INSTANT HDR-NERF: FAST LEARNING OF HIGH DYNAMIC RANGE VIEW SYNTHESIS WITH UNKNOWN EXPOSURE SETTINGS

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Nam Nguyen

June 2024
COMMITTEE MEMBERSHIP

TITLE: Instant HDR-NeRF: Fast Learning of High Dynamic Range View Synthesis With Unknown Exposure Settings

AUTHOR: Nam Nguyen

DATE SUBMITTED: June 2024

COMMITTEE CHAIR: Jonathan Ventura, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Lubomir Stanchev, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: John Seng, Ph.D.
Professor of Computer Science
ABSTRACT

Instant HDR-NeRF: Fast Learning of High Dynamic Range View Synthesis With
Unknown Exposure Settings

Nam Nguyen

We propose Instant High Dynamic Range Neural Radiance Fields (Instant HDR-
NeRF), a method of learning high dynamic range (HDR) view synthesis from a set
of low dynamic range (LDR) views with unknown and varying exposure and white
balance in as little as minutes. Our method can render novel HDR views without
ground-truth supervision, and novel LDR views in different exposure settings, includ-
ing those that match the ground-truth LDR views. The key to our method is to
model the physical process of the camera with two implicit MLPs: a radiance field
and a monotonically increasing tone-mapper. Built upon Instant Neural Graphics
Primitives (Instant-NGP), the radiance field encodes the scene geometry and radi-
ance (from 0 to $\infty$), and outputs the densities and the radiance at locations along the
camera ray. The monotonically increasing tone-mapper models the camera response
function (CRF) where the radiance hits on the camera sensor and becomes a pixel
value (from 0 to 255). The radiance at each location is combined with the learnable
exposure parameters, optimized separately for each color band and for each image. A
quantitative evaluation on benchmark datasets shows that our method outperforms
prior HDR novel view synthesis methods in LDR rendering quality and training speed.
To best of our knowledge, our method is also the first HDR radiance field that suc-
cessfully recovers the ground-truth CRF with a low average error rate of 3.70%, while
co-learning geometry, radiance, and exposures all at the same time through implicit
functions. In practical applications, our method can produce high-fidelity 3D re-
construction of real-world scenes from images of varying exposure settings, which is
particularly useful for casual capturing, where fixed settings aren’t guaranteed. The
tone-mapper MLP can be easily controlled to simulate auto-exposure effects, making it useful in filming and video games. Furthermore, the HDR radiance maps produced by our method can be edited and tone-mapped according to user preferences.
ACKNOWLEDGMENTS

Thanks to:

• My parents, who has been supporting me throughout my academic career in a new country. I wouldn’t be here without them.

• Professor Jonathan Ventura, for all his advice, time, and patience on this thesis and beyond. You ignited my passion for doing research in computer vision, gave me many opportunities to shine and expand my knowledge of the field. Your mentoring and constant support have impacted me not just as a researcher, but also as a person. Anytime you need someone to collaborate on your big ideas, I am always here, just message me.

• My committee members, Professor Lubomir Stanchev, and Professor John Seng for their time and wisdom.

• My friend Edward, who has helped me with the baseline comparisons of this paper, and collecting real-world data.

• My family and friends who motivate and support me throughout my journey.

