018].

This distinction between display-referred and scene-referred representations underscores the fundamental difference in how LDR and HDR images encode and preserve scene information. HDR images retain high precision and fidelity, allowing for accurate optimization at the display stage, while LDR images inherently sacrifice scene accuracy to accommodate device limitations. By employing higher bit-depths and linear relationships with luminance, HDR imaging provides a closer approximation to real-world luminance and chromaticity, making it better suited for representing high-contrast scenes and intricate details in brightness and color. It is precisely these advantages of HDR that allow it to provide users with a superior visual experience while also facilitating subsequent image editing, such as image-based lighting [Miller et al., 1999] and depth-of-field rendering [Mildenhall et al., 2022; Zhang et al., 2019c]. Table 2.1 provides a detailed comparison between LDR and HDR image characteristics.

Feature	LDR	HDR
Peak Brightness	up to 500 cd/m^2	2,000–10,000 cd/m^2
linearity	Non-linear	Linear
Display Contrast	1:1,000	1:1,000,000
Quantization	8 to 16 bit	Floating-point or variable
Image Representation	Display-referred	Scene-referred
Color Gamut	sRGB	BT.2020
Format	JPG, PNG	OpenEXR, Radiance HDR

TABLE 2.1: *Comparison of LDR and HDR imaging characteristics. Information adapted from Mantiuk et al. [2015a], with gratitude to the original contributors.*

2.2 Acquisition and Display of HDR Images

The acquisition of HDR content has been a long-standing challenge, with three primary classes of methods commonly identified. These include specialized HDR cameras, single-shot methods that expand the dynamic range of a single LDR input through techniques such as inverse tone mapping (ITM), and multiple exposure fusion methods that merge differently exposed LDR images into an HDR output. Each of these approaches offers distinct advantages and trade-offs, addressing the

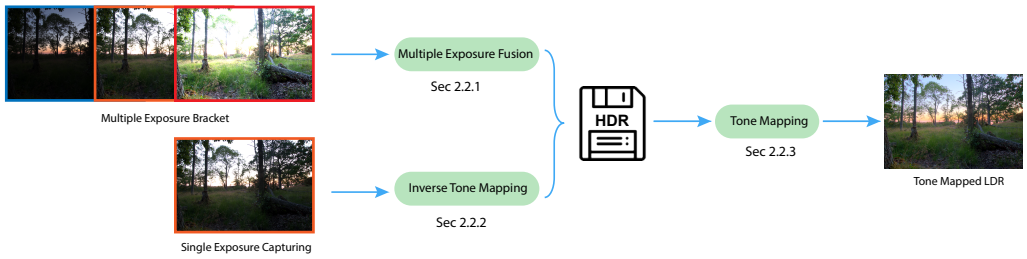


FIGURE 2.1: *The figure illustrates the overview of this section. Sections 2.2.1 and 2.2.2 introduce two common approaches for acquiring HDR images: multi-exposure fusion and inverse tone mapping, respectively. Section 2.2.3 discusses tone mapping techniques for dynamic range compression.*

challenges of HDR image construction from different perspectives. Specialized HDR cameras, despite their effectiveness, have seen limited practical adoption due to high costs and inherent limitations in sensor technology [Eilertsen, 2018]. Single-shot methods, such as inverse tone mapping [Banterle et al., 2017], present a promising alternative for processing legacy LDR images, offering compatibility with machine learning techniques for tasks such as denoising underexposed regions and inpainting overexposed details. However, the fidelity of these methods remains significantly lower compared to that achieved by true multi-exposure techniques [Debevec and Malik, 2023].

2.2.1 Multiple Exposure Fusion

The limitations of traditional camera sensors prevent the full capture of the complete dynamic range of the real world, often leading to phenomena such as overexposure or underexposure [Eilertsen, 2018]. A common approach to address this limitation is to take multi-exposure LDR images of the same scene and then fuse these exposures to generate an HDR image.

Debevec and Malik [2023] propose a method that uses an optimization-based method grounded in the principle of reciprocity to recover the Camera Response Function (CRF) and further employs a weighted fusion approach to combine radiance values from multi-exposure images, producing an HDR radiance map. Mertens et al. [2007] propose using contrast, saturation, and exposure to generate weight maps, which are applied in a Laplacian Pyramid-based method to fuse multi-exposure images, bypassing CRF calibration and enhancing computational efficiency. Although these methods have demonstrated effectiveness, their focus is primarily on static scenes. However, when utilizing handheld cameras, both object and camera motion are unavoidable challenges. Sen et al. [2012] propose a method that formulates HDR reconstruction as an optimization problem, integrating alignment and merging through a patch-based approach that directly optimizes local similarity between the output and all LDR sources, effectively handling motion and creating high-fidelity HDR images.

Kalantari et al. [2017] pioneer the application of deep learning techniques in HDR reconstruction. They develop a customized dataset with three different exposure LDR images as inputs and well-aligned HDR images as ground truth, aligning the inputs using optical flow Liu et al. [2009] and training a convolutional neural network (CNN) [LeCun et al., 2015] to fuse them into HDR images. Subsequent studies build upon the work of [Kalantari et al., 2017], exploring various network architectures. Wu et al. [2018] are the first to treat this problem as an image translation task, proposing a U-Net-based [Ronneberger et al., 2015] non-flow-based network that implicitly aligns the input LDR images and directly translates them into an HDR image. Yan et al.

[2019] integrated an attention module into the network, where the learned attention map is employed to suppress harmful components caused by misalignment and saturation while emphasizing useful details, thereby enhancing the quality of HDR synthesis. Yan et al. [2020] embed the non-local module [Wang et al., 2018a] within the latent space to capture long-range dependencies and suppress artifacts.

With the rapid development of Vision Transformers, their larger receptive field compared to CNNs and superior awareness of global features have led to their adoption in the field of HDR reconstruction. Chen et al. [2023b] propose the HDR Fusion Transformer, a hybrid CNN-Transformer architecture that integrates a Pyramid Fusion Modulefor multi-scale feature fusion, a Channel Attention Dilated Block to address ghosting artifacts, and a Transformer component for global context modeling, delivering efficient and high-quality HDR fusion. Liu et al. [2022] propose HDR-Transformer, which employs a dual-branch architecture to capture global and local dependencies, effectively addressing ghosting artifacts and achieving high-quality HDR reconstruction with reduced computational costs. Tel et al. [2023] propose SCTNet, which incorporates spatial and semantic-consistent cross-frame attention to jointly model foreground motion and static context, enhancing HDR fusion and merging. Yan et al. [2023a] propose a dual-network architecture combining an alignment module for patch aggregation and ghost removal with a Swin-Transformer-based [Liu et al., 2021] fusion network to adaptively merge information from differently exposed regions, achieving high-quality HDR image synthesis.

Further work explores the integration of generative models, Niu et al. [2021] present HDR-GAN, a pioneering GAN-based framework for multi-exposure HDR reconstruction, featuring a residual merging block for motion alignment and a deep supervision strategy to minimize artifacts. Yan et al. [2023b] present the first DDPM-based [Ho et al., 2020] HDR reconstruction method for multi-exposed LDR images, incorporating spatial affine transformations from implicitly aligned features, sliding-window noise estimation, and a pixel-space loss to ensure reliable content generation and perception-distortion balance, even under significant object motion. To enhance the utilization of intrinsic image information and reduce reliance on training datasets, several works investigate semi- and self-supervised methodologies, Prabhakar et al. [2021] first introduce a two-stage zero-shot and few-shot learning framework for HDR deghosting, leveraging unlabeled dynamic exposure stacks to address labeled data constraints and achieve performance comparable to or surpassing supervised methods with minimal labeled samples. Yan et al. [2023c] employs a semi-supervised two-stage framework for few-shot HDR imaging, utilizing a Saturated Mask AutoEncoder to reconstruct non-saturated HDR content and an iterative semi-supervised learning strategy with adaptive pseudo-label selection to mitigate ghosting. Zhang et al. [2024b] propose SelfHDR, a self-supervised HDR reconstruction framework that reduces reliance on labeled data by decomposing latent ground truth into color and structure components, leveraging aligned multi-exposure images for color supervision and a structure-focused network guided by the color component and reference input.

2.2.2 Inverse Tone Mapping

An alternative to multi-exposure fusion is to recover HDR images from a single-exposure LDR input, a task known as Inverse Tone Mapping (ITM). This process primarily involves de-quantization, de-contouring, linearization, dynamic range expansion, completion of overexposed and underexposed regions, and color gamut extension [Banterle et al., 2017]. Before the advent of deep learning, ITM methods

[Meylan et al., 2006; Meylan et al., 2007; Banterle et al., 2006; Akyüz et al., 2007; Banterle et al., 2007; Rempel et al., 2007; Banterle et al., 2008; Didyk et al., 2008; Kovaleski and Oliveira, 2009; Masia et al., 2017] primarily concentrated on linearization and dynamic range expansion.

These model-driven methods primarily aim to visualize LDR content on HDR displays and perform adequately when the input LDR is of high quality, with minimal saturation or quantization artifacts. However, they face challenges in addressing missing information in overexposed and underexposed regions. In contrast, with the advancements in deep learning, data-driven approaches have made substantial progress in recovering missing details in these problematic areas. Learning-based methods can be categorized into two approaches: direct methods, where the network directly reconstructs HDR images from a single-exposure LDR input; and indirect methods, where the network generates multiple exposure brackets first and subsequently merges them using conventional techniques discussed in Section 2.2.1.

For the direct methods, Eilertsen et al. [2017] pioneer the field by developing a VGG-based [Simonyan and Zisserman, 2014] network with a luminance and reflectance decomposition loss function to improve HDR reconstruction. Marnerides et al. [2018] propose a multi-branch network to capture features at local, medium, and global scales, effectively reducing artifacts introduced by the U-Net structure. Santos et al. [2020] draw inspiration from the partial convolution technique in inpainting [Liu et al., 2018a] and introduce a feature mask mechanism to emphasize unsaturated information in the feature map, thereby enhancing the completion of overexposed regions. Liu et al. [2020b] decompose the ITM process into subproblems, including de-quantization, linearization, and hallucination, and design dedicated sub-networks to address each subproblem individually, reducing the overall learning complexity of the network. Yu et al. [2021b] design a two-stream network to effectively balance the influence of saturated and unsaturated regions on the final output. Dille et al. [2025] utilize intrinsic image decomposition to first separate the input LDR into shading and albedo components, then perform HDR reconstruction on each component individually before fusing them, ensuring color fidelity in the final HDR.

For the indirect methods, Endo et al. [2017a] employ a 2D U-Net encoder combined with a 3D U-Net [Çiçek et al., 2016] decoder to generate bracketed images from a single LDR input, which are then merged via [Debevec and Malik, 2023] to achieve HDR reconstruction. Lee et al. [2018a] take inspiration from the super-resolution architecture EDSR [Lim et al., 2017a] and propose a cascaded network that creates multiple exposure LDR images from the input. Building on this, Lee et al. [2018b] design a recursive network structure that reuses sub-networks for exposure adjustment, reducing the number of network parameters. Le et al. [2023] extract features from differently exposed images and use a representation loss to enforce consistent representations, training exposure adjustment sub-networks to generate the brackets from the LDR input. Zhang et al. [2023b] leverage Swin-Transformer to establish long-range connections and design an end-to-end fusion module, eliminating the need for post-merging.

Also, the extension to video ITM has been explored extensively, Kim et al. [2019] generate paired training data from YouTube HDR videos via degradation and proposed a two-branch network to jointly address inverse tone mapping and super-resolution by separately reconstructing low- and high-frequency information. The following work [Kim et al., 2020] introduces the GAN regularization to further enhance the synthesis of realistic details. He et al. [2022b] develop a context-aware hierarchical module to more effectively model the correspondence between global and local features, thereby enhancing the fidelity of ITM results. Kalantari and

Ramamoorthi [2019] treated adjacent LDR video frames as exhibiting implicit multi-exposure effects and proposed a fusion-based framework for HDR reconstruction, employing a pyramid-based optical flow estimation to enhance warping accuracy. SelfVideo-ITM [Banterle et al., 2022] presents a method that exploits inter-frame similarity to recover details from non-saturated frames and transfer them to saturated frames within a video sequence.

All the above learning-based methods adopt regression approaches, which, while achieving notable results and partially addressing hallucination issues, are inherently constrained by the nature of regression tasks. These methods tend to learn mean-seeking solutions, often resulting in blurred details in the completed regions. Moreover, due to the limited availability of HDR datasets, the trained networks often struggle to generalize effectively to diverse real-world scenarios.

Addressing these challenges, this thesis introduces **GlowGAN** (Chapter 5), a framework that learns HDR content generation from unlabeled LDR images via a differentiable camera model. Furthermore, utilizing the GAN-Inversion [Roich et al., 2022] technique, the trained model functions as a prior for ITM tasks. This method reduces the dependency on extensive HDR training datasets and achieves significant improvements in the completion of overexposed regions.

This approach of leveraging generative model priors through inversion for ITM has also influenced subsequent works. Bemana et al. [2024] and Goswami et al. [2024] extend this concept to diffusion models, optimizing pre-trained diffusion models in a tune-free manner to generate multi-exposure LDR images, which are then fused to produce HDR outputs.

While these methods achieve some promising results, they still face challenges in image reconstruction quality, generalization ability, and inference time. Inspired by image fusion techniques, this thesis introduces **LEDiff** (Chapter 6), which fine-tunes pre-trained Stable Diffusion models using limited HDR data to enable high-quality HDR generation and ITM.

2.2.3 Tone Mapping

Given that a large number of current display devices still only support LDR feature, it is necessary to compress the dynamic range of HDR content for proper visualization on such devices. This process is referred to as tone mapping (TM). In general, tone mapping can be categorized into model-driven traditional approaches and data-driven learning-based approaches.

Traditional Tone Mapping The traditional tone mapping approaches generally fall into two main categories. *Global methods* apply a global mapping function to ensure uniform transformation across all pixels [Drago et al., 2003; Jack and Holly, 1993; Larson et al., 1997; Mantiuk et al., 2008]. In contrast, *local methods* adapt the mapping function spatially, allowing it to vary per pixel based on local neighborhood information [Fattal et al., 2002; Reinhard et al., 2002; Durand and Dorsey, 2002; Mantiuk et al., 2006]. Extensive reviews of both categories can be found in prior literature [Reinhard et al., 2010; Banterle et al., 2017; Mantiuk et al., 2015b].

Building upon these foundations, recent studies have focused on adaptive or automated selection of tone mapping operators (TMOs) to better suit diverse image characteristics or application scenarios. Mantiuk and Seidel [2008] introduce a unified parametric framework that models both global and local operators, facilitating standardized evaluation and backward-compatible HDR compression. To enhance rendering efficiency on LDR displays, Banterle et al. [2016] partition HDR images into

luminance regions and assign the most suitable TMO to each segment. Meanwhile, Debattista [2018] explores the use of genetic programming to evolve custom TMOs tailored for multi-tasks such as visualization, feature extraction, and compression.

Alternative approaches, such as exposure fusion [Mertens et al., 2007; Raman and Chaudhuri, 2009], generate high-quality low dynamic range (LDR) images by directly combining multiple bracketed exposures, thereby circumventing the explicit reconstruction of an HDR image.

A number of recent model-based tone mapping algorithms have demonstrated improved performance through more sophisticated formulations. For instance, Shan et al. [2009] propose a method that processes images using overlapping local windows, optimizing dynamic range compression globally while simultaneously enforcing local constraints. This approach enables the preservation of both fine-grained details and large-scale structures. Ma et al. [2015] introduce an iterative optimization framework that directly enhances the tone-mapped output by maximizing structural fidelity and statistical naturalness, guided by a tone mapping quality index (TMQI)-based metric. To mitigate common artifacts such as haloing and contrast over-enhancement, Liang et al. [2018] design a hybrid multi-scale decomposition model. Their method separates the image into a base layer, regularized with an ℓ_1 sparsity term to enforce piecewise smoothness, and a detail layer, constrained by an ℓ_0 prior to preserve structural edges. In another line of work, Li et al. [2018] propose a clustering-based approach where HDR images are segmented into color patches based on statistical properties. Each cluster is then projected into a compact domain using principal component analysis (PCA), allowing for the application of distinct tone mapping curves per cluster. While these methods achieve compelling results, they are typically model-driven and rely on handcrafted priors. Moreover, their performance often depends on careful parameter tuning, which may limit accessibility and usability for non-expert users.

Learning-based Tone Mapping Given the increasing impact of deep learning on image processing, numerous learning-based tone mapping operators have emerged in recent years. The majority of these approaches are *supervised*, requiring paired HDR-LDR images for training.

Patel et al. [2017] adopt a generative adversarial network (GAN) to learn an ensemble of traditional tone mapping operators that generalize well across diverse scenes. During training, the target LDR image is selected as the one achieving the highest score under the TMQI metric [Yeganeh and Wang, 2012] from a pool of tone-mapped candidates generated by various traditional methods. Rana et al. [2019] follow a similar supervision strategy but adopt a multi-scale conditional GAN architecture to enhance local detail preservation.

Instead of relying solely on objective metrics, Zhang et al. [2019b] introduce a loss function tailored to encourage the resulting tone-mapped images to lie closer to the natural image manifold. Their training targets are manually curated by professional photographers using tools such as Photomatix2 [1] and HDRToolbox [Banterle et al., 2017]. Extending this idea, Su et al. [2021] propose an explorable tone mapping network based on BicycleGAN [Zhu et al., 2017b], where target images are generated using LuminanceHDR [2] and selected via TMQI-based ranking.

To address the challenge of acquiring suitable supervision, Panetta et al. [2021] utilize low-light images as proxies for HDR data, arguing that their under-exposed

[1] https://www.hdrsoft.com/

[2] https://sourceforge.net/projects/qtpfsgui/

regions statistically resemble HDR image distributions, especially on LDR displays. Meanwhile, Yang et al. [2018] adopt a two-stage pipeline: they first restore HDR images from corrupted LDR inputs using an end-to-end network, then apply tone mapping with supervision derived from Photoshop-generated dataset.

More recently, semi-supervised strategies have been explored. Zhang et al. [2021] combine unsupervised objectives with a supervised loss term, requiring only a small set of HDR-LDR pairs. For supervision, they reuse the tone-mapped images from [Zhang et al., 2019b]. Similarly, Guo and Jiang [2021] propose a semi-supervised approach inspired by image quality assessment metrics, employing Photoshop-fine-tuned bracketed exposures as pseudo ground truth.

Despite their promising results, these learning-based methods are inherently constrained by the quality and diversity of their training data. Using tone-mapped outputs from existing operators as ground truth enables network training but also imposes an upper bound on achievable quality. In contrast, we propose a self-supervised framework, **SelfTMO**, in Chapter 7, that directly learns a tone mapping operator from HDR inputs alone. Our method leverages a perceptually informed loss function that incorporates contrast masking characteristics of the Human Visual System (HVS), enabling high-quality tone mapping without reliance on external ground truth.

2.3 Neural Rendering

Over the past few years, neural rendering has become an influential paradigm in computer graphics. This section introduces two representative categories: implicit neural representations (Section 2.3.1) and point-based representations (Section 2.3.2).

2.3.1 Implicit Neural Representations

The core concept of implicit neural representations lies in utilizing MLPs as a universal function approximator [Hornik et al., 1989], employing optimization techniques to fit a wide range of signals, including audio [Gao et al., 2022], shapes [Genova et al., 2019; Genova et al., 2020], objects [Park et al., 2019; Atzmon and Lipman, 2020; Michalkiewicz et al., 2019; Gropp et al., 2020], and scenes [Sitzmann et al., 2019; Jiang et al., 2020; Chabra et al., 2020], thus distinguishing them from explicit models. Sitzmann et al. [2020] propose SIREN, which utilize periodic activation functions to effectively model fine details and spatial-temporal derivatives. Tancik et al. [2020] introduce positional encoding techniques to improve training convergence and better capture high-frequency details. Lindell et al. [2022] present BACON, which enable frequency control and Fourier-based analysis to address aliasing and scale-awareness issues.

A groundbreaking application emerged with the introduction of NeRF [Mildenhall et al., 2020], which takes 5D inputs consisting of spatial coordinates and viewing directions, employs an MLP as a continuous representation, and combines it with volume rendering. By training on multi-view images, this method allows the MLP to capture a complete 3D scene, representing a major step forward in 3D reconstruction. The following work on NeRF, focusing on improving the quality and speed the rendering time. To accelerate NeRF, several approaches replace the MLP with more compact representations, significantly improving training and rendering speed. Plenoxels [Yu et al., 2023] utilizes a sparse 3D grid with spherical harmonics to directly represent the scene, eliminating the need for an MLP and achieving substantial speedups. Similarly,

NSVF [Liu et al., 2020a] leverages a sparse voxel octree to organize voxel-bounded implicit fields, modeling local scene properties while skipping irrelevant voxels via differentiable ray-marching, enhancing efficiency and scalability. SNeRG [Hedman et al., 2021] takes a different approach by first training a NeRF and then precomputing and storing a Sparse Neural Radiance Grid, where learned feature vectors represent the scene. This pre-baking strategy retains the MLP-based representation's quality while significantly accelerating rendering.

PlenOctrees [Yu et al., 2021a] accelerates rendering by using NeRF-SH, a modified NeRF model predicting spherical harmonic coefficients for view-dependent effects. The trained NeRF-SH is densely sampled to construct a PlenOctree storing density and coefficients, enabling fast rendering with optional fine-tuning for improved quality. DVGO [Sun et al., 2022] accelerates training and improves volumetric scene reconstruction by directly optimizing a dense voxel grid with priors to avoid suboptimal geometry, introducing post-activated voxel-grid interpolation for sharp boundary modeling, and significantly reducing grid resolution compared to previous methods. Chen et al. [2022a] present TensoRF, an innovative radiance field representation that utilizes tensor decomposition to efficiently encode volumetric features, achieving substantial reductions in memory usage and reconstruction time compared to NeRF, while preserving high rendering quality. Müller et al. [2022] present InstantNGP, which uses a multiresolution hash table of trainable feature vectors to augment a small neural network, enabling significantly faster training and real-time rendering without compromising quality.

Another line of research focuses on enhancing the rendering quality of NeRF. Mip-NeRF [Barron et al., 2021] leverages integrated positional encoding to prefilter radiance fields across scales, addressing aliasing and blurring in multi-resolution scenes while reducing errors. Building on Mip-NeRF, Barron et al. [2022] propose Mip-NeRF 360, extending it to unbounded scene reconstruction by integrating smooth volumetric parameterization, introducing a dual-MLP architecture for efficient density computation during ray tracing, and incorporating a distortion-based regularization to mitigate artifacts like floaters and background collapse. Grid-based methods excel in rendering efficiency, while pure-MLP-based approaches achieve superior quality. Some techniques aim to combine these strengths for both speed and fidelity. Leveraging multisampling and conical frustum feature representations, Zip-NeRF [Barron et al., 2023] integrates grid-based efficiency with mip-NeRF 360's anti-aliasing, effectively mitigating artifacts, enhancing rendering quality, and accelerating training. Building on this, MERF [Reiser et al., 2023] employs an NGP hash grid for compact parameterization and differentiable sparsification during optimization, enabling rapid convergence. After optimization, MERF encodes the radiance field into a voxel-triplane representation with a binary occupancy grid, supporting real-time rendering while maintaining high reconstruction accuracy.

2.3.2 Point Based Representations

Point-based rendering [Gross and Pfister, 2011] has undergone long-term development and is a naturally suitable rendering approach, as points can be easily obtained from RGB-D cameras or SFM [Schonberger and Frahm, 2016] and conveniently reprojected into image space. With advances in deep learning, recent works integrate point-based rendering with neural networks. Aliev et al. [2020] use raw point clouds as geometry and a U-net to learn latent descriptors, enabling photorealistic rendering without meshing or surface estimation.

Yifan et al. [2019] present Differentiable Surface Splatting (DSS), integrating EWA [Zwicker et al., 2001b] splatting into a differentiable rendering pipeline to achieve smooth, hole-free rendering and improved geometric detail preservation. A breakthrough in this direction comes from [Kerbl et al., 2023], who model scenes with 3D Gaussians as primitives (3DGS), embedding attributes such as spherical harmonics, density, rotation, translation, scale, and anisotropic covariance. These Gaussians are rendered using EWA splatting into image space, achieving high-quality, real-time novel view synthesis. Subsequent works, Mip-Splatting [Yu et al., 2024] and multi-scale 3DGS [Yan et al., 2024] mitigate aliasing artifacts in the original 3DGS. Mip-Splatting enhances image quality in zoomed-in and zoomed-out views through 3D size constraints and 2D Mip filters, while multi-scale 3DGS maintains Gaussians at different scales, using smaller Gaussians for high resolution and larger ones for low resolution.

All methods mentioned above typically assume high-quality, artifact-free inputs. Recent works have explored more realistic distortions, such as motion/defocus blur, noisy and multiple exposure condyions. Ma et al. [2022] enhance NeRF with a Deformable Sparse Kernel (DSK) to reconstruct sharp 3D scenes. Lee et al. [2023] improve geometric and appearance consistency in blurred images using physical priors and adaptive weighting. Wang et al. [2023c] address motion blur in NeRF by modeling camera motion trajectories with a linear SE(3) model, while Wu et al. [2022] extend NeRF to handle shallow depth-of-field inputs via a differentiable circle of confusion. In point-based methods, Lee et al. [2024] propose Deblurring Splatting, refining 3D Gaussian covariance and point cloud sparsity to handle motion and defocus blur inputs. Zhao et al. [2024] introduce BAD-Gaussians, modeling camera trajectories during exposure to enable real-time rendering from motion-blurred images.

Huang et al. [2022] reconstruct HDR radiance fields from LDR images with varying exposures, integrating the physical imaging process into NeRF to generate novel HDR and LDR views. Jun-Seong et al. [2022] adopt a plenoptic function as a scene representation for HDR radiance fields, eliminating the reliance on exposure information during training. HDR-GS [Cai et al., 2024a] extends HDR-NeRF to 3D Gaussian Splatting, significantly improving training and inference efficiency. RawNeRF [Mildenhall et al., 2022] trains directly on linear raw images, preserving the full dynamic range for high-fidelity HDR view synthesis. Xu et al. [2023b] introduce a perceptually linear mapping of the full dynamic range, constraining HDR values to the unit range for improved radiance field learning. Rückert et al. [2022] propose a point-based differentiable neural rendering framework that optimizes scene parameters, enabling high-quality novel view synthesis with fast and high-fidelity rendering of large-scale point clouds. Other applications include image alignment and layering [Nam et al., 2022], video fitting with latent codes [Feng et al., 2022], video editing with layered 2D atlases parameterized by MLPs [Kasten et al., 2021], and enhancing mixed reality rendering coherence through learned camera characteristics [Mandl et al., 2021].

2.4 Generative Models

This section introduces generative models, focusing on two of the most popular types: Generative Adversarial Networks (GANs) in Section 2.4.1 and Diffusion Models in Section 2.4.2.

2.4.1 Generative Adversarial Network (GAN)

Generative Adversarial Network (GAN) [Goodfellow et al., 2014] employ a generator G to map noise to images and a discriminator D to estimate the likelihood ratio, implicitly optimizing maximum likelihood by minimizing the Jensen-Shannon divergence for realistic image generation. However, the original GAN suffered from instability and mode collapse. Radford et al. [2015] improve the architecture by substituting fully connected layers with convolutional layers, utilizing batch normalization for stabilized training, and adopting leaky ReLU activation to improve gradient flow, making GANs more stable and enhancing the diversity. Arjovsky et al. [2017] address the mode collapse and vanishing gradients by using the Wasserstein distance instead of the Jensen-Shannon divergence, providing a more meaningful loss metric for measuring the difference between real and generated distributions. Gulrajani et al. [2017] further improve WGAN by introducing a gradient penalty to enforce the Lipschitz continuity constraint, eliminating the need for weight clipping and leading to more stable training dynamics. To improve the resolution,

Karras et al. [2017] introduce a novel training strategy where low-resolution layers are trained first, and higher-resolution layers are added progressively. This method significantly improved the stability of training and enabled the generation of high-resolution (1024×1024) images for the first time. Karras et al. [2019] introduce an adaptive style-based synthesis architecture, where the generator does not directly take a latent vector as input but instead applies adaptive instance normalization (AdaIN) at multiple layers to control style at different scales, allowing finer control over generated images, and leading to higher realism and better disentanglement of attributes. Karras et al. [2020b] address artifacts present in StyleGAN by replacing AdaIN with a more direct weight demodulation approach, improving normalization, and redesigning progressive growing into a more stable architecture. Karras et al. [2020a] introduce adaptive discriminator augmentation, which applies stochastic data augmentations dynamically based on discriminator feedback, which enables StyleGAN to train effectively on extremely limited data while reducing overfitting. Karras et al. [2021] further resolve aliasing artifacts and enforced strict spatial consistency in the generator, making it more robust for applications like video synthesis and 3D-aware image generation. Sauer et al. [2022] optimize generator scaling, regularization, and training objectives for efficient large-scale, high-resolution synthesis (1024×1024 and beyond), and enable better generalization across diverse datasets like ImageNet.

GANs have been extensively used in numerous applications, such as image translation, image restoration, 3D generation from 2D observations. cGAN [Mirza and Osindero, 2014] extended standard GANs by incorporating conditional inputs, enabling controlled generation based on auxiliary labels, thereby establishing the foundation for structured image generation. Pix2Pix [Isola et al., 2017] utilizes conditional GANs to perform image-to-image translation by learning a direct mapping between paired and spatially aligned images, which proves effective for tasks such as translating semantic layouts into photorealistic images. To address the limitation of requiring paired training data, CycleGAN [Zhu et al., 2017a] introduce cycle consistency loss, ensuring that translating an image to another domain and back reconstructs the original input, enabling unsupervised domain adaptation and style transfer without requiring aligned datasets. Furthermore, by conditioning on domain labels, StarGAN [Choi et al., 2018] supports multi-domain image-to-image translation with a single network, avoiding the need for training domain-specific models and improving scalability.

SRGAN [Ledig et al., 2017] incorporates perceptual loss alongside GAN loss to enhance visual realism in super-resolution results. Building on SRGAN, Wang et al. [2018b] introduce a relativistic discriminator and residual-in-residual dense blocks, achieving sharper and more detailed images. Yu et al. [2018] develop a GAN-based inpainting model with contextual attention, enabling the generator to reconstruct missing regions by referencing known textures. GANs have also been extensively applied to other image restoration tasks, including denoising [Cha et al., 2019; Zamir et al., 2020], deblurring [Kupyn et al., 2018; Kupyn et al., 2019], and deraining [Zhang et al., 2019a; Yasarla et al., 2020].

The aforementioned methods rely on paired data for training, while an alternative approach leverages GAN inversion (tune-free) techniques to exploit the generative prior of GANs for image restoration. GAN inversion aims to project an image back into the latent space, facilitating manipulation in a more structured manner. Current approaches fall into three main groups: (1) Optimization-based approaches, which refine the latent code [Creswell and Bharath, 2018; Albright et al., 2019] or jointly optimize both the latent code and the generator via gradient backpropagation [Pan et al., 2021b; Roich et al., 2022]; (2) Encoder-based methods, which train an additional encoder to directly map images to their latent codes [Zhu et al., 2016; Donahue et al., 2016]; and (3) Hybrid approaches, which first obtain a coarse latent representation via an encoder and subsequently refine it through optimization [Zhu et al., 2020]. Building on GAN inversion, Pan et al. [2021b] propose DGP, which utilizes the prior of a pre-trained GAN [Brock et al., 2018] for image restoration. Inspired by Deep Image Prior (DIP) [Ulyanov et al., 2018], which models image priors solely from the image itself using a CNN, DGP instead leverages the richer information encoded in a pre-trained GAN. This approach eliminates the need for paired training data while enabling applicability across multiple downstream tasks.

Another limitation of vanilla GANs is their reliance on extensive high-quality datasets, such as FFHQ [Karras et al., 2019], CelebA [Liu et al., 2018b], and SUN [Xiao et al., 2010]. However, in many scenarios, collecting such high-quality data is challenging, particularly for domains like 3D data. AmbientGAN [Bora et al., 2018] addresses this issue by learning the full signal distribution from lossy measurements. It models lossy data as a degraded version of the full signal and designs differentiable degradation models tailored for various degradations (e.g., noise, blur). This approach enables GANs to learn the full signal distribution directly from lossy measurements, significantly reducing dependence on high-quality datasets. This idea has inspired numerous subsequent works on learning 3D from 2D [Wu et al., 2020; Pan et al., 2020; Pan et al., 2022; Pan et al., 2021a; Chan et al., 2022], as 2D images can be viewed as projections of a 3D scene under different camera poses.

2.4.2 Diffusion Model

With diffusion models becoming a strong contender to GANs, they offer a more stable alternative by mitigating issues like mode collapse and training instability. Unlike GANs, diffusion models generate images through a learned denoising process, progressively refining outputs to achieve high-quality synthesis with enhanced diversity and reliability.

Early work on diffusion-based generative models was formalized by Denoising Diffusion Probabilistic Models (DDPM) [Ho et al., 2020], which introduces a stochastic forward process that incrementally corrupts the image with noise, along with a learned reverse process to recover it. Subsequent improvements, such as Improved DDPM [Nichol and Dhariwal, 2021], enhanced sample efficiency and introduced

classifier guidance to improve conditional generation. To further accelerate inference, Denoising Diffusion Implicit Models (DDIM) [Song et al., 2020a] propose a deterministic formulation of the reverse process, which allows for fewer sampling steps without compromising generation quality. Subsequently, Latent Diffusion Models (LDMs) [Rombach et al., 2022] propose applying diffusion process in a lower-dimensional latent domain rather than directly in pixel space, drastically improving computational efficiency while preserving high-fidelity synthesis.

The success of diffusion models has led to breakthroughs in text-to-image synthesis. GLIDE [Nichol et al., 2021] demonstrates the effectiveness of diffusion for conditional image generation, later extended by Imagen [Saharia et al., 2022b] and Stable Diffusion [Rombach et al., 2022], both of which leverage large-scale training to improve text alignment and image quality. DALL·E 2 [Ramesh et al., 2022] introduce a hierarchical approach by conditioning the diffusion process on CLIP embeddings, achieving fine-grained text-image alignment. Beyond generation, recent work has explored the use of diffusion models in image editing. Palette [Saharia et al., 2022a] explored image-to-image translation tasks, such as colorization and super-resolution, while InstructPix2Pix [Brooks et al., 2023a] enables flexible modifications guided by natural language instructions. Diffusion models have also been extended beyond static images to more complex domains. Make-A-Video [Singer et al., 2022] generates video through text-to-image diffusion model, producing temporally consistent motion. In the realm of 3D generation, DreamFusion [Poole et al., 2022] leverages the combination of NeRF and text-to-image diffusion models to produce 3D representations guided by text prompts.Additionally, 3DiM [Watson et al., 2022] and LION [Vahdat et al., 2022] apply diffusion models to 3D shape generation, offering improvements over traditional voxel-based or GAN-based approaches.

While standard diffusion models generate images in an uncontrolled manner, many tasks require conditional control to guide the generative process. Recent advancements have introduced techniques to incorporate structured conditioning, enabling fine-grained control over diffusion-based synthesis.

One prominent approach to spatial control is ControlNet [Zhang et al., 2023a], which extends pretrained Stable Diffusion by introducing trainable branches that process structural conditions such as edge maps, depth maps, and pose keypoints, ensuring faithful adherence to input constraints while preserving the original model's generative capabilities. Similarly, T2I-Adapter [Mou et al., 2024] introduces a lightweight conditioning framework that augments text-to-image diffusion models with segmentation masks and additional spatial priors without modifying the base model, offering flexible control with minimal computational overhead. Meanwhile, DreamBooth [Ruiz et al., 2023] enables personalized generation by fine-tuning diffusion models on a small set of reference images, preserving identity and visual coherence in new contexts. Marigold [Ke et al., 2024] conditions the latent space by using an RGB image as input while applying noise and denoising on the corresponding depth map, resulting in a high-quality single image to depth framework. A similar approach has also been applied to image-to-normal estimation [Fu et al., 2024] and image matting [Wang et al., 2024e].

Chapter 3

An Implicit Neural Representation for the Image Stack

This chapter introduces a unified approach for reconstructing all-in-focus high-dynamic-range (HDR) images from an image stack by combining neural fields with an expressive camera model. We leverage a set of specialized implicit neural representations, each addressing a key sub-problem: neural implicits predict flow to correct misalignments from lens breathing and depth variations, model depth of field for all-in-focus reconstruction, and apply tone mapping to handle sensor responses and saturation. These components are trained within a physically inspired framework, centered on a differentiable thin lens model. A key advantage of our method is its ability to handle these tasks jointly, enabling flexible post-editing operations such as refocusing and exposure adjustment. By systematically sampling focal distance, aperture, and exposure time, we gain valuable insights into their influence on reconstruction quality. Extensive validation demonstrates that our approach outperforms existing methods in depth-from-defocus and all-in-focus image reconstruction, while also showing promising results across all three photographic dimensions, offering enhanced image quality and post-processing control.

3.1 Introduction

Capturing and reproducing the real world is a fundamental goal of image processing. However, due to imaging device limitations, capturing all relevant information in a single shot is often challenging. Image fusion techniques improve image quality by combining multiple images. Traditionally, all-in-focus (AiF) image reconstruction [Li et al., 2013; Zhang and Levine, 2016; Liu et al., 2017; Suwajanakorn et al., 2015; Si et al., 2023; Bouzos et al., 2019], depth estimation [Favaro, 2010; Lin et al., 2013; Moeller et al., 2015; Suwajanakorn et al., 2015; Hazirbas et al., 2019; Maximov et al., 2020; Won and Jeon, 2022; Yang et al., 2022; Si et al., 2023], and high-dynamic-range (HDR) reconstruction [Debevec and Malik, 2023; Mertens et al., 2007] are treated separately, requiring multiple images with varying focus and exposure settings. This process is time-consuming and leads to redundant image capture.

In this work, we address these challenges by focusing on the reconstruction of AiF HDR radiance maps from a sparse image burst, where each image may contain entangled information from both defocus blur and different exposure conditions. Our approach simplifies image capture, improves fusion efficiency, and enables post-editing capabilities such as refocusing, bokeh reproduction, and exposure adjustment. We combine an implicit neural representation with an expressive camera model to reconstruct both the AiF HDR image and the corresponding depth from the sparse image burst without the need for ground-truth supervision.

The problem we address is notably challenging and requires a comprehensive approach due to the interplay of three crucial elements in the image capture process: *focal distance*, *aperture*, and *exposure time*, which we model in our approach to simulate the imaging pipeline (Section 6.2). *Focal distance* determines which parts of the scene will appear sharp and in focus. During focal distance adjustment in a focal stack, lens breathing usually appears, leading to pixel misalignment and ghosting artifacts in captured images. The *aperture* controls lens opening, affecting both the amount of light reaching the sensor and depth of field. Our framework incorporates a differentiable thin lens formulation to model both focal distance and aperture. Additionally, we introduce a novel approach to mitigate lens breathing using an implicit flow, which enables warping each image in the focal stack to a canonical view, eliminating the need for a reference image and preserving all pixel information. While few existing learning-based methods have attempted to address lens breathing, they focus on a single specific problem such as depth from defocus (DfD) [Won and Jeon, 2022; Suwajanakorn et al., 2015; Moeller et al., 2015] and often result in a loss of pixel information. Our framework also tackles the problem of DfD by leveraging a thin-lens formulation as a physical constraint to model shallow depth of field in the input images, outperforming the state of the art in both depth estimation and AiF image reconstruction (Section 3.4.1). Finally, *exposure time* controls the light exposure of the image sensor. Longer exposures capture more light, resulting in a brighter image. However, extended exposures can also saturate the sensor and lead to clipping. To model this process, we introduce an implicit tone mapper that supports the reconstruction of the HDR radiance map from the different exposures in the low-dynamic-range (LDR) input images.

Relying on these components, our framework achieves high flexibility. In contrast to previous approaches [Debevec and Malik, 2023; Mertens et al., 2007; Hasinoff and Kutulakos, 2007; Won and Jeon, 2022; Si et al., 2023; Yang et al., 2022; Suwajanakorn et al., 2015], it supports as input any set of images with different unstructured combinations of focal distance, aperture, and exposure to learn an AiF HDR image and its corresponding depth in an unsupervised manner. We term this stack of unstructured images *Time-Aperture-Focus stack* (TAF). Our approach can handle an arbitrary number of input images and produces outstanding results with only five images. To achieve this, the estimated depth serves as an input to our lens model, which produces a differentiable disk kernel. This kernel is applied to blur the AiF image, effectively modeling the combined effects of focus and aperture. In a final step, the blurred image is tone mapped to produce a defocused LDR image. Then, a reconstruction loss is computed by comparing this output image and the input image from the TAF stack, guiding the iterative refinement of the predicted AiF image and depth map. After optimization, our framework further enables post-editing capabilities such as focus, aperture, and exposure editing (Figure 3.1).

Further, we extensively investigate the impact of different combinations of focal distance, aperture, and exposure on the overall reconstruction performance of our method (Section 3.4.2), revealing two effective strategies for improving AiF HDR image reconstruction and depth estimation. On one hand, utilizing a moderate aperture size and multiple exposure times with focal sweeping yields better AiF HDR image reconstruction. On the other hand, maintaining a fixed exposure time while varying the aperture size during focal sweeping leads to improved depth estimation.

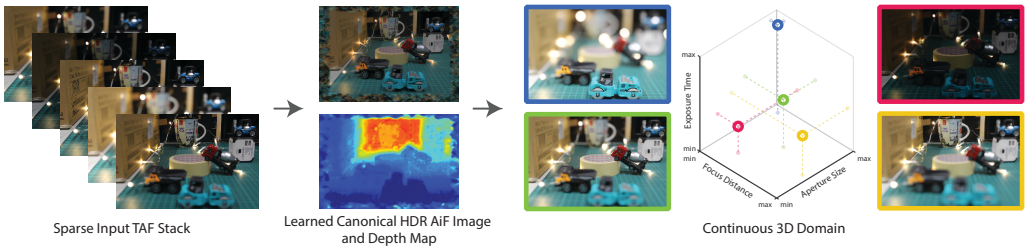


FIGURE 3.1: *Our approach is capable of accurately reconstructing a canonical all-in-focus (AiF), high-dynamic-range (HDR) radiance map alongside with depth information, using a sparse Time-Aperture-Focus (TAF) stack as input. This gives full post-processing control over focus, aperture, and exposure conditions.*

3.2 Background

We first discuss previous work addressing the problem of obtaining depth and AiF from the focal stack, which is one of the main applications of our framework. Then, based on the summary in Section 2.3 regarding implicit neural representations and their recent applications related to our task, we further explain the motivation behind adopting implicit neural representations in our work.

3.2.1 Depth and AiF From Focal Stacks

Traditional depth estimation from a focal stack can be categorized into depth from focus (DfF) and depth from defocus (DfD) methods. DfF approaches face challenges in determining a suitable focus criterion that accurately captures the focus information across various scenes and conditions, and effectively detecting the highest focus values and accurately localizing the corresponding depths. DfD approaches, on the other hand, face challenges in accurately capturing depth information from defocused images and balancing depth estimation accuracy with image resolution. Despite notable advancements [Moeller et al., 2015; Suwajanakorn et al., 2015; Favaro, 2010; Lin et al., 2013], the quality of depth estimation in traditional methods still falls short of achieving satisfactory results. With the advent of deep learning, the differences between DfF and DfD have diminished, as both approaches usually utilize focal stacks as input and ground-truth depth as supervision to train end-to-end models [Maximov et al., 2020; Hazirbas et al., 2019; Yang et al., 2022; Won and Jeon, 2022]. Wang et al. [2021] presented a novel approach for jointly estimating depth and an AiF image from an input focal stack by designing a shared common network that can be trained either supervised with ground-truth depth maps or unsupervised with only ground-truth AiF images. Recently, in concurrent work, Si et al. [2023] proposed a self-supervised framework for DfD which also considers a thin lens model for predicting both depth and AiF images from focal stacks while being supervised through input reconstruction. However, their end-to-end network is trained on a synthetic focal stack dataset, which may not generalize well to real-world scenarios. Additionally, they do not consider lens breathing and struggle with a small number of input images. We show and discuss the limitations of their approach through quantitative and qualitative comparisons. Despite significant advancements in learning-based approaches, multiple problems remain unsolved. For instance, obtaining real focal stacks with ground-truth depth for training is challenging due to factors like lens breathing. Although synthetic datasets can circumvent this issue, they often introduce domain gaps [Yang et al., 2022; Si et al., 2023]. Moreover, while many learning-based methods treat DfD and DfF as a regression problem, defocus

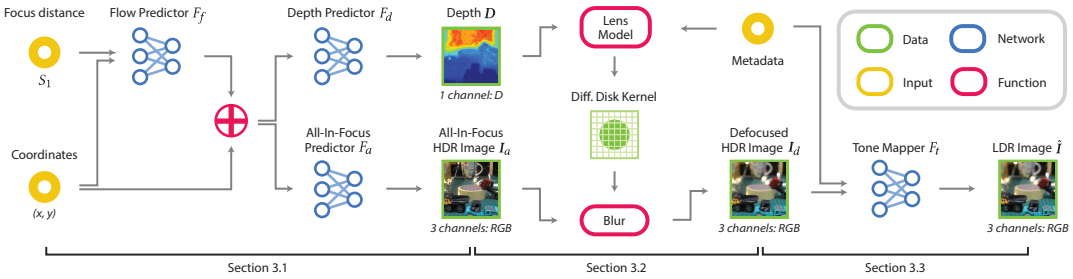


FIGURE 3.2: *Pipeline of our method. We represent a depth map and AiF HDR image via coordinate-based MLPs. An implicit flow predictor is adopted to model the lens breathing effect. Given the depth map, AiF HDR image, and the metadata of the camera, we calculate the defocused HDR image via a novel thin-lens model that is differentiable w.r.t. the depth map. Finally, the defocused HDR image is projected to LDR via an implicit tone mapper, again modeled by an MLP.*

blur can be more effectively modeled as a physical phenomenon that holds implicit cues for direct depth inference. Our approach does not rely on ground-truth data for supervision during training and, unlike previous works, is highly flexible as it can handle diverse combinations of focus, aperture and exposure in the TAF stack as input to reconstruct the AiF HDR image and corresponding depth. Further, our approach outperforms state-of-the-art approaches in reconstructing depth from the focal stack.

3.2.2 Implicit Neural Representations and Their Applications

Section 2.3 highlights the potential of fully connected networks for building memory-efficient and continuous implicit neural representations (INRs). Implicit neural representations are first successfully applied to image representation [Sitzmann et al., 2020] and later extended to 3D scene modeling [Mildenhall et al., 2021; Mildenhall et al., 2022; Wu et al., 2022; Ma et al., 2022]. Subsequent work has explored their use in tasks such as image alignment and layering [Nam et al., 2022], video fitting with latent codes [Feng et al., 2022], video editing via layered 2D atlases parameterized by MLPs [Kasten et al., 2021], and mixed reality rendering [Mandl et al., 2021]. Motivated by their strong representational power and flexibility, we adopt INR as the representation for the TAF image stack.

3.3 Method

Our goal is to design an implicit neural representation that, by learning from an image stack, models the AiF HDR image and depth map of the scene. The pipeline of our approach is shown in Figure 3.2. We first introduce how we represent the depth and the AiF HDR image (Section 3.3.1). Importantly, we present an implicit flow model to compensate for the lens breathing effect in the image stack. Next, we describe our differentiable thin lens model that renders a defocused HDR image from the depth and AiF image (Section 3.3.2). We then introduce an implicit tone mapper that maps the defocused HDR image to an LDR image conditioned on the level of exposure (Section 3.3.3). Finally, we discuss the loss functions used to train the implicit neural representation (Section 3.3.4). After training, the representation can be used to synthesize images under new focal distances, apertures, and exposures.

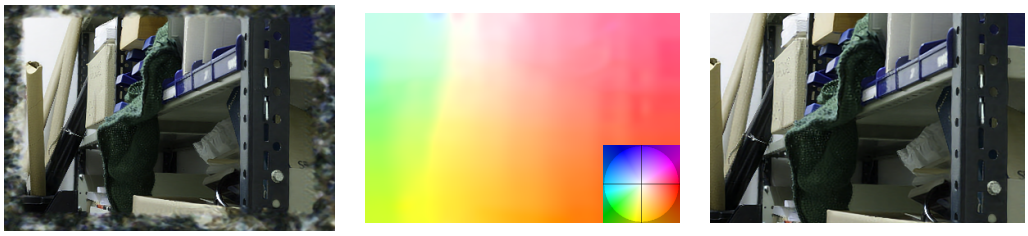


FIGURE 3.3: *Visualization of our learned flow. We warp the learned canonical view to images corresponding to different focal distances, simulating lens breathing.*

3.3.1 Depth and Image Representation with Implicit Flow

Input to our system is an image stack $\{\boldsymbol{I}_i\}_{i=1}^n$. For each image $\boldsymbol{I}_i$, we also have its corresponding camera metadata including focal distance, aperture, and exposure, which differ between images in the stack. We aim to learn a shared depth map $\boldsymbol{D}$ and AiF HDR image $\boldsymbol{I}_a$ from the image stack.

In our pipeline, we represent the depth map $\boldsymbol{D}$ and AiF HDR image $\boldsymbol{I}_a$ using coordinate-based MLPs. A simple design is to use two MLPs that take pixel coordinates (x, y) as input and output the depth $S_2 \in \mathbb{R}^+$ and RGB color $\boldsymbol{C} \in \mathbb{R}^3$ respectively. However, this design does not take into account lens breathing, the phenomenon where changing the focus distance causes a slight change in the field of view (FoV) [Gross, 2005]. This effect occurs due to the movement of the lens elements inside the lens barrel to adjust the focus distance and can be observed even when the camera is stationary. Consequently, it introduces misalignments between the same pixel coordinates of images with different focus distances.

Previous techniques for addressing lens breathing typically use the smallest FoV image in a focal stack as a reference and align all other images in the stack to it by cropping and resizing using feature extraction and matching [Suwajanakorn et al., 2015; Won and Jeon, 2022]. However, this approach may cause a loss of information in the focal stack and can only generate a single FoV AiF image. In this paper, we propose a novel approach to address lens breathing by learning a canonical view from the focal stack that encompasses all pixel information. This approach employs an implicit flow to project the canonical view onto each FoV in the stack, instead of aligning to a specific reference image. To account for lens breathing, we embed the focus distance S_1, which is the key factor in lens breathing, into the coordinate input, and use an MLP F_f as the flow predictor to predict an offset in the x and y coordinates. This continuous offset compensates for the lens breathing effect and warps pixel locations to canonical coordinates:

$$(\triangle x, \triangle y) = F_f\left(x, y, S_1\right), \tag{3.1}$$

$$\left(x', y'\right) = \left(x + \triangle x, y + \triangle y\right). \tag{3.2}$$

This way, pixels corresponding to different focus distances are aligned in the canonical space. An example of the learned flow is shown in Figure 3.3.

The canonical coordinates are then used to query the depth and HDR color with two MLPs: $F_d : (x', y') \mapsto S_2$ and $F_a : (x', y') \mapsto \boldsymbol{C}$. Thus, the full depth map $\boldsymbol{D}$ and AiF image $\boldsymbol{I}_a$ can be obtained by querying the MLPs using the coordinates grid of all pixels, as illustrated in the left half of Figure 3.2.

3.3.2 Differentiable Thin Lens Model

Given the AiF HDR image and depth map, our next step is to estimate the defocused HDR image. Defocus blur is a prevalent phenomenon in everyday photography that arises when a camera lens fails to focus light rays onto a single point on the image sensor. The thin-lens model [Potmesil and Chakravarty, 1981] can be used to explain defocus blur in an image, where the camera lens is parameterized by a focal length f, focus distance S_1, and aperture N (i. e., F-number). When an object is in focus, the distance between the lens and the image sensor is adjusted to ensure that the object's image is formed on the image sensor. However, if the object is not at the focal distance, the lens will produce a disk on the sensor rather than a single point. This disk is referred to as the circle of confusion (CoC), and its diameter d can be computed as

$$d = \frac{|S_2 - S_1|}{S_2} \frac{f^2}{N\,(S_1 - f)}. \tag{3.3}$$

Once we have obtained d, we can simulate the defocus blur via a spatially-varying convolution. Let $\boldsymbol{I}_a(x, y)$ be the pixel value of $\boldsymbol{I}_a$ at (x, y) and $W(u, v, x, y)$ be the weight of a spatially-varying kernel, where x and y represent the current position of the convolution in image space, u and v is the location within the kernel. The defocused image $\boldsymbol{I}_d$ of the spatially-varying convolution is given by

$$\boldsymbol{I}_d(x, y) = \iint \boldsymbol{I}_a(x - u, y - v) W(u, v, x, y)\, \mathrm{d}u\, \mathrm{d}v. \tag{3.4}$$

Following the discussion above, we model W using a disk of unit energy with diameter d. Notice that a disk kernel is a closer approximation to the defocus operation than a commonly used Gaussian kernel [Potmesil and Chakravarty, 1981]. However, a conventional disk kernel is discontinuous and therefore not differentiable with respect to its diameter d [Bangaru et al., 2021]. Consequently, the gradients cannot be back-propagated to the depth S_2 and thus the depth map cannot be optimized. In this paper, we propose an approach inspired by [Bruckstein et al., 2012], to address this issue by introducing a differentiable disk kernel. This kernel enables smooth optimization while accurately representing the defocus blur effect. To connect the continuous CoC diameter d to the discrete kernel grid, we introduce a soft boundary of the kernel as follows:

$$\hat{W}(u, v, x, y) = \begin{cases} 1, & m \leq \frac{d-1}{2} \\ \frac{d+1}{2} - m, & \frac{d-1}{2} < m \leq \frac{d+1}{2} \\ 0, & \frac{d+1}{2} < m \end{cases} \tag{3.5}$$

$$W(u, v, x, y) = \frac{\hat{W}(u, v, x, y)}{\sum_{u,v} \hat{W}(u, v, x, y)} \tag{3.6}$$

where m represents the distance from (u, v) to the center of the disk, and (u, v) are evaluated on the discrete pixel grid. Notice how this is a simple and efficient approximation of anti-aliasing, which does not require integrating over pixel area. This allows the gradients to be back-propagated through d, ensuring smooth optimization of the depth map. Figure 3.4 illustrates our kernels.

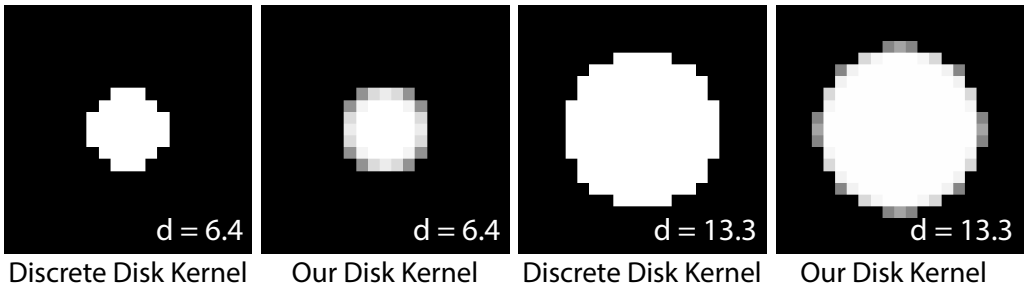


FIGURE 3.4: *Discrete disk kernels vs. our differentiable versions.*

3.3.3 Implicit Tone Mapper

In order to infer HDR information from an LDR image stack with different levels of exposure, we require a tone mapping operation that projects HDR to LDR.

Typically, the following function is used for tone mapping [Liu et al., 2020b; Wang et al., 2023a]:

$$\hat{\boldsymbol{I}}(x, y, EV) = \gamma \left(\text{clip} \left(\boldsymbol{I}_d(x, y) \cdot 2^{EV} \right) \right), \tag{3.7}$$

where $\hat{\boldsymbol{I}}$ is an LDR image, and EV quantifies the level of exposure, i. e., , the amount of light that reaches the sensor, depending on aperture and exposure time. The clip operation is employed to constrain the range of values within the interval [0, 1], and $\gamma(\cdot)$ is the gamma correction. However, there is still a gap between this approximation and real camera response curves, and we found that directly incorporating this explicit model into our framework produces artifacts (Section 3.4.3).

To tackle this challenge, we use an implicit tone mapper F_t, again represented by an MLP:

$$\hat{\boldsymbol{I}}_j(x, y, EV) = F_t \left(\boldsymbol{I}_{d,j}(x, y) \cdot 2^{EV} \right), \tag{3.8}$$

where subscripts j denote color channels, i. e., , each channel is processed independently using the same network. To ensure an LDR output, we employ a hyperbolic tangent as the last layer of the MLP and linearly remap its outputs to [0, 1].

3.3.4 Loss

We jointly optimize all four MLPs using the loss function

$$\mathcal{L} = \sum_{i=1}^{n} \left(\mathcal{L}_{\text{rec}}(\hat{\boldsymbol{I}}_i, \boldsymbol{I}_i) + \alpha \mathcal{L}_{\text{VGG}}(\hat{\boldsymbol{I}}_i, \boldsymbol{I}_i) + \beta \mathcal{L}_{\text{reg}} + \gamma \mathcal{L}_{\text{TV}} \right), \tag{3.9}$$

where $\mathcal{L}_{\text{rec}}$ is the MSE loss, $\mathcal{L}_{\text{VGG}}$ is a perceptual loss [Johnson et al., 2016], $\mathcal{L}_{\text{reg}}$ is a regularization term for flow, and $\mathcal{L}_{\text{TV}}$ is a total variation regularization penalty for the depth map. α, β and γ are weights to balance the components, which are set to $\alpha = 0.01$, $\beta = 0.5$, and $\gamma = 0.05$ in all our experiments. The combination of MSE and perceptual loss is a common practice [Feng et al., 2022; Liu et al., 2020b]. For the perceptual loss $\mathcal{L}_{\text{VGG}}$, we compute squared differences of intermediate feature maps of a pre-trained VGG16 [Simonyan and Zisserman, 2014] network (layers 5, 10, and 16). To enforce smoothness in our implicit flow, we include the regularization term

$$\mathcal{L}_{\text{reg}} = \sum \left\| J_{F_f}(x, y, f_d) \right\|_1, \tag{3.10}$$

where J_{F_f} is a Jacobian matrix assembled from the gradients of F_f with respect to all input coordinates. Finally, to encourage smoothness in the estimated depth values, we

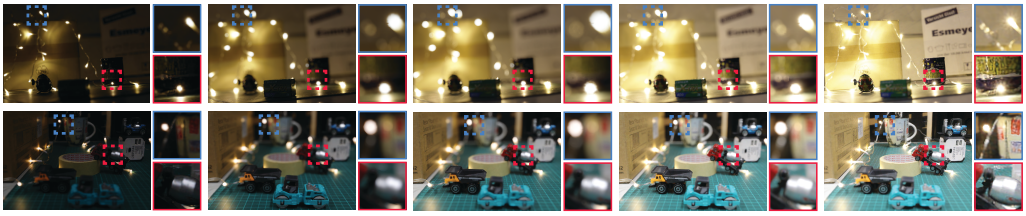


FIGURE 3.5: *Novel views with varying exposure time, aperture and focus distance synthesized by our method*

apply the total variation loss $\mathcal{L}_{\text{TV}}$ on the depth map $\boldsymbol{D}$, penalizing depth differences of neighboring pixels in both the horizontal and vertical direction.

After training, we can synthesize new images by feeding any focus distance, aperture, and exposure into the pipeline. Results are presented in Figure 3.5.

3.4 Evaluation

In this section we demonstrate the advantages of our method in a series of experiments. In Section 3.4.1 we evaluate depth-from-defocus (DfD) and all-in-focus (AiF) reconstruction. In Section 3.4.2 we extensively investigate the three factors focal distance, aperture, and exposure time, which our approach for the first time allows to consider in a unified framework. We conduct ablation studies in Section 3.4.3, and give an overview of runtime performance in Section 3.4.4.

3.4.1 Depth-From-Defocus and All-In-Focus Reconstruction

Here, we consider the one-dimensional problem of DfD reconstruction by comparing to the state-of-the-art methods DFFMobile [Suwajanakorn et al., 2015], DFFWild [Won and Jeon, 2022], DEReD [Si et al., 2023], and DFFV [Yang et al., 2022]. Among these methods, DFFMobile and DEReD are also able to reconstruct AiF images, so we consider them as baselines for this task as well. As there is no variation in exposure for both tasks, we do not use our tone mapper in this set of experiments.

To facilitate this comparison, we consider three datasets: *(a)* A novel rendered dataset, consisting of 10 scenes, generated using path tracing, providing ground-truth data for quantitative evaluation; *(b)* a synthetic dataset generated from the RGBD NYUv2 corpus [Silberman et al., 2012] in conjunction with the camera simulator from DFFWild [Won and Jeon, 2022], consisting of 100 scenes; *(c)* a novel real dataset for qualitative analysis, captured using a Canon RP camera with EF 50mm/F1.8, RF 24-105mm/F4.0 and RF 85mm/F2.0 lenses, as well as a Canon 6D2 camera with an EF 24-105mm F/3.5-5.6 lens, consisting of 25 scenes. All datasets contain lens breathing. For each experiment, we select five images as input to the methods, each with a different focus distance.

Most of our competitors take into account lens breathing by choosing the input image with the smallest effective FoV as reference and performing alignment with respect to it, cropping away pixels close to the image boundaries in all other images. Our approach is markedly different in this regard and does not require any cropping. However, to facilitate meaningful comparisons, in the following we consider only pixels that are visible for all methods.

We use the mean absolute error (MAE), the mean squared error (MSE) and the absolute relative distance (Abs-Rel) as the metrics for DfD evaluations, which are

TABLE 3.1: *Evaluation of DfD on the rendered dataset.*

Method	MAE (↓)	MSE (↓)	Abs-Rel (↓)
DFFMobile	0.31 ± 0.14	0.13 ± 0.08	2.05 ± 2.74
DFFWild	0.30 ± 0.14	0.15 ± 0.11	0.39 ± 0.23
DEReD	0.35 ± 0.24	0.21 ± 0.23	0.80 ± 0.29
DFFV	0.37 ± 0.18	0.22 ± 0.14	0.51 ± 0.26
Ours	$\mathbf{0.22 \pm 0.11}$	$\mathbf{0.09 \pm 0.06}$	$\mathbf{0.33 \pm 0.25}$

TABLE 3.2: *Evaluation of DfD on the NYU dataset.*

Method	MAE (↓)	MSE (↓)	Abs_rel (↓)
DFFMobile	0.24 ± 0.13	0.10 ± 0.09	0.96 ± 0.01
DFFWild	0.24 ± 0.09	0.09 ± 0.07	0.62 ± 0.7
DEReD	0.21 ± 0.10	0.08 ± 0.07	0.83 ± 0.06
DFFV	0.33 ± 0.18	0.17 ± 0.16	0.65 ± 0.67
Ours	$\mathbf{0.14 \pm 0.06}$	$\mathbf{0.03 \pm 0.02}$	$\mathbf{0.09 \pm 0.06}$

commonly used in this field [Eigen et al., 2014; Laina et al., 2016; Li et al., 2017]. For the evaluation of AiF results, we utilize PSNR and SSIM [Wang et al., 2004] metrics.

Based on the results presented in Tables 3.1 and 3.2, our method demonstrates superior performance compared to other approaches on the DfD task. Similarly, from the results shown in Table 3.3, our method outperforms the baselines in the AiF task.

In Figures 3.6 and 3.7, we showcase qualitative results for DfD and AiF, confirming our numerical evaluation. We observe that DFFWild and DFFV, which require full supervision, suffer from a domain gap: While they can generate visually plausible results on the rendered dataset, their performance deteriorates significantly on other datasets. On the other hand, we see that label-free methods like DFFMobile and DEReD generally struggle to produce high-quality results across the spectrum.

Regarding the AiF task, our results appear sharper and more natural compared to other methods. DFFMobile's cropping-based approach to account for lens breathing produces undesirable black boundaries, while DEReD, which does not account for lens breathing, tends to produce noticeable ghosting artifacts.

We are further interested in the performance of all methods when data is provided that does not exhibit lens breathing. To this end, we re-ran all evaluations on a variant of the NYU dataset that does not contain this effect. On this simpler dataset, our competitors significantly increase their performance – in particular DFFMobile and DFFWild. The performance of our method increases only marginally, but still outperforms all competitors, indicating robustness.

3.4.2 Exploration of TAF Sampling Strategies

Here, we investigate different sparse sampling strategies of the TAF cube. We are interested in devising strategies that result in highest-quality reconstruction and

TABLE 3.3: *Evaluation of AiF image reconstruction on the rendered and NYU dataset.*

Method	Rendered		NYU	
	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)
DFFMobile	16.8 ± 6.5	0.61 ± 0.23	31.9 ± 5.3	0.95 ± 0.07
DEReD	17.7 ± 6.4	0.63 ± 0.25	20.3 ± 3.2	0.71 ± 0.15
Ours	$\mathbf{21.7 \pm 7.9}$	$\mathbf{0.76 \pm 0.27}$	$\mathbf{32.6 \pm 4.7}$	$\mathbf{0.96 \pm 0.04}$

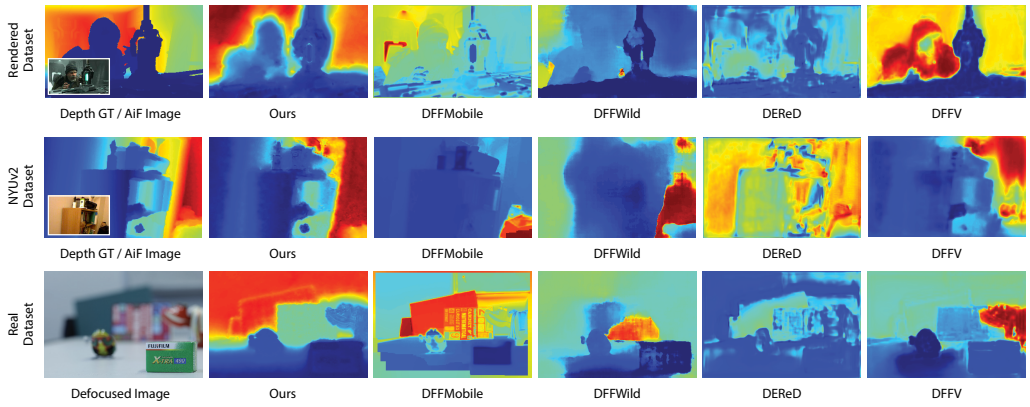


FIGURE 3.6: *Depth reconstruction results for the three datasets used in our evaluation.*

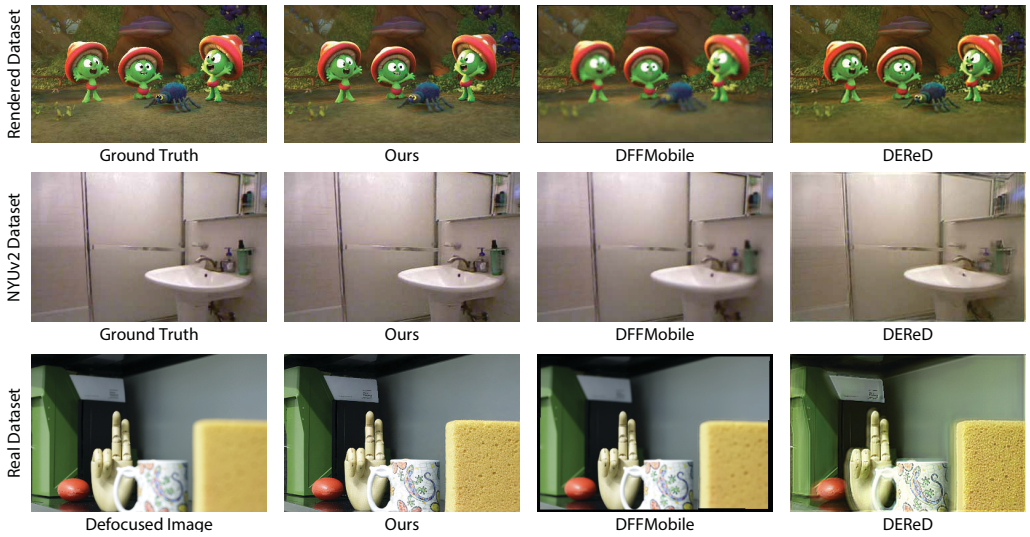


FIGURE 3.7: *All-in-focus image reconstruction for the three datasets used in our evaluation.*

disentanglement, providing guidelines for capturing TAF stacks.

The final photometric exposure H of an image is determined by both aperture and exposure time:

$$H \propto \frac{t}{N^2}. \tag{3.11}$$

Our TAF formulation allows to disentangle these factors [Jacobson et al., 2013], while previous methods often struggle with this decomposition [Huang et al., 2022; Wang et al., 2023a]. In our exploration setup, we simplify Eq. 3.11 by switching to the log domain, resulting in the linear relationship (Figure 3.8a)

$$T = 2 \cdot log_2(N) + EV, \tag{3.12}$$

where $EV = log_2(H)$, and $T = log_2(t)$. Based on this parameterization, we conduct a comprehensive set of experiments. We base our evaluation on five path-traced scenes. For each scene, we render $5 \times 5 \times 5 = 125$ samples on a parallelepiped in the linearized TAF space (Figure 3.8b).

In the 2D case, we manually select four distinct cross-sections. Three of these cross-sections represent scenarios where the aperture is fixed at small, medium, and large values (denoted as F_S, F_M, and F_L), while varying the exposure time and focus distance. The fourth cross-section represents a scenario where exposure time is fixed, while aperture and focus distance vary (denoted as F_V). Within each cross-section,

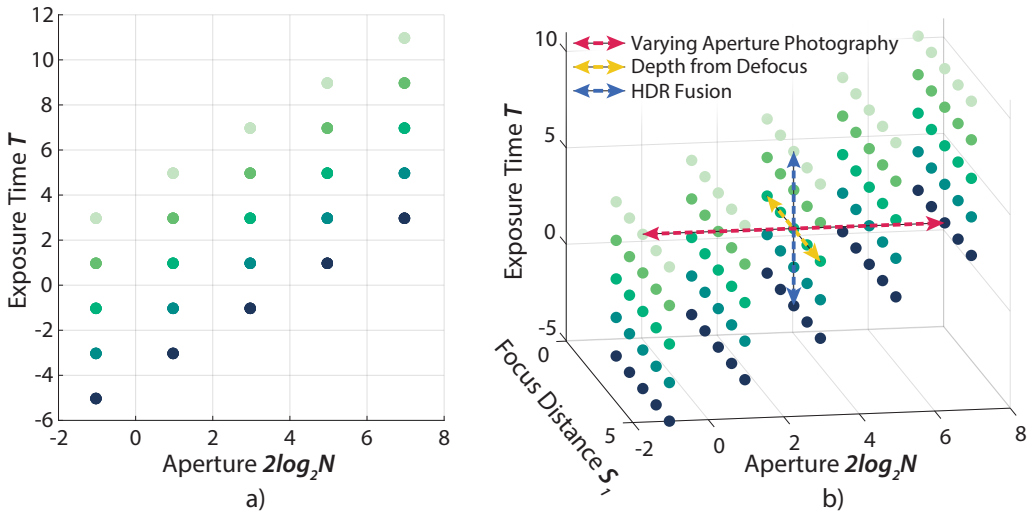


FIGURE 3.8: *a) The parallelogram visually depicts the relationship between F-number, exposure time, and photometric exposure (EV) in the logarithmic domain. Dots with the same color signify an equal exposure intensity. b) Focus distance is incorporated, resulting in the construction of a 3D parallelepiped. Previous approaches only deal with a single dimension (arrows), while we consider the entire 3D space.*

TABLE 3.4: *Explorations of different combinations.*

Method	Dim.	HDR-VDP3 (↑)	PU21-SSIM (↑)	Abs-Rel (↓)
F_S	2D	9.503 ± 0.229	0.937 ± 0.031	0.405 ± 0.239
F_M	2D	9.827 ± 0.031	0.964 ± 0.007	0.335 ± 0.159
F_L	2D	9.863 ± 0.042	0.966 ± 0.011	0.555 ± 0.145
F_V	2D	9.807 ± 0.031	0.962 ± 0.007	0.294 ± 0.128
n-rooks	3D	9.805 ± 0.049	0.960 ± 0.009	0.301 ± 0.143

we design four sampling patterns: two diagonals, a vertical cross, and a horizontal cross. In the 3D case, we produce random instances of n-rooks sampling for all three dimensions.

Our main focus lies in exploring the strategies in 2D and 3D, which are in principle capable of reconstructing and disentangling all information, while the 1D results are naturally limited. We evaluate the results in terms quality for AiF HDR as well as depth map reconstruction. We use HDR-VDP-3.0.6 [Mantiuk et al., 2023] and PU21-SSIM [Azimi et al., 2021] to measure AiF HDR results. For depth map evaluation, we again employ the Abs-Rel metric.

Table 3.4 provides the results of this analyis. We observe that our method demonstrates robustness and achieves good performance under various sampling strategies.

In the 2D evaluation, F_S exhibits the lowest performance in AiF HDR reconstruction, as expected due to significant defocus blur. Conversely, F_L demonstrates lower performance in depth estimation, primarily attributed to a wider depth-of-field and limited blur cues available. F_M offers a trade-off, showcasing satisfactory results in both AiF HDR reconstruction and depth estimation.

Remarkably, within the fixed exposure time cross section, F_V demonstrated competitive performance in AiF HDR reconstruction and improved performance in depth estimation. This can be attributed to the fact that, in this cross section, the variation in exposure primarily arises from changes in the aperture size. Consequently, this variation introduces both exposure variation and defocus blur, resulting in a wider range of defocus blur types and richer cues for depth estimation. Both focus distance and aperture size contribute to the strength of the defocus blur, explaining the observed performance variations.

TABLE 3.5: *Ablation Study*

Method	PSNR (↑)	SSIM (↑)
Ours	**33.6 ± 3.0**	**0.95 ± 0.03**
w/o Flow	28.5 ± 3.1	0.89 ± 0.07
Polynomial Flow (k=1)	32.3 ± 3.8	0.93 ± 0.07
Polynomial Flow (k=3)	32.8 ± 3.8	0.94 ± 0.06
Gaussian Kernel	32.0 ± 2.9	0.93 ± 0.05
Explicit Tone Mapper	27.8 ± 2.0	0.89 ± 0.09
Tone Mapper w/o Weight Sharing	31.7 ± 4.1	0.92 ± 0.07
w/o VGG loss	31.5 ± 3.6	0.90 ± 0.11
w/o Jacobian loss	30.2 ± 2.5	0.91 ± 0.05
w/o TV loss	33.2 ± 2.3	0.94 ± 0.05

We observe similar trends in the 3D n-rooks sampling scenario, although the performance is marginally inferior to F_V, with a notable increase in standard deviation indicating higher instability. The good quality of depth estimation in n-rooks sampling further verifies our conjecture that jointly changing the focus distance and aperture size is beneficial for accurate depth estimation.

In conclusion, for capturing we recommend using either a moderate aperture size with multiple exposure times or fixing the exposure time while changing the aperture size, along with focal sweeping. These combinations provide a good balance between capturing high-quality images and maintaining flexibility in adjusting the depth and HDR information.

While the results in this section indicate that for TAF stack *capture* it suffices to consider a 2D subspace, we want to emphasize that TAF space *synthesis* is fully 3D, and that our method for the first time enables this kind of systematic analyis.

3.4.3 Ablations

In this section, we ablate selected components of our pipeline. Numerical results for all ablations are provided in Table 6.3, based on our real dataset. We observe that removing any component decreases the performance of our method. We elaborate on the components in the following paragraphs.

Implicit Flow Predictor We choose to model lens breathing using a neural network. We study alternative models in Table 6.3 and Figure 3.9. First, we consider removing flow prediction from our pipeline and observe severe ghosting artifacts due to the misalignment between the input images. Second, we consider a polynomial model inspired by established practices in lens distortion correction [Weng et al., 1992]. Specifically, we model radial distortion in polar coordinates, i. e., , as a function of radius $r = \sqrt{x^2 + y^2}$ via

$$r'(S_1) = \sum_{i=1}^{k} \alpha_i(S_1) r^i, \tag{3.13}$$

where r' is the remapped radius, and the coefficients α_i are 3rd-order polynomials of the focal distance S_1. During training, we optimize for the coefficients of the latter polynomials. We study the cases $k = 1$, i. e., , simple linear magnification, and $k = 3$, which includes non-linear distortions. We observe that neither model reaches the quality of our neural flow predictor.

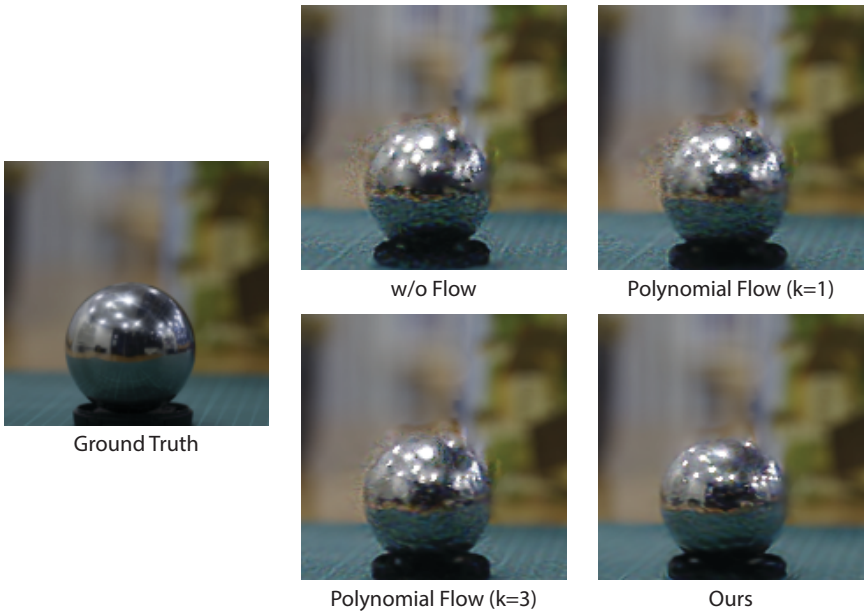


FIGURE 3.9: *Comparison of novel-view image quality for different flow predictors. Only our implicit flow-based approach gets rid of ghosting artifacts arising from misalignment caused by lens breathing.*

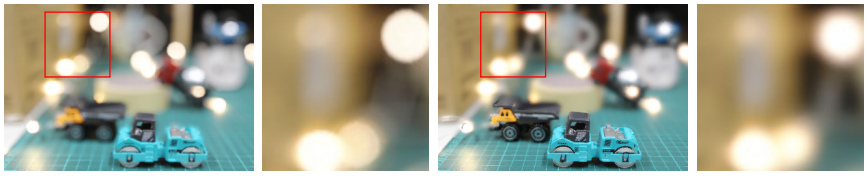


FIGURE 3.10: *Comparisons of bokeh generation between a disk and a Gaussian kernel for focus editing. The disk kernel demonstrates a more natural bokeh.*

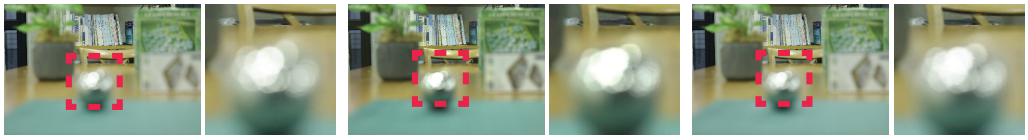


FIGURE 3.11: *Comparing LDR outputs of an explicit tone mapper to our implicit approach.The explicit solution struggles with the accurate reproduction of out-of-focus highlights.*

Disk vs. Gaussian Kernel Previous DfD methods [Favaro, 2010; Si et al., 2023] commonly employ a Gaussian kernel to simulate defocus blur. Figure 3.10 demonstrates that a disk kernel produces more realistic bokeh.

Implicit Tone Mapper We compare our implicit tone mapper to two alternatives. First, we consider an explicit tone mapping function as per Eq. 3.7 in Figure 3.11, which struggles to reproduce highlights. Second, we consider an alternative implicit tone mapper without weight sharing between the color channels in Figure 3.12, which leeds to color shifts. Table 6.3 reveals that our choice of implicit tone mapping delivers the highest-quality results.

In all our experiments, we have used JPEG images as input. Alternatively, linear/RAW images could be considered, for which a way simpler tone mapper would suffice, at the expense of additional processing steps [Mildenhall et al., 2022]. We emphasize that linear images still are not HDR, therefore requiring at least a clipping operation for "tone mapping".

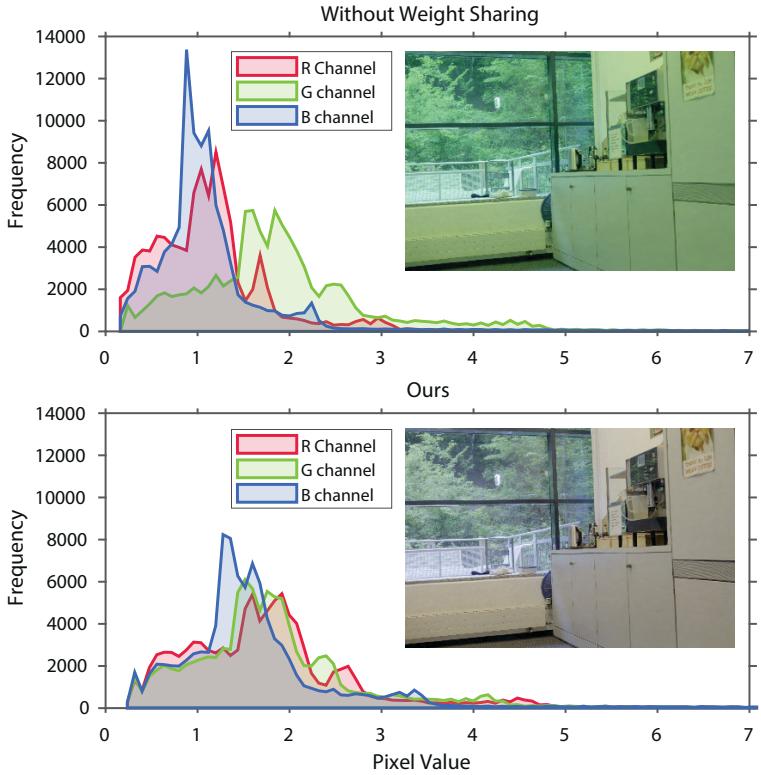


FIGURE 3.12: *Comparisons of reconstructed HDR using a tone mapper without and with (Ours) weight sharing, revealing that weight sharing is essential in preventing hue shifts.*

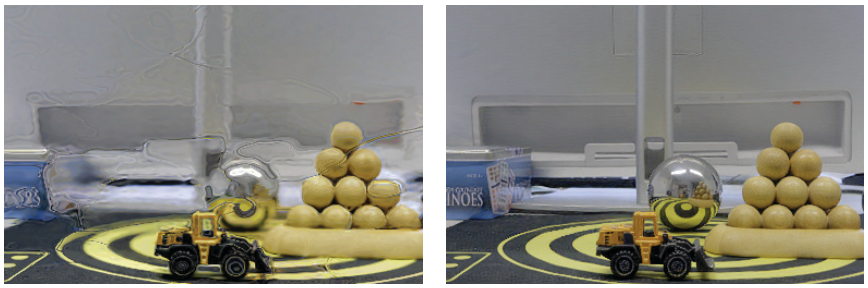


FIGURE 3.13: *Comparisons of the reconstructed HDR without and with the Jacobian regularization (Ours). The introduction of Jacobian regularization allows the method to recover artifact-free images.*



FIGURE 3.14: *Comparisons of the reconstructed HDR without and with the VGG loss (Ours). Our method synthesizes sharper details.*

Loss Terms Here we investigate the effect of our loss terms. Figure 3.14 reveals that the VGG-loss is important to reproduce high-frequency details, avoiding blurry

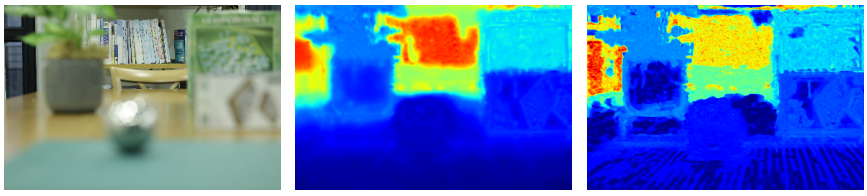


FIGURE 3.15: *Comparisons of estimated depth map with (Ours) and without the TV loss. This loss allows the method to reconstruct smoother depth maps and reduces local fluctuations caused by textures of the captured objects.*

reconstructions. The Jacobian regularization helps to smooth the implicit flow and thus prevents deformation artifacts, as shown in Figure 3.13. The TV loss enhances smoothness and continuity of the depth map by reducing noise and abrupt pixel transitions, illustrated in Figure 3.15. Again, corresponding numerical results are provided in Table 6.3.

Number of input images

We investigate depth reconstruction quality as a function of the number of input images to our system, evaluated on the NYU dataset. In Figure 3.16 we plot errors relative to our solution which uses five images. We see that less images give inferior results, while more images tend to only marginally improve the reconstruction. Even though these results tend to vary across scenes, they indicate that five input images are a reasonable default choice.

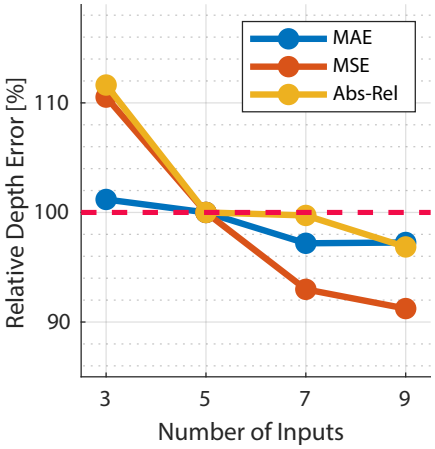


FIGURE 3.16: *Relative depth reconstruction error for different stack sizes.*

3.4.4 Runtime

Our method runs at interactive rates, but the inference time highly depends on the aperture size during post editing, as it determines the blur kernel size. When a small aperture is used, one frame takes as little as 22 ms per frame for a resolution of 384 $\times$ 256 pixels, and 85 ms for a resolution of 768 $\times$ 512 pixels on an Nvidia RTX 3090. In the worst-case scenario, when the aperture is big and larger kernel size has to be used, our method provides results within 62 ms and 788 ms per frame for the lower and higher resolution, respectively. These results can be further improved by replacing the differentiable blur kernel with a faster approximation at inference time. Training time is substantial. It takes around 20 minutes for low resolution and 4 hours for the higher resolution on a GPU RTX 8000.