• Anyone who are interested in my work, who may or may not come to my defense. I am happy to receive any feedback, comments, and questions from you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Related Work &amp; Background</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Novel View Synthesis</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Neural Radiance Field</td>
<td>6</td>
</tr>
<tr>
<td>2.3 High Dynamic Range Imaging</td>
<td>8</td>
</tr>
<tr>
<td>2.4 High Dynamic Range Novel View Synthesis</td>
<td>9</td>
</tr>
<tr>
<td>2.5 NeRF Acceleration Techniques</td>
<td>10</td>
</tr>
<tr>
<td>2.5.1 Instant Neural Graphic Primitives</td>
<td>10</td>
</tr>
<tr>
<td>2.5.2 Importance Sampling</td>
<td>11</td>
</tr>
<tr>
<td>3 Methodologies</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Scene Representation</td>
<td>12</td>
</tr>
<tr>
<td>3.2 Learned Tone-Mapping and Exposures</td>
<td>13</td>
</tr>
<tr>
<td>3.3 Neural Rendering</td>
<td>14</td>
</tr>
<tr>
<td>3.4 Optimization</td>
<td>15</td>
</tr>
<tr>
<td>3.4.1 Color Reconstruction Loss</td>
<td>15</td>
</tr>
<tr>
<td>3.4.2 Unit Exposure Loss</td>
<td>15</td>
</tr>
<tr>
<td>3.5 Initialization of Exposure Parameters</td>
<td>16</td>
</tr>
<tr>
<td>4 Experiments</td>
<td>17</td>
</tr>
<tr>
<td>4.1 Implementation Details</td>
<td>17</td>
</tr>
</tbody>
</table>
4.2 Methods Compared ............................................. 18
4.3 Evaluation Datasets ......................................... 18
4.4 Evaluation Metrics ........................................... 19
   4.4.1 LDR View Synthesis Quality ............................. 19
   4.4.2 HDR View Synthesis Quality ............................. 21
   4.4.3 CRF Estimation Error ................................. 21
5 Evaluation Results ............................................. 22
   5.1 HDR-NeRF Real Dataset .................................. 22
   5.2 HDR-Plenoxels Real and Synthetic Datasets ................. 26
6 Ablation Studies ............................................... 30
   6.1 Monotonic Constraint on the Tone Mapper ................. 30
   6.2 Explicit Versus Implicit CRF Representation ............... 31
   6.3 Tone Mapping Before Versus After Volume Rendering .... 31
7 Real World Applications ....................................... 34
   7.1 Scenes of Varying White-Balance .......................... 34
   7.2 Scenes of Varying Exposure and White-Balance ............ 35
   7.3 Simulation of Auto-Exposure Effects ....................... 37
8 Limitations ..................................................... 39
   8.1 Extreme Radiance Values .................................. 39
   8.2 Learned Exposure Parameters with Nondeterministic Scale .... 40
9 Future Work .................................................... 41
   9.1 Local Tone Mapping ........................................ 41
   9.2 HDR Novel View Synthesis from Photo Collections in the Wild ... 42
   9.3 Integration to Nerfstudio ................................ 42
10 Conclusion .................................................... 44
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Results of quantitative evaluation on the real scenes from HDR-NeRF. Values are the average of the metrics for the test data of each scene. We color code the cells as best and second best.</td>
<td>23</td>
</tr>
<tr>
<td>5.2</td>
<td>Results of quantitative evaluation on the synthetic scenes from HDR-Plenoxels.</td>
<td>26</td>
</tr>
<tr>
<td>5.3</td>
<td>Results of quantitative evaluation on the real scenes from HDR-Plenoxels.</td>
<td>26</td>
</tr>
<tr>
<td>6.1</td>
<td>Quantitative results of LDR rendering accuracy and CRF estimation error of our ablation studies. Although, on average, our method achieves second-best in LDR rendering accuracy on the HDR-NeRF real dataset, our HDR-tonemapped views (Fig. 6.1b) look better as the CRFs are modeled more accurately.</td>
<td>32</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>When capturing under auto exposure and auto white-balance, input training views can have inconsistent exposure settings, resulting in color artifacts on the output novel view from NeRF.</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Our method recovers high dynamic range radiance fields from (a) LDR views with unknown and varying exposure and white balance. Our method can render (a) novel LDR views with different exposure settings and (b) novel HDR views.</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>Novel view synthesis involves two stages: (1) building a scene representation from the input images and their poses, and (2) rendering novel views from new poses using the built representation. Figure is adapted from [36].</td>
<td>6</td>
</tr>
<tr>
<td>2.2</td>
<td>NeRF rendering pipeline. The method samples a set of point of the ray R, and obtain the predicted radiances and densities of these points using a neural network. Then, NeRF applies a differentiable volumetric rendering function on these predictions to get the final radiance and density of ray R. Figure is adapted from [30].</td>
<td>7</td>
</tr>
<tr>
<td>3.1</td>
<td>Our model of the camera pipeline consists of two modules: a HDR radiance field models the scene radiance and a monotonic tone-mapper MLP models the CRF.</td>
<td>12</td>
</tr>
<tr>
<td>3.2</td>
<td>Without initializing exposure parameters to plausible solutions, our method can sometimes interpret the change of exposure as view-dependent specularities. We enhance the image contrast to make this issue easier to visualize.</td>
<td>16</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparison of LDR renderings from the computer scene from the HDR-NeRF real dataset. The results from our method have sharper details and more closely resemble the ground truth.</td>
<td>23</td>
</tr>
<tr>
<td>5.2</td>
<td>Ground-truth LDR views of varying exposure and our rendered LDR views on HDR-NeRF real scene. Our PSNR values are on the bottom right of the error maps.</td>
<td>24</td>
</tr>
</tbody>
</table>
5.3 By sampling within the range of the learned exposure parameter vectors, our method can render LDR novel views under different exposure times (a-f) that are not included in the training set.

5.4 Ground-truth LDR views (top row) and our tone-mapped HDR views (bottom row) from the HDR-NeRF dataset. Our tone-mapped HDR views reveal under-exposed and over-exposed regions of the scenes.

5.5 Comparisons between our estimated CRFs and the ground-truth CRF calibrated using Debevec and Malik method [9] on the HDR-NeRF real dataset.

5.6 Comparison of LDR renderings from the room scene from the HDR-Plenoxels synthetic dataset.

5.7 Ground-truth LDR views and our rendered LDR views on HDR-Plenoxels synthetic scene. Our PSNR values are on the bottom right of the error maps.

5.8 By sampling within the range of the learned exposure parameter vectors, our method can render LDR novel views under different exposures and white-balance (a-f) that are not included in the training set.

5.9 Ground truth LDR views (top row) and our tone-mapped HDR views (bottom row) from the HDR-Plenoxels dataset.

5.10 Comparisons between our estimated CRFs and the ground-truth CRF calibrated using Debevec and Malik method [9] on the HDR-Plenoxels real dataset.

5.11 Training time comparisons between our method and (a) HDR-NeRF and (b) HDR-Plenoxels. Our method shows the fastest training time among the three.

6.1 Qualitative results on the computer scene with/without enforcing monotonic constraint on the tone mapper. The ground-truth CRF was calibrated using Debevec and Malik method [9].

6.2 Qualitative comparisons among different model configurations mentioned in our ablation studies.

7.1 LDR spiral views of the dresser scene between NeRF (top) and our method (bottom).
7.2 Results for downtown scene, captured with a DSLR with exposure and white-balance bracketing. Our HDR results show details in both the bright coffee logo and dark coffee tables. .................................................................................. 35

7.3 Results for window seat scene, captured with a smartphone in auto-exposure mode. Our HDR results show the details of the bright paper and dark book cover. .................................................................................. 36

7.4 Our method’s simulation of auto-exposure effects. When the camera traverses to over-exposed regions of the downtown scene, our method can dynamically decrease the exposure of the LDR view to reveal the details of these regions. .................................................................................. 38
Chapter 1

INTRODUCTION

Virtual Reality (VR) is becoming increasingly popular, but we still lack realistic and high-quality 3D content to fully leverage the immersive experiences of the VR headsets. Generating 3D reconstructions using hardware systems is expensive and time-consuming, so the computer vision community has been exploring algorithmic solutions to generate 3D content from 2D images and video, the vast majority we possess. These algorithmic solutions are all solve a common problem in computer vision: novel view synthesis. Novel view synthesis is the process of using images and their camera poses to build a scene representation that can render novel views.

Recently, deep learning technologies have been widely applied to novel view synthesis problems [30, 32, 20, 3]. One pioneering method in this domain is neural radiance field (NeRF) [30] where scene geometry and volume are encoded in an implicit neural network. While producing realistic 3D scene reconstruction from posed 2D images, NeRF has several limitations. First, NeRF does not adapt to different variations of exposure settings among input viewpoints, which commonly happens in casual data capture, where input images are taken under auto exposure and auto white balance mode. An illustration of this limitation is shown in Fig. 1.1, where the output novel view from NeRF shows incorrect color artifacts when the input views have different white-balance settings. Secondly, NeRF produced novel views limited to a low dynamic range. Real-world scenes exceed the dynamic range of the camera, so it is desirable to reconstruct HDR scenes from the LDR views.
These two limitations of NeRF serve as the main motivations for our work. We are seeking to develop a novel view synthesis method that (1) handles inconsistencies of exposure and white balance in casual capturing and (2) learns the original scene radiance and camera exposure parameters to allow HDR rendering. To handle viewpoints of inconsistent exposure settings, we need to learn the exposure settings per viewpoint and take these settings into account in the rendering process. To reconstruct HDR views, we need to recover the camera response function (CRF) that maps the brightness of the pixel to the original radiance value of the scene. The CRF should be monotonically increasing, meaning that a higher image pixel corresponds to a higher radiance. Previous HDR view synthesis methods using LDR views, including HDR-NeRF [16], and HDR-Plenoxels [17], attempt to recover the CRF without guaranteeing that it is monotonically increasing. In our work, we constrain the CRF to be monotonic to ensure that the appearance of the HDR reconstruction matches that of the LDR views.

HDR-NeRF[16], while generating high-quality view synthetic results, is slow to train as it is built upon the vanilla NeRF[30] architecture. On the other hand, HDR-
Plenoxels[17] archives faster training time by extending from a well-optimized voxel-based radiance field method called Plenoxels[12]. However, Plenoxels [12] has been outperformed in both training time and rendering accuracy by other recent methods including Instant-NGP [32].