3.5 Conclusions and Limitations

In this work, we have proposed a novel approach utilizing implicit neural fields and a differentiable thin-lens model to represent the Time-Aperture-Focus (TAF) stack. This representation allows us to extract complete information from the stack, including depth, all-in-focus, and high-dynamic-range data. Our method achieves state-of-the-art performance in Depth-from-Defocus tasks, and allows a faithful exploration of the

three-dimensional space of imaging dimensions. One notable feature of our method is the ability to perform flexible post-editing. After fitting the model, users can adjust parameters such as focus distance, aperture, and exposure, enabling effective disentanglement.

Our method does not account for the sensitivity of the sensor (ISO), which affects exposure and noise levels. Additionally, we primarily focus on static scenes and do not address motion blur caused during image capture. Our implementation is not performance optimized, and we expect that leveraging recent advances in neural field training and inference [Müller et al., 2022] will significantly boost the performance. As our approach provides just one depth value per pixel, we are not currently able to handle transparent objects. We hope our method inspires future work on full-dimensional image recovery from sparse measurements.

Chapter 4

A Real-Time HDR Radiance Fields with Depth of Field

This chapter introduces a novel approach for reconstructing high-dynamic-range (HDR) radiance fields from multi-view low-dynamic-range (LDR) images by leveraging 3D Gaussian Splatting within an expressive camera model. We address key challenges in HDR reconstruction and depth of field modeling by integrating analytical convolutions of Gaussians with a differentiable thin-lens camera model. Our method effectively compensates for exposure variations, aperture differences, and depth-dependent defocus, enabling flexible refocusing during novel-view synthesis. A dedicated tone mapping module further refines the HDR representation, ensuring realistic rendering across diverse lighting conditions. By systematically varying exposure time, aperture, and focal distance, we gain insights into their impact on reconstruction quality. Extensive validation demonstrates that our approach outperforms existing methods in cinematic rendering, achieving superior HDR synthesis with real-time performance and enhanced post-processing control.

4.1 Introduction

As summarized in Section 2.3, Neural Radiance fields [Mildenhall et al., 2021] remain the most prevalent representation for novel-view synthesis applications due to their outstanding ability to produce high-quality renderings of scenes from multi-view photos. 3D Gaussian representations [Kerbl et al., 2023] have further advanced the field and now routinely deliver real-time rendering performance. Most works in this space are concerned with accurate, all-in-focus reconstructions of well-lit scenes. However, for radiance fields to become versatile tools for storytelling and artistic exploration, i. e., to become *cinematic*, they must simultaneously handle high-contrast illumination and depth of field. To achieve this, we propose a method for reconstructing high-dynamic-range (HDR) radiance fields with real-time re-exposure and re-focusing capabilities based on multi-view low-dynamic-range (LDR) images with varying exposures, apertures, and focal distances.

Considering the significance of the task, naturally, individual aspects of our problem setting have already been investigated in the literature. On the one hand, the limited dynamic range of typical cameras often leads to overexposed and underexposed regions, causing saturation (clamping to white) and noise, respectively [Reinhard et al., 2010]. These limitations directly carry over to radiance fields based on such photos and significantly restrict post-editing capabilities. Addressing this issue, several works have considered reconstructing HDR radiance fields from multi-view exposure brackets [Huang et al., 2022; Jun-Seong et al., 2022; Rückert et al., 2022] or RAW data [Mildenhall et al., 2022]. On the other hand, defocus blur in input images has

been regarded as a degradation to overcome for all-in-focus reconstruction [Wu et al., 2022; Ma et al., 2022; Lee et al., 2023; Lee et al., 2024], with synthetic depth of field as an occasional afterthought. In contrast to all these works, we embrace both HDR and depth of field as expressive tools and set out to develop a radiance field reconstruction approach that enables real-time rendering and high-quality post-editing of exposure, aperture, and focal distance.

Input to our method is a set of LDR images that can be captured with a consumer-grade camera. Different from typical radiance field reconstruction settings, the input images vary not only in extrinsic camera parameters, but also in exposure time, aperture size, and focal distance. This diverse input sample distribution allows our method to build an all-in-focus HDR 3D Gaussian representation that is optimized for a subsequent application of lens blur. Depth of field is modeled as a highly efficient convolution of projected scene Gaussians with a Gaussian aperture function [Kriv´anek et al., 2003], based on a thin-lens model. The convolved Gaussians are then rasterized and projected to LDR using a dedicated tonemapping module. Once trained, our model enables novel-view synthesis with arbitrary combinations of camera and lens parameters, facilitating flexible post-editing in terms of exposure and depth of field in real time. We show that our approach can synthesize high-quality and editable imagery, including appealing bokeh effects that require the joint treatment of HDR and depth of field. We also demonstrate that our system is on par with and often outperforms the state of the art in the individual disciplines of HDR and all-in-focus reconstruction. This evaluation is based on multiple new datasets we introduce, including both synthetic and real data.

Our contributions can be summarized as follows: We propose a unified framework for reconstructing HDR radiance fields while accounting for depth of field, enabling high-quality HDR synthesis with flexible refocusing. Our method provides powerful post-editing capabilities, allowing real-time exposure selection and arbitrary refocusing. Extensive validation demonstrates state-of-the-art performance in HDR and all-in-focus reconstruction, while also introducing new challenging benchmark datasets to advance the field.

4.2 Background

In this section, we first discuss previous work on depth-of-field (DoF) rendering. We then review research on radiance field methods that support DoF and HDR effects, drawing from Section 2.3.

4.2.1 Depth-Of-Field Rendering

Depth-of-field (DoF) is the range within a scene where objects appear in focus, with objects outside this range appearing blurred. This effect plays an important role in photography and cinematography, helping to guide viewer attention and enhance depth perception [Demers, 2004]. Computer graphics has long sought to emulate depth-of-field effects of real-world camera systems: Cook et al. [1984] employ ray tracing to achieve accurate depth-of-field by casting rays across the lens. Haeberli and Akeley [1990] use the accumulation-buffer technique, simulating distributed ray tracing by rendering scenes multiple times from different lens positions and blending them. Levoy and Hanrahan [2023] apply the same concepts in light field rendering. However, the quality obtained with these methods is tied to the number of samples taken, implying a high computational effort to produce artifact-free images.

To trade visual accuracy for efficiency, Scofield [1992] apply a layered technique, segmenting scenes into depth-ordered layers for simplified blurring. This method achieves depth-of-field effects while avoiding blurred edges but struggles with depth-spanning objects. Kriv´anek et al. [2003] propose a surface splatting method suitable for real-time rendering; occasional artifacts are caused by approximate surface reconstructions. Potmesil and Chakravarty [1982] utilize forward mapping to render sprites for depth-of-field effects and blending pixels as circles in the frame buffer, with a final renormalizing pass for producing accurate alpha values. This post-processing method for simulating depth-of-field has influenced many subsequent works [Won and Jeon, 2022; Si et al., 2023; Pidhorskyi et al., 2022; Wang et al., 2023b], which calculate the circle of confusion (CoC) size based on depth and use spatially varying convolution to perform weighted averaging.

4.2.2 Depth-Of-Field and High Dynamic Range Radiance Fields

In most radiance field research as stated in Section 2.3, it is assumed that input images are consistently in focus and well-exposed, either the Implicit Neural Representations [Mildenhall et al., 2021; Hedman et al., 2021; Verbin et al., 2022; Barron et al., 2021; Barron et al., 2022; Barron et al., 2023; Müller et al., 2022] or Point Based Representations [Kerbl et al., 2023; Yu et al., 2024; Yan et al., 2024].

While some works have made progress in handling defocus blur [Wu et al., 2022; Ma et al., 2022; Lee et al., 2023; Lee et al., 2024; Darmon et al., 2024] and varying exposure conditions [Huang et al., 2022; Cai et al., 2024b] to recover all-in-focus and HDR results from degraded inputs, they typically address only one of these issues at a time. This limitation becomes problematic in casual captures, where scenes often include both near and far objects requiring different focus levels, along with high contrast lighting that spans from dark shadows to bright highlights.

As shown by our previous work TAF [Wang et al., 2023b], a stack of images with different exposure/aperture/focus can be employed for all-in-focus HDR image reconstruction of particularly demanding scenes for a specific camera pose. Our goal is to generalize this approach to multiple camera poses within a novel-view synthesis framework based on 3DGS for efficient rendering. Capturing a complete image stack for each input camera pose would be impractical; we aim to sample the space of camera parameter setups between neighboring camera poses so that, as in traditional NeRF and 3DGS methods, a single image per pose is effectively captured. This way, we enable high-quality reconstruction of scenes with high variation of depth and contrast ranges, and their rendering with arbitrary camera parameters for any view. Our approach directly benefits from the more realistic depth-of-field simulations when performed using HDR images as the input [Zhang et al., 2019c].

4.3 Method

Our approach takes as input a set of LDR multi-view images $\{I_k\}_{k=1}^{N}$ of a scene, captured with randomly selected exposure times t, aperture sizes a, and focal distances d_f (Figure 4.1, left). Since exposure and lens settings are typically stored in EXIF headers, we assume this information is available for each image. Using this input, our method reconstructs an HDR radiance field that allows for post-editing of exposure and depth of field.

Our approach is based on the 3D Gaussian Splatting representation [Kerbl et al., 2023], which we briefly review in Section 4.3.1. The first step in our pipeline

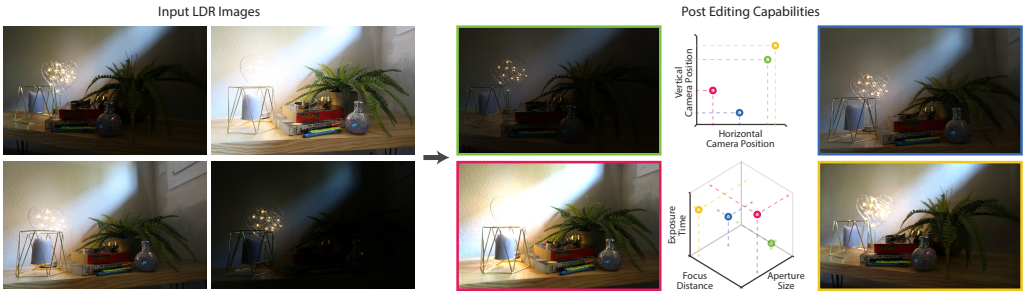


FIGURE 4.1: *Based on sparse images captured from various viewpoints and with different focus distances, apertures, and exposure times* (left)*, our method generates an HDR 3D Gaussian representation that can be rendered with arbitrary depth of field. Our representation supports rendering images in a six-dimensional space of camera and lens parameters* (right, five dimensions are shown)*.*

is calibrating the input cameras, detailed in Section 4.3.2. We then introduce our model, which features an HDR 3D Gaussian scene representation with an integrated defocus blur simulator (Section 4.3.3). A tonemapping module allows us to train the representation on LDR input only (Section 4.3.4).

4.3.1 Prerequisite

Radiance fields based on 3D Gaussians [Kerbl et al., 2023] represent a scene using a Gaussian mixture model. Each primitive has a mean $\boldsymbol{\mu} \in \mathbb{R}^3$, a symmetric covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$, an opacity $o \in \mathbb{R}$, as well as a set of spherical harmonics (SH) coefficients encoding view-dependent RGB color $\mathbf{c} \in \mathbb{R}^3$. To render an image using this representation [Zwicker et al., 2001a], the primitives are projected to 2D screen space using a pinhole camera model to obtain a projected mean $\hat{\boldsymbol{\mu}} \in \mathbb{R}^2$ and projected covariance $\hat{\Sigma} \in \mathbb{R}^{2\times2}$. The final pixel color is obtained using front-to-back compositing:

$$C(\mathbf{x}) = \sum_{i\in\mathcal{N}(\mathbf{x})} \mathbf{c}_i \alpha_i(\mathbf{x}) \prod_{j=1}^{i-1} \left(1 - \alpha_j(\mathbf{x})\right), \tag{4.1}$$

where $\mathcal{N}$ is the sequence of depth-ordered primitives overlapping the pixel at location $\mathbf{x} \in \mathbb{R}^2$, and

$$\alpha_i(\mathbf{x}) = o_i \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T \hat{\Sigma}_i^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i)\right) \tag{4.2}$$

is the opacity of the primitives. Since the rendering function in Eq. 4.1 is differentiable, the model can be trained from posed input images via gradient-based optimization.

4.3.2 Camera Calibration

We require posed input images to build our model. The standard approach for this is Structure from Motion (SfM) [Schonberger and Frahm, 2016], which we run on our input data. Unfortunately, such photometric calibration leaves the scene's global scale ambiguous. As various lens parameters are provided in absolute millimeters, we need to scale the scene accordingly. To achieve this, we leverage the defocused images and their respective lens configurations. Specifically, we run an image blur detector [Alireza Golestaneh and Karam, 2017] on all input images to obtain per-pixel sharpness estimates. By thresholding these sharpness maps, we identify the image space locations of in-focus scene elements. The known focal distance d_f per input image provides ground-truth depth values for these pixels. Next, we train a vanilla 3D Gaussian Splatting model on the input images after rescaling pixel intensities to a

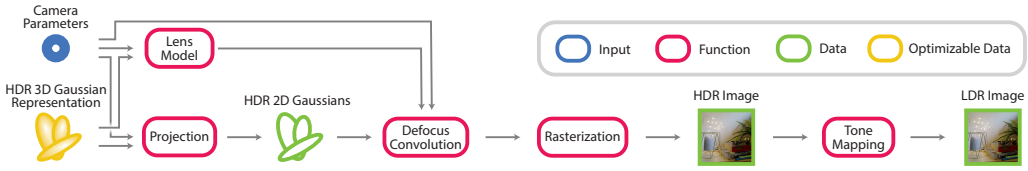


FIGURE 4.2: *Our pipeline for rendering HDR radiance fields with controllable depth of field.*

common exposure. Using this model, we render a depth map for each input view. As expected, the renderings from this model are of low quality given our inputs, but we find the overall scene configuration precise enough for our calibration. Using the depth values of the in-focus pixels, we perform a least-squares fit of the scene scale to match the rendered depth values with the ground truth.

After calibration, we have the full set of camera and lens parameters per input view at our disposal.

4.3.3 HDR Radiance Field with Depth of Field

Here we describe our pipeline, which consists of an HDR 3D Gaussian scene representation, a defocus convolution module, as well as a tonemapper that projects HDR to LDR content. An overview of our model is shown in Figure 4.2.

Our HDR Gaussians differ from conventional LDR Gaussians in one crucial property: the per-primitive SH coefficients encoding view-dependent color $\mathbf{c}$ exhibit a significantly larger dynamic range. We denote this corresponding HDR color as $\mathbf{c}^{\text{HDR}}$. Notably, no architectural changes compared to an LDR representation are necessary at this stage; the high dynamic range emerges from our carefully designed pipeline, as detailed in the remainder of this section.

The next stage of our model is concerned with depth of field (DoF) synthesis, which occurs at the level of 2D Gaussians projected into screen space. Considering a thin-lens model [Potmesil and Chakravarty, 1981], a projected point with depth d is spread over a disk – typically referred to as the circle of confusion – with radius

$$r(d) = \frac{\left|d - d_f\right| \cdot f^2}{2 \cdot a \cdot d \cdot \left(d_f - f\right)}, \tag{4.3}$$

where f is the focal length, a is the aperture size, and d_f is the focal distance. We seek to enlarge each projected scene Gaussian to reflect the circle of confusion [Kriváanek et al., 2003], where the mean $\boldsymbol{\mu}$ determines depth d. We approximate the circle of confusion using an aperture Gaussian with covariance $\hat{\Sigma}^a = \text{diag}\left(3r(d), 3r(d)\right) \in \mathbb{R}^{2\times 2}$ to allow for efficient convolutions with the scene Gaussians. This convolution can be carried out in closed form [Bromiley, 2003; Celarek et al., 2022], resulting in an update of Eq. 4.2:

$$\alpha_i^{\text{DoF}}(\mathbf{x}) = o_i \cdot \beta_i \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T (\hat{\Sigma}_i + \hat{\Sigma}_i^a)^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i)\right), \tag{4.4}$$

with the energy-preserving normalization factor

$$\beta_i = \sqrt{\frac{\det\left(\hat{\Sigma}_i\right)}{\det\left(\hat{\Sigma}_i + \hat{\Sigma}_i^a\right)}}. \tag{4.5}$$

Importantly, each projected scene Gaussian is convolved with an aperture Gaussian reflecting the depth of the scene Gaussian and the lens configuration, producing

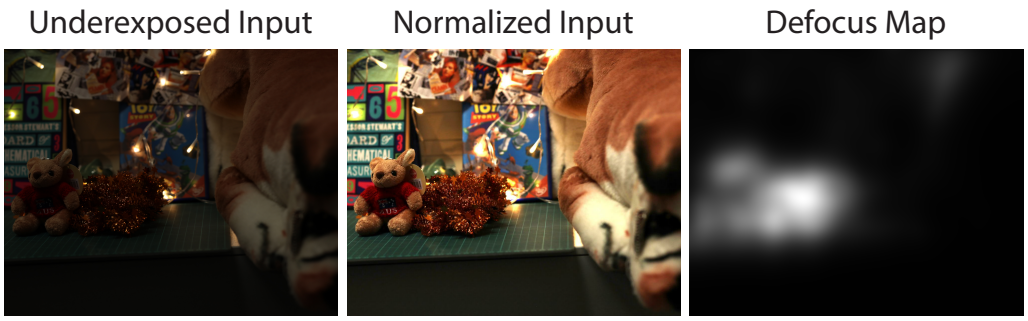


FIGURE 4.3: *Our training data includes underexposed and overexposed images (*left*). Our loss incorporates both the original data (Eq. 4.9) and a normalized version of it (Eq. 4.10). To manage defocus blur, a third loss component (Eq. 4.11) incorporates a defocus map.*

spatially-varying defocus blur. An HDR image with defocus blur can now be rasterized using the quantities developed above:

$$C_{\text{HDR+DoF}}(\mathbf{x}) = \sum_{i \in \mathcal{N}(\mathbf{x})} \mathbf{c}_i^{\text{HDR}} \alpha_i^{\text{DoF}}(\mathbf{x}) \prod_{j=1}^{i-1} \left(1 - \alpha_j^{\text{DoF}}(\mathbf{x})\right). \tag{4.6}$$

Naturally, the HDR content rendered in the previous step has an unbounded range of radiance values. Both training and final visualization require values in the range $[0, 1]$. Therefore, we project HDR values to LDR using the tone mapping function

$$\bar{C}(\mathbf{x}) = \left(\min\left(\frac{t}{a^2} \cdot C_{\text{HDR+DoF}}(\mathbf{x}), 1\right)\right)^{1/\gamma}, \tag{4.7}$$

simulating the per-pixel non-linear response of a real-world camera [Allen and Triantaphillidou, 2011]. Here, the final exposure is determined using exposure time t and aperture size a, before the intensities are clipped to the target domain. Finally, γ performs a gamma correction from linear radiance to pixel intensities.

Eq. 4.7 produces images that the user can flexibly edit across multiple dimensions: In addition to free viewpoint selection, the user can adjust the aperture size a, exposure time t, and focal distance d_f in real-time, providing a wide range of artistic control.

4.3.4 Training

We train our model using the loss function

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{exp}} \mathcal{L}_{\text{exp}} + \lambda_{\text{foc}} \mathcal{L}_{\text{foc}}. \tag{4.8}$$

All three terms make use of the photometric consistency measure Δ that evaluates a sum of ℓ_1-distance and DSSIM [Wang et al., 2004] metric. Yet, each term applies Δ to different versions of the data. The first term simply measures the consistency between our rendered images and the input data for all pixels $\mathbf{x}$ and training views k:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{x},k}\left[\Delta\left(\bar{C}_k(\mathbf{x}), I_k(\mathbf{x})\right)\right]. \tag{4.9}$$

To handle the vastly different exposures in our inputs, the second loss term normalizes the images to a medium exposure by scaling pixel intensities before computing

discrepancies [Liu et al., 2020b; Wang et al., 2022a] (Figure 4.3, left and center):

$$\mathcal{L}_{\text{exp}} = \mathbb{E}_{\mathbf{x},k}\left[\Delta\left(\frac{\bar{C}_k(\mathbf{x})}{2\cdot\mathbb{E}_{\mathbf{x}}\left[\bar{C}_k(\mathbf{x})\right]}, \frac{I_k(\mathbf{x})}{2\cdot\mathbb{E}_{\mathbf{x}}\left[I_k(\mathbf{x})\right]}\right)\right]. \tag{4.10}$$

Finally, due to the prominent occurrence of defocus blur in our inputs, we found it beneficial to include a loss term that explicitly addresses sharpness. Specifically, we use an off-the-shelf differentiable image blur detector g [Alireza Golestaneh and Karam, 2017] that outputs a scalar defocus map (Figure 4.3, right) and compute

$$\mathcal{L}_{\text{foc}} = \mathbb{E}_{\mathbf{x},k}\left[\Delta\left(g\left(\bar{C}_k(\mathbf{x})\right), g\left(I_k(\mathbf{x})\right)\right)\right]. \tag{4.11}$$

We set $\lambda_{\text{exp}} = 0.25$ and $\lambda_{\text{foc}} = 1$ in all our experiments and train our models using the Adam [Kingma, 2014] optimizer with default parameters.

4.4 Results and Evaluation

In this section, we demonstrate the effectiveness of our method through a comprehensive set of experiments. We first describe our experimental dataset in Section 4.4.1, introducing our dataset, which consists of both synthetic and real captured images used for evaluation, followed by the implementation details (Section 4.4.2) used for all results. Then, in Section 4.4.3, we evaluate separately the capability of our model to reconstruct all-in-focus and HDR images, comparing our results to those of specialized state-of-the-art solutions. After this, we present in Section 4.4.4 results for our complete pipeline, showcasing HDR radiance field reconstructions as well as post-editing applications. Lastly, in Section 4.4.5 we provide ablation studies of the components of our approach.

4.4.1 Dataset

Our dataset includes four synthetic rendered scenes and four real captured scenes. The synthetic scenes provide control and ground truth for evaluation, while the real captured scenes demonstrate the practical performance of our approach.

Synthetic dataset. Our synthetic dataset is composed of four diverse indoor scenes rendered with Cycles (path tracer) in Blender 4.1. The scenes encompass challenging scenarios with high dynamic range of illumination (bright light sources and dark regions illuminated by bounced light) and spacious layouts allowing for multiple focal planes (examples shown in Figures 4.6 and 4.7) All scenes have been rendered with high number of samples (ranging from 4096 to 16384) to prevent undesired distortions in the defocused regions caused by denoising. teFor each scene we generated 960 HDR images (8×8 views sampled at a regular grid wrapped to the sphere surface, 5 uniformly sampled focus distances and 3 different apertures). For each view we generated all-in-focus ground truth image. This results in 1024 images per scene. For evaluation purposes we generated another set of 7×7 views located between the training views. All images are accompanied by a file containing parameters used for rendering (camera transformation matrix, focal length, focus distance, aperture, etc.). In addition to the HDR content we have modified the scenes for the AiF reconstruction comparison, producing pleasant LDR images with reduced dynamic range without clipped regions (examples shown in Figure 4.5). The camera



FIGURE 4.4: *All-in-focus HDR reconstruction. Our method can successfully reconstruct sharp HDR images from defocused images captured with different exposure times.*



FIGURE 4.5: *Comparison of all-in-focus reconstruction. Our method provides better or comparable results to Deblur-Splatting method.*



FIGURE 4.6: *Comparison of HDR reconstruction. Left-most column presents input LDR images for a given view. The following columns depict results acquired by HDR-NeRF, Ground Truth, and results of our method, respectively. Please note, that our method properly reconstructs the shape of the flame and reproduces successfully high-frequency content, while these details are missing in the results generated by HDR-NeRF. To depict the accuracy of both methods we provide the HDR-VDP-3 error map in the top right corner of the images.*

locations, apertures, and focus distances are sampled in a same manner as in the case of the HDR subset.

Captured dataset. Our captured dataset is composed of four real scenes. Some examples are shown in Figures 4.1, 4.4 and 4.7. These scenes were captured using Canon EOS RP and Canon EOS Rebel T6i cameras, outfitted with RF 85mm F2.0, and EF-S 18-55mm F3.5-5.6 lenses. Each scene was captured by moving the camera to follow an approximately 9×9 grid pattern, resulting in 81 images per scene. At each camera position in this grid, we varied the exposure time, aperture, and focus distance to ensure a diverse set of configurations.

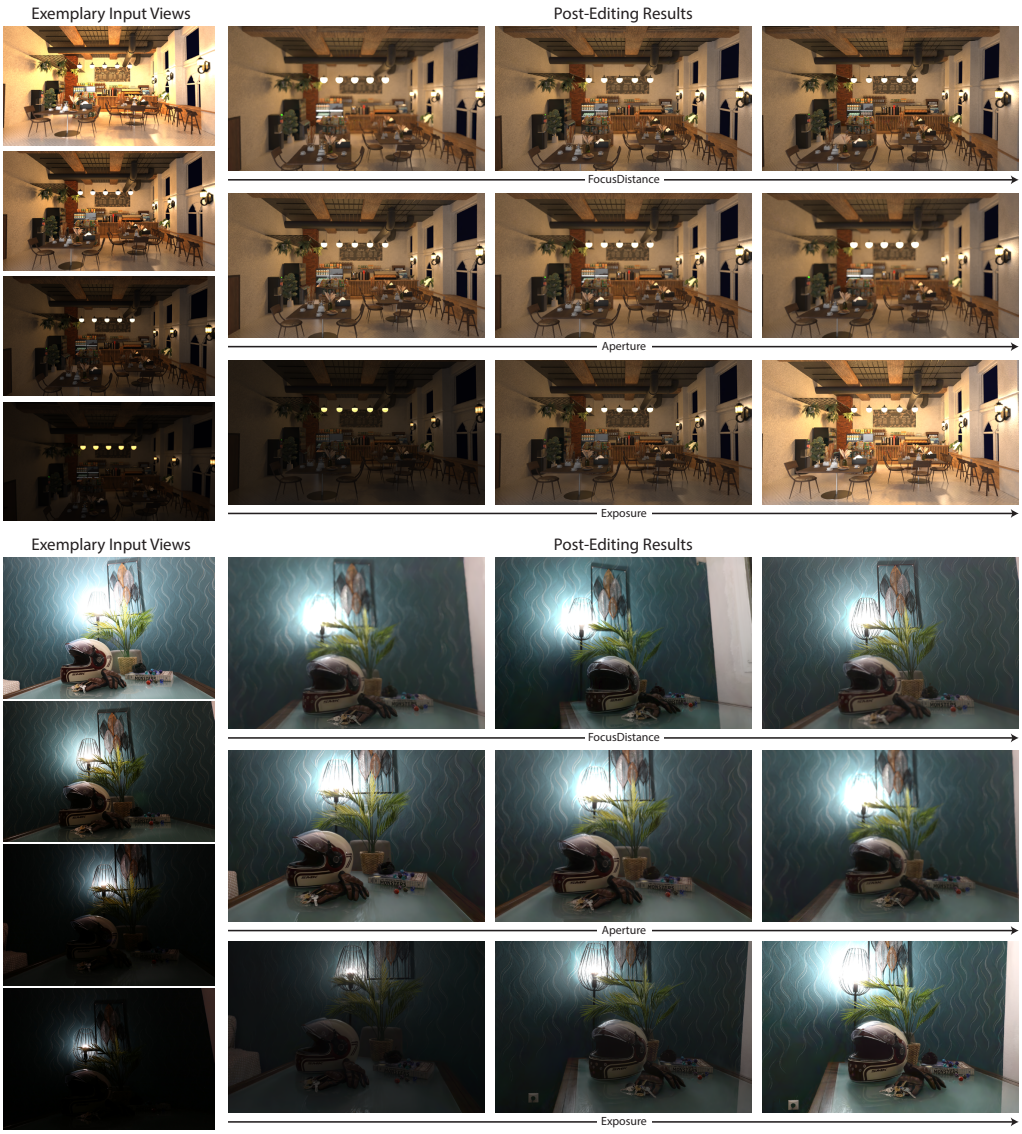


FIGURE 4.7: *Post-editing results. The leftmost column shows the selected input views. The 3 × 3 grid to the right displays novel views, with each row illustrating the manipulation capabilities of one parameter: focus distance, aperture size, and exposure.*

4.4.2 Implementation

We use the 3DGS framework [Kerbl et al., 2023] as our baseline, integrating a thin-lens model and a tone mapping module as described in Section 4.3.3. The implementation is done in PyTorch and CUDA. We will release all source code on publication.

Additionally, we incorporate some practical design choices to tackle our task more effectively. Given that our rendering dataset starts from random noise without an SfM point cloud initialization, we employ a coarse-to-fine optimization strategy. During the initial 7,000 iterations, we exclude tone mapping and depth of field simulation, training on an exposure-aligned dataset to generate an initial point cloud. Following this, we refine the optimization using the coarse result as a starting point. The iteration process consists of 40,000 steps, taking approximately 21 minutes on an RTX A40 GPU, using a training dataset with a resolution of 1200 × 675.

TABLE 4.1: *Comparison of all-in-focus reconstruction performance metrics (PSNR, SSIM, and LPIPS) for our method and Deblur-Splatting using our synthetic dataset.*

Scene	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Car	Deblur-Splatting	27.23 ± 3.21	0.86 ± 0.07	0.18 ± 0.04
	Ours	**28.05 ± 0.26**	**0.89 ± 0.01**	**0.11 ± 0.01**
Attic	Deblur-Splatting	31.86 ± 0.23	0.91 ± 0.01	0.13 ± 0.01
	Ours	**32.90 ± 0.18**	**0.93 ± 0.00**	**0.09 ± 0.00**
Café	Deblur-Splatting	30.24 ± 0.42	**0.87 ± 0.01**	**0.12 ± 0.01**
	Ours	**30.50 ± 0.36**	0.85 ± 0.01	0.14 ± 0.00
Office	Deblur-Splatting	**34.59 ± 0.75**	**0.95 ± 0.01**	0.08 ± 0.01
	Ours	31.70 ± 0.80	0.92 ± 0.01	**0.05 ± 0.01**

4.4.3 Comparisons

In this section we evaluate two important features of our method: all-in-focus reconstruction and HDR reconstruction. We compare our method against specialized state-of-the-art approaches designed to address these specific problems. To facilitate this comparison, we leverage our synthetic dataset, which provides access to ground truth images for numerical evaluations. While these specialized methods excel in their respective domains, it is important to note that they are limited to solving only one of the problems—those optimized for all-in-focus reconstruction cannot handle effectively HDR, and vice versa. Our approach, however, integrates both capabilities and demonstrates comparable, and in some cases superior performance to these specialized solutions.

All-in-Focus Reconstruction For the all-in-focus reconstruction task, we compare our results against the concurrent work Deblur-Splatting [Lee et al., 2024]. We use the four scenes from our synthetic dataset for this evaluation. In particular, for each of the scenes, we use 64 (8×8) viewpoints for training, each exhibiting varying degrees of defocus blur, and 49 (7×7) novel viewpoints not seen during training for the evaluation. Table 4.1 presents the numerical results for PSNR, SSIM, and LPIPS, commonly used for evaluating image quality. Additionally, Figure 4.5 shows illustrative comparisons to the ground truth. Although our method is not specifically designed for deblurring, it achieves results comparable to state-of-the-art method. Regarding rendering time, both methods are similarly efficient, achieving approximately 110 frames per second (FPS) on an RTX A40 GPU.

HDR Reconstruction For HDR reconstruction, we compare our results against HDR-NeRF [Huang et al., 2022] using various metrics specifically targeted towards evaluating HDR quality, including HDR-VDP-3 [Mantiuk et al., 2023], PU-PSNR, and PU-SSIM [Azimi et al., 2021], as summarized in Table 4.2. Visual comparisons with the ground truth are depicted in Figure 4.6. Again, we use for evaluation the four scenes of our synthetic dataset, with 64 (8×8) views for training and 49 (7×7) novel ones for testing. To simulate LDR training views with varying exposures, starting from the HDR renders from our synthetic dataset, we sample a single exposure for each view. For this, we first determine the start and end boundaries that emcompass the entire dynamic range of the scene [Andersson et al., 2021]. Then, we divide this range into eight intervals, within which we sample exposures as described in previous work [Huang et al., 2022]. Our method consistently outperforms HDR-NeRF in almost all metrics, demonstrating superior reconstruction capabilities. Notably,

TABLE 4.2: *Comparison of HDR reconstruction performance metrics (HDR-VDP-3, PU-PSNR, and PU-SSIM) for our method and HDR-NeRF using our synthetic dataset.*

Scene	Method	HDR-VDP-3 ↑	PU-PSNR ↑	PU-SSIM ↑
Car	HDR-NeRF	9.70 ± 0.15	**29.78 ± 1.72**	0.91 ± 0.02
	Ours	**9.75 ± 0.09**	29.67 ± 4.63	**0.95 ± 0.01**
Attic	HDR-NeRF	9.92 ± 0.03	**39.04 ± 0.71**	0.96 ± 0.04
	Ours	**9.98 ± 0.01**	38.97 ± 1.01	**0.98 ± 0.00**
Café	HDR-NeRF	9.16 ± 0.22	26.97 ± 2.56	0.76 ± 0.11
	Ours	**9.93 ± 0.01**	**30.02 ± 0.51**	**0.91 ± 0.00**
Office	HDR-NeRF	9.36 ± 0.18	25.13 ± 1.97	**0.90 ± 0.01**
	Ours	**9.50 ± 0.09**	**30.11 ± 0.22**	0.71 ± 0.05

we observed that HDR-NeRF tends to overfit to the LDR images, resulting in lower quality HDR image generation. Moreover, our approach achieves real-time rendering capabilities, significantly outperforming HDR-NeRF. We deliver approximately 110 FPS at a resolution of 1200 x 675, in contrast to HDR-NeRF's rate of approximately 0.021 FPS for rendering a single image, both using a GPU RTX A40.

4.4.4 HDR Radiance Field Reconstruction and Post-Editing

In this section, we showcase the full capabilities of our model, specifically its ability to reconstruct HDR radiance fields, which enable the generation of all-in-focus HDR images. Figure 4.4 presents results from various viewpoints, illustrating how our approach effectively reconstructs detailed scenes with consistent sharpness and dynamic range from multi-view LDR defocused inputs. Furthermore, we demonstrate the post-editing capabilities enabled by our method. Our proposed approach reconstructs HDR radiance fields with real-time re-exposure and re-focusing capabilities. Users have the flexibility to adjust aperture size, exposure time, and focal distance in real-time, as demonstrated in Figure 4.7.

4.4.5 Ablations

We evaluate the impact of the defocus and exposure loss terms by conducting ablation studies on both our real and synthetic datasets.

We conduct this evaluation differently for our synthetic and real datasets. For the synthetic dataset, where ground truth images are available, we evaluate our all-in-focus HDR reconstructions against ground truth views not included in the training set. Specifically, we utilize 64 (8×8) views for training and 49 (7×7) novel views for testing. This comparison is performed in HDR space, therefore we use PU-PSNR, PU-SSIM and HDR-VDP-3 as metrics. We show these results in Table 4.3. On the other hand, for the captured dataset, where ground truth is not available, we adopt a different approach. We use 74 captured views for training and leave out 7 for testing.

Then, we synthesize the testing views with our model using the original aperture, focal distance, and exposure settings corresponding to each captured view, and compare these synthesized views against the actual captures. This comparison is conducted in LDR space, thus we employ traditional image quality metrics, in particular, PSNR, SSIM, and LPIPS. Quantitative results are presented in Table 4.4, while qualitative results are illustrated in Figure 4.8 for the defocus loss term and in Figure 4.9 for the exposure loss term.



FIGURE 4.8: *The defocus loss provides higher quality reconstruction for high-frequency patterns.*



FIGURE 4.9: *The exposure loss helps to resolve artifacts in the underexposed regions, as the one depicted in the green frame.*

4.5 Conclusions and Limitations

We have presented an approach for reconstructing HDR radiance fields with depth of field. Unlike previous works that treat only exposure variations or defocus blur as a degradation to overcome, we recognize the significance of exposure selection and depth of field for crafting cinematic imagery. Our pipeline mirrors the physical processes of a camera, enabling us to train a radiance field model using LDR images with various exposure and lens configurations. We found that our method not only allows powerful and appealing post-editing in real time, but also achieves state-of-the-art quality for the isolated tasks of HDR and all-in-focus reconstruction. We hope that our approach moves radiance fields a step closer to becoming expressive tools for artists and visual media producers.

However, our approach does not come without limitations. Our defocus model is a two-fold approximation of the actual physical processes in a real camera lens. First, we use a thin-lens model [Potmesil and Chakravarty, 1981], which, despite being commonly used, is a simplified representation of defocus. More sophisticated models have been developed in the forward-rendering literature [Kolb et al., 1995] and related work on defocus modeling in the context of reconstruction from image stacks [Wang et al., 2023b] has shown that advanced effects, such as lens breathing, can significantly improve accuracy. Second, we approximate the thin-lens circle of confusion with a Gaussian function to achieve real-time rendering performance. While this is a commonly used design decision [Favaro, 2010; Si et al., 2023], it results in a slight softening of bokeh due to the deviation from a disk kernel. We believe that moving towards more realistic lens models has the potential to further increase result

TABLE 4.3: *Evaluation of the impact of the defocus and exposure loss terms in the synthetic dataset.*

Methods	PU-PSNR(↑)	PU-SSIM(↑)	HDR-VDP(↑)
Ours	**34.04 ± 1.67**	**0.95 ± 0.00**	**9.83 ± 0.05**
w/o $\mathcal{L}_{\text{exp}}$	33.95 ± 1.59	0.94 ± 0.00	9.80 ± 0.06
w/o $\mathcal{L}_{\text{foc}}$	33.63 ± 1.54	0.94 ± 0.00	9.78 ± 0.05

TABLE 4.4: *Evaluation of the impact of the defocus and exposure loss terms in the captured dataset.*

Methods	PSNR(↑)	SSIM(↑)	LPIPS(↓)
Ours	**25.69 ± 2.71**	**0.86 ± 0.07**	**0.10 ± 0.02**
w/o $\mathcal{L}_{\text{exp}}$	23.95 ± 3.37	0.82 ± 0.10	0.11 ± 0.04
w/o $\mathcal{L}_{\text{foc}}$	24.35 ± 1.06	0.85 ± 0.06	0.11 ± 0.03

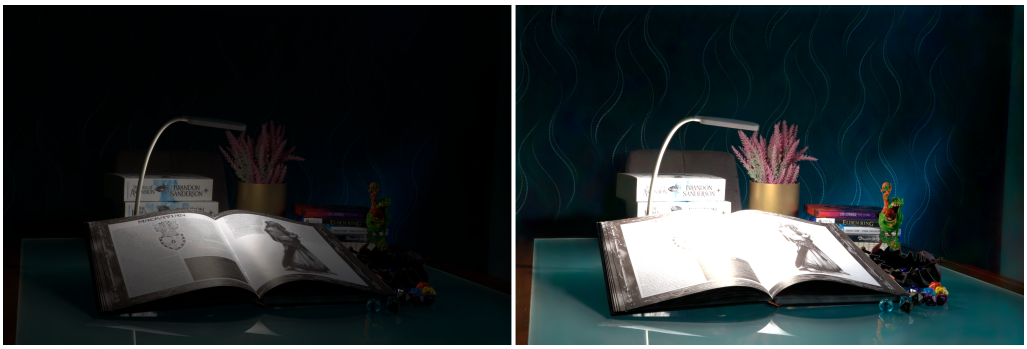


FIGURE 4.10: *Limitation. All-in-focus HDR reconstruction of a scene view with extremely high dynamic range, featuring tone-mapped low* (left) *and high* (right) *exposures. In regions with pronounced darkness and notable noise in the input images, such as the rear wall of the room, artifacts may arise due to information loss.*

quality and cinematic appeal.

In scenes characterized by extremely high dynamic range, where very dark regions may contain significant noise across most input images, artifacts may occur in the reconstruction of these dark areas. These artifacts are pronounced when selecting very high exposures (Figure 4.10). This is expected, since there is insufficient useful information, with noise dominating instead in these areas. Nevertheless, these artifacts are noticeable only under extreme exposure settings in the reconstruction, while our method remains capable of delivering pleasing results for the rest of the scene.

Chapter 5

An Unsupervised HDR Generation Model from In-the-Wild LDR

This chapter introduces a generative approach for synthesizing high-dynamic-range (HDR) images from in-the-wild low-dynamic-range (LDR) collections, leveraging an unsupervised learning framework. We propose a generative adversarial network (GlowGAN) that models the underlying HDR image distribution by learning to project HDR outputs into LDR representations under varying exposure conditions. A physically motivated camera model captures the stochasticity of exposure and camera response functions, enabling realistic LDR synthesis that aligns with real-world observations. A key advantage of our method is its ability to generate photorealistic HDR images across diverse and challenging scenarios, such as landscapes, lightning, and window reflections, where existing supervised methods often fail due to overexposure. Furthermore, we introduce GlowGAN-ITM, an unsupervised inverse tone mapping approach that searches the HDR image manifold modeled by GlowGAN to reconstruct HDR images directly from LDR inputs. Unlike prior methods that progressively infer missing details, GlowGAN-ITM performs global optimization over the HDR space, achieving superior reconstruction of overexposed regions without requiring paired multi-exposure training data. Extensive validation demonstrates that our method surpasses state-of-the-art supervised models, setting a new paradigm for HDR synthesis and inverse tone mapping.

5.1 Introduction

As presented in Section 2.1, High Dynamic Range (HDR) images [Reinhard et al., 2010] are capable of capturing and displaying much richer appearance information than Low Dynamic Range (LDR) images, thus playing an important role in image representation and visualization. The most popular method to acquire HDR images is multiple exposure blending, which requires capturing a set of LDR images of the same scene with different exposures [Debevec and Malik, 2023; Mitsunaga and Nayar, 1999; Robertson et al., 2003]. However, this is time and effort intensive and only suitable for static scenes. Due to this limitation, existing HDR image datasets only cover limited scene categories and have much fewer images than LDR datasets. Thus, supervised learning methods [Eilertsen et al., 2017; Marnerides et al., 2018; Endo et al., 2017a; Lee et al., 2018a; Liu et al., 2020b; Lee et al., 2020; Santos et al., 2020; Zhang and Aydın, 2021; Yu et al., 2021b; Jo et al., 2021] that reconstruct an HDR image from an LDR image are constrained by the HDR datasets and cannot extend to cases where no HDR training data is available, e.g., lightnings or campfires.

While HDR images are hard to collect, it is much easier to collect a large number of LDR images from the Internet. This motivates us to investigate a new unsupervised

learning problem: *Can we learn to reconstruct HDR images from in-the-wild LDR images?* The LDR images do not need to depict the same scene, it is enough if they contain a roughly similar class of scenes (e. g., , landscapes) with various exposures. This weak multi-exposure assumption is often naturally satisfied for in-the-wild LDR datasets as images can come from different camera parameters or different adjustments of the auto-exposure mode. This problem is challenging as only a single exposure is available for each scene; therefore, a way to merge the multi-exposure information spread across different scenes is required.