We propose a novel method called Instant HDR-NeRF to recover the HDR neural radiance field with unknown exposures and varying from LDR images. In this work, the camera processing pipeline is modeled by two implicit MLPs: a radiance field that encodes the scene radiance, and a tone-mapper that models the CRF, mapping the brightness of the pixel to the original radiance value of the scene. The scene radiance at each location is combined with the learnable exposure parameters, optimized separately for each color and for each viewpoint. In this work, we leverage the idea of monotonic neural networks [15] to ensure the learned CRF is monotonically increasing. We incorporate Instant-NGP primitives [32] and NerfAcc[22] for even faster training and rendering time than prior state-of-the-art.

To evaluate our method, we use datasets from HDR-NeRF [16] and HDR-Plenoxels [17]. We compare our method against HDR-NeRF on scenes from HDR-NeRF datasets, and we compare our method against HDR-Plenoxels on scenes from HDR-Plenoxels datasets. We provide quantitative and qualitative results and ablation studies to justify our main technical contributions. In LDR view rendering, our method outperformed baseline methods on average. In HDR view rendering, our method can reveal the under- and over-exposed regions in the scenes.

Our contributions are summarized as follows:

1. We propose a novel method to learn HDR neural radiance fields from LDR images of unknown and varying exposure and white balance.
2. We introduce a method to ensure the learned tone-mapping function is monotonically increasing.

3. We outperform previous methods in LDR rendering accuracy.

4. We correctly models the ground-truth CRF calibrated by method of Debevec and Malik with the low error rate of 3.70%, while co-learning geometry, radiance, and exposures all at the same time through implicit functions.

![Figure 1.2: Our method recovers high dynamic range radiance fields from (a) LDR views with unknown and varying exposure and white balance. Our method can render (a) novel LDR views with different exposure settings and (b) novel HDR views.](image)
Instant HDR-NeRF combines different concepts from multiple areas. Our method builds on optimized primitives from Instant-NGP [32] and efficient sampling from NerfAcc[22] for fast and high-quality view synthesis. We incorporate HDR-NeRF’s [16] idea of using an MLP to model the CRF for HDR scene reconstruction. We cover relevant works in each of these areas.

### 2.1 Novel View Synthesis

Instant HDR-NeRF is a method for novel view synthesis. The novel view synthesis problem is defined as, given a series of images capturing different perspectives of scenes or objects, each paired with a specific camera pose (camera position & direction), the goal is to reconstruct a scene representation that can render new views (or novel views) from new camera poses. An illustration of this problem is shown in Fig. 2.1. Novel view synthesis is commonly applied for virtual reality applications, where we can generate 3D reconstruction of scenes or objects when the only available information is pictures taken from different points of view.

Given a set of dense images, novel views can be rendered using light field sample interpolation [7, 21] Advancements in novel view synthesis have focused on reducing the number of views required for the reconstruction tasks. Mesh-based approaches [4, 42, 10, 43] represent the scene as a mesh and predict the underlying geometry and appearance through image reprojection. Volumetric approaches [35, 38] use volumetric representation to encode the volume density at different scene regions. Mesh-based
Figure 2.1: Novel view synthesis involves two stages: (1) building a scene representation from the input images and their poses, and (2) rendering novel views from new poses using the built representation. Figure is adapted from [36]

approaches are well-suited for real-time rendering due to its simplicity and compactness, but the rendering quality is limited by the reconstructed geometry. Volumetric representations, while having higher rendering quality, are more expensive to render. Advancements in gradient-descent techniques in deep learning have also been applied to both the mesh-based [24, 6, 25] and volumetric-based [19, 11, 29] novel view synthesis methods.

2.2 Neural Radiance Field

The primary motivation behind our method is to overcome the limitations of the neural radiance field (NeRF) approach. Specifically, we aim to address two key challenges: (1) effectively handling images with varying exposure and white balance during casual capture, and (2) supporting HDR reconstruction. In this section, we include the background of NeRF and discuss specific limitations that have driven the development of our work.

Recently, there has been a surge in using neural networks to learn the implicit volumetric function for view synthesis tasks. Among the pioneering methods in this domain is the Neural Radiance Field (NeRF) [30]. NeRF [30] represents a 3D scene
Figure 2.2: NeRF rendering pipeline. The method samples a set of point of the ray $R$, and obtain the predicted radiances and densities of these points using a neural network. Then, NeRF applies a differentiable volumetric rendering function on these predictions to get the final radiance and density of ray $R$. Figure is adapted from [30]

by an implicit continuous volumetric function represented by a fully connected (non-convolutional) neural network. This function maps a 3D position and 2D ray direction to color and density, in which a pixel is synthesized by integrating over samples along the ray. Specifically, given a single 5D coordinate of a ray passing through a pixel of the target image (spatial location $(x, y, z)$ and viewing direction $(\theta, \phi)$), NeRF first samples a set of points along the ray. Then, it feeds these points into an MLP to predict their radiance and density.

After obtaining predicted densities and colors of the points along the ray, NeRF applies a differentiable volumetric rendering function to these predictions to compute the final radiance and density of the ray. Finally, it optimizes the neural network parameters by minimizing the loss between the predicted radiance value of the ray and the ground-truth radiance value from the image using stochastic gradient descent. We show an illustration of the overall process in Figure 2.2.

NeRF has several limitations. First, NeRF does not adapt to the changes in the exposure settings among different viewpoints. Second, NeRF produced high-quality novel views limited to a low dynamic range [30]. Real-world scenes exceed the dynamic range of the camera, so it is desirable to reconstruct HDR scenes from the LDR views.
To make NeRF adapted to change in the exposure settings among different viewpoints, NeRF-W [26] uses a per-image embedding vector to represent the changes in scene appearance, effectively handling variation in both exposure settings and illumination, a technique has been adopted in many subsequent methods (Cf. [39, 28]). However, since NeRF-W does not utilize the change of exposure settings to model the camera processing pipeline, it does not have the ability to reconstruct HDR scenes.

2.3 High Dynamic Range Imaging

Our approach, along with existing HDR novel view synthesis methods [16, 17] draws inspiration from traditional HDR imaging techniques to achieves HDR reconstruction. In this section, we explain key aspects of these techniques, with a particular focus on the CRF modeling process.

Traditional HDR imaging methods calibrate the CRF from a stack of LDR images under different exposures within the same pose [9]. The CRF models the imaging process of the physical camera where the scene radiance hits the sensor and becomes the pixel value after a series image processing steps. Without taking white-balance, compression step and quantization into account, the CRF can be modeled as:

\[ Z = f(E \Delta t) \]  

(2.1)

where \( E \) is the radiance value, the total amount of light that hits the sensor, \( \Delta t \) is the exposure time determined by the shutter speed, and \( Z \) is the final pixel value.

Using a calibrated CRF, we can perform an inverse mapping where we convert each pixel value in the LDR image (from 0 to 255) to the original scene radiance (from 0
to $\infty$). This inverse mapping process produces an HDR radiance map that contains the full dynamic range of the scene.

To achieve HDR reconstruction, our method, and pre-existing HDR novel view synthesis methods [16, 17] model the CRF through separate module(s) in addition to the radiance filed representation. Compared to traditional HDR imaging techniques, these novel view synthesis methods can synthesize HDR views in novel camera poses, and do not require a stack of LDR images within a fixed camera pose.

### 2.4 High Dynamic Range Novel View Synthesis

Built upon NeRF[30], HDR-NeRF [16] attempts to capture high dynamic ranges of the real world from LDR views of varying exposure times. HDR-NeRF explicitly models the camera processing pipeline by adding a learnable tone-mapper MLP that models the CRF. HDR-NeRF assumes that the CRF is monotonic, without enforcing this constraint in their architecture design. HDR-NeRF uses ground-truth information about exposure time to render HDR radiance maps. However, the exposure times of input viewpoints are not always available, such as when using photos from the Internet, or frames extracted from a video. In addition, HDR-NeRF cannot handle varying white balance, which is often observed in casually-captured images and videos with auto-white balance enabled. Another NERF-based method, RawNeRF [28] recovers HDR radiance fields from noisy RAW images, which are not as commonly available as post-processed images that our method focuses on.

One of the main weaknesses of NeRF-based approaches[26, 16] is the long training and rendering time. Voxel-based approaches, including Plenoxels [12], and Instant-NGP [32] store scene features in a highly optimized sparse grid data structure that
allows faster runtime performance. Prior benchmarks [32, 12, 8] show that Instant NGP outperforms Plenoxels in both rendering accuracy and training time.

Built upon Plenoxels[12], HDR-Plenoxels [12] recovers HDR radiance from LDR views of varying exposure and white balance. By learning the CRF and per-image exposure parameters during training, HDR-Plenoxels does not require ground-truth information about the camera settings. Similar to HDR-NeRF[17], HDR-Plenoxels does not guarantee that the learned CRF is monotonically increasing.

2.5 NeRF Acceleration Techniques

In this work, we integrate optimized primitives from Instant-NGP [32] and efficient sampling from NerfAcc[22] for accelerating the training and rendering time of our method. This section covers a brief overview of Instant-NGP and NeRF efficient sampling techniques.