In this work, we address this challenge via *GlowGAN*, which, to our knowledge, is the first method to learn an HDR generative model from in-the-wild LDR image collections in a fully unsupervised manner. GlowGAN uses adversarial training of an HDR generator with a discriminator that operates merely in LDR. Specifically, the generator produces an HDR image, which is projected to LDR via a camera model and is then sent to the discriminator as a fake image for adversarial training. The camera model consists of multiplying the HDR sample with an exposure value, clipping the dynamic range, and applying a camera response function (CRF). Importantly, during training, we use a randomly sampled exposure from a prior Gaussian distribution when projecting HDR to LDR. This requires the generated HDR images to be realistic under any possible exposure, thus only valid HDR samples would satisfy this "multi-exposure constraint". Furthermore, we also model the stochasticity in the non-linear camera response by randomly sampling CRFs according to a well-established parametric distribution [Eilertsen et al., 2017; Grossberg and Nayar, 2003]. During inference, we can disable the camera model so that the generator produces HDR imagery directly.

We conduct extensive experiments on several datasets collected from the Internet, including landscapes, windows, lightning, fireplaces, and fireworks. By training on these LDR images, GlowGAN successfully learns to generate high-quality HDR images that capture rich appearance information. These images can be presented on HDR displays, or via suitable tone mapping to create appealing imagery. In contrast, previous unconditional GANs tend to miss information in over- or under-exposed regions.

By modeling a distribution of HDR images, GlowGAN paves the way for new applications such as unsupervised inverse tone mapping (GlowGAN-ITM). ITM aims to reconstruct an HDR image from a single-exposure LDR input, where a key challenge is to restore the flat-white overexposed regions [Eilertsen et al., 2017; Santos et al., 2020]. We can use a pre-trained GlowGAN as a prior and apply GAN inversion to optimize latent code and exposure, making the model generate the corresponding HDR image for the input LDR image. An exciting result is that our method can, without using any HDR imagery or paired multi-exposure data, reconstruct starkly more plausible information for large overexposed regions than other supervised learning approaches trained on such data. Our contributions are summarized as follows: We present the first unsupervised approach for learning HDR images from in-the-wild LDR images, eliminating the need for ground truth HDR data, which is challenging to acquire. To achieve this, we propose GlowGAN, a generative model that bridges HDR and LDR spaces through a learned camera model, enabling the synthesis of diverse high-quality images with a significantly higher dynamic range than standard GANs. Leveraging GlowGAN as a prior, we introduce GlowGAN-ITM, an unsupervised inverse tone mapping method that outperforms state-of-the-art supervised approaches in reconstructing large overexposed regions, setting a new benchmark for HDR image restoration.

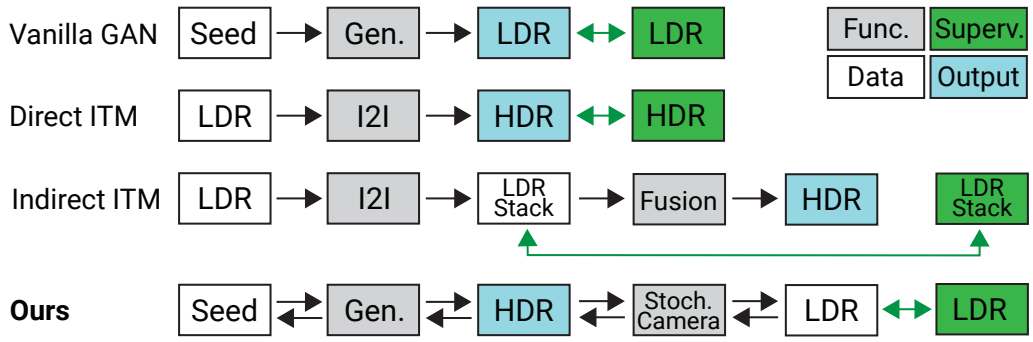


FIGURE 5.1: *Vanilla GANs are trained on unstructured LDR data and can only generate LDR. Direct ITM methods perform image-to-image (I2I) translation, but need to be trained on paired LDR–HDR data. Indirect ITM methods use I2I to produce an LDR exposure stack to be fused into an HDR result and require supervision in the form of LDR exposure stacks. In contrast, our approach allows to generate HDR images (left to right), using a stochastic camera model to allow supervision via unpaired in-the-wild LDR images. ITM (right to left) is performed using GAN inversion. Only our method yields HDR from unstructured LDR training data.*

5.2 Background

5.2.1 High Dynamic Range Imaging

Methods	Training Dataset	HDR[1]	Unsupervised [2]	Generative[3]	Hallucination[4]
Vanilla GAN	Unstructured LDRs	✗	✗	✓	✗
Direct ITM	Paired LDR-HDR images	✓	✗	✗	Weak
Indirect ITM	Paired LDR exposure stacks	✓	✗	✗	Weak
GlowGAN (Ours)	Unstructured LDRs	✓	✓	✓	Strong

[1] Capable of generating HDR images or performing inverse tone mapping.
[2] Capable of learning to produce HDR images from unstructured and unpaired LDR data.
[3] Capable of generating new images.
[4] Capable of hallucinating overexposed areas with plausible content, which is one of the most challenging problems in ITM.

TABLE 5.1: *Contributions of GlowGAN in comparison to other method classes.*

As reviewed in Section 2.2, HDR imaging [Reinhard et al., 2010] is crucial for creating immersive viewing experiences that portray the vast dynamic range of the real world. Despite recent progress made by learning-based methods, the majority of those techniques require supervision from paired LDR–HDR training data or exposure stacks. The need for training data constitutes a fundamental problem: HDR image data is hard to obtain and therefore naturally scarce. Further, most HDR capture techniques require static content, which significantly restricts the applicability of learning-based methods to arbitrary scenes. In contrast, we are the first to train an *unsupervised* HDR image generator from a fully unstructured set of LDR images. We believe this is an important step towards solving two main problems in HDR imaging: First, our generator can synthesize an abundance of HDR samples, alleviating the scarcity of HDR content. Second, our system allows to perform ITM with a significant improvement in the reconstruction of overexposed regions. We summarize the difference between our method and previous methods in Table 5.1 and Figure 5.1.

5.2.2 Lossy Generative Adversarial Networks (GAN)

As disscussed in Section 2.4.1, GAN [Goodfellow et al., 2014] is very successful in modeling distributions of images with high visual fidelity. The StyleGAN family [Karras et al., 2019; Karras et al., 2020b; Karras et al., 2020a; Karras et al., 2021; Sauer et al.,

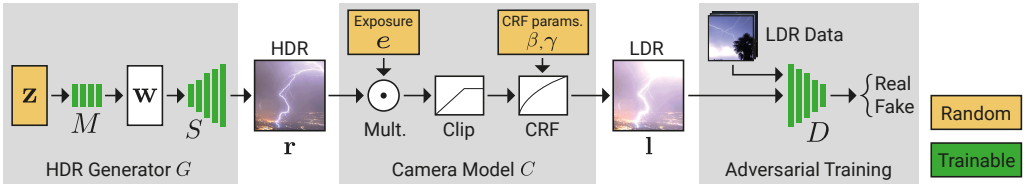


FIGURE 5.2: *Overview. The generator generates an HDR image* $\mathbf{r}$ *from a random noise* $\mathbf{z}$*. Then a camera model C projects* $\mathbf{r}$ *to an LDR image* $\mathbf{l}$ *with a random exposure and CRF. The model is trained on in-the-wild LDR images in an adversarial manner.*

2022] marks the current state of the art, scaling to high resolutions.

However, (unconditional) GANs are typically trained on LDR images, as acquiring high-quality HDR data at the scale required for effective training remains a significant challenge. In contrast, LDR images are abundant and easy to collect. Moreover, LDR and HDR images share strong semantic similarities, and LDR can even be viewed as a quantized projection of HDR in the luminance domain. This raises an important and promising question: can we recover the distribution of HDR content from partial LDR observations using a GAN framework?

This problem falls under the category of lossy generative modeling. AmbientGAN [Bora et al., 2018] demonstrates that it is possible to train a generative model using lossy observations, as long as the degradation process is known and differentiable. This principle has been applied to learn clean image generators from noisy data [Kaneko and Harada, 2020], and to recover all-in-focus images from inputs with shallow depth of field [Kaneko, 2021]. Notably, it has also enabled 3D generative modeling from 2D images by incorporating a projection model with known camera extrinsics [Henzler et al., 2019; Chan et al., 2021; Liao et al., 2020; Nguyen-Phuoc et al., 2019; Schwarz et al., 2020; Szabó et al., 2019].

Following this paradigm, we embed a physically motivated differentiable degradation model into the GAN training pipeline, converting HDR radiance into LDR pixel intensities. This allows the generator to learn an HDR distribution from lossy LDR inputs in a principled and physically consistent manner.

5.3 Method

We train a GAN [Goodfellow et al., 2014] to capture the distribution of HDR images in a domain (e.g., landscapes) by combining it with a stochastic camera model that transforms the generated HDR images into their LDR counterparts. The discriminator is only fed LDR images, which allows the system to be trained on an easily accessible in-the-wild LDR image dataset. Our camera model can be inserted into any GAN model to yield HDR outputs as long as the LDR training dataset consists of images exhibiting different exposures across samples. We consider this a mild assumption, which in particular in-the-wild photo datasets easily satisfy. See Figure 5.2 for an overview of our system.

In the following, we describe our pipeline in detail (Section 5.3.1), before turning to our main application of unsupervised inverse tone mapping GlowGAN-ITM (Section 5.3.2).

5.3.1 GlowGAN

We seek to capture the unknown true distribution of HDR images p_{HDR} from samples of the distribution of LDR images p_{LDR}. Similar to the standard GAN setup, we



(a) LDR Dataset (b) Generated HDR Images (c) Inverse Tone Mapping

FIGURE 5.3: *(Typical image datasets have a low dynamic range (LDR), resulting in over- and underexposed pixels (*a*). We devise a high-dynamic-range (HDR) generator trained only on in-the-wild LDR data. Our samples (*b*, tonemapped) exhibit details at all brightness levels. Our model can be used for inverse tone mapping to recover details in large-scale saturated regions (*c*).*

achieve this by training a generator G which turns a random latent vector $\mathbf{z} \in \mathbb{R}^k$ into an HDR sample $\mathbf{r} \in \mathbb{R}_{\geq 0}^{H \times W \times 3}$, an RGB image with $H \times W$ pixels and no upper restriction on the value range. To train G, we inject a camera model $C \in \mathbb{R}_{\geq 0}^{H \times W \times 3} \rightarrow [0,1]^{H \times W \times 3}$ into the adversarial training pipeline, turning the unbounded HDR image into an LDR image with values between 0 and 1. C captures the distribution p_{cam} of pixel-wise image transformations typically applied in a digital camera to convert incoming radiance to pixel intensities, including varying exposures, clipping, and varying non-linearities arising from the camera response function (CRF). The result of this process is an LDR image $\mathbf{l} = C(\mathbf{r}) = C(G(\mathbf{z}))$. The discriminator D is tasked with differentiating the fake samples $\mathbf{l}$ from samples from the distribution of true LDR images p_{LDR}. Since the samples $\mathbf{r}$ undergo stochastic projections from HDR to LDR, G is forced to produce valid HDR images, resulting in the distribution p_{HDR}^G of generated HDR images approaching the true distribution p_{HDR} [Bora et al., 2018].

As our generator and discriminator backbone, we choose the current state-of-the-art model StyleGAN-XL [Sauer et al., 2022]. This model has been demonstrated to yield excellent image quality on diverse datasets. The generator model consists of two stages (left block in Figure 5.2): First, a mapping network M_{θ_M} in the form of an MLP with parameters θ_M turns the initial random vector $\mathbf{z} \in \mathbb{R}^k$ into a more disentangled latent feature representation $\mathbf{w} \in W \subset \mathbb{R}^k$. Second, $\mathbf{w}$ is fed into a synthesis CNN S_{θ_S} with parameters θ_S to yield the final HDR output $\mathbf{r} = S_{\theta_S}(M_{\theta_M}(\mathbf{z}))$. Notice that the particular choice of the generator and discriminator networks is orthogonal to our approach. Except for the camera model introduced in the next paragraphs, we use StyleGAN-XL without any modifications in architecture or training hyper-parameters. We verify this in Section 5.4.2.

At the core of our method is the stochastic camera model C (central block in Figure 5.2) which projects the HDR image $\mathbf{r}$ onto an LDR counterpart $\mathbf{l}$. It is designed to model the distribution of typical processing steps in a digital camera:

$$C(\mathbf{r}) = \text{CRF}_{\beta,\gamma}\left(\min(2^{\frac{e}{2}} \cdot \mathbf{r}, 1)\right). \tag{5.1}$$

In the first step, we multiply each pixel of $\mathbf{r}$ with a single global exposure value. The exposure is parameterized by the random variable e, capturing the exposure distribution of typical cameras arising from the combined effect of aperture, shutter speed, and sensor sensitivity (ISO). We do not have access to the true distribution

of exposure parameters e, as images from in-the-wild image collections frequently do not have EXIF headers that would contain this information. Since the exposure is a combined effect of multiple factors (aperture, shutter, ISO, time of day, etc.), we choose to model e using a normal distribution, i.e., $e \sim \mathcal{N}(0, \sigma_e^2)$. The exposure variance σ_e^2 is the only hyper-parameter in our system and is analyzed in Section 5.4.2.

After applying a random exposure, the min operator in Eq. 5.1 clips large radiance values to 1, effectively removing all contrast in overexposed regions. Notice that this loss of information is precisely the reason why C is not invertible for individual images: By selecting an exposure, we only observe a bracket of radiance values. In contrast, our method seeks to invert C over the distribution of images and camera models [Bora et al., 2018].

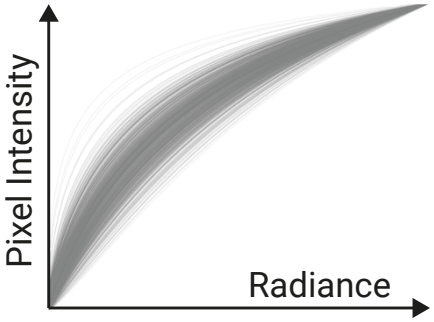


FIGURE 5.4: *CRF Distribution.*

The final component of our model in Eq. 5.1 is the non-linear sensor response. The CRF describes the mapping from radiance values arriving at the sensor to pixel intensities stored in the final image. We follow the established distribution of Eilertsen et al. [Eilertsen et al., 2017]:

$$\mathrm{CRF}_{\beta,\gamma}(x) = \frac{(1+\beta)x^{\gamma}}{\beta + x^{\gamma}},$$

with $\beta \sim \mathcal{N}(0.6, 0.1)$ and $\gamma \sim \mathcal{N}(0.9, 0.1)$ as obtained from the analysis of a large dataset [Grossberg and Nayar, 2003]. We visualize the distribution of CRFs arising from this model in Figure 5.4.

With the full stochastic camera model in place, our system can be trained in an adversarial fashion from scratch without further modifications to the standard GAN pipeline to yield an HDR generator trained only on an LDR image dataset.

5.3.2 GlowGAN-ITM

In addition to producing HDR image samples, a trained GlowGAN can be used to perform unsupervised inverse tone mapping. While existing approaches typically rely on supervision from LDR–HDR image pairs or exposure stacks, our method enables the recovery of HDR imagery from LDR inputs without HDR data using GAN inversion [Zhu et al., 2016; Xia et al., 2022]. This allows us to hallucinate plausible content in over-exposed image regions, which is one of the most challenging problems in ITM and, hence, a particular focus of our approach.

To obtain high-quality ITM results and to facilitate multi-modality, we choose a per-image optimization-based approach to perform the inversion [Creswell and Bharath, 2018; Abdal et al., 2019]. Given an LDR image $\hat{\mathbf{I}}$, we consider the following optimization objective:

$$[e^*, \mathbf{w}^*, \theta_S^*] = \underset{e,\mathbf{w},\theta_S}{\arg\min}\, \Phi\left(C(S_{\theta_S}(\mathbf{w})), \hat{\mathbf{I}}\right). \tag{5.2}$$

Here, we jointly optimize over exposure e and the latent code $\mathbf{w}$ while fine-tuning the synthesis network parameters θ_S to obtain a faithful match between the target LDR image $\hat{\mathbf{I}}$ and its reconstruction. Φ denotes the discrepancy measure between the two images. Using the Adam optimizer [Kingma, 2014] with standard parameters,

we proceed in two stages: In the first stage, we exclude the generator weights θ_S from the optimization and measure image discrepancy Φ using the LPIPS perceptual distance [Zhang et al., 2018]. In the second stage, we only optimize (fine-tune) θ_S using the pivotal tuning technique of Roich et al. [Roich et al., 2022], with Φ being the sum of a pixel-wise ℓ_2 loss and the LPIPS perceptual distance. We relax $\mathbf{w}$ to explore the extended latent space W^+ [Abdal et al., 2019]. We did not find it necessary to optimize over the CRF parameters β, γ for high-quality results and consequently fix them to their mean values. Upon completion of the optimization, we obtain the HDR version of $\hat{\mathbf{I}}$ via $\mathbf{r}^* = S_{\theta_S^*}(\mathbf{w}^*)$.

Following the lossy projection in Eq. 5.1, the mapping from LDR to HDR images is not unique: Overexposed regions in the LDR image can be explained by many different HDR images. Our system allows capturing this multi-modality by running the optimization of Eq. 5.2 multiple times with different parameter initializations for $\mathbf{w}$ and e. Specifically, we initialize each optimization run with $\mathbf{w} = M_{\theta_M}(\mathbf{z})$, where $\mathbf{z}$ is a normally distributed random vector. This allows us to obtain multiple plausible HDR solutions, which almost exclusively differ in the overexposed regions.

Obtaining accurate GAN inversion results is challenging [Blau and Michaeli, 2018; Abdal et al., 2019]. Fortunately, in most cases, we are only interested in hallucinating content in saturated regions, while well-exposed pixels can be re-used after linearization. Therefore, optionally, we diminish potential distortions arising from the inversion by blending the linearized original LDR image $\hat{\mathbf{I}}$ with the reconstructed HDR result $\mathbf{r}^*$ [Eilertsen et al., 2017; Santos et al., 2020]. Referring to the camera model in Figure 5.2, this blending happens in HDR space between *Mult.* and *Clip.* and is computed as follows:

$$\mathbf{r}^*_{\text{blend}} = e^* \cdot (\mathbf{m} \odot \mathbf{r}^*) + (1 - \mathbf{m}) \odot \text{CRF}^{-1}(\hat{\mathbf{I}}).$$

Here, $\odot$ denotes the Hadamard product and $\mathbf{m}$ is a soft mask, indicating saturated pixels in $\hat{\mathbf{I}}$, which we compute for each pixel i following [Eilertsen et al., 2017]:

$$\mathbf{m}_i = \frac{\max\left(0, \max_c \hat{\mathbf{I}}_{i,c} - \tau\right)}{1 - \tau},$$

where $\hat{\mathbf{I}}_{i,c}$ denotes the LDR image with pixel index i and color channel c. We set the threshold $\tau = 0.97$ in all our experiments, resulting in a short ramp toward saturation.

5.4 Experiments

We have conducted the following experiments to demonstrate the effectiveness of our approach. Section 5.4.1 shows the generated HDR data and compares it with an LDR equivalent; Section 5.4.2 explores the influences of the exposure distribution, model backbone, and sampled camera response curve and Section 5.4.3 presents the results of our GlowGAN-ITM. We use the tone mapper of [Mantiuk et al., 2008] to display HDR content in this paper.

5.4.1 Generation of HDR Images

We show in Figure 5.5 a variety of samples generated with GlowGAN and a comparison with a vanilla StyleGAN-XL. Generated samples from the vanilla GAN often bear overexposed regions similar to those in the LDR training images. In contrast, samples from GlowGAN preserve detailed appearance information even for bright



FIGURE 5.5: *(a) Dynamic range comparison between our GlowGAN (top) and a vanilla GAN (bottom) when varying the exposure values (EV) of the image. (b) HDR samples generated by our GlowGAN for the five tested datasets (please see Figure 5.3 for more samples). Our method generates high-quality images with much higher dynamic range and without overexposure.*

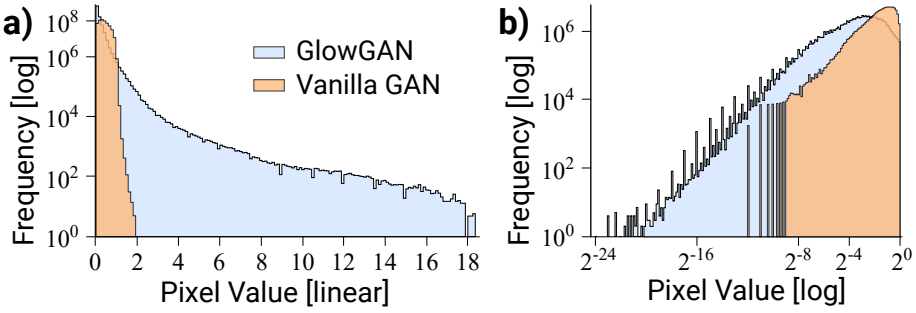


FIGURE 5.6: *Distribution of pixel values from our model and a vanilla GAN. The linear-scale histogram* (a) *shows that we significantly extend the dynamic range for high pixel values. To demonstrate our extension also in the low-intensity regime, we provide a truncated log-scale histogram* (b)*, where only values up to 1 are shown.*

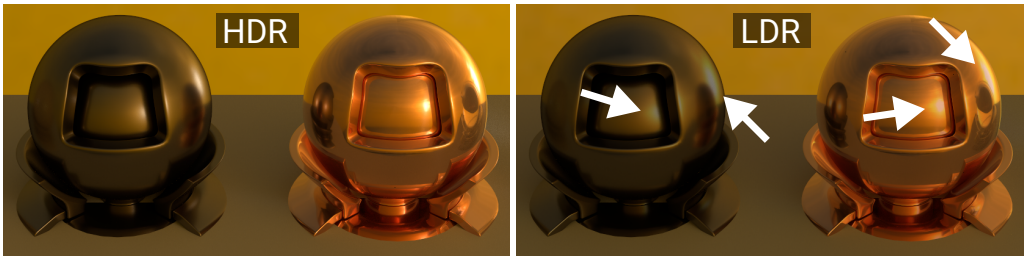


FIGURE 5.7: *Image-based lighting using a generated HDR environment map (left) vs. a plain LDR equivalent (right). Arrows mark regions where the differences are most pronounced. The HDR illumination leads to high-contrast and high-frequency reflections in the specular materials (right object) and produces a more realistic bloom effect in the glossy material (left object).*

objects such as the sun, as they have more extensive dynamic ranges than those from the vanilla GAN.

To analyze dynamic range, we plot the image brightness histogram of the two models in Figure 5.6, where each histogram is computed from 500 randomly sampled images. It can be seen that the histogram of the vanilla GAN exhibits an early cutof, while GlowGAN clearly avoids pixel intensity clamping (Figure 5.6a) and has a much wider histogram, also in the dark regions (Figure 5.6b). These results satisfy our expectation of learning HDR information from LDR data.

GlowGAN thus opens up a new avenue for getting cheap abundant HDR data that can be used in several applications, e.g., for augmenting training data for ITM methods or for creating environment maps for image-based lighting (IBL) [Debevec, 2006], as demonstrated in the Figure 5.7.

Model	σ_e^2	FID↓	KID($\times 10^4$)↓	DR50	DR90
SG-XL[1]	–	3.48	2.69	11.9	12.3
Ours	0.1	3.54	3.03	15.3	19.9
Ours	1.0	3.61	3.07	15.4	20.2
Ours	3.0	3.87	4.04	16.2	20.7
Ours	5.0	4.00	4.78	16.5	20.8
SG2[2]	–	9.2	23.37	12.9	16.9
Ours w/ SG2[3]	1.0	10.02	28.41	16.3	22.9

[1] Refers to a vanilla StyleGAN-XL model.
[2] Refers to a vanilla StyleGAN2-ADA model.
[3] Refers to our approach with a StyleGAN2-ADA backbone.

TABLE 5.2: *The effect of generator backbone and exposure variance σ_e^2 on image quality and dynamic range.*

5.4.2 Ablations

Exposure Distribution. Our method assumes the exposure e follows a Gaussian distribution with variance σ_e^2. Here we study how σ_e^2 impacts the generated image quality and the dynamic range. We employ the commonly used FID [Heusel et al., 2017] and KID [Bińkowski et al., 2018] scores to evaluate image quality on the Landscapes dataset. As we do not have ground truth HDR images, the scores are computed between the generated LDR images (i.e., the output of the camera model) and the LDR training images. We also compute the dynamic range (DR) for each generated HDR image $\mathbf{r}$ as $\text{DR} = \log_2\left(\mathbf{r}_{\text{max}}/\mathbf{r}_{\text{min}}\right)$, where $\mathbf{r}_{\text{max}}$ and $\mathbf{r}_{\text{min}}$ are the max and min values of the image, respectively. We report the median and the 90^{th} percentile of DR computed over 50k images, referred to as DR50 and DR90, respectively. From Table 5.2, we can observe a trade-off between image quality and the dynamic range, i.e., increasing σ_e^2 leads to a higher dynamic range (with diminishing returns for high σ_e^2) but slightly worse FID and KID scores. To understand the positive correlation between σ_e^2 and DR, suppose that $\mathbf{r}$ has a low dynamic range, then with a very small or large exposure e (which is more likely to happen for large σ_e^2), it would produce an out-of-distribution LDR image $\mathbf{l}$ that is overly dark or bright. In other words, only a valid high dynamic range $\mathbf{r}$ can yield realistic LDR images when processed with different exposures. On the other hand, as σ_e^2 increases, the camera model interferes more with the image generation process, which may increase the training difficulty as the Gaussian distribution is only an approximation to the underlying exposure distribution. In practice, users can choose a suitable σ_e^2 depending on their goal. In most of our experiments, we use $\sigma_e^2 = 1$ as it already removes overexposure while featuring good quality, and it also obtains better scores in the inverse tone mapping application. From Table 5.2, we can further see that a vanilla StyleGAN-XL exhibits slightly higher image quality, but a substantially smaller (log-scale) dynamic range.

Generator backbone. We further test our method based on StyleGAN2-ADA [Karras et al., 2020a]. As Table 5.2 shows, our method also successfully synthesizes images with high dynamic range, albeit the final image quality directly depends on the baseline generative model.

Stochastic CRF. We further study the effects of the stochastic CRF sampling process in our model. Table 5.3 compares our stochastic CRF sampling with a fixed CRF, using $\beta = 0.6$ and $\gamma = 0.9$. Modelling the stochasticity leads to a clear improvement in FID and KID scores. This is because the in-the-wild LDR images used for training

Dataset	CRF	FID↓	KID($\times 10^4$)↓
Landscapes	Fixed	3.89	3.80
	Stochastic	**3.61**	**3.07**
Lightning	Fixed	3.40	4.77
	Stochastic	**3.29**	**4.57**

TABLE 5.3: *Comparing quality with fixed and stochastic CRFs.*

are captured with different cameras with diverse CRFs, which can be better modelled via stochastic CRF sampling.

5.4.3 Application: Inverse Tone Mapping

A potential application of our approach is GlowGAN-ITM. We show both quantitatively and through a user study that our method outperforms previous approaches in hallucinating content in large overexposed regions, effectively recovering HDR content from a single LDR image.

Objective comparisons. We compare our GlowGAN-ITM to six state-of-the-art fully-supervised ITM methods, which we abbreviate as HDRCNN [Eilertsen et al., 2017], MaskHDR [Santos et al., 2020], SingleHDR [Liu et al., 2020b], ExpandNet [Marnerides et al., 2018], ReHDR [Lee et al., 2018a], and LANet [Yu et al., 2021b]. Following the work of Hanji et al. [2022] on quality assessment of single image HDR reconstruction methods, we select their three recommended full-reference metrics (PU21-PSNR, PU21-VSI, and HDR-VDP3) and their recommended non-reference metric (PU21-PIQE). As test set for the reference metrics we use as HDR ground truth a set of 62 images collected from existing datasets SIHDR [Hanji et al., 2022; Panetta et al., 2021] and generate the corresponding LDR input images following the pipeline proposed by [Eilertsen et al., 2017]. For the non-reference metric, we use an extended set of 100 LDR images obtained from the Internet which we use directly as input with 15% to 45% of the pixels saturated. For fairness in these comparisons, both sets are composed of landscape images, since fully-supervised approaches are typically trained with datasets mainly containing this type of content.

We show in Table 5.4 the results of these metrics and in Figure 5.8 visual comparisons for our five datasets. Note that, since our method focuses on hallucinating content in completely saturated regions, it is highly unlikely that this content fully matches that of the original ground truth image, therefore full-reference metrics are not well suited for assessing the quality of our reconstructions. Nevertheless, GlowGAN-ITM is on par with fully-supervised approaches for the reference metrics, while it excels in the non-reference metric, showing that our hallucinated content is more plausible in terms of naturalness. Additionally, previous methods can generate only one potential reconstruction given an input LDR image, while our approach allows the generation of multiple results with different but plausible semantic information, as Figure 5.9 shows.

We perform a subjective study to further assess the quality of our generated results for the ITM application. We include 20 scenes (four for each of our five datasets). For each scene, HDR results obtained with each of the seven methods were shown on a single screen, and participants were asked to rank the seven images from 1 (most preferred) to 7 (least preferred). The presentation order of the scenes and methods was randomized. The images were displayed in an HDR display Dell UP3221Q (3840×2160 resolution) in a standard office room with natural illumination, and

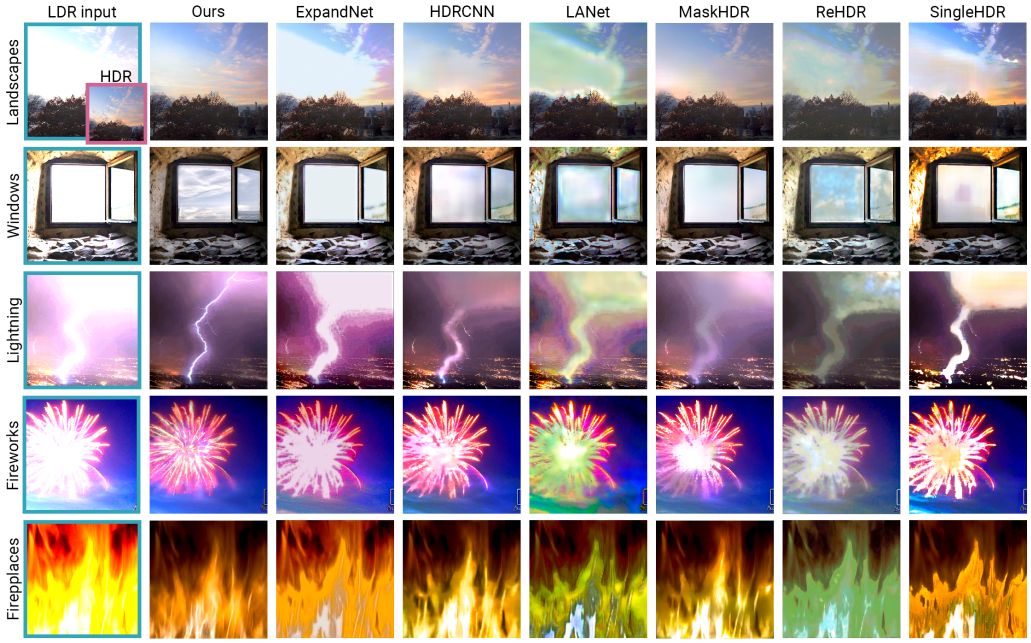


FIGURE 5.8: *Results for the ITM application and comparisons to six state-of-the-art methods for different datasets. Given a single LDR image as input, our method can produce an HDR image that contains plausible and realistic content in the previously overexposed regions, while previous methods tend to produce blurred results or even noticeable artifacts in such regions. In the first row (Landscapes dataset), it can be seen that recovering the ground truth content present in the original HDR scene is not possible in fully saturated regions, however, our method is able to produce plausible results that are consistent with the scene.*



FIGURE 5.9: *Our method can generate different but plausible HDR images from a single overexposed LDR input image.*

Method	Unsupervised	Reference HDR-VDP3 ↑	Reference PU21-VSI ↑	Reference PU21-PSNR ↑	Non-Reference PU21-PIQE ↓
HDRCNN	✗	7.42 ± 1.02	0.961 ± 0.034	32.1 ± 5.1	36.1 ± 5.3
MaskHDR	✗	7.60 ± 0.93	0.962 ± 0.032	32.4 ± 5.1	33.3 ± 6.5
SingleHDR	✗	7.01 ± 1.17	0.956 ± 0.031	30.1 ± 4.5	40.3 ± 6.2
ExpandNet	✗	6.66 ± 1.61	0.957 ± 0.033	30.7 ± 4.2	43.2 ± 6.6
ReHDR	✗	7.06 ± 1.31	0.953 ± 0.035	30.3 ± 4.2	39.7 ± 4.7
LANet	✗	6.94 ± 0.98	0.956 ± 0.031	29.0 ± 3.6	40.6 ± 6.5
Ours	✓	7.44 ± 0.94	0.961 ± 0.032	31.8 ± 4.4	31.8 ± 5.1

TABLE 5.4: *Evaluation of the ITM application. We achieve on-par quality with the best-performing supervised methods in reference metrics while outperforming other approaches in the non-reference metric that evaluates the naturalness and quality of the image. Note that reference metrics are not best suited for this evaluation, since the strength of our method lies in the reconstruction of overexposed regions in which the hallucinated content does not match that of the original HDR scene.*

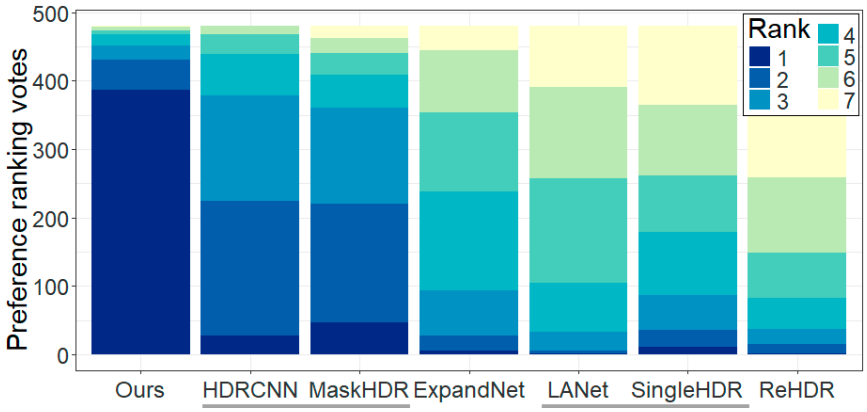


FIGURE 5.10: *Preference rankings for ITM methods aggregated across participants and scenes. Different colors indicate the rankings. Methods marked in the same set (gray underline) are statistically indistinguishable, while all others present statistically significant differences in their distributions.*

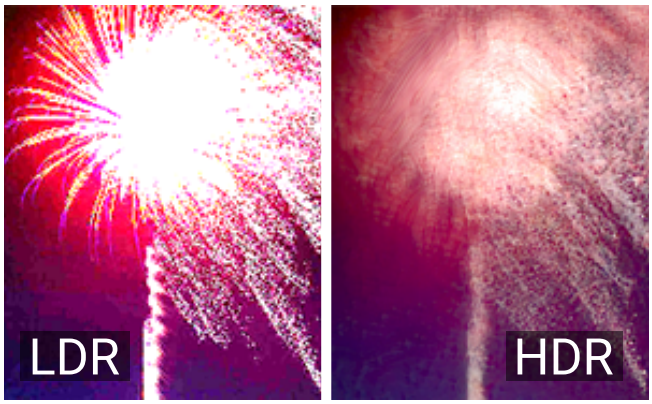


FIGURE 5.11: *Failure case.*

participants sat at a distance of 0.5 meters from the display. A total of 24 participants (38% female, aged 22 to 37 years old with normal or corrected-to-normal vision) participated in the study. We show in Figure 5.10 the preference rankings for each method, aggregated for all scenes and participants. For the analysis, we use pairwise Kruskal-Wallis tests adjusted by Bonferroni correction for multiple comparisons since the rankings do not follow a normal distribution. Results reveal that our method was ranked significantly higher than all others ($p < 0.001$), and it was selected as the top performing method in over 80% of the trials.

5.5 Conclusions and Limitations

We have introduced GlowGAN, a novel paradigm for learning HDR imagery from LDR data. Our method is orthogonal to other advances in generative adversarial learning and can be easily incorporated into any GAN-based pipeline. A trained GlowGAN acts as a strong prior, producing starkly more plausible inverse tone mapping results than previous approaches. Our inverse tone mapping method builds on GAN inversion via optimization, which can suffer from low-quality results, especially for high-frequency content (Figure 5.11) – a problem that is orthogonal to our approach. We rely on training data with a diverse exposure distribution. While this assumption is naturally satisfied for in-the-wild photo datasets, the dynamic range we can obtain is tightly linked to the exposure variance in the dataset. We

hope that our approach inspires future work on learning rich models from casually captured images.

Chapter 6

A Latent Exposure Fusion Diffusion Model for HDR Generation

This chapter introduces LEDiff, a generative approach for high-dynamic-range (HDR) content creation that extends the capabilities of existing diffusion models through latent space fusion. While modern displays support over 10 stops of dynamic range, most image assets—including internet photography and AI-generated content—remain constrained to 8-bit low-dynamic-range (LDR), limiting their applicability in HDR workflows. Existing LDR-to-HDR conversion techniques often fail to recover photorealistic details or physically plausible dynamic range in clipped highlights and shadows. To address these challenges, LEDiff integrates a small HDR dataset with a pretrained diffusion model, enabling high-fidelity HDR synthesis while preserving fine details in overexposed and underexposed regions. Inspired by image-space exposure fusion techniques, our approach not only facilitates HDR image generation but also enhances existing LDR-to-HDR conversion, expanding dynamic range without requiring paired multi-exposure training data. By leveraging LEDiff, generative models gain HDR capabilities, supporting applications such as photorealistic image synthesis, HDR environment map generation for image-based lighting, and advanced photographic effects like depth of field simulation, where linear HDR data is essential for realism. Extensive validation demonstrates that LEDiff outperforms existing methods, setting a new benchmark for generative HDR content creation.

6.1 Introduction

Generative models today can produce highly realistic and creative visual content, yet they remain largely confined to 8-bit, low dynamic range (LDR) representations, as demonstrated in Section 2.4. This limitation causes bright highlights and deep shadows to be clipped, preventing the full capture of light and color variation present in real-world scenes. As a result, these models restrict further photographic editing, as stronger tone mapping cannot reveal any additional details. Additionally, LDR limits downstream applications, such as image-based lighting and depth of field rendering, where realistic visual quality depends on HDR data. The need for HDR content has also grown with the rise of consumer displays that support HDR, often with over 10 stops of dynamic range, making the screen feel like a window looking out at the real scene [Benz, 2022; Chan, 2023]. The longstanding popularity of LDR content — optimized for traditional display purposes through clipping, gamma correction, and quantization — may no longer hold as HDR displays become increasingly prevalent.

HDR reconstruction has been widely studied in recent years and generally falls into two main categories. The first approach focuses on multi-exposure fusion,

which assumes a set of exposure-bracketed photos as input and primarily addresses alignment issues among these images [Debevec and Malik, 2023; Sen et al., 2012; Kalantari et al., 2017; Wu et al., 2018; Yan et al., 2019; Yan et al., 2020; Chen et al., 2023b; Liu et al., 2022; Tel et al., 2023; Niu et al., 2021; Xiong and Chen, 2021; Yan et al., 2023c; Prabhakar et al., 2021; Song et al., 2022; Kong et al., 2024; Ye et al., 2021]. The second approach centers on inverting the tone mapping applied to the input LDR image, aiming to recover missing details in clipped regions [Eilertsen et al., 2017; Marnerides et al., 2018; Liu et al., 2020b; Santos et al., 2020; Endo et al., 2017a; Yu et al., 2021b; Zhang et al., 2023b; Lee et al., 2018a]. With the growth of generative models, approaches to HDR generation also emerged.

GlowGAN [Wang et al., 2023a] is the first attempt to generate HDR content from scratch and can extend existing LDR images through GAN inversion. However, its effectiveness is limited by the class-specific training and generation constraints inherent to GANs.

Our goal is to repurpose a pre-trained latent diffusion model (e.g., Stable Diffusion [Rombach et al., 2022]) to generate a closely related yet distinct distribution: linear HDR content. To maintain the generative capabilities of the pre-trained model without compromising it by training on limited HDR data, we aim to retain the powerful latent space of the original latent diffusion model. Enabling HDR capabilities in a pre-trained generative model requires 1) hallucinating clipped area in highlights and shadows, 2) linearizing the image and expanding its dynamic range.

Interestingly, we observe that the latent space of a diffusion model strongly correlates with the image space in terms of clipping and pixel intensity; pixels that are clipped in the image space are similarly clipped in the latent space. This motivates our approach of aiming for a clipping-free latent space. We then fine-tune the decoder to linearize and expand the dynamic range, generating an HDR image in the image space. To get a clipping-free latent code, we draw inspiration from image-space exposure fusion, applying this concept in latent space by merging latent code that represents different exposure-level captures. To generate these bracketed codes, we train a highlight generator that takes as input a latent code with highlight clipping (representing an over-exposed image) and produces a lower-exposed latent code without clipping. Similarly, we train a shadow generator that takes an under-exposed latent code as input and generates a higher-exposed latent code free from shadow clipping. We train the highlight and shadow generators using a relatively small HDR dataset, relying on the preserved latent space to offload intensive generation tasks to the pre-trained model. Once we have these exposure-bracketed latent codes, a learnable fusion module merges them into a single, clipping-free latent code.

In summary, our approach involves training highlight and shadow generators to hallucinate missing details in over- and under-exposed areas of the latent space and fine-tuning the decoder to linearize and expand dynamic range during decoding.

This approach offers several advantages. Leveraging the *original pre-trained latent space* allows our method to act as a plug-and-play solution, converting any latent code to HDR, whether for generated content or real images encoded into the latent space. Using a *learnable latent space exposure fusion*, combined with a *learnable linearization decoder*, eliminates the need for hand-crafted merging features [Mertens et al., 2007] and avoids the necessity of estimating exposure parameters and camera response curves [Bemana et al., 2024] required by traditional exposure fusion algorithms.

Our method equips generative models with HDR synthesis capabilities, significantly broadening their applicability across various domains. We introduce a fine-tuning approach that preserves the pre-trained LDR latent space while leveraging a small HDR dataset to recover details in both deep shadows and bright highlights.