2.5.1 Instant Neural Graphic Primitives

At a high level, Instant-NGP [32] is NeRF[30] like: A pixel is rendered by casting a ray from the camera, and features are sampled along the ray. These features are passed into a neural network that outputs radiance and density. The main difference is, Instant-NGP stored the scene features in a multi-level voxel grid, which is encoded by a hash table and queried by a lightweight neural network. Instant NGP exploits the sparsity of the scene and skips the empty voxels during ray marching, which reduces the computational cost and memory usage. With the highly-optimized grid structure, Instant NGP achieves fast training time of less than 15 minutes, and can render a novel view in as fast as 1 second.
2.5.2 Importance Sampling

NeRF[30], and HDR-NeRF [16] use different sampling techniques for the coarse and the fine level. At the coarse level, these methods use uniform sampling, meaning that every point along the rays contributes equally to the rendering. At the fine level, these methods distribute the samples according to the probability density function (PDF). Vanilla NeRF and HDR-NeRF train the coarse MLP using volumetric rendering loss to obtain the densities to generate the piece-wise PDF. Mip-NeRF 360 [2] improves PDF estimation by introducing proposal networks to guide sampling and supervising them with a PDF matching loss. Voxel-based neural rendering methods, including Plenoxels [12] and Instant NGP [32] use spatial-skipping sampling, meaning that empty regions are identified and skipped during sampling. The Instant NGP implementation in NerfAcc [22] combines the occupancy grid from Instant NGP [32] with the proposal network concept from Mip-NeRF 360 [2] to further improve sampling efficiency and reduce training time.
In this section, we explain our method Instant HDR-NeRF for reconstructing high dynamic range neural radiance fields from multi-view images of unknown and varying exposure and white balance. Our approach models the camera’s physical behavior using two implicit functions: a radiance field and a monotonically increasing tone-mapper. The radiance field encodes scene geometry and radiance, outputting radiance, and densities along the camera ray. Meanwhile, the monotonically increasing tone-mapper models the CRF on the camera sensor into pixel values. The predicted radiance is combined with the exposure parameters, optimized separately per color channel and per image. We include an illustration of our pipeline in Fig. 3.1

3.1 Scene Representation

Our scenes are represented as a radiance field $F$ within unbounded 3D volumes. We use Instant-NGP [32] from NerfAcc [22] as our neural radiance field representation.

Figure 3.1: Our model of the camera pipeline consists of two modules: a HDR radiance field models the scene radiance and a monotonic tone-mapper MLP models the CRF.
For efficient sampling, this representation uses a sequence of two proposal networks represented as multi-scale occupancy grids to guide the final NeRF sampling pass. To support unbounded scenes, the coordinates queried into the occupancy grid are mapped from the unbounded space into a finite volume using the non-linear contraction function from Mip-NeRF 360 [1, 23]. However, while Instant NGP’s radiance field $F$ outputs the color and density at each point, our radiance field $F$ outputs the radiance $e$ and density $\sigma$ of the given point $r(s)$ at position $s$ along the ray $r$:

$$(e(r(s)), \sigma(r(s))) = F(r(s)) \quad (3.1)$$

### 3.2 Learned Tone-Mapping and Exposures

We use a multi-layer MLP $g$ to estimate the camera response function (CRF) of a camera. Our method represents exposure parameters as a learnable vector of three coefficients corresponding to the three color channels R, G, and B. Specifically, we assume that the exposure time, aperture, ISO gain, and white-balance of a view can all be modeled by a per-channel multiplier to the radiance $e$. Let $\lambda \in \mathbb{R}^3$ be the exposure parameters of a view. The tone-mapping function $f$ maps the radiance $e$ of point $r(s)$ to into the colors $c$ given $\lambda$:

$$c(r(s), \lambda) = f(\text{diag}(\lambda)e(r(s)) \quad (3.2)$$

Following the CRF calibration method by Debevec and Malik [9], we optimize our tone-mapping function in the logarithmic radiance domain:

$$c(r(s), \lambda) = g(\ln e(r(s)) + \ln \lambda) \quad (3.3)$$
To ensure that the learned function $g$ is monotonic and invertible, we replace the MLP from HDR-NeRF with a monotonic MLP. To constrain an MLP to be monotonic, it is sufficient to use strictly positive weights and strictly monotonic activation functions [15].