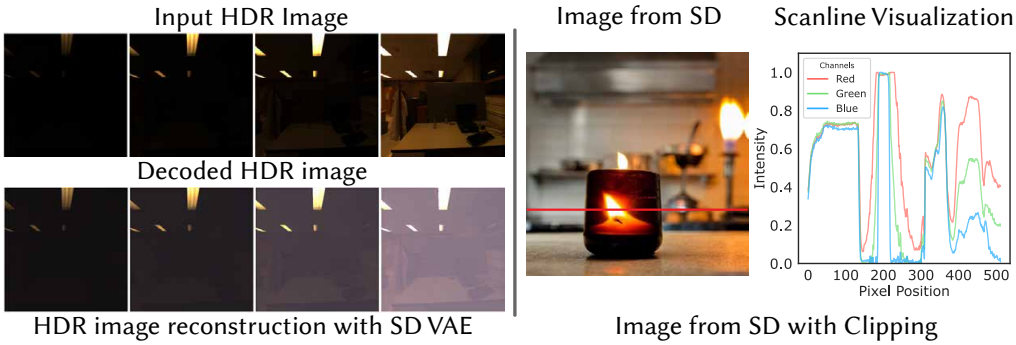


FIGURE 6.1: *Limitations of the vanilla Stable Diffusion (SD) in generating HDR content.* ***Left:*** *The limitation of the SD VAE in encoding and decoding an HDR image, visualized in multiple exposure levels, which reveals a significant fidelity loss, especially in the shadow.* ***Right:*** *An image generated with SD, along with a scanline that shows pixel clipping in highlight and shadow regions.*

Additionally, LEDiff enables HDR generation in existing image and video generative models, achieving state-of-the-art performance in inverse tone mapping by effectively linearizing LDR inputs and reconstructing photorealistic details in clipped regions. Through extensive validation, we demonstrate the effectiveness of our approach in enhancing HDR content generation and expanding the dynamic range of real-world LDR images.

6.2 Method

6.2.1 Preliminaries

A latent diffusion model (LDM), such as Stable Diffusion [Rombach et al., 2022], consists of a variational autoencoder (VAE), a denoising UNet, and a text encoder. The VAE encoder $\mathcal{E}$ compresses an image into a low-dimensional latent space. The denoising UNet then iteratively refines this noisy latent code into its clean version, which is subsequently decoded by the VAE decoder $\mathcal{D}$ into an LDR image.

LDMs are inherently limited to generating LDR images for two reasons. First, the VAE is trained exclusively on LDR images, constraining its ability to represent the dynamic range for HDR content. Second, the denoising UNet is designed to produce latent codes that encode only LDR images. These limitations are demonstrated in Figure 6.1. While diffusion models can plausibly hallucinate details in clipped regions — similar to their performance in inpainting tasks [Corneanu et al., 2024; Xie et al., 2023; Lugmayr et al., 2022] — the challenge lies in guiding them to generate images with true high dynamic range and to preserve details in both highlights and shadows.

A straightforward approach to adapting LDMs for HDR generation is to fine-tune both the VAE and the UNet with HDR data. However, this requires a significant amount of HDR data and does not fully utilize the generative priors of the pre-trained model. Instead, we propose to preserve the latent space of a pre-trained latent diffusion model as is, generating LDR latent codes at multiple exposure levels — a "latent exposure bracket" consisting of shorter and longer exposure variants. These latent codes are then fused in the latent space using a learnable lightweight fusion module. This merged HDR latent code is subsequently decoded by a fine-tuned VAE decoder to generate an HDR image.

In Section 6.2.2, we describe merging exposure brackets within the latent space and fine-tuning a specialized HDR decoder. In Section 6.2.3, we outline the approach for generating the latent exposure brackets.

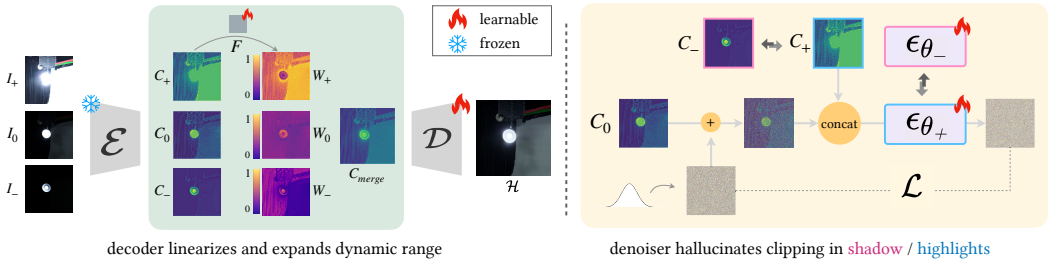


FIGURE 6.2: *Finetuning scheme of decoder and denoiser. (Left: fine-tuning the decoder) Exposure bracketed images* (I_+, I_0, I_-) *are encoded via the pre-trained encoder to generate corresponding latent codes. These latent codes are fused using a learnable fusion module* $(\mathcal{F})$ *to produce a latent code* (C_{merge}) *free of clipping, which is then decoded into an HDR image* $\mathcal{H}$ *through the finetuned decoder. (Right: fine-tuning the denoiser) The model takes as input the latent code* C_+ *for training highlight denoiser* ϵ_{θ_-} *or* C_- *for training shadow denoiser* ϵ_{θ_+}*, along with a* C_0 *corrupted by randomly sampled noise.*

6.2.2 Merging Latent Exposure Brackets

A key idea in our method is to preserve the priors learned by a pre-trained latent diffusion model by restricting it to generating latent codes for LDR images – the domain in which the model has been extensively trained on using billions of samples. The main challenge, then, is to effectively merge these LDR latent codes and decode them into an HDR image. To achieve this, we propose merging the latent LDR exposure bracket directly in the latent space and only fine-tuning the VAE decoder on a small set of HDR images to reconstruct the final HDR output, as shown in Figure 6.2-left.

We simulate exposure bracketed images from real HDR images. For each HDR image $\mathcal{H}$, we use the method from [Andersson et al., 2021] to calculate the lower (E_-) and upper (E_+) exposure bounds needed to cover the entire scene's dynamic range. We then project $\mathcal{H}$ into three exposure intervals, $\{E_-, E_0, E_+\}$, where $E_0 = (E_- + E_+)/2$, and randomly sample camera response curves (CRFs) from [Eilertsen et al., 2017] to introduce non-linearity. This generates an exposure bracket of low-, medium-, and high-exposure LDR images: I_-, I_0, I_+, as follows:

$$I_i = \frac{(1+\beta)\min(E_i, 1.0)^\gamma}{\beta + \min(E_i, 1.0)^\gamma}, \tag{6.1}$$

with $\beta \sim \mathcal{N}(0.6, 0.1)$ and $\gamma \sim \mathcal{N}(0.9, 0.1)$. We incorporate this randomization into the CRF sampling to improve the training data diversity. We use the pre-trained encoder $\mathcal{E}$ to map these LDR images into the latent codes: C_-, C_0, C_+.

Visualizing these latent codes (Figure 6.2) reveals a strong correlation between the latent space and image space, where intensity levels in over- and under-exposed regions are reflected similarly in the latent space. Thus, analogous to image-space HDR merging techniques that rely on brightness, color, and saturation to identify valid pixels [Mertens et al., 2007], we hypothesize that an efficient merging of the latent exposure bracket can be achieved by deriving a per-pixel weight map for each channel based on local regions of the latent codes.

To implement this, we introduce a straightforward yet effective learnable fusion module, $\mathcal{F}$. Each latent code is processed through a depth-wise convolution [Chollet, 2017] to produce an initial weight map, which is then normalized using a softmax function to ensure the merging weights sum to one across each pixel and channel. This merged latent code C_{merge}, combining information from multiple exposures, retains scene details without clipping. We then decode C_{merge} into an HDR image in

log space, fine-tuning the decoder with the original VAE's loss functions [Rombach et al., 2022], including both a reconstruction loss and a GAN loss that compares $\mathcal{D}(C_{\text{merge}})$ with $\mathcal{H}$ in log space:

$$\mathcal{L}_{\mathcal{F},\mathcal{D}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{GAN}} \tag{6.2}$$

6.2.3 Generating Latent Exposure Brackets

With a trained $\mathcal{F}$ and a fine-tuned $\mathcal{D}$, the remaining step is to generate an exposure bracket in the latent space, Figure 6.2-right.

We assume a single latent code is given, which may originate from either 1) a real image encoded into the latent space or 2) a text prompt. This allows LEDiff to naturally handle generated contents (i.e., Text-to-HDR generation) and real photos (i.e., LDR-to-HDR conversion).

When given a single LDR latent code C, we need to expand it into an exposure bracket $\{C_-, C_0, C_+\}$, which involves hallucinating details in over-exposed highlights and under-exposed shadows. We handle the hallucination of highlights and shadows separately but with similar processes, which is inspired by the conditioning strategy of [Ke et al., 2024]. For simplicity, we describe the highlight hallucination process: generating a lower-exposure latent code, C_- from C_0, and C_0 from C_+. This method implicitly learns to reduce the exposure of the conditioned image, effectively restoring details in clipped highlights.

Denote the highlight hallucination denoiser as $\epsilon_{\theta_-}(\cdot)$ with parameters θ_-. We fine-tune $\epsilon_{\theta_-}(\cdot)$ from the pretrained Stable Diffusion denoiser to generate C_- conditioned on the input latent C_0 and generate C_0 conditioned on the input latent C_+. The conditioning is achieved by concatenating the input condition with the noisy input before feeding it into the denoiser, with a modified first layer to handle change in channels. The rest of the network remains initialized with the pretrained weights. Fine-tuning uses a standard denoising diffusion loss.

To train $\epsilon_{\theta_-}(\cdot)$, we first randomly sample an exposure bracket from the HDR dataset, following the process outlined in Section 6.2.2. These samples are encoded using the pretrained VAE encoder, producing a latent exposure bracket $\{C_-, C_0, C_+\}$. We then corrupt the underexposed latent C_- and C_0 with Gaussian noise at a randomly sampled timestamp t. The task of the denoising network, $\epsilon_{\theta_-}(\cdot)$ is to predict the noise $\hat{\epsilon}_{C_-,C_0} = \epsilon_{\theta_-}(C_0, C_-^t, t)$ and $\hat{\epsilon}_{C_0,C_+} = \epsilon_{\theta_-}(C_+, C_0^t, t)$ using the objective:

$$\mathcal{L}_C = \mathbb{E}_{C,\epsilon_{C_-,C_0}\sim\mathcal{N}(0,\mathcal{I}),t\sim\mathcal{U}(T)} \left\| \epsilon_{C_-,C_0} - \hat{\epsilon}_{C_-,C_0} \right\|_2^2 \tag{6.3}$$

$$+ \mathbb{E}_{C,\epsilon_{C_0,C_+}\sim\mathcal{N}(0,\mathcal{I}),t\sim\mathcal{U}(T)} \left\| \epsilon_{C_0,C_+} - \hat{\epsilon}_{C_0,C_+} \right\|_2^2 \tag{6.4}$$

During inference, given an LDR image, we iteratively apply the fine-tuned denoiser to generate its corresponding lower-exposure latent codes.

For generating a higher-exposure latent code C_+, C_0 that hallucinates shadow details, we follow the same procedure to fine-tune a conditional denoiser $\epsilon_{\theta_+}(\cdot)$ with parameters θ_+, conditioned on C_0 and C_- respectively.

After generating the latent exposure bracket, we merge it using $\mathcal{F}$, then decode the merged latent code into an HDR image with the fine-tuned HDR decoder.

6.3 Dataset and Training

Datasets We gather HDR images from multiple sources [Liu et al., 2020b; Gardner et al., 2017; Bolduc et al., 2023; Kalantari et al., 2017; Tel et al., 2023; Panetta et al., 2021; Fairchild, 2007; "Poly Haven" 2024]. For the panorama datasets [Gardner et al., 2017; Bolduc et al., 2023; "Poly Haven" 2024], we create regular images by projecting the panoramas from random camera viewpoints. In total, we obtain 36,000 HDR images, and then apply the method mentioned in Section 6.2.2 to generate the training pairs $\{I_-, I_0, I_+\}$ and $\{\mathcal{H}\}$.

Training We use this dataset to train $\mathcal{F}$ and to fine-tune $\mathcal{D}$, $\epsilon_{\theta_-}(\cdot)$ and $\epsilon_{\theta_+}(\cdot)$. $\mathcal{F}$ and $\mathcal{D}$ are trained with $\{I_-, I_0, I_+\}$ as inputs and $\{\mathcal{H}\}$ as the ground truth. For $\epsilon_{\theta_-}(\cdot)$ training, we use the pairs $\{I_0, I_+\}$ and $\{I_-, I_0\}$, where the first image serves as the conditioning input and the second is used as the input (to be applied with noise) and ground truth. Similarly, for $\epsilon_{\theta_+}(\cdot)$ training, we use the pairs $\{I_0, I_-\}$ and $\{I_+, I_0\}$. We fine-tune the VAE decoder for 200,000 steps with a learning rate of 10^{-6} and the denoiser for 400,000 steps with a learning rate of 10^{-5}. All fine-tuning processes use the Adam optimizer [Kingma, 2014].

6.4 Applications and Evaluations

We evaluate LEDiff across various HDR content generation tasks: text-to-image, text-to-panorama, and image-to-video (Section 6.4.1), and use it for HDR reconstruction from an LDR image (inverse tone mapping) in Section 6.4.2. Our results cover diverse scenes, demonstrating the generalization ability of LEDiff and its seamless integration into applications benefiting from HDR capabilities.

6.4.1 HDR Content Generation

Text to HDR Image We compare the text-to-image of LEDiff with that of Stable Diffusion (SD). Starting with a generated latent code C_+, we apply diffusion with the denoiser $\mathbf{ffl}_{\theta_-}$ to generate two lower exposure latent codes, C_0 and C_-. The fusion module $\mathcal{F}$ merges these latent codes into $C_{\text{merge}} = \mathcal{F}(C_-, C_0, C_+)$, which is then decoded into an HDR image via $\mathcal{D}$. Figure 6.3-left shows that LEDiff achieves a higher dynamic range without clipping, while SD output suffers from clipping. We also demonstrate the effectiveness of HDR for further image editing, such as applying synthetic defocus, where linear HDR content is crucial for realistic depth of field rendering, especially for bokeh simulation [Mildenhall et al., 2022; Zhang et al., 2019c].

Text to HDR Panorama Our method seamlessly integrates with existing SD-based panorama generation models. Using MVDiffusion [Tang et al., 2023] as the baseline, LEDiff enables HDR panoramas, which are essential for image-based lighting. As shown in Figure 6.3-right, using the HDR panorama as the environment light map enhances contrast and produces more realistic highlights in the rendered image.

Image to HDR Video Our method can be incorporated into SD-based image-to-video models [Blattmann et al., 2023], facilitating the generation of HDR video from a single LDR image. Examples of this application are included Figure 6.4.

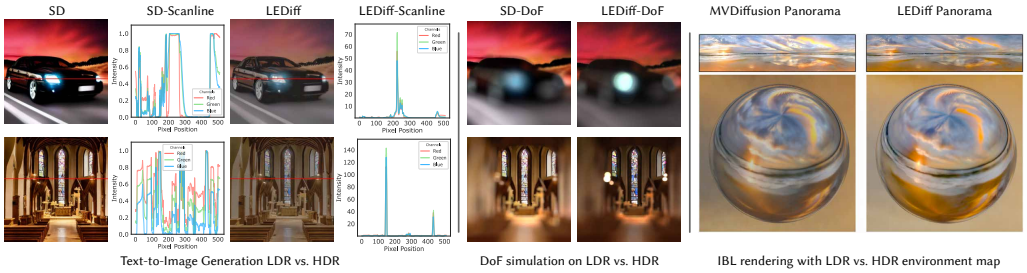


FIGURE 6.3: *textbfLeft: Text to HDR image generation using prompts: (1)* ***"Bright car headlights on a narrow street at night"*** *and (2)* ***"A grand church interior with tall stained glass windows, intricate wooden arches, and warm lighting"****. We compare the images generated by SD and LEDiff. We plot a scanline passing through the car headlights and the bright luminaries to demonstrate our ability to produce a wide dynamic range with photorealistic detail. We also show an application of synthetic depth of field (****DoF****), where a linear HDR image is crucial for rendering realistic defocus effects.* ***Right:*** *Comparisons are made between the panoramas produced by MVDiffusion [Tang et al., 2023] and our approach with a prompt* ***"A peaceful beach at sunset with soft clouds in the sky"****, along with their respective image-based lighting results. Our method leads to higher contrast and an overall more realistic appearance.*

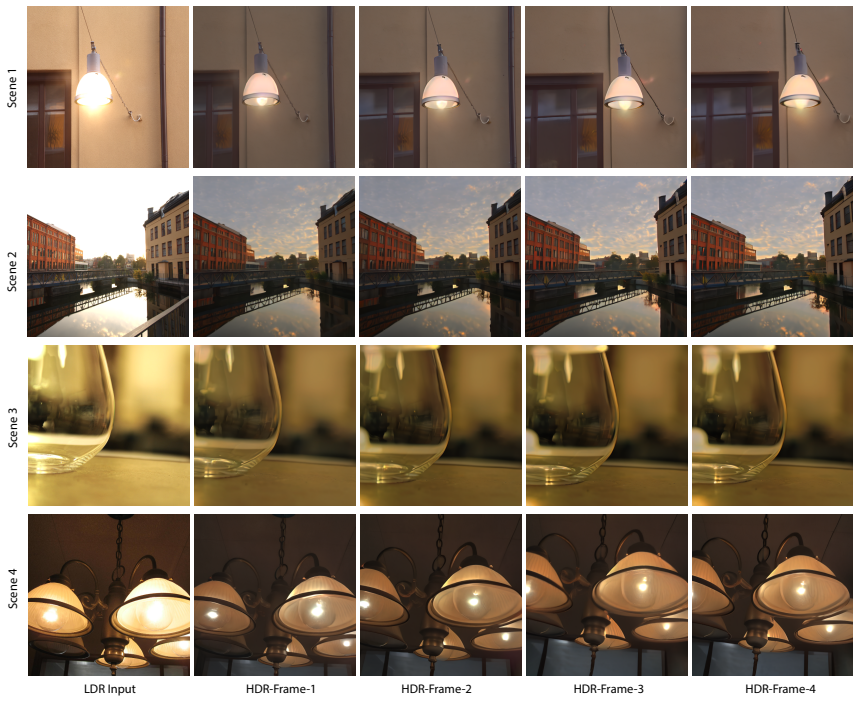


FIGURE 6.4: *Results of image-to-HDR video. Our method enables the baseline model, SVD, to generate HDR video from a single LDR image.*

6.4.2 LDR to HDR Image Reconstruction

Another promising application of LEDiff is inverse tone mapping. Given an LDR image I_+ with potential clipping, we obtain its latent code $C_+ = \mathcal{E}(I_+)$. Next, we generate lower exposure latents C_0 and C_-, which are fused into a merged latent code $C_{\text{merge}} = \mathcal{F}(C_-, C_0, C_+)$. Finally, the HDR reconstruction $\hat{\mathcal{H}}$ is obtained through the decoder $\mathcal{D}$.

We evaluate our method using standard image quality metrics. However, most established metrics are optimized for LDR content. To better evaluate the HDR outputs, we also conduct a user study in a controlled environment, displaying the

content on an HDR display to assess human perception of the HDR contents.

We compare our method with five inverse tone mapping methods: HDRCNN [Eilertsen et al., 2017], MaskHDR [Santos et al., 2020], SingleHDR [Liu et al., 2020b], ExpandNet [Marnerides et al., 2018], and GlowGAN [Wang et al., 2023a]. Since GlowGAN handles only class-specific inputs, we quantitatively evaluate it alongside our method on a subset of SI-HDR [Hanji et al., 2022], but exclude it from the user study.

Image Quality Metrics In our quantitative evaluation, we evaluate on images from SI-HDR [Hanji et al., 2022], which are excluded from the training dataset. We use the full-reference HDR-VDP3 [Mantiuk et al., 2023] and no-reference PU21-PIQE [Hanji et al., 2022] metrics for quality evaluation. Since standard metrics struggle to assess hallucinated details in clipped regions, we also use the FID score [Heusel et al., 2017] to evaluate image distributions. For HDR evaluation with FID, we tone-map both generated and ground-truth images to the LDR domain using Reinhard [Reinhard and Devlin, 2005], Durand [Durand and Dorsey, 2002], and Liang [Liang et al., 2018]. Following [Chai et al., 2022], we generate 60 random 128 $\times$ 128 pixel crops per image, resulting in 10k patches for FID computation. While tone-mapped images have different feature statistics from typical LDR images, we found that our FID scores correlate well with perceived quality and maintain consistent rankings across tone-mapping methods.

Previous works [Eilertsen et al., 2017; Santos et al., 2020; Liu et al., 2020b] primarily focus on hallucinating highlight regions, with limited attention to shadows. By contrast, our approach enables effective hallucination across both highlight and shadow areas. For highlight hallucination, we follow [Eilertsen et al., 2017; Santos et al., 2020; Wang et al., 2023a] to blend the generated content with the inputs. The process models HDR luminance in the non-overexposed regions of the LDR image, aligning it with the corresponding HDR area. Figure 6.5 shows that our method produces more natural hallucinated content within clipped highlight and shadow regions, while other methods produce low-quality artifacts and blur. This is supported by the quantitative results in Table 6.1.

Method	Highlights					Shadow				
	HDR-VDP3 ↑	PU21-PIQE ↓	FID-R ↓	FID-D↓	FID-L↓	HDR-VDP3 ↑	PU21-PIQE ↓	FID-R ↓	FID-D↓	FID-L↓
HDRCNN	6.90 ± 1.10	49.43 ± 6.91	13.39	16.95	16.67	6.56 ± 1.10	70.95 ± 7.84	32.55	44.20	26.41
MaskHDR	6.47 ± 1.06	49.38 ± 6.95	12.83	13.85	15.21	6.49 ± 1.24	70.98 ± 8.64	32.54	44.43	26.10
SingleHDR	6.13 ± 0.87	49.74 ± 7.29	28.68	34.53	29.50	7.47 ± 0.78	46.04 ± 9.41	39.99	57.28	37.17
ExpandNet	5.23 ± 1.61	52.53 ± 6.50	18.85	25.49	21.34	5.40 ± 0.93	60.01 ± 8.37	32.60	45.50	29.10
Ours	6.16 ± 0.97	48.46 ± 7.04	12.70	13.08	13.73	6.98 ± 0.83	43.37 ± 8.04	20.93	25.54	24.26
GlowGAN	5.67 ± 1.24	46.18 ± 8.22	27.44	40.43	43.85	–	–	–	–	–
Ours∗	5.92 ± 1.17	45.29 ± 7.27	15.62	17.89	16.19	–	–	–	–	–

TABLE 6.1: *Quantitative evaluation on content hallucination (Ours∗ denotes the test results of our method on a subset of the test set, selected to meet GlowGAN's requirement for class-specific inputs). Full-reference metrics are less suited to generative models due to natural deviations in hallucinated content from the original HDR scene. Our method achieves competitive performance on the full-reference HDR-VDP3 metric and outperforms alternative methods on the no-reference PU21-PIQE metric, which better captures image naturalness and perceptual fidelity. To complement these metrics, we use FID [Heusel et al., 2017] to evaluate the distribution of generated and ground-truth HDR images. Since FID is optimized for LDR, we tone map all HDR images using three tone mapping operators, resulting in three FID scores: FID-R [Reinhard and Devlin, 2005], FID-D [Durand and Dorsey, 2002], and FID-L [Liang et al., 2018]. Across all tone-mapping methods, our method consistently achieves lower FID scores, demonstrating closer alignment with the HDR ground-truth distribution and superior overall quality.*

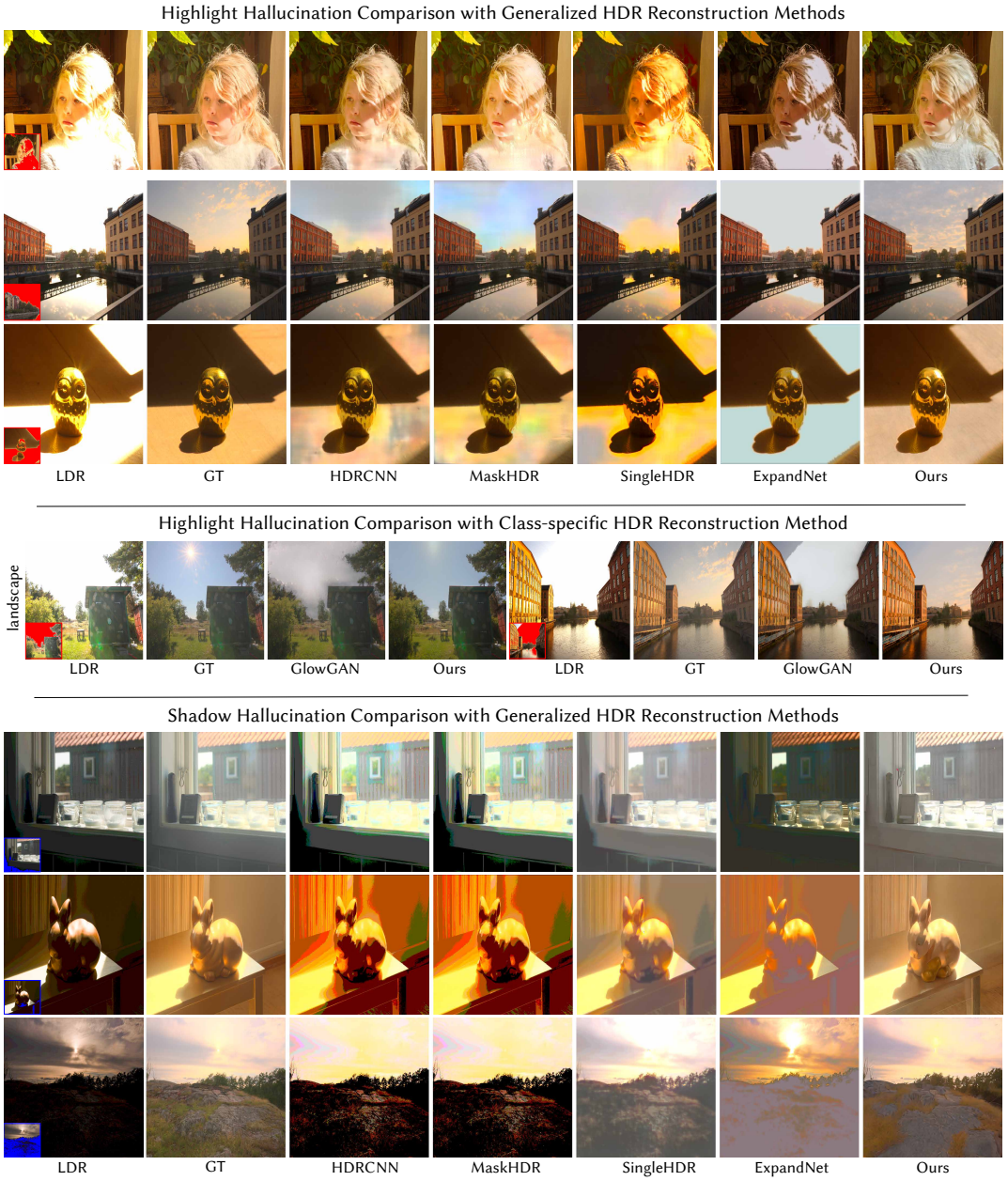


FIGURE 6.5: *LDR-to-HDR image reconstruction comparisons. Our method effectively hallucinates details in both over- and under-exposed regions, while previous approaches [Eilertsen et al., 2017; Santos et al., 2020; Liu et al., 2020b; Marnerides et al., 2018; Wang et al., 2023a] struggle to produce plausible results, especially in shadow regions that they do not address (e.g., HDRCNN and MaskHDR yield identical results for shadow hallucination, as both methods process non-clipped regions in the same way.). Images are tone-mapped for visualization. Best viewed in HDR on an HDR display.*

User Study To comprehensively evaluate the quality of our generated results for the ITM task, we conducted a subjective study. The evaluation included 60 scenes, with half drawn from the SIHDR dataset and the remaining half comprising LDR images collected from online sources or real-world captures. For each scene, a competing method was randomly selected and compared to our approach, with participants tasked to express their preferences. The study was carried out using an HDR display (Dell UP3221Q, 3840×2160 resolution) in a standard office environment with natural lighting, with participants positioned 0.5 meters from the screen. The study engaged

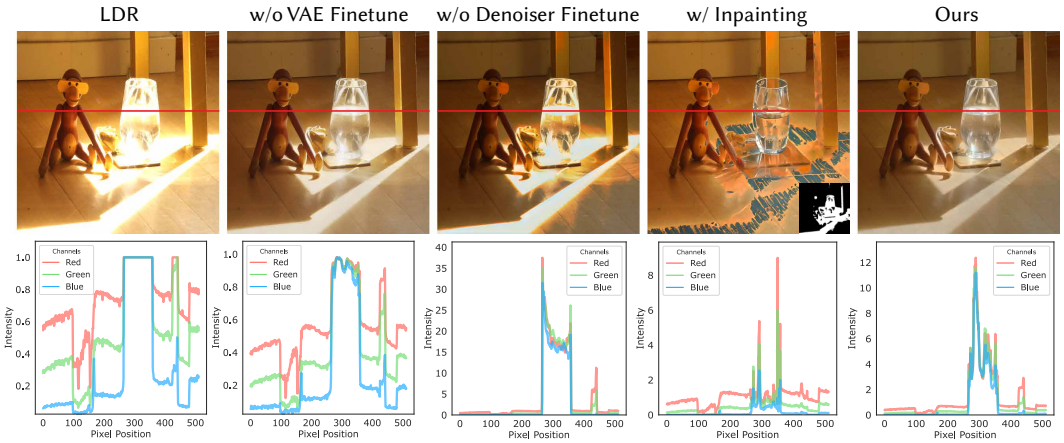


FIGURE 6.6: *Ablation study. The first row displays the input LDR image alongside the reconstructed outputs; the second row plots the scanline. The mask used for inpainting is shown in the inset. Fine-tuning both the decoder and denoiser is essential for achieving hallucination and dynamic range extension.*

20 participants, all possessing normal or corrected-to-normal vision, resulting in a total of 1200 pairwise comparisons. The outcome of this assessment is detailed in Table 6.2. The results indicate that our method significantly outperforms other approaches, demonstrating a substantial margin of superiority.

Total	Ours	HDRCNN	Total	Ours	MaskHDR
310	84.19%	15.81%	297	88.22%	11.78%
Total	**Ours**	**SingleHDR**	**Total**	**Ours**	**ExpandNet**
283	89.40%	10.60%	310	94.52%	5.48%

TABLE 6.2: *User study comparisons with other methods. Item **Total** shows the total number of comparisons, item **Ours** reflects the percentage of users who preferred our results, and the other item indicates the percentage of users who chose other methods. In all cases, we observe a statistically significant preference for our method, as confirmed by a binomial test ($p < 0.01$).*

6.4.3 Ablations

We conduct the following experiments to illustrate the contribution of each component in our method, by excluding: 1) VAE decoder finetuning and 2) denoiser finetuning. We also use SD-based inpainting as an alternative approach for hallucinating clipped areas. We follow Section 6.4.2 for evaluation. Since the inpainting needs an extra mask as input, we extract the clipped region with a threshold.

Results are shown in Figure 6.6, with quantitative evaluation in Table 6.3. As illustrated in Figure 6.6, omitting the decoder leads to hallucinations in overexposed regions, but restricts the content to low dynamic range (LDR). In contrast, without denoiser fine-tuning, the dynamic range is extended, but hallucinations within clipped regions are not achieved. Additionally, using an SD-based inpainting method proves suboptimal, as it relies on irregular masks that differ significantly from those seen during training, resulting in unnatural results.

Method	HDR-VDP3 ↑	PU21-PIQE ↓
wo VAE	4.67 ± 0.87	48.60 ± 7.02
wo denoiser	5.59 ± 1.12	50.25 ± 6.55
inpainting	3.65 ± 1.20	50.18 ± 7.53
Ours	6.16 ± 0.97	48.46 ± 7.04

TABLE 6.3: *Ablation study on individual components. The table shows that fine-tuning both the VAE decoder and the denoiser yields the highest performance.*

We choose to use an exposure bracket with three exposure levels, as it is a common practice in image-based exposure fusion [Kalantari et al., 2017; Wu et al., 2018; Chen et al., 2023b; Liu et al., 2022; Song et al., 2022; Tel et al., 2023].

6.5 Conclusions and Limitations

We propose LEDiff, a method for HDR content generation that captures the full dynamic range of real-world scenes. LEDiff operates within the latent space of a pre-trained diffusion model, leveraging its generative capabilities while enabling HDR by reconstructing details in over- and under-exposed regions and performing dynamic range extension. This is accomplished by fine-tuning the VAE decoder and denoiser on a relatively small HDR dataset. LEDiff is agnostic to how the LDR latent code is generated, making it adaptable as an LDR-to-HDR converter for any LDR image encoded into latent space. LEDiff unlocks a range of HDR applications, including HDR panorama generation for realistic image-based lighting and HDR video generation.

However, our approach has limitations. By fine-tuning Stable Diffusion, LEDiff inherits the generation limitation of SD. Additionally, we do not yet simulate degradations such as compression artifacts (e.g., ringing and blocking) or noise in the input LDR, which could improve the model's robustness to real-world images. Addressing these limitations could improve generalization and is a promising direction for future work.

As HDR displays become more widely available on consumer devices, we anticipate that the demand for HDR content will continue to grow, and believe that LEDiff's potential to expand HDR content creation will help meet the increasing demand for photorealistic HDR contents.

Chapter 7

A Self-supervised Tone Mapping Operator

This chapter introduces a self-supervised approach for tone mapping high-dynamic-range (HDR) images, trained adaptively at test time without requiring labeled data. We formulate tone mapping as a perceptual optimization problem, leveraging a novel loss function rooted in contrast perception principles to directly compare HDR and tone-mapped images. Our method redefines classic VGG feature maps into feature contrast maps, normalizing local differences based on their neighborhood context to account for contrast masking effects. A key advantage of our approach is its independence from preexisting tone mapping operators, overcoming the limitations of dataset-driven methods constrained by traditional operator quality. By optimizing the tone mapping function specifically for each HDR input, we achieve flexible adaptation across diverse scenes while preserving local contrast and fine structural details. Extensive ablation studies and parameter exploration validate that our method consistently outperforms existing techniques with a single set of fixed parameters, as confirmed by both objective evaluations and perceptual user studies.

7.1 Introduction

As discussed in Section 2.1, High Dynamic Range (HDR) images can reproduce real-world appearance by encoding wide luminance ranges. With the fast-paced developments in capturing devices, access to HDR image and video is becoming commonplace. Nevertheless, the majority of widespread displays are still limited in the dynamic range they can reproduce, preventing direct reproduction of HDR content. Therefore, the use of tone mapping techniques is yet needed in order to adapt such content to current display capabilities.

Different tone mapping techniques have been developed for decades [Reinhard et al., 2010; Banterle et al., 2017], however the performance of even the most prominent techniques strongly depends on the HDR image content and specific parameter settings [Zhang et al., 2019b; Guo and Jiang, 2021; Panetta et al., 2021]. Subjective evaluations of these different techniques indicate that both specific algorithms as well as default parameter settings, as often proposed by their respective authors, do not generalize well across scenes [Ledda et al., 2005; Yoshida et al., 2005; Čadík et al., 2008]. In these cases, manual selection of a suitable algorithm and experience in fine-tuning its parameters might be required for high-quality results. This hinders smooth adoption of HDR technology and is the key obstacle in developing machine learning solutions, as we discuss next.

Recently, an increasing number of learning-based tone mapping methods have been proposed that show huge potential in terms of generality and quality with

respect to their traditional algorithm-based predecessors. However, there are still some shortcomings. First, most learning-based models regard tone mapping as an image restoration task and optimize the network in a fully supervised manner, which requires high-quality paired HDR and tone mapped training data. Given a set of HDR scenes as training set, the Tone Mapping Quality Index (TMQI) [Yeganeh and Wang, 2012] is typically used to select the best tone mapped image for each scene from a preexisting set of tone mapped results, limiting the quality of newly generated results to that of preexisting methods [Zhang et al., 2019b; Rana et al., 2019; Panetta et al., 2021]. And second, since such methods treat tone mapping as an image restoration task instead of an information reduction process, they usually involve large-scale networks. Motivated by these observations, in this work we seek an alternative solution that allows for reducing the network size and optimizes information reduction for a given HDR image content.

Since image contrast is arguably one of the key cues in the Human Visual System (HVS) while seeing and interpreting images [Palmer, 1999], we aim at reproducing perceived contrast in HDR scenes while also ensuring structural fidelity by reproducing visible image details. To this end, we propose a simple image-specific, self-supervised tone mapping network that is trained at test time and does not require any data labeling. The only training data is the input HDR image, and the key novelty in our approach is the loss function that directly compares the content in the HDR and tone mapped images. Since the compared signals present different luminance and contrast ranges, a direct computation of the loss in the feature space, as in e.g., perceptual VGG loss [Simonyan and Zisserman, 2014; Johnson et al., 2016], leads to sub-optimal results as we demonstrate in Section 7.5. To mitigate this problem, we first propose an adaptive μ-law compression that accounts for the scene brightness and brings HDR image histograms closer to those of low dynamic range images (Figure 7.4). Then, motivated by contrast perception literature [Legge and Foley, 1980; Foley, 1994; Watson and Solomon, 1997; Daly et al., 2000], we introduce in our loss function a non-linearity considering both the HVS response for stronger contrasts and visual neighborhood masking, and we model it in the network's feature space. To this end, we first formulate a local contrast measure in the feature space, normalizing local feature differences by their average magnitude in a local neighborhood. This also allows for further abstraction from the magnitude difference between HDR and LDR signals. Then, we introduce a compressive non-linearity as a function of feature contrast magnitude for the HDR signal, so that for higher magnitudes any departs in feature contrast are less strongly penalized in the loss function. This directly translates into compressing higher contrasts while preserving image details in the tone mapped images generated by our network. Finally, we also introduce feature contrast neighborhood masking, so that the loss function penalizes more weakly changes in feature contrast when similar features are present in the spatial neighborhood of the image. We perform extensive ablation studies and exploration of parameters and demonstrate that our solution outperforms existing approaches for a single set of parameters, as confirmed by both objective and subjective metrics.

7.2 Background

7.2.1 Contrast Perception Modeling

Visual sensitivity is affected by a number of key image dimensions such as luminance level [Ahumada Jr and Peterson, 1992], spatial and temporal frequency [Robson,

1966], or local image contrast [Watson, 1989], as well as their interactions. In particular, changes in sensitivity as a function of local image contrast are generally termed masking effects. There are two main masking effects related to spatial contrast perception that have been widely studied and applied to computer graphics applications: contrast self-masking and visual contrast masking. Contrast self-masking [Daly et al., 2000] is characterized by a strong non-linearity that allows for stronger absolute changes for higher supra-threshold contrasts than for lower near-threshold contrasts before such changes become noticeable [Kingdom and Whittle, 1996]. Visual masking (also called neighborhood masking) [Legge and Foley, 1980; Foley, 1994; Watson and Solomon, 1997; Daly et al., 2000] is a phenomenon in which sensitivity is locally reduced with increases in image local contrast [Daly, 1992]. When contrasts with similar spatial frequencies are present in a close neighborhood, the thresholds for detection of lower contrasts and for change discrimination of higher contrast rise. To model this effect, the input image contrast is decomposed into frequency bands using a filter bank such as a Laplacian pyramid [Peli, 1990; Mantiuk et al., 2021], a cortex transform [Daly, 1992], wavelets [Daly et al., 2000], or discrete cosine transforms (DCTs) [Watson, 1993], and then the visual masking is modeled for each frequency band.

Modeling contrast self-masking and neighborhood masking has been proven to be beneficial for several applications including image [Daly et al., 2000] or video [Yi et al., 2021] compression, image quality evaluation [Lubin, 1995; Mantiuk et al., 2021], rendering [Bolin and Meyer, 1998; Ramasubramanian et al., 1999], and foveated rendering [Tursun et al., 2019]. Existing works apply such contrast perception models in the primary image contrast domain (or disparity domain [Didyk et al., 2010]), and employ predefined filter banks. Instead, we use a neural network and compute per-channel contrast signals over feature maps, where optimal filters are learned for the task at hand, and we formulate a novel loss function that models contrast self-masking and neighborhood masking in the feature contrast domain.

7.2.2 Tone Mapping Techniques

As summarized in Section 2.2.3, tone mapping approaches are commonly divided into traditional and learning-based methods. Traditional methods are typically model-driven, relying on handcrafted rules or perceptual heuristics. In contrast, learning-based methods adopt data-driven models and often require paired HDR–LDR training data, which are either manually curated or generated using existing tone mapping operators.

Thus, these learning methods are intrinsically limited by the input data they see during training. Therefore, using images tone mapped with existing methods as target, although allows for training new models with better generalization, fundamentally limits the quality that such new learned models can achieve. In contrast, we propose a self-supervised network that only takes as input the original HDR image for training, and learns a tone mapping operator relying on a carefully designed loss function that takes into account contrast masking effects present in the Human Visual System. To our knowledge, the only work that does not need carefully selected HDR-LDR image pairs is the work of [Hou et al., 2017], however, they rely on combining feature losses from different layers chosen empirically and only demonstrate their approach for two selected images.

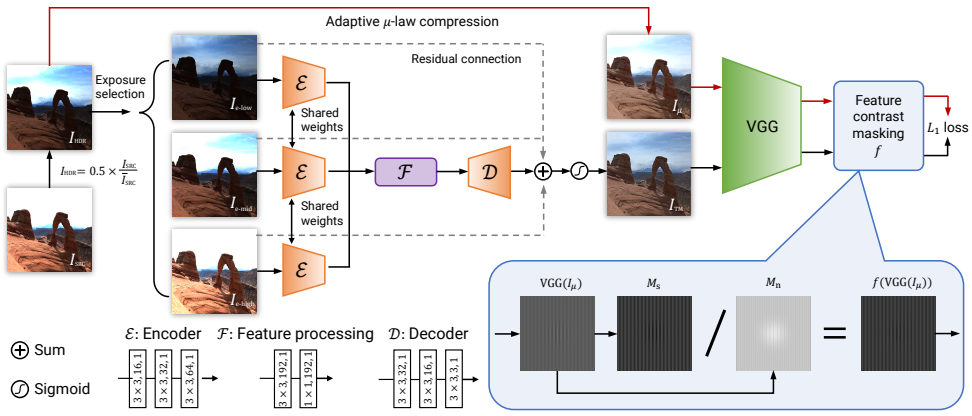


FIGURE 7.1: *Overview of our method. The input HDR image I_{SRC} is first normalized into I_{HDR}, and then three exposures $I_{e\text{-}low}$, $I_{e\text{-}mid}$, and $I_{e\text{-}high}$ are selected. A network with learnable HDR-image-specific weights is used to encode ($\mathcal{E}$) each exposure into feature space ($\mathcal{F}$), where a corrective residual is derived, that after decoding ($\mathcal{D}$) is summed up with all three exposures. The resulting image values are compressed into the range of [0,1] by a sigmoid function, and the output tone mapped image I_{TM} is obtained. The appearance and contrast of I_{TM} is directly guided by the normalized HDR I_{HDR} that is further processed into I_{μ} using an adaptive μ-law compression. Both I_{TM} and I_{μ} are transformed into feature space (VGG) and compared taking into account our novel contrast masking model f. As shown in the bottom-right inset I_{μ} is transformed into the feature space $\mathrm{VGG}(I_{\mu})$, where the ratio between feature contrast self-masking M_s and feature contrast neighborhood masking M_n define the feature contrast masking model $f(\mathrm{VGG}(I_{\mu}))$. Finally, the L_1 loss is computed between $f(\mathrm{VGG}(I_{\mu}))$ and $f(\mathrm{VGG}(I_{TM}))$. The inset in the bottom-left shows the structure of decoder ($\mathcal{E}$), feature space residual correction ($\mathcal{F}$), and decoder ($\mathcal{D}$) networks.*

7.3 Method

In this section we present our image-specific, self-supervised tone mapping network, whose structure is shown in Figure 7.1. The input HDR image is first normalized, and then decomposed into three differently exposed LDR images (Section 7.3.1). The three exposures are first transformed into feature space by an encoder with shared weights. Then, they are fused in this feature space to produce a corrective residual that is then decoded, and finally added to the three input exposures to generate the output tone mapped image (Section 7.3.2). To compute the training loss, first the normalized HDR image is processed by an adaptive μ-law compression that brings its distribution closer to typical LDR image histogram distribution (Section 7.3.3). Then, the processed HDR image along with its tone mapped counterpart go through a VGG network to derive their respective feature spaces (we employ VGG19 [Simonyan and Zisserman, 2014], which we denote VGG for brevity). Finally, we compute the L1 loss in feature space, however, instead of the standard perceptual loss between corresponding features [Simonyan and Zisserman, 2014; Johnson et al., 2016], we compute feature contrast and model contrast self-masking and neighborhood masking effects (Section 7.4).