### 3.3 Neural Rendering

Similar to HDR-NeRF, we use a conventional volume rendering technique [18] to render the color and the radiance of each ray. To render HDR radiance map, we skip the tone-mapping operation after obtaining the logarithmic radiance.

\[
\hat{c}(r, \lambda) = \int_{s_n}^{s_f} T(s) \sigma(r(s)) c(r(s), \lambda) \, ds \\
\hat{e}(r) = \int_{s_n}^{s_f} T(s) \sigma(r(s)) e(r(s)) \, ds
\] (3.4)

where $s_n$ and $s_f$ are the near bound and far bound; $\hat{c}$ and $\hat{e}$ are the color and the radiance value of the ray, respectively. $T(s)$ is the accumulated transmittance:

\[
T(s) = \exp \left( - \int_{s_n}^{s_f} \sigma(r(s)) \, ds \right)
\] (3.5)
3.4 Optimization

3.4.1 Color Reconstruction Loss

Similar to the Instant NGP implementation in NerfAcc [22], we minimize the smooth L1 loss between rendered LDR views to ground-truth LDR views:

$$L_c = \sum_{r \in R} \begin{cases} 
\frac{0.5(\Delta c)^2}{\beta} & \text{if } |\Delta c| < \beta \\
|\Delta c| - 0.5 \times \beta & \text{otherwise}
\end{cases} \quad (3.6)$$

where $c$ is the ground-truth color of each ray, and $\hat{c}$ is the predicted color. $\beta$ is the threshold at which to change between L1 and L2 loss, set to 1.0.

3.4.2 Unit Exposure Loss

Because the radiance maps are learned without supervision, they can only be determined up to an unknown scale factor. Following HDR-NeRF, we fix the scale of the learned radiance maps using a point constraint. Specifically we fix the value of $g(0)$ to be 0.5. We define our unit exposure loss to be:

$$L_u = ||g(0) - 0.5||_2^2 \quad (3.7)$$

Finally, our loss function is the combination of the color reconstruction loss and the unit exposure loss:

$$L = L_c + w_u L_u \quad (3.8)$$

where $w_u$ is the weight of the unit exposure loss, which is set to 0.5 in our experiments.
3.5 Initialization of Exposure Parameters

When co-learning scene geometry, radiance, and exposure settings, there arises an exposure/radiance ambiguity. In other words, our method can use both the exposure settings and the radiance values to control brightness. Thus, our method sometimes interprets view-dependent effects such as high-frequency specularities as being caused by changes in exposure, and vice-versa (Fig. 3.2). A study [13] of previous methods [31, 41] highlights that breaking the exposure/radiance ambiguity requires making prior assumptions on the form of the CRF, or starting with the rough estimates on the exposures. Since it is challenging to directly constrain the form of the response curve through our tone mapper MLP, we adapt the idea from HDR-Plenoxels to initialize the exposure parameters $\lambda$ to physically plausible solutions. Let $I^i_c$ represent the values in color band $c$ in training image $i$ and $S_c$ be the values in color band $c$ in all training images. The exposure setting $\lambda^i_c$ for training image $I^i$ in color band $c$ is initialized as:

$$\lambda^i_c = \frac{\text{mean} I^i_c}{\text{mean} S_c}. \quad (3.9)$$

Figure 3.2: Without initializing exposure parameters to plausible solutions, our method can sometimes interpret the change of exposure as view-dependent specularities. We enhance the image contrast to make this issue easier to visualize.
Chapter 4

EXPERIMENTS

This chapter describes our experiment setup to access our method’s ability to synthesize LDR and HDR novel views from a set of LDR views of varying exposure and white balance.

4.1 Implementation Details

Our method is built upon the NerfAcc’s implementation of Instant NGP with proposal networks [22]. This implementation uses a multi-scale hash grid of 16 levels to store the scene features. This hash grid is augmented with a one-layer MLP of 64 channels to predict the color and density at points in the volume. To sample the densities, this implementation use a proposal network of two hash grids. Each hash grid has 5 levels, queried by a one-layer MLP of 64 channels.

We extend this implementation with the following modifications. We remove the sigmoid activation from the color output of the radiance field so that the output is log radiance, and add a tone-mapper which is a two-layer MLP of widths 48 and 32. To enforce the monotonic constraint on the tone-mapper, we take the absolute value of the tone-mapper’s weights and use the ReLU activation function in the tone-mapper MLP\(^1\). We use NerfAcc’s settings for the optimizer and learning rate schedule to train each model. Specifically, we use the Adam optimizer with a starting learning rate of 0.02. In the first 100 iterations, our learning rate decays linearly by a factor of 0.01. Then, our learning rate decays by a factor of 0.33 at 50%, 75% and 90% of

\(^1\)ReLU is not strictly monotonic, but in practice we found it produces the best results.
the training. We optimize each model for 20K iterations on a single NVIDIA Tesla V100 GPU, which takes about 10 minutes per model.