7.3.1 Multiple Exposure Selection

HDR images are stored in linear intensity space and might feature extremely large dynamic range compared to LDR images. Moreover, the distribution of pixel intensities in the HDR image is also unbalanced, where pixels with very high intensity have large values but correspond small image regions, while low-brightness pixels usually cover larger portions of the image [Eilertsen et al., 2017; Panetta et al., 2021].

In neural network applications, to align such HDR image characteristics to those of LDR images, the logarithm of HDR pixel intensities is often applied [Eilertsen et al., 2017; Zhang et al., 2019b; Su et al., 2021]. This way a compressive response of the HVS to increasing luminance values (the Weber-Fechner law) is modeled as it is common in the tone mapping literature [Reinhard et al., 2010]. However, as we detail in Section 7.5.4, we found that our tone mapping network leads to better results when multiple differently exposed images with linear pixel intensity relation are used instead.

As HDR images are typically stored as relative positive values, before choosing the exposures, we first derive the normalized I_{HDR} image, where each pixel value is multiplied by 0.5, and divided by the mean of the original HDR image I_{SRC} intensity [Endo et al., 2017b]. To estimate the exposure range for each exposure we employ an automatic procedure originally proposed for HDR image quality evaluation [Andersson et al., 2021]. This way we obtain the low e_{low} and high e_{high} exposures[1] , and additionally we derive an intermediate third exposure as $e_{\text{mid}} = (e_{\text{low}} + e_{\text{high}})/2$.

An alternative for exposure selection is the work of [Gallo et al., 2012], which provides an optimization method for exposure metering based on the input HDR histogram. This approach guarantees that all pixels will be well represented (avoiding saturation), however, the resulting images are non-linear LDR images quantized in 8-bits. Note that in contrast to standard multi-exposure 8-bit LDR images, our three floating-point exposure selection is sufficient to represent high dynamic range information for tone mapping purposes. As we do not perform pixel quantization, the only information loss is due to clipping higher intensities into the range [0,1] with the clip() function:

$$I_{\text{e-x}} = \text{clip}(2^{e_x} I_{\text{HDR}}), \tag{7.1}$$

where $I_{\text{e-x}}$ represents one of multi-exposure images with the exposure factor e_x.

7.3.2 Tone Mapping Network

The tone mapping network is composed of an encoder $\mathcal{E}$, a feature module $\mathcal{F}$, and a decoder $\mathcal{D}$. All details on the number of layers, as well as the per-layer kernel size, channel number, and stride extent are specified in the bottom-left corner of Figure 7.1. All layers use Relu [Nair and Hinton, 2010] as the activation function except for the last layer using sigmoid. The three input exposures are used as an input to the encoder network $\mathcal{E}$ with shared weights between the exposures. The resulting feature maps are used to derive a corrective residual signal (module $\mathcal{F}$) that after decoding ($\mathcal{D}$) is added up with the three exposures. Finally, the resulting image is fitted into the range [0,1] by a sigmoid function (Figure 7.1) to produce the final tone mapped image I_{TM}. While similar residual corrections have been proposed in the past [He et al., 2016; Xu et al., 2018], our residual has two goals specific to our application. First, I_μ (Section 7.3.3) is outside the range [0,1] as shown in Figure 7.2 (left) so it needs to be compressed.

Through our perceptual loss, the corrective residual is trained such that deviations from the compressed HDR image I_μ in the tone mapped image I_{TM} are more strongly penalized for low contrast image details than for high contrast details, encouraging their compression. This way the desired appearance of the final tone mapped image I_{TM} is achieved, while an exact reconstruction of I_μ is avoided (Figure 7.3 and the

[1]Note that the goal of Andersson et al. [2021] is to get aesthetical results. We divide by two both e_{low} and e_{high}, resulting in less dark and less bright exposures, respectively. Our goal is that all relevant content is within reasonable pixel values (not too dark, not too bright).

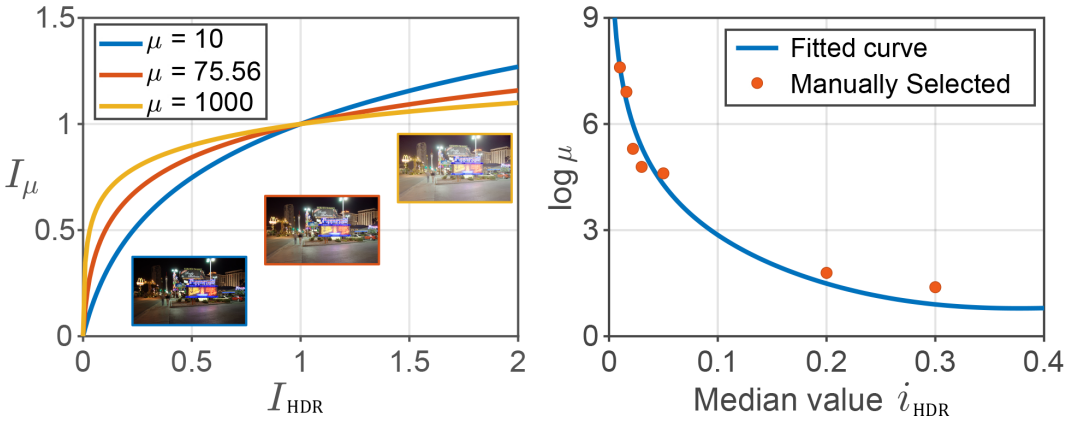


FIGURE 7.2: *Left: Adaptive μ-law compression (Eq. 7.2) for different μ values and the corresponding compressed HDR images. In this case $\mu = 75.56$ has been selected for this image using Eq. 7.4. Right: Experimentally derived μ selection as a function of HDR image median intensity i_{HDR}.*

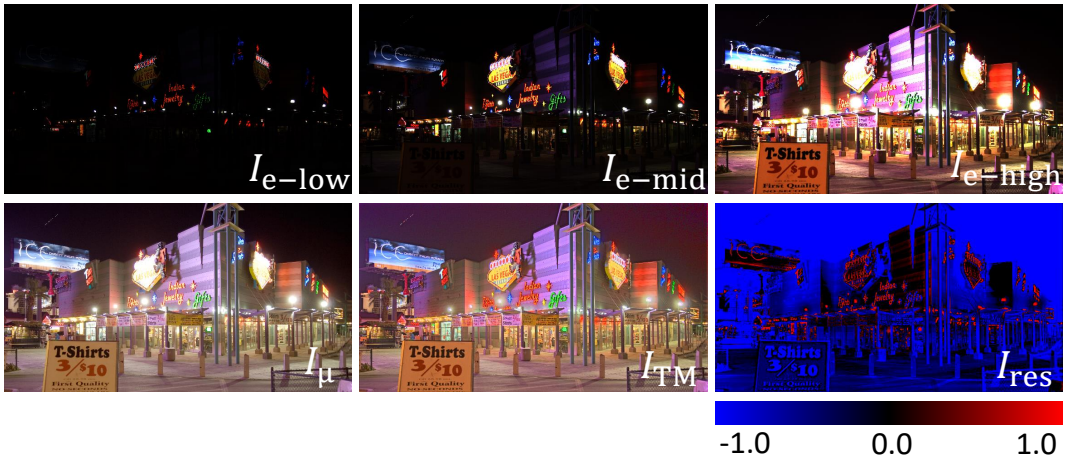


FIGURE 7.3: *Visualization of three input exposures ($I_{e\text{-}low}$, $I_{e\text{-}mid}$, and $I_{e\text{-}high}$), corresponding compressed I_μ and resulting tone mapped image I_{TM}, together with the residual image I_{res} displayed in false color. For visualization, positive (red) and negative (blue) values in the residual are normalized to the range [-1, 1]. As the sum of the three exposures falls into the range [0, 3], and the argument space of a sigmoid function spans an $[-\infty, +\infty]$ input range, large positive residual values are expected in bright regions to push output values in I_{TM} towards 1. Similarly, large negative residual values are expected in dark regions to push output values in I_{TM} towards 0.*

insets in Figure 7.2 show the relatively poor quality of I_μ). Second, during our exposure selection, darker pixels in I_{HDR} (Figure 7.1) are linearly scaled according to the three selected exposures and brighter pixels may be affected in some cases by clipping in $I_{\text{e-high}}$ or $I_{\text{e-mid}}$ (Eq. 7.1). The residual can be seen as a correction that, when added to the three exposures and after applying the sigmoid function, brings each pixel to the right intensity value while accounting for the clipping non-linearity and assuring perceptually plausible contrast processing (following the loss guided by I_μ). In Figure 7.3 we show an example of the residual together with the three exposures that it corrects.

7.3.3 Adaptive μ-Law Compression

Our image-specific tone mapping network is self-supervised by the input HDR image I_{SRC} that is transformed by the VGG network into feature space (Figure 7.1). We first adapt the range of intensity values by converting I_{SRC} into I_{HDR} as discussed in

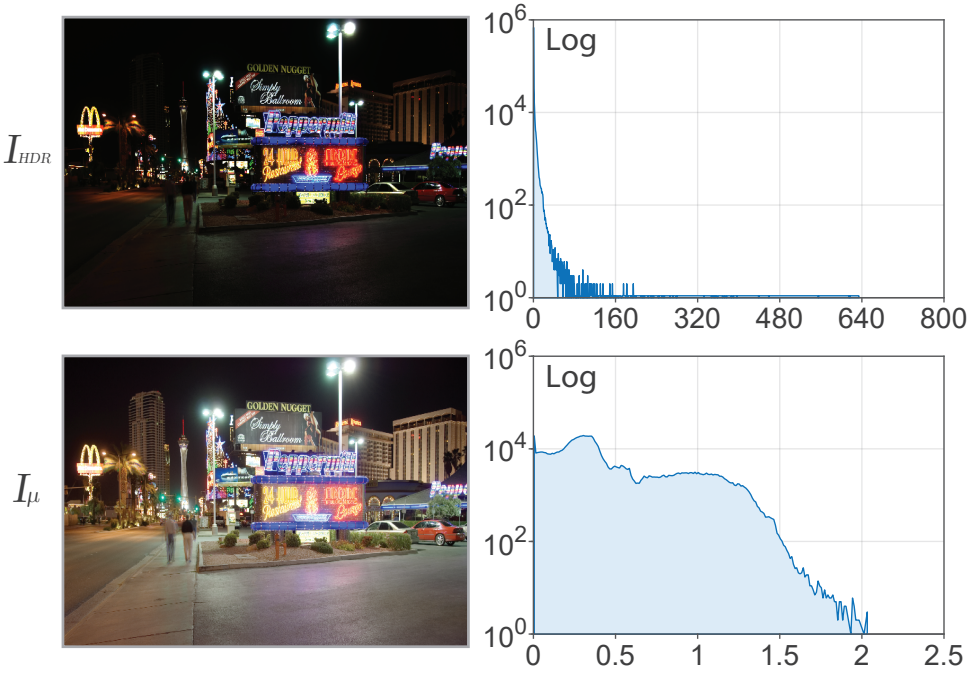


FIGURE 7.4: *Image appearance and logarithmic histogram of pixel intensities for a transformed HDR image. Top Row: I_{HDR} with the mean normalized to 0.5. Bottom row: I- derived using the adaptive μ-law compression (Eqs. 7.2 and 7.4). Notice the dramatically different scales on the x-axis.*

Section 7.3.1. As shown in Figure 7.4 (top row), I_{HDR} still has a strong skew of its histogram towards low intensity values (typical for HDR images [Eilertsen et al., 2017; Panetta et al., 2021]), which strongly differs from LDR image characteristics. Since LDR images are typically used for training VGG, we resort to a μ-law compression to correct the histogram of the image I_{HDR}:

$$I_\mu = \frac{\log(1+\mu I_{\text{HDR}})}{\log(1+\mu)}, \tag{7.2}$$

This algorithm is widely used in HDR image coding [Jinno et al., 2011] and inverse tone mapping [Wu et al., 2018; Santos et al., 2020] to re-arrange pixel intensity distribution. Figure 7.4 (bottom row) shows an example of this transformation, where image details become more visible and the long tail in the histogram is corrected.

$$\mu = \lambda_1 (i_{\text{HDR}})^{\gamma_1} + \lambda_2 (i_{\text{HDR}})^{\gamma_2}, \tag{7.3}$$

Typically a fixed μ is used to derive the transformed image I_μ, but as we show in the insets in Figure 7.2 (left), the resulting image appearance strongly depends on the selected μ value. We observe that the choice of μ also affects greatly the performance of the VGG-based loss that drives our tone mapping network, where larger μ values are required for darker HDR images. We propose an adaptive μ-law compression, where the μ value changes with the median pixel intensity value i_{HDR} that is computed for the I_{HDR} image:

$$\mu = \lambda_1 (i_{\text{HDR}})^{\gamma_1} + \lambda_2 (i_{\text{HDR}})^{\gamma_2}, \tag{7.4}$$

with fitted parameter values $\lambda_1 = 8.759$, $\gamma_1 = 2.148$, $\lambda_2 = 0.1494$, and $\gamma_2 = -2.067$. We derive this function experimentally for a number of representative HDR images featuring different appearances as well as different i_{HDR} (i.e., brightness). We use the TMQI quality metric [Yeganeh and Wang, 2012] to select the best performing

μ values, and then by visual inspection we confirm this selection. Figure 7.2 (right) shows the fitted curve based on this procedure.

7.4 Feature Contrast Masking Loss

In this section we propose the feature contrast masking (FCM) loss that guides our tone mapping network to reproduce image details and overall perceived contrast. To this end, we first model feature contrast, and then introduce self contrast masking and neighborhood masking for such feature contrast inspired by their analogues in image domain described in Section 7.2.1.

Feature contrast. While the HDR image representation I_μ, which results from the adaptive μ-law compression (Section 7.3.3), greatly facilitates its meaningful processing by the VGG network, still significant intensity differences might exist with respect to its tone mapped version I_{TM}. Such intensity differences translate into corresponding differences in the feature magnitude when I_μ and I_{TM} are transformed by the VGG network (Figure 7.1). To further reduce such feature magnitude differences we compute a per-channel local feature contrast:

$$C_p = \frac{f_\text{p} - \bar{f}_\text{p}}{|\bar{f}_\text{p}| + \epsilon}, \tag{7.5}$$

where f_p denotes the feature magnitude at pixel p, $\bar{f}_\text{p}$ is the Gaussian-filtered feature value computed for the patch $\mathcal{P}$ (centered at p), and ϵ is a small constant to avoid division by zero. We experimentally set the patch size $\mathcal{P}$ to 13×13 pixels. To compute this contrast, the feature difference with respect to its local neighborhood is first computed and then normalized. This normalization enables further reduction of the impact of the differences in absolute feature magnitudes between I_μ and I_{TM}. Note that Eq. 7.5 is aligned with contrast definitions for images that also use Laplacian and Gaussian filter responses in the numerator and denominator, respectively [Peli, 1990]. More complex filter banks such as Laplacian pyramids [Mantiuk et al., 2021], cortex transforms [Daly, 1992], wavelets [Daly et al., 2000], or discrete cosine transforms [Watson, 1993] are often used for advanced contrast processing operations.

Feature contrast self-masking. An important property of contrast perception is a higher sensitivity to contrast discrimination for lower contrasts than for suprathreshold contrasts [Kingdom and Whittle, 1996], this is termed contrast self-masking [Daly et al., 2000]. In the context of tone mapping this means that even small contrast changes can be perceived in low-contrast image regions, which are often neglected in global tone mapping operators [Mantiuk et al., 2015b]. Conversely, for larger image contrast, even strong contrast compression might remain undetected. Many tone mapping operators take advantage of this effect [Durand and Dorsey, 2002; Fattal et al., 2002; Mantiuk et al., 2006]. Different from conventional methods, we do not model contrast self-masking in the image contrast domain, instead we model it in the feature contrast domain C defined in Eq. 7.5 (we skip the pixel index p for brevity):

$$M_\text{s} = \text{sign}(C)|C|^\alpha, \tag{7.6}$$

where M_s denotes a non-linear response to the feature contrast magnitude controlled by the compressive power factor α. The function $\text{sign}(x) = \frac{x}{|x|+\epsilon}$ preserves the feature



FIGURE 7.5: *Feature contrast masking visualization of VGG feature maps (1st layer, 18th channel). From left to right: tone mapped image I_{TM}, original VGG feature map* VGG(I_μ)*, feature contrast self-masking M_s, neighborhood masking M_n, and final f(VGG(I_μ)) masking terms.*

contrast polarity in M_s. We experimentally set $\alpha_{I_\mu} = 0.5$ when processing I_μ while we keep $\alpha_{I_{TM}} = 1.0$ for I_{TM}, allowing for visually pleasant low-contrast detail enhancement. Figure 7.5 shows a feature map VGG(I_μ) for a selected VGG channel, as well as the response M_s to feature contrast C. As shown in this figure, M_s vividly responds for low-contrast details in the sky and rocks that are instead strongly suppressed in VGG(I_μ). Note that when VGG(I_μ) is directly used in the perceptual VGG loss driving the tone mapping operation $\mathcal{L}_{VGG} = ||\text{VGG}(I_\mu) - \text{VGG}(I_{TM})||_1$ [Johnson et al., 2016], such details are likely to be neglected in the resulting I_{TM} due to the low penalty in the loss.

Feature contrast neighborhood masking. Inspired by successful applications of neighborhood masking, as discussed in Section 7.2.1, we also model feature contrast neighborhood masking. Image contrast neighborhood masking is performed selectively for different spatial frequency bands; this requires image decomposition by a filter bank [Daly et al., 2000; Lubin, 1995]. We approximate this process by modeling feature contrast neighborhood masking per channel, where features with similar frequency characteristics are naturally isolated. Our goal is to suppress the magnitude of M_s when there is a high variation of feature magnitudes f_p in the local neighboring of pixel p that we measure as:

$$M_n = \frac{\sigma_b}{|\mu_b| + \epsilon}, \tag{7.7}$$

where μ_b and σ_b denote the mean and standard deviation of feature magnitude f_p in the patch $\mathcal{P}$ that is centered at pixel p. Again, we experimentally set the patch $\mathcal{P}$ size to 13×13 pixels. Finally, our feature contrast masking is calculated as the ratio of self- and neighborhood masking:

$$f(\text{VGG}(I)) = \frac{M_s}{1 + M_n} \tag{7.8}$$

As can be seen in Figure 7.5, M_n vividly responds in the regions with high local feature variation as seen in VGG(I_μ), so that the final feature contrast masking measure f(VGG(I_μ)) is strongly suppressed in such regions. In particular, this means that in regions with strong image contrast, such as the horizon line, f(VGG(I_μ)) is relatively much smaller with respect to the orginal VGG(I_μ) feature magnitudes. Consequently, when including f(VGG(I_μ)) into the loss computation that drives the tone mapping operation, the penalty for any distortion of such high contrast is much smaller than in the perceptual VGG loss $\mathcal{L}_{VGG}$ that directly employs VGG(I_μ). Effectively, this gives the tone mapping network more freedom for compressing image contrast in such regions. We compute our feature contrast masking (FCM) loss $\mathcal{L}_{FCM}$ as the L_1 loss between the masked feature maps $f()$ for the transformed input

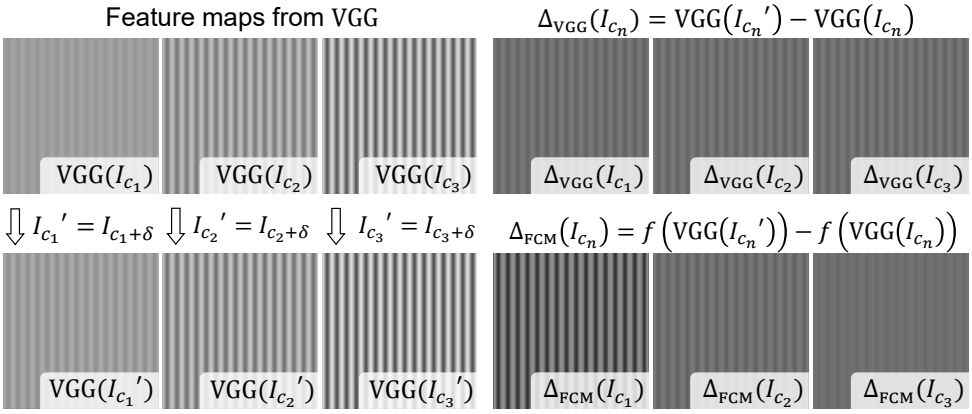


FIGURE 7.6: *Left: VGG feature maps (1st layer, 2nd channel) for three input sinusoids of increasing image contrast ($c_1 < c_2 < c_3$) (upper row) that are distorted by further increasing their amplitudes by the same factor δ (bottom row). Right: the differences Δ_{VGG} between the VGG feature maps, which are computed for each sinusoid and its distorted version, show a weak dependence to the input sinusoid contrast (upper row), while the corresponding differences Δ_{FCM} resulting from our feature contrast masking model penalize more strongly the distortion for smaller contrast sinusoids.*

HDR image I_μ and the output tone mapped image I_{TM}:

$$\mathcal{L}_{FCM} = \left|\left|f(\text{VGG}(I_\mu)) - f(\text{VGG}(I_{TM}))\right|\right|_1 \tag{7.9}$$

To further illustrate the behavior of our loss, In Figure 7.6 we consider simple sinusoid patterns with three different contrasts (c_1, c_2, and c_3). The corresponding feature maps of the VGG network are shown in the top-left image row. We distort each sinusoid by increasing their respective amplitudes by the same factor δ and their corresponding feature maps are shown in the bottom-left row.

We compare the feature map difference Δ_{VGG} used in the VGG loss $\mathcal{L}_{VGG}$ [Johnson et al., 2016] (top-right row) and the corresponding Δ_{FCM} used in our FCM loss $\mathcal{L}_{FCM}$ (bottom-right row). As can be seen, Δ_{FCM} yields a higher penalty when the distortion δ is added to the lowest contrast pattern. This forces our tone mapping, driven by $\mathcal{L}_{FCM}$, to reproduce image details in low-contrast areas. The VGG loss $\mathcal{L}_{VGG}$ remains similar irrespectively on the input sinusoid contrast. This puts equal pressure on the tone mapping network to reproduce image details for large contrast regions that cannot be perceived, and low-contrast regions where they are clearly visible.

Finally, we perform color correction in I_{TM} following Tumblin and Turk [Tumblin and Turk, 1999]:

$$C_{out}^{\{R,G,B\}} = L_{out}\left(\frac{C_{in}^{\{R,G,B\}}}{L_{in}}\right)^s \tag{7.10}$$

$$L_{in/out} = 0.2126 * R + 0.7152 * G + 0.0722 * B \tag{7.11}$$

where L_{in} and L_{out} are the luminance of the input HDR image I_{HDR} and the output tone mapped image I_{TM} respectively, and s is the saturation control parameter (we use $s = 0.6$ in all examples). Since we apply color correction as a post-processing operation, other algorithms [Mantiuk et al., 2009; Artusi et al., 2018] could be also applied.

7.5 Results and Comparisons

In this section we first describe the implementation details of our approach. Then, we provide objective and subjective comparisons including both traditional methods and state-of-the-art learning-based approaches. Finally, we perform an ablation study showing how each of the components of our approach contributes to achieving the final quality of our results.

implementation We adopt an online training strategy and train a model for each HDR image at test time. Our model is implemented on TensorFlow 1.10 and the results reported in the paper are computed with a RTX 8000 GPU. We use the Adam optimizer with an initial learning rate of 2×10^{-4} and an exponential decay factor of 0.9 every ten epochs. The training converges after 400 epochs, which translates into around 583 ± 6.62 seconds, for an image resolution of 768×384. We fix a single set of parameters for all our experiments and results. As discussed in the previous section, we set $\mathcal{P} = 13 \times 13$, $\alpha_{I_\mu} = 0.5$, and $\alpha_{I_{\text{TM}}} = 1.0$. We compute our loss function based on the first three layers of VGG.

7.5.1 Competitors and Dataset

We include in our comparisons fourteen tone mapping operators including ten traditional methods which for simplicity we refer to as: Mantiuk [Mantiuk et al., 2006], Shan [Shan et al., 2009], Durand [Durand and Dorsey, 2002], Drago [Drago et al., 2003], Mertens [Mertens et al., 2007], Reinhard [Reinhard et al., 2002], Ma [Ma et al., 2015], Liang [Liang et al., 2018], Shibata [Shibata et al., 2016], and Li [Li et al., 2018]; and four recent learning-based methods: Guo [Guo and Jiang, 2021], Zhang [Zhang et al., 2021], DeepTMO [Rana et al., 2019] and TMO-Net [Panetta et al., 2021]. We use the publicly available implementations of these methods or if not available, their implementation in HDRToolBox [Banterle et al., 2017]. For the case of DeepTMO [Rana et al., 2019] and TMO-Net [Panetta et al., 2021], we were not able to access their implementations, therefore we directly use the results provided in their works for comparisons.

We use a large test set of 275 images gathered from the Fairchild dataset [Fairchild, 2007], Poly Haven[2], the Laval Indoor HDR Database [Gardner et al., 2017], and the LVZ-HDR benchmark dataset [Panetta et al., 2021], which cover various indoors, outdoors, bright and dark scenes.

7.5.2 Objective Evaluation

For objective comparisons we adopt as metrics the Tone Mapped Image Quality Index (TMQI) [Yeganeh and Wang, 2012], the Blind Tone Mapped Quality Index (BTMQI) [Gu et al., 2016] and the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [Mittal et al., 2012]. The former two metrics are widely used for evaluating tone mapping operators [Guo and Jiang, 2021; Zhang and Aydın, 2021; Liang et al., 2018], while the latter is typically used as a blind metric for evaluating contrast [Jiang et al., 2019; Lim et al., 2017b; Song et al., 2020b]. For completeness, we briefly discuss here these metrics.

TMQI is a full-reference tone mapping image quality metric, which consists of two main terms assessing the structural fidelity (TMQI_S) and the naturalness (TMQI_N) of the tone mapped image. BTMQI is a no-reference tone mapping metric, which is

[2] https://polyhaven.com/hdris

Table 7.1: *Mean and standard deviation for TMQI (including* TMQI_S *and* TMQI_N*), BTMQI and BRISQUE computed for the 275 images in our test set, and average computing time for an image with 768 × 384 pixels resolution.*

Methods	TMQI (↑)	TMQI_S (↑)	TMQI_N (↑)	BTMQI (↓)	BRISQUE (↓)	time (s)
Ours	**0.9248 ± 0.0432**	**0.8938 ± 0.0582**	**0.7062 ± 0.2369**	**2.9065 ± 1.0481**	**21.2208 ± 8.5981**	583.26
Guo	0.8883 ± 0.0354	0.8166 ± 0.0762	0.5975 ± 0.2002	3.7110 ± 1.0239	23.9528 ± 8.2566	0.0283
Zhang	0.8767 ± 0.0600	0.8343 ± 0.0872	0.5118 ± 0.2686	3.7440 ± 1.4209	22.0520 ± 8.9580	0.0021
Liang	0.8964 ± 0.0490	0.8534 ± 0.0737	0.5194 ± 0.2605	3.5333 ± 1.0915	25.1433 ± 8.5510	1.1266
Shibata	0.7689 ± 0.0506	0.7498 ± 0.0858	0.1203 ± 0.1719	4.3470 ± 0.7469	32.2273 ± 8.9383	11.655
Li	0.8480 ± 0.0612	0.8301 ± 0.0743	0.3716 ± 0.2939	4.3670 ± 1.1621	24.0348 ± 8.5411	3.5321
Shan	0.8301 ± 0.0732	0.7458 ± 0.1448	0.4174 ± 0.2792	4.0685 ± 0.9839	22.8230 ± 8.5913	50.287
Durand	0.8719 ± 0.0669	0.8375 ± 0.0989	0.4824 ± 0.2685	3.6537 ± 1.1016	22.0285 ± 8.4576	0.0665
Drago	0.8794 ± 0.0537	0.8600 ±0.0803	0.4840 ± 0.2558	4.0213 ± 1.2733	22.6600 ± 8.2405	0.0974
Mertens	0.8425 ± 0.0717	0.8403 ± 0.0923	0.3373 ± 0.2848	4.8981 ± 1.6026	23.2085 ± 8.8589	1.4282
Reinhard	0.8506 ± 0.0533	0.8176 ± 0.0864	0.3903 ± 0.2347	4.2774 ± 1.4580	25.5317 ± 7.7396	0.2393
Mantiuk	0.8529 ± 0.0753	0.8903 ± 0.0849	0.3238 ± 0.3050	4.5339 ± 1.4086	21.2943 ± 8.8302	1.8733
Ma	0.8782 ± 0.0698	0.8644 ± 0.1043	0.4782 ± 0.2458	3.6444 ± 1.6857	24.9172 ± 8.0956	3225.4

composed of three terms accounting for entropy (richness of information), naturalness, and presence of structural details. BRISQUE is a well-known no-reference image quality metric based on natural scene statistics that quantifies the naturalness of an image and considers distortions such as noise, ringing, blur, or blocking artifacts.

We show in Table 7.1 the results of these objective metrics together with computation times for eleven of the tested methods with our testing set, while Table 7.2 shows the results for DeepTMO and TMO-Net, for which we use their provided test sets and results. We include examples of our results compared to the seven best performing methods in Figure 7.7, and compared to DeepTMO and TMO-Net in Figure 7.8. Additionally, we show in Table 7.3: time resolution running times for our method with Standard Definition and High Definition images. Since we adopt an online training strategy, our network has to be optimized for every different HDR image, resulting in moderately high running times.

Our proposed approach outperforms previous methods in different aspects. We discuss now more in detail results for the best performing approaches. In general, all approaches except DeepTMO yield low TMQI_N values, indicating that they fall short in preserving the naturalness of the image. For the case of DeepTMO, the low performance in the BRISQUE metric indicates that the tone mapped images do not preserve natural image statistics (Figure 7.8, first three columns). We can also see that TMO-Net additionally produces saturated results in the brightest regions of the image (Figure 7.8, last three columns). Looking into Figure 7.7 we can observe that Guo over-enhances dark regions, sometimes resulting in heavy artifacts (*garage* scene), while Zhang tends to produce overly dark results in regions with low brightness. This is in agreement with a relatively low score in structural fidelity TMQI_S, indicating that the tone mapped images differ from the HDR in terms of conveyed structural information. While Liang and Ma perform well in terms of TMQI_S, they tend to produce under-saturated results with lower contrast (e.g., *garage* and *store* scenes), which is in agreement with a relatively low performance in the BRISQUE metric. Drago and Durand also perform well in terms of TMQI_S. However, we can see that Drago produces blurry results and fails to reproduce fine details, such as the floor tiling in the *station* scene or the highlights of the bottles in the *store* scene. Durand presents good scores in terms of BRISQUE score which means the tone mapped images do align with natural image statistics, however in some cases it over-enhances contrast, producing artifacts such as those around the car windows in the *garage* scene or those in the luminous numbers of the *station* display sign. Mantiuk has the lowest TMQI_N of the methods included in Figure 7.7. This operator tends to produce very

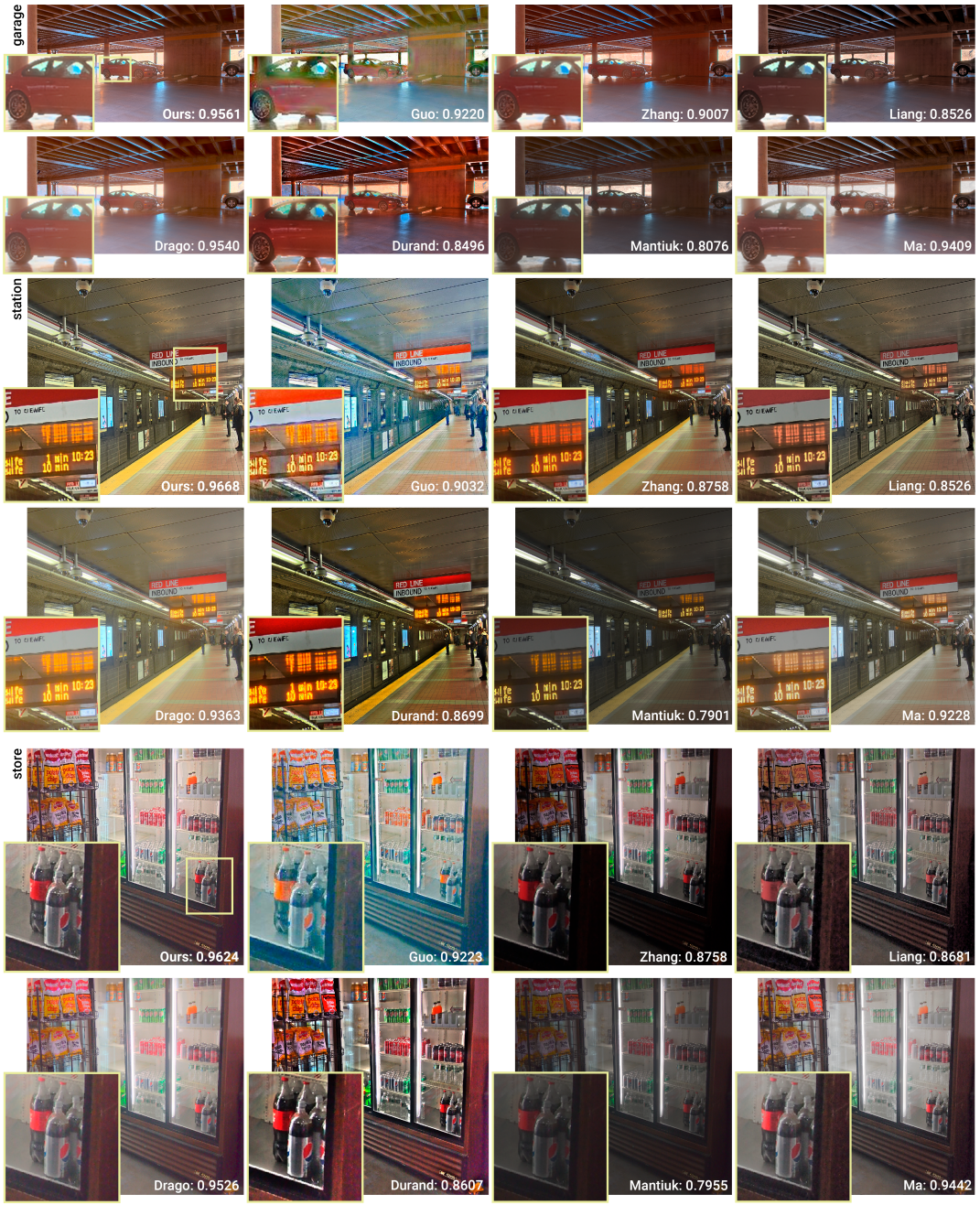


FIGURE 7.7: *Visual comparisons for the best performing methods and their TMQI scores. Overall, our results achieve good contrast reproduction while preserving the details (highlights of the bottles in* store*, floor in* station*) and avoiding visual artifacts (display sign in* station*, car edges and windows in* garage*). Please refer to the text for an in-depth discussion of the observed differences.*

dark images with low contrast. Our method outperforms existing approaches for all tested metrics, exhibiting a good contrast reproduction while preserving the details present in the HDR images. Our results also produce natural images without visible artifacts.

7.5.3 Subjective Evaluation

To further validate the performance of our approach we additionally performed a subjective study. We included the six most commonly used, best-performing methods



FIGURE 7.8: *Visual comparisons for DeepTMO [Rana et al., 2019] and TMO-Net [Panetta et al., 2021] on images from their respective test set. DeepTMO tends to produce over-enhanced contrast and saturated results that do not preserve natural image statistics. TMO-Net can not handle some challenging scenarios, producing saturated results in very bright regions and overly dark pixels in dark regions.*

Table 7.2: *Mean and standard deviation for TMQI (including* TMQI_S *and* TMQI_N*), BTMQI and BRISQUE computed for the DeepTMO [Rana et al., 2019] and TMO-net [Panetta et al., 2021] test sets. The former contains 100 images from the Fairchild dataset [Fairchild, 2007] while the latter contains 457 captured images from their own dataset.*

Methods	TMQI_Q (↑)	TMQI_S (↑)	TMQI_N (↑)	BTMQI (↓)	BRISQUE (↓)
Ours	**0.9106 ± 0.0511**	**0.8987 ± 0.0664**	**0.6052 ± 0.2807**	**3.3420 ± 1.0686**	**19.5406 ± 9.1347**
DeepTMO	0.9052 ± 0.0619	0.8810 ± 0.0717	0.6015 ± 0.2679	3.4230 ± 1.1502	27.5489 ± 7.6865
Ours	**0.9073 ± 0.0541**	**0.8939 ± 0.0551**	**0.6020 ± 0.3126**	**3.3069 ± 1.2266**	**23.1010 ± 8.5232**
TMO-Net	0.8609 ± 0.0594	0.8066 ± 0.0825	0.4723 ± 0.2871	3.9633 ± 1.2477	26.6078 ± 8.1895

in terms of average TMQI score according to our previous objective evaluation, in particular: Guo [Guo and Jiang, 2021], Zhang [Zhang et al., 2021], Liang [Liang et al., 2018], Drago [Drago et al., 2003], Durand [Durand and Dorsey, 2002], and ours. The study was approved by the Ethical Review Board of the Computer Sciences department at Saarland University, and participants provided written consent for participating in the study. A total of twelve participants (aged 25 to 34 years old with normal or corrected-to-normal vision) voluntarily took part in the study. We included fifteen scenes covering different scenarios and, for each scene, we showed the six tone mapped images at a random order. We asked the participants to rank them from 1 (preferred) to 6 (least preferred). Before the study, participants underwent a simple training with a few images showing common artifacts in tone mapped images such as over-saturated colors, over- or under-exposure, contrast loss and contrast over-enhancement. The study was conducted in a room with natural illumination and participants sat at a distance of 0.5 meters from the display. The images were displayed in a DELL UltraSharp U2421E monitor (1920 × 1200 resolution, 60 HZ refresh rate).

We show in Figure 7.9 the preference rankings for each method, aggregated for all participants and scenes. We use Kruskal-Wallis (non-parametric extension of ANOVA) for analyzing the rankings, since these do not follow a normal distribution [Jarabo et al., 2014; Rubinstein et al., 2010]. We then compute post-hocs using pairwise

Table 7.3: *Computation times for our method with Standard Definition (SD, 720 × 480) and High Definition (HD, 1080 × 720) images for our baseline method and our two proposed techniques for speeding-up the computation: depth-wise convolutions [Chollet, 2017] and progressive training.*

Methods	Time-SD (s)	Time-HD (s)
Ours	654.72	1245.7
depth-conv	223.74	486.57
depth-conv & pro-training	126.28	251.36

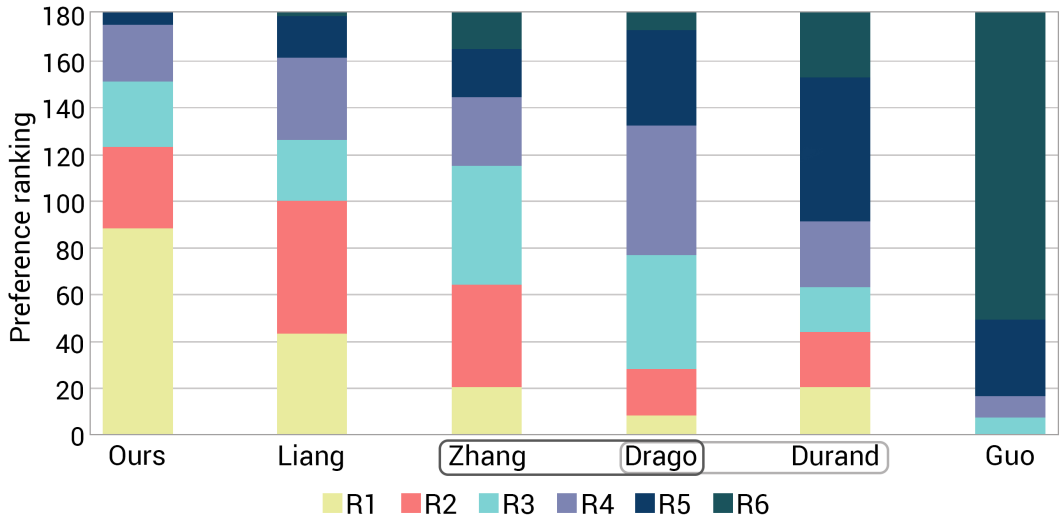


FIGURE 7.9: *Preference rankings for each method aggregated across the twelve participants and fifteen scenes. Different colors indicate the received rankings (from R1 to R6). Pairwise comparisons between methods reveal that preference rankings are significantly different, except those methods marked in the same set (gray squares), which are statistically indistinguishable.*

TABLE 7.4: *Mean and standard deviation for TMQI (including* TMQI_S *and* TMQI_N*), BTMQI and BRISQUE computed for different variations of components in our pipeline.*

Input	$I_{\text{HDR}} \mapsto I_\mu$	Loss	TMQI (↑)	TMQI_S (↑)	TMQI_N (↑)	BTMQI (↓)	BRISQUE(↓)
log	log	$\mathcal{L}_{\text{VGG}}$	0.5785 ± 0.1073	0.3759± 0.1965	0.0066± 0.0578	6.6324 ± 0.5731	42.5287 ± 7.4740
MEI	linear	$\mathcal{L}_{\text{FCM}}$	0.8776 ± 0.0610	0.8763± 0.0966	0.4555 ± 0.2642	3.9174 ± 1.4472	21.4104 ± 8.9001
MEI	Ada μ	$\mathcal{L}_{\text{VGG}}$	0.9178 ± 0.0462	0.8913 ± 0.0607	0.6596 ± 0.2470	3.1934 ± 1.1220	22.1773 ± 8.6294
linear	Ada μ	$\mathcal{L}_{\text{FCM}}$	0.8522 ± 0.0793	0.8166± 0.1206	0.4173 ± 0.2902	4.5342± 1.6754	30.2186 ± 13.7742
log	Ada μ	$\mathcal{L}_{\text{FCM}}$	0.8166 ± 0.0563	0.8921 ± 0.0774	0.0844 ± 0.0741	5.0949 ± 1.1779	22.4499 ± 8.3301
MEI	Ada μ	$\mathcal{L}_{\text{FCM}}$	**0.9248 ± 0.0432**	**0.8938 ± 0.0582**	**0.7062 ± 0.2369**	**2.9065 ± 1.0481**	**21.2208 ± 8.5981**

Kruskal-Wallis tests adjusted by Bonferroni correction for multiple test. Results reveal a statistically significant difference in the rankings for the different methods ($p < 0.001$), with our approach being ranked significantly higher than all others.

7.5.4 Ablations

In this section we evaluate the importance of each of the components in our method for achieving the final quality of the results. Table 7.4 shows the results of the objective metrics for different combinations of (i) input: linear HDR [Rana et al., 2019], log HDR [Zhang et al., 2019b; Su et al., 2021] or our multiple exposure inputs (*MEI*); (ii) HDR compression algorithm for computing the loss: linear, log or our adaptive μ-law compression (*Ada* μ); and (iii) loss function: $\mathcal{L}_{\text{VGG}}$ or our $\mathcal{L}_{\text{FCM}}$. Figure 7.10 shows the corresponding visual results.

In general, we can see that our loss $\mathcal{L}_{\text{FCM}}$, which considers masking effects, plays an important role in emphasizing the local contrast, especially in large contrast regions, such as the clouds in the sky. Compared with our MEI, the logarithm input leads to overall darker results (brightness distortion), and the linear input can cause overexposure with missing information in the highlight regions. Finally, we can see that the Ada μ-law compression is important for overall image contrast, when linear or logarithmic transformations are applied instead, the resulting images are flatter in terms of contrast.

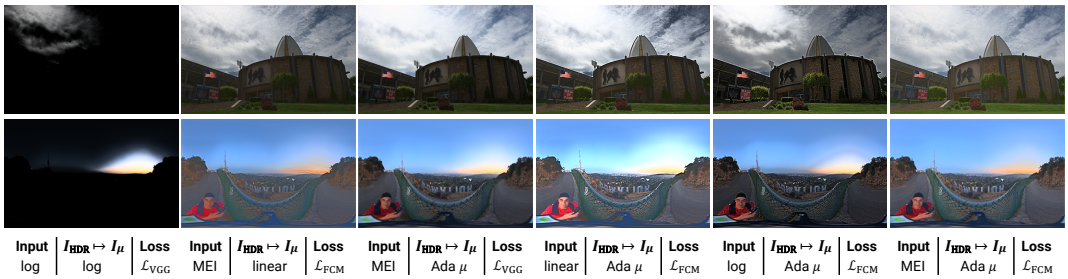


FIGURE 7.10: *Example visualizations of our ablation study. Better contrast, brightness and detail reproduction is achieved with our full pipeline including multiple exposure inputs* (MEI)*, adaptive μ-law compression* (Ada μ)*, and $\mathcal{L}_{FCM}$ loss.*



FIGURE 7.11: *Example exposure of the original HDR image (left) and our tone mapped result (right). The inset shows a failure case in which a soft halo appears around the edge of the mountain.*

7.6 Conclusions and Limitations

In this work we propose an image-specific self-supervised tone mapping approach that leads to consistent high-quality results for a large variety of HDR scenes. Previous learning-based approaches present two main limitations: (i) the variety of HDR content they can adequately tone map is limited by the images used during training, and (ii) most of these approaches are supervised, i.e., they need HDR-LDR image pairs for training. These LDR images are obtained either from tone mapped results from previous methods, or manually tone mapped images. Therefore, the quality of the learned tone mapper is limited to that of the selected training pairs. In contrast, our learned tone mapping operator is guided by a novel feature contrast masking loss that allows to represent in feature space important image contrast perception characteristics of the Human Visual System.

The loss gives the network more freedom for compressing higher contrast while enhancing weak contrast, as it is often desirable for high quality HDR scene reproduction and an overall pleasant appearance, all in the context of local image content as modeled by neighborhood masking.

In some rare cases our method may produce soft halos at high contrast edges as shown in Figure 7.11. In future work we would like to experiment with edge stopping filters while deriving feature contrast and neighborhood masking for avoiding this issue. Nevertheless, these soft artifacts are not present in most of our results, and whenever present, they are not obviously visible as confirmed by both objective and subjective evaluations. As discussed in [Trentacoste et al., 2012], unsharp masking or weak counter-shading effects, similar to these soft halos, may be an effective way of enhancing perceived image contrast due to the Cornsweet illusion and are often employed for image enhancement.

Chapter 8

Conclusion

This thesis introduces several deep-learning-based HDR methods, encompassing the complete workflow of reconstruction, generation, and display. Chapter 3 presents an implicit architecture (TAF) for reconstructing all-in-focus HDR images from LDR inputs with arbitrary combinations of aperture size, focus distance, and exposure time, enabling flexible post-editing. This approach uniquely integrates HDR fusion, depth-from-defocus, and varying-aperture photography into a unified framework. Chapter 4 extends the scene representation to incorporate camera pose variations by integrating the thin-lens model into 3D Gaussian Splatting (CinematicGS). This enables the reconstruction of all-in-focus HDR radiance fields from multi-view LDR defocused inputs, while also supporting real-time rendering effects such as depth-of-field and exposure changes. Chapter 5 introduces GlowGAN, an HDR generation model that learns the HDR distribution solely from in-the-wild LDR images, addressing the scarcity of HDR data. The pretrained model can also serve as a prior to guide LDR-to-HDR conversion (inverse tone mapping), demonstrating exceptional hallucination capabilities. Chapter 6 introduces LEDiff, a latent exposure fusion model that fine-tunes a pretrained LDR Stable Diffusion model, addressing the limitations of GAN-based methods, including class-specific training requirements and limited diversity. Additionally, LEDiff functions as a plug-and-play module, seamlessly integrating into any Stable Diffusion-based model to enhance HDR capabilities, enables tasks such as panoramic HDR generation and image-to-HDR video synthesis. Chapter 7, inspired by visual masking, introduces a self-supervised tone mapping approach, SelfTMO. Specifically, it designs a feature contrast masking loss in feature space as a self-supervised guidance signal, enabling a lightweight CNN-based model to effectively map HDR to LDR while preserving contrast as accurately as possible.

Building upon the contributions of this thesis, the following sections analyze key findings and propose future research directions for each chapter, maintaining a focused exploration of HDR content reconstruction, generation, and display technologies.

8.1 HDR Content Reconstruction From Photographs

Capturing high-quality images in a single exposure is fundamentally limited by the physical constraints of camera lenses and sensors, often resulting in saturation, defocus blur, and other degradations. Traditional fusion-based techniques, such as multi-exposure [Debevec and Malik, 2023; Mertens et al., 2007; Li et al., 2013; Zhang and Levine, 2016] and multi-focus fusion [Li et al., 2013; Zhang and Levine, 2016; Liu et al., 2017; Suwajanakorn et al., 2015; Si et al., 2023; Bouzos et al., 2019], attempt to address these issues by combining image sequences, but typically handle only one type of artifact at a time. In practice, such artifacts often occur simultaneously, and users often need to adjust focus or exposure after capture.

To address these challenges, we introduce TAF (Time-Aperture-Exposure Time) in Chapter 3, a 2D reconstruction framework that leverages implicit neural representations to jointly model exposure, focus, and aperture variations in image stacks. The system integrates flow estimation, depth prediction, and HDR reconstruction within a unified pipeline, allowing the recovery of high-quality all-in-focus HDR images while supporting flexible editing. The use of a differentiable lens model and implicit tone mapping further enables photorealistic blur simulation and faithful LDR rendering.

While TAF is effective for 2D stacks, practical image capturing often involves varying camera poses, leading to multi-view observations. To extend HDR reconstruction into the 3D domain, we propose CinematicGS in Chapter 4, which adopts 3D Gaussian Splatting [Kerbl et al., 2023] as the scene representation due to its ability to achieve both high-quality and efficient rendering. In addition, since both the scene and the aperture are modeled as Gaussians, defocus blur can be simulated via closed-form convolution, significantly improving the efficiency of depth-of-field rendering over traditional accumulation buffur based methods.

Despite the achievements presented in TAF and CinematicGS on HDR content reconstruction, there remains substantial room for promising future work.

For the 2D scenario discussed in TAF, beyond the aspects such as ISO and motion blur covered in Chapter 3, we believe that future directions, especially with recent technological advances, may focus on two main areas: efficiency and accuracy.

First, since the entire pipeline is built upon Implicit Neural Representations (INRs), despite their success in capturing high-frequency details, often rely on MLPs, leading to long training times, high memory consumption, and slow decoding speeds. These issues limit their practicality, especially in resource constrained devices.

Although introducing a hash grid-based structure (NGP) [Müller et al., 2022] as mentioned in Chapter 3 can provide some improvements, it is not a fundamental solution, and better alternatives may emerge as the technology evolves. One promising direction is the use of 2D Gaussian Splatting (2DGS) as a representation method [Huang et al., 2024]. Originally proposed for extracting high-quality 3D scene meshes, 2DGS has recently shown significant improvements in tasks such as image compression [Zhang et al., 2024a] and image super-resolution [Hu et al., 2025], particularly in terms of rendering efficiency. Additionally, as discussed in CinematicGS, using Gaussians as image representations enables a closed-form solution for defocus simulation, which further enhances computational efficiency.

To improve accuracy, we first focus on the flow predictor. Inspired by the concept of optical flow, the flow predictor is designed to correct misalignments between images caused by lens breathing effects. However, in our current setup, the flow predictor is only supervised by the per-image loss, which may limit its performance. Drawing inspiration from optical flow estimation, incorporating a warping loss [Brox et al., 2004; Wang et al., 2022b] that enforces consistency between neighboring images could further enhance the accuracy of the predictions.

On the other hand, in terms of defocus simulation, TAF adopts a spatially-varying convolution approach using isotropic disk kernels. While effective in producing smooth blur, this design can cause foreground and background intensities to blend at depth boundaries, leading to intensity leakage [Hazirbas et al., 2019; Zhao et al., 2022; Sheng et al., 2024]. Hazirbas et al. [2019] mitigate intensity leakage by leveraging end-to-end learning on focal stacks, enabling the network to learn focus-aware features that distinguish foreground and background. Building on this idea, Zhao et al. [2022] address the blending artifacts at defocus boundaries by explicitly separating focused and defocused regions and guiding the deblurring process with self-referenced

features. Most recently, Dr. Bokeh [Sheng et al., 2024] further improves boundary rendering by introducing differentiable occlusion-aware synthesis, which explicitly separates foreground and background contributions during aperture-based blur generation, effectively eliminating depth-related color bleeding. These methods represent promising directions for addressing the problem. Considering the characteristics of our approach, we believe that incorporating the concept of bilateral filtering [Durand and Dorsey, 2002] by introducing spatially-aware weights into the convolution process can serve as an edge-preserving strategy that helps mitigate intensity leakage to some extent.

For the 3D scene reconstruction, beyond addressing limitations related to aperture simulation accuracy and noise in dark regions discussed in Chapter 4, there are several other directions that deserve further exploration and refinement.

CinematicGS relies on relatively dense input, similar to that required in light field rendering, which inevitably increases the burden of data acquisition. Reducing the number of input images while maintaining reconstruction quality represents a highly promising direction for future research. We identify several key challenges in addressing sparse-input scenarios. The first concerns the generation of the initial point cloud. In the current 3D Gaussian Splatting (3DGS) pipeline, the initial point cloud is typically obtained using SfM [Schonberger and Frahm, 2016]. However, reducing the number of input images inevitably degrades the quality and completeness of the reconstructed point cloud. Recent neural approaches offer promising alternatives. Wang et al. [2024c] propose a new paradigm, DUSt3R, for dense 3D reconstruction, leveraging a sparse voxel-grid representation trained on indoor RGB video data. Their method enables the recovery of high-quality dense point clouds and mesh representations from sparse RGB inputs, significantly improving reconstruction completeness and robustness. It further offers strong real-time performance and generalization to unseen scenes, overcoming the typical failure modes of COLMAP in low-texture or low-overlap conditions. The following work [Tang et al., 2024] introduces a multi-view decoder architecture that jointly processes multiple unposed RGB images in a single forward pass, eliminating the need for pairwise matching and global optimization required by DUSt3R. By incorporating cross-reference-view attention blocks, it enhances robustness to reference view selection and significantly improves reconstruction accuracy and efficiency. In addition to enhancing the quality of the initial point cloud, another effective strategy involves reinforcing internal constraints among multi-view images. SCGaussian [Peng et al., 2024] employs matching priors [Shen et al., 2024] to identify correspondences between rays from different viewpoints, effectively constraining the positions of Gaussian primitives to converge at surface points where these rays intersect and leading to improved reconstruction quality. Another valuable direction for improving quality under sparse input conditions is to integrate depth regularization into the 3D Gaussian Splatting pipeline, as demonstrated by DepthGS [Chung et al., 2024].

Beyond the challenge of sparse-view inputs, CinematicGS assumes that all LDR images can be modeled using a single global tone mapper. While this is generally a reasonable assumption, it cannot capture more complex or view-dependent image processing. A promising idea we can draw inspiration from is the Differentiable 3D Bilateral Grid proposed by Wang et al. [2024d], which extends the bilateral grid framework [Chen et al., 2007] by learning grid coefficients to approximate various complex image operations. This method offers a more accurate, efficient, and generalizable modeling approach.

8.2 Advanced HDR Content Generation Models

TAF and CinematicGS focus on recovering HDR information from real-world captured images. With the advancement of generative models, generating content from latent codes is becoming increasingly promising. With the rapid development of generative models, synthesizing realistic images directly from latent codes has become increasingly feasible. However, existing generative models remain constrained to LDR content, mainly due to the limited availability of HDR data. To address this limitation, we propose two strategies for enabling HDR generation.

Chapter 5 (GlowGAN) tackles the problem from a distributional perspective. Instead of relying on paired HDR data or multi-exposure sequences, we leverage unstructured LDR images to recover the underlying HDR distribution. This is achieved by embedding a differentiable camera imaging model, which includes exposure sampling, intensity clipping, and tone mapping, into a GAN framework. By simulating the HDR-to-LDR degradation process during training, the model learns to generate physically plausible HDR content that aligns with partial LDR observations. Compared to vanilla GANs, GlowGAN produces images with significantly higher dynamic range and fewer saturation artifacts.

While GlowGAN performs well in reconstructing HDR content, its reliance on separate training for each scene limits its scalability. To overcome this, we further introduce LEDiff in Chapter 6, which builds upon Stable Diffusion [Rombach et al., 2022], that offer strong generation quality and semantic control. Given the challenge of training such models from scratch using limited HDR data, we instead fine-tune a pre-trained LDR diffusion model. By analyzing multi-exposure LDR-HDR pairs, we observe that the latent codes encodes brightness-related semantics. Motivated by this, we propose a latent exposure-based fine-tuning scheme that updates both the VAE and denoiser components. This enables the model to decode HDR content and infer missing information in over- or underexposed regions.

GlowGAN and LEDiff demonstrate two viable directions for HDR synthesis. The former focuses on learning HDR distributions from unstructured LDR data, while the latter leverages the strengths of large-scale diffusion models with minimal HDR supervision.

While GlowGAN and LEDiff have made significant progress, we believe the following directions are still highly worth exploring.

Regarding the improvements in generative capabilities, the most prominent generative models today include GANs, diffusion models, and autoregressive models [Cao et al., 2024]. Even the most advanced GANs [Kang et al., 2023] have fallen behind diffusion and autoregressive approaches in terms of image quality and diversity. Given that LEDiff is primarily developed based on Stable Diffusion [Rombach et al., 2022], we remain interested in further exploring diffusion models.

Stable Diffusion adopts a CNN-based U-Net as its denoiser. However, with recent architectural advances, transformer-based diffusion models, Diffusion Transformer (DiT) [Peebles and Xie, 2023], are becoming increasingly popular. A key advantage of DiT lies in its ability to capture long-range dependencies and global context through self-attention. Unlike U-Net-based denoisers, DiT divide the input latent representations into patches and process them with transformer blocks, offering better scalability and improved performance, particularly as model size increases. Since the latent space utilized by DiT differs from that of Stable Diffusion, it is crucial to develop a comprehensive understanding of how DiT's latent codes correspond to input images, and to further investigate methods for fine-tuning a pretrained DiT

model using limited HDR data, which presents a promising avenue for enhancing the quality of HDR image generation.

Another promising direction is HDR video generation. While LEDiff has made some progress on image-to-HDR Video generation based on SVD (Stable Video Diffusion) [Blattmann et al., 2023], several challenges remain. Since the model is only fine-tuned on the image dataset, the generated videos often suffer from limited visual quality and poor temporal consistency. A straightforward and effective reference for improving performance is VideoDepth [Shao et al., 2024], where paired RGB-Depth video data is used as supervision to fine-tune SVD, along with a temporal inpainting mechanism to enhance temporal consistency. Inspired by this, we could construct an HDR-LDR video dataset and apply a similar strategy. However, hallucinating overexposed or underexposed regions would require leveraging information from neighboring frames [Kim et al., 2019; Kim et al., 2020; He et al., 2022a; He et al., 2022b].

Moreover, working in a linear space is crucial for many tasks [Xu et al., 2023a; Zhang et al., 2019c; Zhou et al., 2023]. However, LEDiff assumes a global degradation model when constructing the HDR-LDR dataset. In practice, the training data of Stable Diffusion includes many locally tone mapped images, making it difficult for the model to generate accurate linear HDR outputs. This also poses challenges in recovering true linear content from LDR images that have undergone local tone mapping during the LDR-to-HDR image reconstruction process. A feasible direction is to explicitly include locally tone mapped images when constructing the training dataset, so that the model can implicitly learn the corresponding relationships through fine-tuning.

Reconstruction, which can be seen as image-based rendering, and generation are becoming increasingly complementary in the context of content creation. Physical constraints from reconstruction can be leveraged to guide higher-quality generation [Yuan et al., 2023; Voynov et al., 2024; Fortes et al., 2025], while generative models can serve as powerful priors to hallucinate missing information in reconstruction [Wewer et al., 2024; Wang et al., 2024b].

8.3 Unsupervised and Controllable Tone Mapper

Although HDR content offers a superior visual experience and is becoming increasingly available, tone mapping remains an important research problem due to the need to maintain compatibility with the vast number of existing LDR displays. To this end, we introduce SelfTMO in Chapter 7, an image-specific, self-supervised tone mapping method that avoids the need for paired HDR-LDR datasets. Traditional learning-based approaches often rely heavily on curated training data, which limits their generalization and scalability. In contrast, SelfTMO is guided by a feature contrast masking loss that models human visual perception in feature space. This design allows for flexible contrast compression and enhancement, enabling more perceptually faithful HDR-to-LDR conversion. By incorporating neighborhood masking, the model further adapts to local image structures, preserving important details.

As noted in Chapter 7, tone mapping is fundamentally a compression task. Unlike image restoration, which focuses on recovering lost information, tone mapping aims to preserve perceptual quality by ensuring that the compressed result on LDR displays closely matches human visual perception of the real world [Reinhard et al., 2010; Mantiuk et al., 2015b; Banterle et al., 2017]. We believe that a good tone mapper should have two key properties: the ability to compress HDR content with high fidelity, and

the flexibility to adjust its output. Since perception is inherently subjective, training an end-to-end network through regression on a labeled dataset may not be the optimal solution for tone mapping.

While SelfTMO demonstrates promising potential, there remain other interesting directions worth exploring.

First, it optimizes a CNN through a test-time training pipeline, requiring separate optimization for each image, which makes the process time-consuming. We believe it is worthwhile to explore replacing the CNN architecture with Lookup Tables (LUTs), as demonstrated in related work [Ke et al., 2023; Li et al., 2025]. As mentioned above, tone mapping does not require complex operations, and LUTs offer higher efficiency, faster inference, and a much smaller model size, making them more suitable for lightweight deployment.

On the other hand, SelfTMO adopts a model-driven approach by manually designing handcrafted features to guide the tone mapping process. A promising alternative is to learn such constraints in a data-driven manner from unstructured HDR and LDR image collections. Vinker et al. [2021] demonstrates this possibility by achieving unsupervised tone mapping through adversarial training on unpaired HDR and LDR images, using a content-preserving loss and shallow discriminators to produce realistic and artifact-free LDR outputs. Building on this idea, recent work extends the approach to diffusion models. Based on ControlNet [Zhang et al., 2023a], Zhu et al. [2024] propose a zero-shot HDR tone mapping framework that decomposes images into structure and tone components, and leverages shared structural cues across HDR and LDR domains to guide the diffusion model, resulting in perceptually natural and structurally faithful LDR outputs without paired training data.

While these tone mapping methods produce high-quality results, they lack controllability, which makes post-editing less convenient for users.

A bold yet promising direction is to design a prompt-based tone mapping adjustment pipeline. Given that Zhu et al. [2024] has already demonstrated the feasibility of diffusion-based tone mapping, integrating prompts as a natural form of control could offer both interpretability and user interactivity. With recent advances in prompt-based image editing [Brooks et al., 2023b], language has emerged as a flexible and user-friendly interface for controlling visual outputs, often more intuitive than manually adjusting sliders. In particular, this approach allows for targeted manipulation of specific regions within an image, such as the sky or grass. Compared to the high-level semantic editing in [Brooks et al., 2023b], tone mapping involves low-level but more fine-grained adjustments. A relevant reference is the prompt-based material editing method proposed in [Sharma et al., 2024], which enables precise control over material properties such as roughness, metallicity, albedo, and transparency in real images. This is achieved by fine-tuning a pre-trained text-to-image diffusion model on a synthetic dataset containing physically-based materials. Inspired by this, we can similarly define common tone mapping parameters, such as contrast, saturation, brightness, and detail enhancement, as editable attributes. A synthetic dataset can be constructed for fine-tuning, allowing the model to learn how to interpret and apply these operations, and enabling continuous and controllable editing after training.

Bibliography - Own Work

Chen, Bin, Akshay Jindal, Michal Piovarči, Chao Wang, Hans-Peter Seidel, Piotr Didyk, Karol Myszkowski, Ana Serrano and Rafał K Mantiuk [2023a]. "The effect of display capabilities on the gloss consistency between real and virtual objects". In: *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11.

Chen, Bin, Michal Piovarči, Chao Wang, Hans-Peter Seidel, Piotr Didyk, Karol Myszkowski and Ana Serrano [2022b]. "Gloss management for consistent reproduction of real and virtual objects". In: *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9.

Chen, Bin, Chao Wang, Michal Piovarči, Hans-Peter Seidel, Piotr Didyk, Karol Myszkowski and Ana Serrano [2021]. "The effect of geometry and illumination on appearance perception of different material categories". In: *The Visual Computer* 37, pp. 2975–2987.

Serrano, Ana, Bin Chen, Chao Wang, Michal Piovarči, Hans-Peter Seidel, Piotr Didyk and Karol Myszkowski [2021]. "The effect of shape and illumination on material perception: model and applications". In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–16.

Wang, Chao, Bin Chen, Hans-Peter Seidel, Karol Myszkowski and Ana Serrano [2022a]. "Learning a self-supervised tone mapping operator via feature contrast masking loss". In: *Computer Graphics Forum*. Vol. 41. 2. Wiley Online Library, pp. 71–84.

Wang, Chao, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt and Thomas Leimkühler [2023a]. "Glowgan: Unsupervised learning of hdr images from ldr images in the wild". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10509–10519.

Wang, Chao, Ana Serrano, Xingang Pan, Krzysztof Wolski, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt and Thomas Leimkühler [2023b]. "An Implicit Neural Representation for the Image Stack: Depth, All in Focus, and High Dynamic Range". In: *ACM Transactions on Graphics* 42.6, pp. 1–11.

Wang, Chao, Krzysztof Wolski, Bernhard Kerbl, Ana Serrano, Mojtaba Bemana, Hans-Peter Seidel, Karol Myszkowski and Thomas Leimkühler [2024a]. "Cinematic Gaussians: Real-Time HDR Radiance Fields with Depth of Field". In: *Computer Graphics Forum*. Wiley Online Library, e15214.

Wang, Chao, Zhihao Xia, Thomas Leimkühler, Karol Myszkowski and Xuaner Zhang [2025]. "LEDiff: Latent Exposure Diffusion for HDR Generation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Ning, Yang Zhao, Chao Wang and Ronggang Wang [2021]. "A real-time semi-supervised deep tone mapping network". In: *IEEE Transactions on Multimedia* 24, pp. 2815–2827.

Bibliography

Abdal, Rameen, Yipeng Qin and Peter Wonka [2019]. "Image2stylegan: How to embed images into the stylegan latent space?" In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4432–4441.

Ahumada Jr, Albert J and Heidi A Peterson [1992]. "Luminance-model-based DCT quantization for color image compression". In: *Human Vision, Visual Processing, and Digital Display III*. Vol. 1666. Proc SPIE, pp. 365–374.

Akyüz, Ahmet Oğuz, Roland Fleming, Bernhard E Riecke, Erik Reinhard and Heinrich H Bülthoff [2007]. "Do HDR displays support LDR content? A psychophysical evaluation". In: *ACM Transactions on Graphics (TOG)* 26.3, 38–es.

Albright, Michael, Scott McCloskey and ACST Honeywell [2019]. "Source Generator Attribution via Inversion." In: *CVPR workshops*. Vol. 8, p. 3.

Aliev, Kara-Ali, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov and Victor Lempitsky [2020]. "Neural point-based graphics". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, pp. 696–712.

Alireza Golestaneh, S and Lina J Karam [2017]. "Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5800–5809.

Allen, Elizabeth and Sophie Triantaphillidou [2011]. *The manual of photography*. Taylor & Francis.

Andersson, Pontus, Jim Nilsson, Peter Shirley and Tomas Akenine-Möller [2021]. "Visualizing errors in rendered high dynamic range images". In.

Arjovsky, Martin, Soumith Chintala and Léon Bottou [2017]. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

Artusi, Alessandro, Tania Pouli, Francesco Banterle and Ahmet Oğuz Akyüz [2018]. "Automatic saturation correction for dynamic range management algorithms". In: *Signal Processing: Image Communication* 63, pp. 100–112.

Atzmon, Matan and Yaron Lipman [2020]. "Sal: Sign agnostic learning of shapes from raw data". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2565–2574.

Azimi, Maryam et al. [2021]. "PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR". In: *2021 Picture Coding Symposium (PCS)*. IEEE, pp. 1–5.

Bangaru, Sai Praveen, Jesse Michel, Kevin Mu, Gilbert Bernstein, Tzu-Mao Li and Jonathan Ragan-Kelley [2021]. "Systematically differentiating parametric discontinuities". In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–18.

Banterle, Francesco, Alessandro Artusi, Kurt Debattista and Alan Chalmers [2017]. "Advanced high dynamic range imaging". In.

Banterle, Francesco, Alessandro Artusi, Elena Sikudova, Patrick Ledda, Thomas Bashford-Rogers, Alan Chalmers and Marina Bloj [2016]. "Mixing tone mapping

operators on the GPU by differential zone mapping based on psychophysical experiments". In: *Signal Processing: Image Communication* 48, pp. 50–62.

Banterle, Francesco, Patrick Ledda, Kurt Debattista and Alan Chalmers [2006]. "Inverse tone mapping". In: *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pp. 349–356.

Banterle, Francesco, Patrick Ledda, Kurt Debattista and Alan Chalmers [2008]. "Expanding low dynamic range videos for high dynamic range applications". In: *Proceedings of the 24th Spring Conference on Computer Graphics*, pp. 33–41.

Banterle, Francesco, Patrick Ledda, Kurt Debattista, Alan Chalmers and Marina Bloj [2007]. "A framework for inverse tone mapping". In: *The Visual Computer* 23, pp. 467–478.

Banterle, Francesco, Demetris Marnerides, Kurt Debattista and Thomas Bashford-Rogers [2022]. "Unsupervised HDR Imaging: What Can Be Learned from a Single 8-bit Video?" In: *arXiv preprint arXiv:2202.05522*.

Barron, Jonathan T, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla and Pratul P Srinivasan [2021]. "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5855–5864.

Barron, Jonathan T, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan and Peter Hedman [2022]. "Mip-nerf 360: Unbounded anti-aliased neural radiance fields". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479.

Barron, Jonathan T, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan and Peter Hedman [2023]. "Zip-nerf: Anti-aliased grid-based neural radiance fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19697–19705.

Bemana, Mojtaba, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel and Tobias Ritschel [2024]. "Exposure Diffusion: HDR Image Generation by Consistent LDR denoising". In: *arXiv preprint arXiv:2405.14304*.

Benz, Greg [Aug. 2022]. *Create and edit true HDR (High Dynamic Range) images - Greg Benz Photography*. en-US.

Bińkowski, Mikołaj, Danica J Sutherland, Michael Arbel and Arthur Gretton [2018]. "Demystifying mmd gans". In: *arXiv preprint arXiv:1801.01401*.

Blattmann, Andreas, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts et al. [2023]. "Stable video diffusion: Scaling latent video diffusion models to large datasets". In: *arXiv preprint arXiv:2311.15127*.

Blau, Yochai and Tomer Michaeli [2018]. "The perception-distortion tradeoff". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237.

Bolduc, Christophe, Justine Giroux, Marc Hébert, Claude Demers and Jean-François Lalonde [2023]. "Beyond the pixel: a photometrically calibrated hdr dataset for luminance and color prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8071–8081.

Bolin, Mark R and Gary W Meyer [1998]. "A perceptually based adaptive sampling algorithm". In: *Proc ACM SIGGRAPH*, pp. 299–309.

Bora, Ashish, Eric Price and Alexandros G Dimakis [2018]. "AmbientGAN: Generative models from lossy measurements". In: *International conference on learning representations*.

Bouzos, Odysseas, Ioannis Andreadis and Nikolaos Mitianoudis [2019]. "Conditional random field model for robust multi-focus image fusion". In: *IEEE Transactions on Image Processing* 28.11, pp. 5636–5648.

Brock, Andrew, Jeff Donahue and Karen Simonyan [2018]. "Large scale GAN training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096*.

Bromiley, Paul [2003]. "Products and convolutions of Gaussian probability density functions". In: *Tina-Vision Memo* 3.4, p. 1.

Brooks, Tim, Aleksander Holynski and Alexei A Efros [2023a]. "Instructpix2pix: Learning to follow image editing instructions". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402.

Brooks, Tim, Aleksander Holynski and Alexei A. Efros [2023b]. "InstructPix2Pix: Learning to Follow Image Editing Instructions". In: *CVPR*.

Brox, Thomas, Andrés Bruhn, Nils Papenberg and Joachim Weickert [2004]. "High accuracy optical flow estimation based on a theory for warping". In: *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*. Springer, pp. 25–36.

Bruckstein, Alfred M, Bart M ter Haar Romeny, Alexander M Bronstein and Michael M Bronstein [2012]. *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers*. Vol. 6667. Springer Science & Business Media.

Čadík, Martin, Michael Wimmer, Laszlo Neumann and Alessandro Artusi [2008]. "Evaluation of HDR tone mapping methods using essential perceptual attributes". In: *Computers & Graphics* 32.3, pp. 330–349.

Cai, Yuanhao, Zihao Xiao, Yixun Liang, Minghan Qin, Yulun Zhang, Xiaokang Yang, Yaoyao Liu and Alan Yuille [2024a]. "Hdr-gs: Efficient high dynamic range novel view synthesis at 1000x speed via gaussian splatting". In: *arXiv preprint arXiv:2405.15125*.

Cai, Yuanhao, Zihao Xiao, Yixun Liang, Minghan Qin, Yulun Zhang, Xiaokang Yang, Yaoyao Liu and Alan Yuille [2024b]. "Hdr-gs: Efficient high dynamic range novel view synthesis at 1000x speed via gaussian splatting". In: *arXiv preprint arXiv:2405.15125*.

Cao, Hanqun, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng and Stan Z Li [2024]. "A survey on generative diffusion models". In: *IEEE Transactions on Knowledge and Data Engineering*.

Celarek, Adam, Pedro Hermosilla, Bernhard Kerbl, Timo Ropinski and Michael Wimmer [2022]. "Gaussian mixture convolution networks". In: *arXiv preprint arXiv:2202.09153*.

Cha, Sungmin, Taeeon Park, Byeongjoon Kim, Jongduk Baek and Taesup Moon [2019]. "GAN2GAN: Generative noise learning for blind denoising with single noisy images". In: *arXiv preprint arXiv:1905.10488*.

Chabra, Rohan, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove and Richard Newcombe [2020]. "Deep local shapes: Learning local sdf priors for detailed 3d reconstruction". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, pp. 608–625.

Chai, Lucy, Michael Gharbi, Eli Shechtman, Phillip Isola and Richard Zhang [2022]. "Any-resolution training for high-resolution image synthesis". In: *European Conference on Computer Vision*. Springer, pp. 170–188.

Chan, Eric [Oct. 2023]. *High Dynamic Range Explained*. en-US.

Chan, Eric R, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis

et al. [2022]. "Efficient geometry-aware 3d generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133.

Chan, Eric R, Marco Monteiro, Petr Kellnhofer, Jiajun Wu and Gordon Wetzstein [2021]. "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5799–5809.

Chen, Anpei, Zexiang Xu, Andreas Geiger, Jingyi Yu and Hao Su [2022a]. "Tensorf: Tensorial radiance fields". In: *European conference on computer vision*. Springer, pp. 333–350.

Chen, Jiawen, Sylvain Paris and Frédo Durand [2007]. "Real-time edge-aware image processing with the bilateral grid". In: *ACM Transactions on Graphics (TOG)* 26.3, 103–es.

Chen, Rufeng, Bolun Zheng, Hua Zhang, Quan Chen, Chenggang Yan, Gregory Slabaugh and Shanxin Yuan [2023b]. "Improving dynamic hdr imaging with fusion transformer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 1, pp. 340–349.

Choi, Yunjey, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim and Jaegul Choo [2018]. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797.

Chollet, François [2017]. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258.

Chung, Jaeyoung, Jeongtaek Oh and Kyoung Mu Lee [2024]. "Depth-regularized optimization for 3d gaussian splatting in few-shot images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 811–820.

Çiçek, Özgün, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox and Olaf Ronneberger [2016]. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, pp. 424–432.

Cook, Robert L, Thomas Porter and Loren Carpenter [1984]. "Distributed ray tracing". In: *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pp. 137–145.

Corneanu, Ciprian, Raghudeep Gadde and Aleix M Martinez [2024]. "Latentpaint: Image inpainting in latent space with diffusion models". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4334–4343.

Creswell, Antonia and Anil Anthony Bharath [2018]. "Inverting the generator of a generative adversarial network". In: *IEEE transactions on neural networks and learning systems* 30.7, pp. 1967–1974.

Daly, Scott J [1992]. "Visible differences predictor: an algorithm for the assessment of image fidelity". In: *Human Vision, Visual Processing, and Digital Display III*. Vol. 1666. Proc SPIE, pp. 2–15.

Daly, Scott J, Wenjun Zeng, Jin Li and Shawmin Lei [2000]. "Visual masking in wavelet compression for JPEG-2000". In: *Image and Video Communications and Processing*. Vol. 3974. Proc SPIE, pp. 66–80.

Darmon, François, Lorenzo Porzi, Samuel Rota-Bulò and Peter Kontschieder [2024]. "Robust gaussian splatting". In: *arXiv preprint arXiv:2404.04211*.

Debattista, Kurt [2018]. "Application-Specific Tone Mapping Via Genetic Programming". In: *Computer Graphics Forum*. Vol. 37. 1, pp. 439–450.

Debevec, Paul [2006]. "Image-based lighting". In: *ACM SIGGRAPH 2006 Courses*, 4–es.

Debevec, Paul E and Jitendra Malik [2023]. "Recovering high dynamic range radiance maps from photographs". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 643–652.

Demers, Joe [2004]. "Depth of field: A survey of techniques". In: *Gpu Gems* 1.375, U390.

Didyk, Piotr, Elmar Eisemann, Tobias Ritschel, Karol Myszkowski and Hans-Peter Seidel [2010]. "Apparent display resolution enhancement for moving images". In: *Proc ACM SIGGRAPH*, pp. 1–8.

Didyk, Piotr, Rafal Mantiuk, Matthias Hein and Hans-Peter Seidel [2008]. "Enhancement of bright video features for HDR displays". In: *Computer Graphics Forum*. Vol. 27. 4. Wiley Online Library, pp. 1265–1274.

Dille, Sebastian, Chris Careaga and Yağız Aksoy [2025]. "Intrinsic Single-Image HDR Reconstruction". In: *European Conference on Computer Vision*. Springer, pp. 161–177.

Donahue, Jeff, Philipp Krähenbühl and Trevor Darrell [2016]. "Adversarial feature learning". In: *arXiv preprint arXiv:1605.09782*.

Drago, Frédéric, Karol Myszkowski, Thomas Annen and Norishige Chiba [2003]. "Adaptive logarithmic mapping for displaying high contrast scenes". In: *Computer Graphics Forum*. Vol. 22. 3, pp. 419–426.

Durand, Frédo and Julie Dorsey [2002]. "Fast bilateral filtering for the display of high-dynamic-range images". In: *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 257–266.

Eigen, David, Christian Puhrsch and Rob Fergus [2014]. "Depth map prediction from a single image using a multi-scale deep network". In: *Advances in neural information processing systems* 27.

Eilertsen, Gabriel [2018]. *The high dynamic range imaging pipeline*. Vol. 1939. Linköping University Electronic Press.

Eilertsen, Gabriel, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk and Jonas Unger [2017]. "HDR image reconstruction from a single exposure using deep CNNs". In: *ACM transactions on graphics (TOG)* 36.6, pp. 1–15.

Endo, Yuki, Yoshihiro Kanamori and Jun Mitani [2017a]. "Deep reverse tone mapping". In: *ACM Trans. Graph* 36.6, pp. 1–10.

Endo, Yuki, Yoshihiro Kanamori and Jun Mitani [2017b]. "Deep Reverse Tone Mapping". In: *ACM Trans. Graph (SIGGRAPH Asia 2017)* 36.6.

Fairchild, Mark D [2007]. "The HDR photographic survey". In: *Color and imaging conference*. Vol. 15. Society of Imaging Science and Technology, pp. 233–238.

Fattal, Raanan, Dani Lischinski and Michael Werman [2002]. "Gradient domain high dynamic range compression". In: *ACM Trans. Graph*. Vol. 21. 3, pp. 249–256.

Favaro, Paolo [2010]. "Recovering thin structures via nonlocal-means regularization with application to depth from defocus". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1133–1140.

Feng, Brandon Yushan, Susmija Jabbireddy and Amitabh Varshney [2022]. "Viinter: View interpolation with implicit neural representations of images". In: *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9.

Foley, John M [1994]. "Human luminance pattern-vision mechanisms: masking experiments require a new model". In: *JOSA A* 11.6, pp. 1710–1719.

Fortes, Armando, Tianyi Wei, Shangchen Zhou and Xingang Pan [2025]. "Bokeh Diffusion: Defocus Blur Control in Text-to-Image Diffusion Models". In: *arXiv preprint arXiv:2503.08434*.

Fu, Xiao, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin and Xiaoxiao Long [2024]. "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image". In: *European Conference on Computer Vision*. Springer, pp. 241–258.

Fu, Xiao, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin and Xiaoxiao Long [2025]. "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image". In: *European Conference on Computer Vision*. Springer, pp. 241–258.

Gallo, Orazio, Marius Tico, Roberto Manduchi, Natasha Gelfand and Kari Pulli [2012]. "Metering for exposure stacks". In: *Computer Graphics Forum*. Vol. 31. 2pt2, pp. 479–488.

Gao, Ruohan, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan and Jiajun Wu [2022]. "Objectfolder 2.0: A multisensory object dataset for sim2real transfer". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10598–10608.

Gardner, Marc-André, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné and Jean-François Lalonde [2017]. "Learning to predict indoor illumination from a single image". In: *arXiv preprint arXiv:1704.00090*.

Genova, Kyle, Forrester Cole, Avneesh Sud, Aaron Sarna and Thomas Funkhouser [2020]. "Local deep implicit functions for 3d shape". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4857–4866.

Genova, Kyle, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman and Thomas Funkhouser [2019]. "Learning shape templates with structured implicit functions". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7154–7164.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio [2014]. "Generative adversarial nets". In: *Advances in neural information processing systems* 27.

Goswami, Abhishek, Aru Ranjan Singh, Francesco Banterle, Kurt Debattista and Thomas Bashford-Rogers [2024]. "Semantic Aware Diffusion Inverse Tone Mapping". In: *arXiv preprint arXiv:2405.15468*.

Gropp, Amos, Lior Yariv, Niv Haim, Matan Atzmon and Yaron Lipman [2020]. "Implicit geometric regularization for learning shapes". In: *arXiv preprint arXiv:2002.10099*.

Gross, Herbert [2005]. "Handbook of Optical Systems". In.

Gross, Markus and Hanspeter Pfister [2011]. *Point-based graphics*. Elsevier.

Grossberg, Michael D and Shree K Nayar [2003]. "What is the space of camera response functions?" In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Vol. 2. IEEE, pp. II–602.

Gu, Ke, Shiqi Wang, Guangtao Zhai, Siwei Ma, Xiaokang Yang, Weisi Lin, Wenjun Zhang and Wen Gao [2016]. "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure". In: *IEEE Trans. on Multimedia* 18.3, pp. 432–443.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin and Aaron C Courville [2017]. "Improved training of wasserstein gans". In: *Advances in neural information processing systems* 30.

Guo, Cheng and Xiuhua Jiang [2021]. "Deep tone-mapping operator using image quality assessment inspired semi-supervised learning". In: *IEEE Access* 9, pp. 73873–73889.

Haeberli, Paul and Kurt Akeley [1990]. "The accumulation buffer: Hardware support for high-quality rendering". In: *ACM SIGGRAPH computer graphics* 24.4, pp. 309–318.

Hanji, Param, Rafal Mantiuk, Gabriel Eilertsen, Saghi Hajisharif and Jonas Unger [2022]. "Comparison of single image HDR reconstruction methods—the caveats of quality assessment". In: *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–8.

Hasinoff, Samuel W and Kiriakos N Kutulakos [2007]. "A layer-based restoration framework for variable-aperture photography". In: *2007 IEEE 11th International Conference on Computer Vision*. IEEE, pp. 1–8.

Hazirbas, Caner, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé and Daniel Cremers [2019]. "Deep depth from focus". In: *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, pp. 525–541.

He, Gang, Shaoyi Long, Li Xu, Chang Wu, Jinjia Zhou, Ming Sun, Xing Wen and Yurong Dai [2022a]. "Global priors guided modulation network for joint super-resolution and inverse tone-mapping". In: *arXiv preprint arXiv:2208.06885*.

He, Gang, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen and Yu-Wing Tai [2022b]. "SDRTV-to-HDRTV via hierarchical dynamic context feature mapping". In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2890–2898.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun [2016]. "Deep residual learning for image recognition". In: *CVPR*, pp. 770–778.

Hedman, Peter, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron and Paul Debevec [2021]. "Baking neural radiance fields for real-time view synthesis". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5875–5884.

Henzler, Philipp, Niloy J Mitra and Tobias Ritschel [2019]. "Escaping plato's cave: 3d shape from adversarial rendering". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9993.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler and Sepp Hochreiter [2017]. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30.

Ho, Jonathan, Ajay Jain and Pieter Abbeel [2020]. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851.

Hornik, Kurt, Maxwell Stinchcombe and Halbert White [1989]. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hou, Xianxu, Jiang Duan and Guoping Qiu [2017]. "Deep feature consistent deep image transformations: Downscaling, decolorization and HDR tone mapping". In: *arXiv preprint arXiv:1707.09482*.

Hu, Jintong, Bin Xia, Bin Chen, Wenming Yang and Lei Zhang [2025]. "Gaussiansr: High fidelity 2d gaussian splatting for arbitrary-scale image super-resolution". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 4, pp. 3554–3562.

Huang, Binbin, Zehao Yu, Anpei Chen, Andreas Geiger and Shenghua Gao [2024]. "2d gaussian splatting for geometrically accurate radiance fields". In: *ACM SIGGRAPH 2024 conference papers*, pp. 1–11.

Huang, Xin, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang and Qing Wang [2022]. "Hdr-nerf: High dynamic range neural radiance fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18398–18408.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou and Alexei A Efros [2017]. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

Jack, Tumblin and Rushmeier Holly [1993]. "Tone reproduction for realistic images". In: *IEEE Computer Graphics and Applications* 13.6, pp. 42–48.

Jacobson, Ralph, Sidney Ray, Geoffrey G Attridge and Norman Axford [2013]. *Manual of Photography*. Routledge.

Jarabo, Adrian, Belen Masia, Adrien Bousseau, Fabio Pellacini and Diego Gutierrez [2014]. "How do people edit light fields". In: *ACM Trans. Graph (SIGGRAPH 2014)* 33.4, p. 4.

Jiang, Chiyu, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser et al. [2020]. "Local implicit grid representations for 3d scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6001–6010.

Jiang, Xuesong, Hongxun Yao and Dilin Liu [2019]. "Nighttime image enhancement based on image decomposition". In: *Signal, Image and Video Processing* 13.1, pp. 189–197.

Jinno, Takao, Hironori Kaida, Xinwei Xue, Nicola Adami and Masahiro Okuda [2011]. "μ-Law based HDR coding and its error analysis". In: *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences* 94.3, pp. 972–978.

Jo, So Yeon, Siyeong Lee, Namhyun Ahn and Suk-Ju Kang [2021]. "Deep arbitrary HDRI: Inverse tone mapping with controllable exposure changes". In: *IEEE Transactions on Multimedia* 24, pp. 2713–2726.

Johnson, Justin, Alexandre Alahi and Li Fei-Fei [2016]. "Perceptual losses for real-time style transfer and super-resolution". In: *ECCV*, pp. 694–711.

Jun-Seong, Kim, Kim Yu-Ji, Moon Ye-Bin and Tae-Hyun Oh [2022]. "Hdr-plenoxels: Self-calibrating high dynamic range radiance fields". In: *European Conference on Computer Vision*. Springer, pp. 384–401.

Kalantari, Nima Khademi and Ravi Ramamoorthi [2019]. "Deep HDR video from sequences with alternating exposures". In: *Computer graphics forum*. Vol. 38. 2. Wiley Online Library, pp. 193–205.

Kalantari, Nima Khademi et al. [2017]. "Deep high dynamic range imaging of dynamic scenes." In: *ACM Trans. Graph.* 36.4, pp. 144–1.

Kaneko, Takuhiro [2021]. "Unsupervised learning of depth and depth-of-field effect from natural images with aperture rendering generative adversarial networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15679–15688.

Kaneko, Takuhiro and Tatsuya Harada [2020]. "Noise robust generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8404–8414.

Kang, Minguk, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris and Taesung Park [2023]. "Scaling up gans for text-to-image synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134.

Karras, Tero, Timo Aila, Samuli Laine and Jaakko Lehtinen [2017]. "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196*.

Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen and Timo Aila [2020a]. "Training generative adversarial networks with limited data". In: *Advances in neural information processing systems* 33, pp. 12104–12114.

Karras, Tero, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen and Timo Aila [2021]. "Alias-free generative adversarial networks". In: *Advances in neural information processing systems* 34, pp. 852–863.

Karras, Tero, Samuli Laine and Timo Aila [2019]. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen and Timo Aila [2020b]. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119.

Kasten, Yoni, Dolev Ofri, Oliver Wang and Tali Dekel [2021]. "Layered neural atlases for consistent video editing". In: *ACM Transactions on Graphics (TOG)* 40.6, pp. 1–12.

Ke, Bingxin, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt and Konrad Schindler [2024]. "Repurposing diffusion-based image generators for monocular depth estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502.

Ke, Zhanghan, Yuhao Liu, Lei Zhu, Nanxuan Zhao and Rynson W.H. Lau [2023]. "Neural Preset for Color Style Transfer". In: *Computer Vision and Pattern Recognition Conference (CVPR)*.

Kerbl, Bernhard, Georgios Kopanas, Thomas Leimkühler and George Drettakis [2023]. "3d gaussian splatting for real-time radiance field rendering." In: *ACM Trans. Graph.* 42.4, pp. 139–1.

Kim, Soo Ye, Jihyong Oh and Munchurl Kim [2019]. "Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3116–3125.

Kim, Soo Ye, Jihyong Oh and Munchurl Kim [2020]. "Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07, pp. 11287–11295.

Kingdom, Frederick AA and Paul Whittle [1996]. "Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing". In: *Vision research* 36.6, pp. 817–829.

Kingma, Diederik P [2014]. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling [2014]. "Auto-Encoding Variational Bayes". In: *stat* 1050, p. 1.

Kolb, Craig, Don Mitchell and Pat Hanrahan [1995]. "A realistic camera model for computer graphics". In: *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 317–324.

Kong, Lingtong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu and Jinwei Chen [2024]. "Safnet: Selective alignment fusion network for efficient hdr imaging". In: *European Conference on Computer Vision*. Springer, pp. 256–273.

Kovaleski, Rafael Pacheco and Manuel M Oliveira [2009]. "High-quality brightness enhancement functions for real-time reverse tone mapping". In: *The Visual Computer* 25, pp. 539–547.

Kriv\'anek, Jaroslav, Jiri Zara and Kadi Bouatouch [2003]. "Fast depth of field rendering with surface splatting". In: *Proceedings Computer Graphics International 2003*. IEEE, pp. 196–201.

Kupyn, Orest, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin and Jiří Matas [2018]. "Deblurgan: Blind motion deblurring using conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8183–8192.

Kupyn, Orest, Tetiana Martyniuk, Junru Wu and Zhangyang Wang [2019]. "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8878–8887.

Laina, Iro, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari and Nassir Navab [2016]. "Deeper depth prediction with fully convolutional residual networks". In: *2016 Fourth international conference on 3D vision (3DV)*. IEEE, pp. 239–248.

Larson, Gregory Ward, Holly Rushmeier and Christine Piatko [1997]. "A visibility matching tone reproduction operator for high dynamic range scenes". In: *IEEE Trans Visual and Comp Graph* 3.4, pp. 291–306.

Le, Phuoc-Hieu, Quynh Le, Rang Nguyen and Binh-Son Hua [2023]. "Single-image hdr reconstruction by multi-exposure generation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4063–4072.

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton [2015]. "Deep learning". In: *nature* 521.7553, pp. 436–444.

Ledda, Patrick, Alan Chalmers, Tom Troscianko and Helge Seetzen [2005]. "Evaluation of tone mapping operators using a high dynamic range display". In: *ACM Trans. Graph* 24.3, pp. 640–648.

Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang et al. [2017]. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.

Lee, Byeonghyeon, Howoong Lee, Xiangyu Sun, Usman Ali and Eunbyung Park [2024]. "Deblurring 3d gaussian splatting". In: *European Conference on Computer Vision*. Springer, pp. 127–143.

Lee, Dogyoon, Minhyeok Lee, Chajin Shin and Sangyoun Lee [2023]. "Dp-nerf: Deblurred neural radiance field with physical scene priors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12386–12396.

Lee, Siyeong, Gwon Hwan An and Suk-Ju Kang [2018a]. "Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image". In: *IEEE Access* 6, pp. 49913–49924.

Lee, Siyeong, Gwon Hwan An and Suk-Ju Kang [2018b]. "Deep recursive hdri: Inverse tone mapping using generative adversarial networks". In: *proceedings of the European Conference on Computer Vision (ECCV)*, pp. 596–611.

Lee, Siyeong, So Yeon Jo, Gwon Hwan An and Suk-Ju Kang [2020]. "Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging". In: *IEEE Transactions on Multimedia* 23, pp. 2561–2574.

Legge, Gordon E and John M Foley [1980]. "Contrast masking in human vision". In: *Josa* 70.12, pp. 1458–1471.

Levoy, Marc and Pat Hanrahan [2023]. "Light field rendering". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 441–452.

Li, Hui, Xixi Jia and Lei Zhang [2018]. "Clustering based content and color adaptive tone mapping". In: *CVIU* 168, pp. 37–49.

Li, Jun, Reinhard Klein and Angela Yao [2017]. "A two-streamed network for estimating fine-scaled depth maps from single rgb images". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3372–3380.

Li, Mujing, Guanjie Wang, Xingguang Zhang, Qifeng Liao and Chenxi Xiao [2025]. "D-LUT: Photorealistic Style Transfer via Diffusion Process". In: *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 9206–9214.

Li, Shutao, Xudong Kang and Jianwen Hu [2013]. "Image fusion with guided filtering". In: *IEEE Transactions on Image processing* 22.7, pp. 2864–2875.

Liang, Zhetong, Jun Xu, David Zhang, Zisheng Cao and Lei Zhang [2018]. "A hybrid l1-l0 layer decomposition model for tone mapping". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4758–4766.

Liao, Yiyi, Katja Schwarz, Lars Mescheder and Andreas Geiger [2020]. "Towards unsupervised learning of generative models for 3d controllable image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5871–5880.

Lim, Bee, Sanghyun Son, Heewon Kim, Seungjun Nah and Kyoung Mu Lee [2017a]. "Enhanced deep residual networks for single image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144.

Lim, Jaemoon, Minhyeok Heo, Chul Lee and Chang-Su Kim [2017b]. "Contrast enhancement of noisy low-light images based on structure-texture-noise decomposition". In: *Journal of Visual Communication and Image Representation* 45, pp. 107–121.

Lin, Xing, Jinli Suo, Xun Cao and Qionghai Dai [2013]. "Iterative feedback estimation of depth and radiance from defocused images". In: *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV 11*. Springer, pp. 95–109.

Lindell, David B, Dave Van Veen, Jeong Joon Park and Gordon Wetzstein [2022]. "Bacon: Band-limited coordinate networks for multiscale scene representation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16252–16262.

Liu, Ce et al. [2009]. "Beyond pixels: exploring new representations and applications for motion analysis". PhD thesis. Massachusetts Institute of Technology.

Liu, Guilin, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao and Bryan Catanzaro [2018a]. "Image inpainting for irregular holes using partial convolutions". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 85–100.

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

Liu, Yu, Xun Chen, Hu Peng and Zengfu Wang [2017]. "Multi-focus image fusion with a deep convolutional neural network". In: *Information Fusion* 36, pp. 191–207.

Liu, Yu-Lun, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang and Jia-Bin Huang [2020b]. "Single-image HDR reconstruction by learning to reverse the camera pipeline". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1651–1660.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin and Baining Guo [2021]. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022.

Liu, Zhen, Yinglong Wang, Bing Zeng and Shuaicheng Liu [2022]. "Ghost-free high dynamic range imaging with context-aware transformer". In: *European Conference on computer vision*. Springer, pp. 344–360.

Liu, Ziwei, Ping Luo, Xiaogang Wang and Xiaoou Tang [2018b]. "Large-scale celebfaces attributes (celeba) dataset". In: *Retrieved August* 15.2018, p. 11.

Lubin, Jeffrey [1995]. "A visual discrimination model for imaging system design and evaluation". In: *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*. World Scientific, pp. 245–283.

Lugmayr, Andreas, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte and Luc Van Gool [2022]. "Repaint: Inpainting using denoising diffusion probabilistic models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471.

Ma, Kede, Hojatollah Yeganeh, Kai Zeng and Zhou Wang [2015]. "High dynamic range image compression by optimizing tone mapped image quality index". In: *IEEE Trans on Image Proc* 24.10, pp. 3086–97.

Ma, Li, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang and Pedro V Sander [2022]. "Deblur-nerf: Neural radiance fields from blurry images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12861–12870.

Mandl, David, Peter M Roth, Tobias Langlotz, Christoph Ebner, Shohei Mori, Stefanie Zollmann, Peter Mohr and Denis Kalkofen [2021]. "Neural cameras: Learning camera characteristics for coherent mixed reality rendering". In: *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, pp. 508–516.

Mantiuk, Radoslaw, Rafal Mantiuk, Anna Tomaszewska and Wolfgang Heidrich [2009]. "Color correction for tone mapping". In: *Computer Graphics Forum*. Vol. 28. 2, pp. 193–202.

Mantiuk, Rafał, Scott Daly and Louis Kerofsky [2008]. "Display adaptive tone mapping". In: *ACM SIGGRAPH 2008 papers*, pp. 1–10.

Mantiuk, Rafał, Grzegorz Krawczyk, Dorota Zdrojewska, Radosław Mantiuk, Karol Myszkowski and Hans-Peter Seidel [2015a]. *High dynamic range imaging*. na.

Mantiuk, Rafal, Karol Myszkowski and Hans-Peter Seidel [2006]. "A perceptual framework for contrast processing of high dynamic range images". In: *ACM Trans. on Applied Perception (TAP)* 3.3, pp. 286–308.

Mantiuk, Rafal, Karol Myszkowski and Hans-Peter Seidel [2015b]. *High dynamic range imaging*. Wiley Encyclopedia of Electrical and Electronics Engineering.

Mantiuk, Rafał and Hans-Peter Seidel [2008]. "Modeling a generic tone-mapping operator". In: *Computer Graphics Forum*. Vol. 27. 2, pp. 699–708.

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

Mantiuk, Rafal K, Dounia Hammou and Param Hanji [2023]. "HDR-VDP-3: A multimetric for predicting image differences, quality and contrast distortions in high dynamic range and regular content". In: *arXiv preprint arXiv:2304.13625*.

Marnerides, Demetris, Thomas Bashford-Rogers, Jonathan Hatchett and Kurt Debattista [2018]. "Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content". In: *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library, pp. 37–49.

Masia, Belen, Ana Serrano and Diego Gutierrez [2017]. "Dynamic range expansion based on image statistics". In: *Multimedia Tools and Applications* 76, pp. 631–648.

Maximov, Maxim, Kevin Galim and Laura Leal-Taixé [2020]. "Focus on defocus: bridging the synthetic to real domain gap for depth estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1071–1080.

Mertens, Tom, Jan Kautz and Frank Van Reeth [2007]. "Exposure fusion". In: *15th Pacific Conference on Computer Graphics and Applications (PG'07)*. IEEE, pp. 382–390.

Meylan, Laurence, Scott Daly and Sabine Süsstrunk [2006]. "The reproduction of specular highlights on high dynamic range displays". In: *IS&T/SID 14th Color Imaging Conference (CIC)*.

Meylan, Laurence, Scott Daly and Sabine Süsstrunk [2007]. "Tone mapping for high-dynamic range displays". In: *Human Vision and Electronic Imaging XII*. Vol. 6492. SPIE, pp. 370–381.

Michalkiewicz, Mateusz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh and Anders Eriksson [2019]. "Implicit surface representations as layers in neural networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4743–4752.

Mildenhall, Ben, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan and Jonathan T Barron [2022]. "Nerf in the dark: High dynamic range view synthesis from noisy raw images". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16190–16199.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi and Ren Ng [2020]. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *ECCV*.

Mildenhall, Ben, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi and Ren Ng [2021]. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* 65.1, pp. 99–106.

Miller, Gene, Ken Perlin, Mike Chou and Lance Williams [1999]. "Reflection Mapping-1982". In.

Mirza, Mehdi and Simon Osindero [2014]. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784*.

Mitsunaga, Tomoo and Shree K Nayar [1999]. "Radiometric self calibration". In: *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*. Vol. 1. IEEE, pp. 374–380.

Mittal, Anish, Anush Krishna Moorthy and Alan Conrad Bovik [2012]. "No-reference image quality assessment in the spatial domain". In: *IEEE Trans. on Image Processing* 21.12, pp. 4695–4708.

Moeller, Michael, Martin Benning, Carola Schönlieb and Daniel Cremers [2015]. "Variational depth from focus reconstruction". In: *IEEE Transactions on Image Processing* 24.12, pp. 5369–5378.

Mou, Chong, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi and Ying Shan [2024]. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. 5, pp. 4296–4304.

Müller, Thomas, Alex Evans, Christoph Schied and Alexander Keller [2022]. "Instant neural graphics primitives with a multiresolution hash encoding". In: *ACM transactions on graphics (TOG)* 41.4, pp. 1–15.

Nair, Vinod and Geoffrey E Hinton [2010]. "Rectified linear units improve restricted boltzmann machines". In: *Icml*.

Nam, Seonghyeon, Marcus A Brubaker and Michael S Brown [2022]. "Neural image representations for multi-image fusion and layer separation". In: *European conference on computer vision*. Springer, pp. 216–232.

Nguyen-Phuoc, Thu, Chuan Li, Lucas Theis, Christian Richardt and Yong-Liang Yang [2019]. "Hologan: Unsupervised learning of 3d representations from natural images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7588–7597.

Nichol, Alex, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever and Mark Chen [2021]. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models". In: *arXiv preprint arXiv:2112.10741*.

Nichol, Alexander Quinn and Prafulla Dhariwal [2021]. "Improved denoising diffusion probabilistic models". In: *International conference on machine learning*. PMLR, pp. 8162–8171.

Niu, Yuzhen, Jianbin Wu, Wenxi Liu, Wenzhong Guo and Rynson WH Lau [2021]. "Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions". In: *IEEE Transactions on Image Processing* 30, pp. 3885–3896.

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

Pan, Xingang, Bo Dai, Ziwei Liu, Chen Change Loy and Ping Luo [2020]. "Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans". In: *arXiv preprint arXiv:2011.00844*.

Pan, Xingang, Ayush Tewari, Lingjie Liu and Christian Theobalt [2022]. "Gan2x: Non-lambertian inverse rendering of image gans". In: *2022 International Conference on 3D Vision (3DV)*. IEEE, pp. 711–721.

Pan, Xingang, Xudong Xu, Chen Change Loy, Christian Theobalt and Bo Dai [2021a]. "A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis". In: *Advances in Neural Information Processing Systems* 34, pp. 20002–20013.

Pan, Xingang, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy and Ping Luo [2021b]. "Exploiting deep generative prior for versatile image restoration and manipulation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11, pp. 7474–7489.

Panetta, Karen, Landry Kezebou, Victor Oludare, Sos Agaian and Zehua Xia [2021]. "Tmo-net: A parameter-free tone mapping operator using generative adversarial network, and performance benchmarking on large scale hdr dataset". In: *IEEE Access* 9, pp. 39500–39517.

Park, Jeong Joon, Peter Florence, Julian Straub, Richard Newcombe and Steven Lovegrove [2019]. "Deepsdf: Learning continuous signed distance functions for shape representation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174.

Patel, Vaibhav Amit, Purvik Shah and Shanmuganathan Raman [2017]. "A generative adversarial network for tone mapping HDR images". In: *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*. Springer, pp. 220–231.

Peebles, William and Saining Xie [2023]. "Scalable diffusion models with transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205.

Peli, Eli [1990]. "Contrast in complex images". In: *JOSA A* 7.10, pp. 2032–2040.

Peng, Rui, Wangze Xu, Luyang Tang, Jianbo Jiao, Ronggang Wang et al. [2024]. "Structure consistent gaussian splatting with matching prior for few-shot novel view synthesis". In: *Advances in Neural Information Processing Systems* 37, pp. 97328–97352.

Pidhorskyi, Stanislav, Timur Bagautdinov, Shugao Ma, Jason Saragih, Gabriel Schwartz, Yaser Sheikh and Tomas Simon [2022]. "Depth of field aware differentiable rendering". In: *ACM Transactions on Graphics (TOG)* 41.6, pp. 1–18.

"Poly Haven" [2024]. In: Accessed: 2024-10-21.

Poole, Ben, Ajay Jain, Jonathan T Barron and Ben Mildenhall [2022]. "Dreamfusion: Text-to-3d using 2d diffusion". In: *arXiv preprint arXiv:2209.14988*.

Potmesil, Michael and Indranil Chakravarty [1981]. "A lens and aperture camera model for synthetic image generation". In: *ACM SIGGRAPH Computer Graphics* 15.3, pp. 297–305.

Potmesil, Michael and Indranil Chakravarty [1982]. "Synthetic image generation with a lens and aperture camera model". In: *ACM Transactions on Graphics (TOG)* 1.2, pp. 85–108.

Prabhakar, K Ram, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu and Rama Krishna Sai S Gorthi [2021]. "Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4875–4885.

Radford, Alec, Luke Metz and Soumith Chintala [2015]. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*.

Raman, Shanmuganathan and Subhasis Chaudhuri [2009]. "Bilateral Filter Based Compositing for Variable Exposure Photography." In: *Eurographics (short papers)*, pp. 1–4.

Ramasubramanian, Mahesh, Sumanta N Pattanaik and Donald P Greenberg [1999]. "A perceptually based physical error metric for realistic image synthesis". In: *Proc ACM SIGGRAPH*, pp. 73–82.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu and Mark Chen [2022]. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* 1.2, p. 3.

Rana, Aakanksha, Praveer Singh, Giuseppe Valenzise, Frederic Dufaux, Nikos Komodakis and Aljosa Smolic [2019]. "Deep tone mapping operator for high dynamic range images". In: *IEEE Trans. on Image Processing* 29, pp. 1285–1298.

Reinhard, Erik and Kate Devlin [2005]. "Dynamic range reduction inspired by photoreceptor physiology". In: *IEEE transactions on visualization and computer graphics* 11.1, pp. 13–24.

Reinhard, Erik, Michael Stark, Peter Shirley and James Ferwerda [2002]. "Photographic tone reproduction for digital images". In: *Proc ACM SIGGRAPH*, pp. 267–276.

Reinhard, Erik, Greg Ward, Sumanta Pattanaik, Wolfgang Heidrich, Paul Debevec and Karol Myszkowski [2010]. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Elsevier Science.

Reiser, Christian, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron and Peter Hedman [2023]. "Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes". In: *ACM Transactions on Graphics (TOG)* 42.4, pp. 1–12.

Rempel, Allan G, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead and Greg Ward [2007]. "Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs". In: *ACM transactions on graphics (TOG)* 26.3, 39–es.

Robertson, Mark A, Sean Borman and Robert L Stevenson [2003]. "Estimation-theoretic approach to dynamic range enhancement using multiple exposures". In: *Journal of electronic imaging* 12.2, pp. 219–228.

Robson, John G [1966]. "Spatial and temporal contrast-sensitivity functions of the visual system". In: *JOSA* 56.8, pp. 1141–1142.

Roich, Daniel, Ron Mokady, Amit H Bermano and Daniel Cohen-Or [2022]. "Pivotal tuning for latent-based editing of real images". In: *ACM Transactions on graphics (TOG)* 42.1, pp. 1–13.

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

Ronneberger, Olaf, Philipp Fischer and Thomas Brox [2015]. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, pp. 234–241.

Rubinstein, Michael, Diego Gutierrez, Olga Sorkine and Ariel Shamir [2010]. "A comparative study of image retargeting". In: *Proc ACM SIGGRAPH Asia*, pp. 1–10.

Rückert, Darius, Linus Franke and Marc Stamminger [2022]. "Adop: Approximate differentiable one-pixel point rendering". In: *ACM Transactions on Graphics (ToG)* 41.4, pp. 1–14.

Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein and Kfir Aberman [2023]. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510.

Saharia, Chitwan, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet and Mohammad Norouzi [2022a]. "Palette: Image-to-image diffusion models". In: *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans et al. [2022b]. "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in neural information processing systems* 35, pp. 36479–36494.

Santos, Marcel Santana, Tsang Ing Ren and Nima Khademi Kalantari [2020]. "Single image HDR reconstruction using a CNN with masked features and perceptual loss". In: *arXiv preprint arXiv:2005.07335*.

Sauer, Axel, Katja Schwarz and Andreas Geiger [2022]. "Stylegan-xl: Scaling stylegan to large diverse datasets". In: *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10.

Schonberger, Johannes L and Jan-Michael Frahm [2016]. "Structure-from-motion revisited". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman et al. [2022]. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in neural information processing systems* 35, pp. 25278–25294.

Schwarz, Katja, Yiyi Liao, Michael Niemeyer and Andreas Geiger [2020]. "Graf: Generative radiance fields for 3d-aware image synthesis". In: *Advances in Neural Information Processing Systems* 33, pp. 20154–20166.

Scofield, Cary [1992]. "212-d depth-of-field simulation for computer animation". In: *Graphics Gems III (IBM Version)*. Elsevier, pp. 36–38.

Sen, Pradeep, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman and Eli Shechtman [2012]. "Robust patch-based hdr reconstruction of dynamic scenes." In: *ACM Trans. Graph.* 31.6, pp. 203–1.

Shan, Qi, Jiaya Jia and Michael S Brown [2009]. "Globally optimized linear windowed tone mapping". In: *IEEE Trans Visual and Comp Graph* 16.4, pp. 663–675.

Shao, Jiahao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi and Yiyi Liao [2024]. "Learning temporally consistent video depth from video diffusion priors". In: *arXiv preprint arXiv:2406.01493*.

Sharma, Prafull, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, Bill Freeman and Mark Matthews [2024]. "Alchemist: Parametric Control

of Material Properties with Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24130–24141.

Shen, Xuelun, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen and Cheng Wang [2024]. "Gim: Learning generalizable image matcher from internet videos". In: *arXiv preprint arXiv:2402.11095*.

Sheng, Yichen, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin and Bedrich Benes [2024]. "Dr. bokeh: differentiable occlusion-aware bokeh rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4515–4525.

Shibata, Takashi, Masayuki Tanaka and Masatoshi Okutomi [2016]. "Gradient-domain image reconstruction framework with intensity-range and base-structure constraints". In: *CVPR*, pp. 2745–2753.

Si, Haozhe, Bin Zhao, Dong Wang, Yunpeng Gao, Mulin Chen, Zhigang Wang and Xuelong Li [2023]. "Fully self-supervised depth estimation from defocus clue". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9140–9149.

Silberman, Nathan, Derek Hoiem, Pushmeet Kohli and Rob Fergus [2012]. "Indoor segmentation and support inference from rgbd images". In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. Springer, pp. 746–760.

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

Singer, Uriel, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni et al. [2022]. "Make-a-video: Text-to-video generation without text-video data". In: *arXiv preprint arXiv:2209.14792*.

Sitzmann, Vincent, Julien Martel, Alexander Bergman, David Lindell and Gordon Wetzstein [2020]. "Implicit neural representations with periodic activation functions". In: *Advances in neural information processing systems* 33, pp. 7462–7473.

Sitzmann, Vincent, Michael Zollhöfer and Gordon Wetzstein [2019]. "Scene representation networks: Continuous 3d-structure-aware neural scene representations". In: *Advances in Neural Information Processing Systems* 32.

Song, Jiaming, Chenlin Meng and Stefano Ermon [2020a]. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502*.

Song, Jou Won, Ye-In Park, Kyeongbo Kong, Jaeho Kwak and Suk-Ju Kang [2022]. "Selective transhdr: Transformer-based selective hdr imaging using ghost region mask". In: *European Conference on Computer Vision*. Springer, pp. 288–304.

Song, Wei, Yan Wang, Dongmei Huang, Antonio Liotta and Cristian Perra [2020b]. "Enhancement of underwater images with statistical model of background light and optimization of transmission map". In: *IEEE Trans. on Broadcasting* 66.1, pp. 153–169.

Su, Chien-Chuan, Ren Wang, Hung-Jin Lin, Yu-Lun Liu, Chia-Ping Chen, Yu-Lin Chang and Soo-Chang Pei [2021]. "Explorable Tone Mapping Operators". In: *ICPR*, pp. 10320–10326.

Sun, Cheng, Min Sun and Hwann-Tzong Chen [2022]. "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5459–5469.

Suwajanakorn, Supasorn, Carlos Hernandez and Steven M Seitz [2015]. "Depth from focus with your mobile phone". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3497–3506.

Szabó, Attila, Givi Meishvili and Paolo Favaro [2019]. "Unsupervised generative 3d shape learning from natural images". In: *arXiv preprint arXiv:1910.00287*.

Tancik, Matthew, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron and Ren Ng [2020]. "Fourier features let networks learn high frequency functions in low dimensional domains". In: *Advances in neural information processing systems* 33, pp. 7537–7547.

Tang, Shitao, Fuyang Zhang, Jiacheng Chen, Peng Wang and Yasutaka Furukawa [2023]. "MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion". In: *arXiv*.

Tang, Zhenggang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing and Zhicheng Yan [2024]. "MV-DUSt3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds". In: *arXiv preprint arXiv:2412.06974*.

Tel, Steven, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte and Dominique Ginhac [2023]. "Alignment-free hdr deghosting with semantics consistent transformer". In: *arXiv preprint arXiv:2305.18135*.

Trentacoste, M., R. Mantiuk, W. Heidrich and F. Dufrot [2012]. "Unsharp Masking, Countershading and Halos: Enhancements or Artifacts?" In: *Proc. Eurographics*, to appear.

Tumblin, Jack and Greg Turk [1999]. "LCIS: A boundary hierarchy for detail-preserving contrast reduction". In: *Proc ACM SIGGRAPH*, pp. 83–90.

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

Ulyanov, Dmitry, Andrea Vedaldi and Victor Lempitsky [2018]. "Deep image prior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454.

Vahdat, Arash, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis et al. [2022]. "Lion: Latent point diffusion models for 3d shape generation". In: *Advances in Neural Information Processing Systems* 35, pp. 10021–10039.

Verbin, Dor, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron and Pratul P Srinivasan [2022]. "Ref-nerf: Structured view-dependent appearance for neural radiance fields". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5481–5490.

Vinker, Yael, Inbar Huberman-Spiegelglas and Raanan Fattal [2021]. "Unpaired Learning for High Dynamic Range Image Tone Mapping". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14657–14666.

Voynov, Andrey, Amir Hertz, Moab Arar, Shlomi Fruchter and Daniel Cohen-Or [2024]. "Curved Diffusion: A Generative Model with Optical Geometry Control". In: *European Conference on Computer Vision*. Springer, pp. 149–164.

Wang, NingHsu, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen and Kevin Jou [2021]. "Bridging unsupervised and supervised depth from focus via all-in-focus supervision". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12621–12631.

Wang, Peng, Lingzhe Zhao, Ruijie Ma and Peidong Liu [2023c]. "Bad-nerf: Bundle adjusted deblur neural radiance fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4170–4179.

Wang, Qianqian, Zhengqi Li, David Salesin, Noah Snavely, Brian Curless and Janne Kontkanen [2022b]. "3D moments from near-duplicate photos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3906–3915.

Wang, Ruicheng, Jianfeng Xiang, Jiaolong Yang and Xin Tong [2024b]. "Diffusion Models are Geometry Critics: Single Image 3D Editing Using Pre-Trained Diffusion Priors". In: arXiv: 2403.11503 [cs.CV].

Wang, Shuzhe, Vincent Leroy, Yohann Cabon, Boris Chidlovskii and Jerome Revaud [2024c]. "Dust3r: Geometric 3d vision made easy". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709.

Wang, Xiaolong, Ross Girshick, Abhinav Gupta and Kaiming He [2018a]. "Non-local neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803.

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

Wang, Yuehao, Chaoyi Wang, Bingchen Gong and Tianfan Xue [2024d]. "Bilateral guided radiance field processing". In: *ACM Transactions on Graphics (TOG)* 43.4, pp. 1–13.

Wang, Zhixiang, Baiang Li, Jian Wang, Yu-Lun Liu, Jinwei Gu, Yung-Yu Chuang and Shin'ichi Satoh [2024e]. "Matting by generation". In: *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11.

Wang, Zhou, Alan C Bovik, Hamid R Sheikh and Eero P Simoncelli [2004]. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4, pp. 600–612.

Watson, Andrew B [1989]. "Receptive fields and visual representations". In: *Human Vision, Visual Processing, and Digital Display*. Vol. 1077. Proc SPIE, pp. 190–197.

Watson, Andrew B [1993]. "Visually optimal DCT quantization matrices for individual images". In: *[Proceedings] DCC93: Data Compression Conference*. IEEE, pp. 178–187.

Watson, Andrew B and Joshua A Solomon [1997]. "Model of visual contrast gain control and pattern masking". In: *JOSA A* 14.9, pp. 2379–2391.

Watson, Daniel, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi and Mohammad Norouzi [2022]. "Novel view synthesis with diffusion models". In: *arXiv preprint arXiv:2210.04628*.

Weng, Juyang, Paul Cohen, Marc Herniou et al. [1992]. "Camera calibration with distortion models and accuracy evaluation". In: *IEEE Transactions on pattern analysis and machine intelligence* 14.10, pp. 965–980.

Wewer, Christopher, Kevin Raj, Eddy Ilg, Bernt Schiele and Jan Eric Lenssen [2024]. "latentSplat: Autoencoding Variational Gaussians for Fast Generalizable 3D Reconstruction". In: *arXiv*.

Won, Changyeon and Hae-Gon Jeon [2022]. "Learning depth from focus in the wild". In: *European Conference on Computer Vision*. Springer, pp. 1–18.

Wu, Shangzhe, Christian Rupprecht and Andrea Vedaldi [2020]. "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1–10.

Wu, Shangzhe, Jiarui Xu, Yu-Wing Tai and Chi-Keung Tang [2018]. "Deep high dynamic range imaging with large foreground motions". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 117–132.

Wu, Zijin, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao and Weicai Zhong [2022]. "Dof-nerf: Depth-of-field meets neural radiance fields". In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1718–1729.

Xia, Weihao, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou and Ming-Hsuan Yang [2022]. "Gan inversion: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 45.3, pp. 3121–3138.

Xiao, Jianxiong, James Hays, Krista A Ehinger, Aude Oliva and Antonio Torralba [2010]. "Sun database: Large-scale scene recognition from abbey to zoo". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 3485–3492.

Xie, Shaoan, Zhifei Zhang, Zhe Lin, Tobias Hinz and Kun Zhang [2023]. "Smartbrush: Text and shape guided object inpainting with diffusion model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22428–22437.

Xiong, Pengfei and Yu Chen [2021]. "Hierarchical fusion for practical ghost-free high dynamic range imaging". In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4025–4033.

Xu, Kai, Zhikang Zhang and Fengbo Ren [2018]. "Lapran: A scalable laplacian pyramid reconstructive adversarial network for flexible compressive sensing reconstruction". In: *ECCV*, pp. 485–500.

Xu, Ke, Gerhard Petrus Hancke and Rynson WH Lau [2023a]. "Learning image harmonization in the linear color space". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12570–12579.

Xu, Linning, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič et al. [2023b]. "VR-NeRF: High-fidelity virtualized walkable spaces". In: *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12.

Yan, Qingsen, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun and Yanning Zhang [2023a]. "A unified HDR imaging method with pixel and patch level". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22211–22220.

Yan, Qingsen, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid and Yanning Zhang [2019]. "Attention-guided network for ghost-free high dynamic range imaging". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1751–1760.

Yan, Qingsen, Tao Hu, Yuan Sun, Hao Tang, Yu Zhu, Wei Dong, Luc Van Gool and Yanning Zhang [2023b]. "Towards high-quality hdr deghosting with conditional diffusion models". In: *IEEE Transactions on Circuits and Systems for Video Technology*.

Yan, Qingsen, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi and Yanning Zhang [2020]. "Deep HDR imaging via a non-local network". In: *IEEE Transactions on Image Processing* 29, pp. 4308–4322.

Yan, Qingsen, Song Zhang, Weiye Chen, Hao Tang, Yu Zhu, Jinqiu Sun, Luc Van Gool and Yanning Zhang [2023c]. "Smae: Few-shot learning for hdr deghosting with saturation-aware masked autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5775–5784.

Yan, Zhiwen, Weng Fei Low, Yu Chen and Gim Hee Lee [2024]. "Multi-scale 3d gaussian splatting for anti-aliased rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20923–20931.

Yang, Fengting, Xiaolei Huang and Zihan Zhou [2022]. "Deep depth from focus with differential focus volume". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12642–12651.

Yang, Xin, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei and Rynson WH Lau [2018]. "Image correction via deep reciprocating HDR transformation". In: *CVPR*, pp. 1798–1807.

Yasarla, Rajeev, Vishwanath A Sindagi and Vishal M Patel [2020]. "Syn2real transfer learning for image deraining using gaussian processes". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2726–2736.

Ye, Qian, Jun Xiao, Kin-man Lam and Takayuki Okatani [2021]. "Progressive and selective fusion network for high dynamic range imaging". In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5290–5297.

Yeganeh, Hojatollah and Zhou Wang [2012]. "Objective quality assessment of tone-mapped images". In: *IEEE Trans. on Image Processing* 22.2, pp. 657–667.

Yi, Shinyoung, Daniel S. Jeon, Ana Serrano, Se-Yoon Jeong, Hui-Yong Kim, Diego Gutierrez and Min H. Kim [2021]. "Modeling Surround-aware Contrast Sensitivity". In: *Eurographics Symposium on Rendering*.

Yifan, Wang, Felice Serena, Shihao Wu, Cengiz Öztireli and Olga Sorkine-Hornung [2019]. "Differentiable surface splatting for point-based geometry processing". In: *ACM Transactions On Graphics (TOG)* 38.6, pp. 1–14.

Yoshida, Akiko, Volker Blanz, Karol Myszkowski and Hans-Peter Seidel [2005]. "Perceptual evaluation of tone mapping operators with real-world scenes". In: *Human Vision and Electronic Imaging X*. Vol. 5666. Proc SPIE, pp. 192–203.

Yu, Alex, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht and Angjoo Kanazawa [2023]. "Plenoxels: Radiance fields without neural networks". In.

Yu, Alex, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng and Angjoo Kanazawa [2021a]. "Plenoctrees for real-time rendering of neural radiance fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5752–5761.

Yu, Hanning, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou and Chunxia Xiao [2021b]. "Luminance attentive networks for HDR image and panorama reconstruction". In: *Computer Graphics Forum*. Vol. 40. 7. Wiley Online Library, pp. 181–192.

Yu, Jiahui, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu and Thomas S Huang [2018]. "Generative image inpainting with contextual attention". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514.

Yu, Zehao, Anpei Chen, Binbin Huang, Torsten Sattler and Andreas Geiger [2024]. "Mip-splatting: Alias-free 3d gaussian splatting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456.

Yuan, Ye, Jiaming Song, Umar Iqbal, Arash Vahdat and Jan Kautz [2023]. "Physdiff: Physics-guided human motion diffusion model". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16010–16021.

Zamir, Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang and Ling Shao [2020]. "Cycleisp: Real image restoration via improved data synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2696–2705.

Zeng, Wenjun, Scott Daly and Shawmin Lei [2000]. "Point-wise extended visual masking for JPEG-2000 image compression". In: *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*. Vol. 1. IEEE, pp. 657–660.

Zhang, He, Vishwanath Sindagi and Vishal M Patel [2019a]. "Image de-raining using a conditional generative adversarial network". In: *IEEE transactions on circuits and systems for video technology* 30.11, pp. 3943–3956.

Zhang, Lvmin, Anyi Rao and Maneesh Agrawala [2023a]. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847.

Zhang, Ning, Chao Wang, Yang Zhao and Ronggang Wang [2019b]. "Deep tone mapping network in HSV color space". In: *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, pp. 1–4.

Zhang, Ning, Yuyao Ye, Yang Zhao and Ronggang Wang [2023b]. "Revisiting the stack-based inverse tone mapping". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9162–9171.

Zhang, Qiang and Martin D Levine [2016]. "Robust multi-focus image fusion using multi-task sparse representation and spatial context". In: *IEEE Transactions on Image Processing* 25.5, pp. 2045–2058.

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman and Oliver Wang [2018]. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zhang, Xinjie, Xingtong Ge, Tongda Xu, Dailan He, Yan Wang, Hongwei Qin, Guo Lu, Jing Geng and Jun Zhang [2024a]. "Gaussianimage: 1000 fps image representation and compression by 2d gaussian splatting". In: *European Conference on Computer Vision*. Springer, pp. 327–345.

Zhang, Xuaner, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang and Ren Ng [2019c]. "Synthetic defocus and look-ahead autofocus for casual videography". In: *ACM Transactions on Graphics (TOG)* 38.4, pp. 1–16.

Zhang, Yang and TO Aydın [2021]. "Deep HDR estimation with generative detail reconstruction". In: *Computer graphics forum*. Vol. 40. 2. Wiley Online Library, pp. 179–190.

Zhang, Zhilu, Haoyu Wang, Shuai Liu, Xiaotao Wang, LEI LEI and Wangmeng Zuo [2024b]. "Self-Supervised High Dynamic Range Imaging with Multi-Exposure Images in Dynamic Scenes". In: *The Twelfth International Conference on Learning Representations*.

Zhao, Lingzhe, Peng Wang and Peidong Liu [2024]. "Bad-gaussians: Bundle adjusted deblur gaussian splatting". In: *European Conference on Computer Vision*. Springer, pp. 233–250.

Zhao, Wenda, Fei Wei, You He and Huchuan Lu [2022]. "United defocus blur detection and deblurring via adversarial promoting learning". In: *European Conference on Computer Vision*. Springer, pp. 569–586.

Zhou, Yuyan, Dong Liang, Songcan Chen, Sheng-Jun Huang, Shuo Yang and Chongyi Li [2023]. "Improving lens flare removal with general-purpose pipeline and multiple light sources recovery". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12969–12979.

Zhu, Jiapeng, Yujun Shen, Deli Zhao and Bolei Zhou [2020]. "In-domain gan inversion for real image editing". In: *European conference on computer vision*. Springer, pp. 592–608.

Zhu, Jun-Yan, Philipp Krähenbühl, Eli Shechtman and Alexei A Efros [2016]. "Generative visual manipulation on the natural image manifold". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, pp. 597–613.

Zhu, Jun-Yan, Taesung Park, Phillip Isola and Alexei A Efros [2017a]. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zhu, Jun-Yan, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang and Eli Shechtman [2017b]. "Multimodal Image-to-Image Translation by Enforcing Bi-Cycle Consistency". In: *Advances in neural information processing systems*, pp. 465–476.

Zhu, Ruoxi, Shusong Xu, Peiye Liu, Sicheng Li, Yanheng Lu, Dimin Niu, Zihao Liu, Zihao Meng, Zhiyong Li, Xinhua Chen and Yibo Fan [2024]. "Zero-Shot Structure-Preserving Diffusion Model for High Dynamic Range Tone Mapping". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26130–26139.

Zwicker, Matthias, Hanspeter Pfister, Jeroen Van Baar and Markus Gross [2001a]. "EWA volume splatting". In: *Proceedings Visualization, 2001. VIS'01.* IEEE, pp. 29–538.

Zwicker, Matthias, Hanspeter Pfister, Jeroen Van Baar and Markus Gross [2001b]. "Surface splatting". In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 371–378.