4.2 Methods Compared

We compare our method to the following baseline methods, described briefly here:

- HDR Plenoxels [17] – an extension of Plenoxels [12] including learnable exposure parameters and an explicit CRF representation;
- Ours (MLP) – our method but using vanilla NeRF to represent the radiance field instead of NerfAcc/Instant NGP.

4.3 Evaluation Datasets

We used four real scenes from the HDR-NeRF [16] to evaluate our method’s ability to learn the HDR radiance fields from LDR images of varying exposure. These scenes were captured using a digital SLR camera, using exposure bracketing with five different exposure times. White balance is kept fixed in this dataset.

We used datasets from HDR-Plenoxels [17] to evaluate our method’s ability to learn HDR radiance fields from LDR images of varying exposure and white balance. The HDR-Plenoxel datasets consist of five synthetic scenes generated from Blender and four real scenes captured from a digital SLR camera using exposure and white balance bracketing.
For each group of exposure bracketed images, we estimated a ground truth CRF using the method of Debevec and Malik [9].

4.4 Evaluation Metrics

We evaluate our method on: (1) LDR view synthesis quality (2) HDR view synthesis quality and (3) CRF estimation error. An overview of our evaluation metrics for each category is in section 4.4.1, 4.4.2, and 4.4.3.

4.4.1 LDR View Synthesis Quality

We employ three metrics for our quantitative comparison between LDR synthesized views and the ground-truth views: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [44].

Peak Signal-to-Noise Ratio (PSNR). PSNR is the ratio of the maximum possible value of a signal and the power of corrupting noise that affects the fidelity of its representation. We usually express PSNR as a logarithmic quantity using the decibel scale. In image similarity comparison, PSNR is most easily defined via the mean squared error (MSE), as follow:

\[
PSNR = 10 \log_{10} \left( \frac{R^2}{MSE(\hat{I}, I)} \right)
\]  

(4.1)

where R is the maximum possible value of a pixel (255 for an 8-bit image), \(MSE(\hat{I}, I)\) is the mean-square error between the predicted image \(\hat{I}\) with the target image \(I\). From
the formula, the lower the MSE, the higher the PSNR. Therefore, higher PSNR values are better.

**Structural Similarity Index Measure (SSIM).** SSIM models the human visual perception system by measuring image similarities area by area in three different properties: luminance \( l \), contrast \( c \), and structure \( s \):

\[
\begin{align*}
    l(x, y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\
    c(x, y) &= \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\
    s(x, y) &= \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}
\end{align*}
\]  \hspace{1cm} (4.2)

\[
SSIM(x, y) = l(x, y)^\alpha c(x, y)^\beta s(x, y)^\gamma
\]  \hspace{1cm} (4.3)

where \( x, y \) are two corresponding windows from two images \( I \) and \( \hat{I} \). \( \mu_x, \mu_y \) are the mean pixel value of \( x \) and \( y \) respectively. \( \sigma_x \) and \( \sigma_y \) are the standard deviation of \( x \) and \( y \) respectively. \( \sigma_{xy} \) is the covariance of \( x \) and \( y \). \( c_1, c_2, \) and \( c_3 \) are small constants to prevent division by zero. \( \alpha, \beta \) and \( \gamma \) are weights for \( l, c, \) and \( s \), respectively. The SSIM value is between -1 (very different) to 1 (very similar). Therefore, higher SSIM values are better.

**Learned Perceptual Image Patch Similarity (LPIPS).** LPIPS[44] computes image similarities using pre-trained neural networks that have been shown to match human perception well. A lower LPIPS score means that the images are more perceptual similar.

Finally, our metrics are not calculated on the entire test view. Following the evaluation methodology of HDR-Plenoxels [17], since we cannot predict the exposure parameters on test views, we use the left half of the test image for training and learn-
ing the exposure parameters. Then, we evaluate the performance on the unseen right half.

4.4.2 HDR View Synthesis Quality

Since HDR images are usually displayed after a tone-mapping operation, we evaluate our HDR images qualitatively by tone-mapping the predicted HDR radiance map. Our tone-mapping operator for HDR qualitative results is $\mu$-law, a simple and established tone-mapping operator used by HDR-NeRF [16] and other works [5, 34, 27]. This tone-mapping operator is

$$M(E) = \frac{\log(1 + \mu E)}{\log(1 + \mu)}$$

(4.4)

where $E$ is the HDR pixel value normalized to the range $[0, 1]$, and $\mu$ is the compression factor. To ensure consistent results across all the datasets, $\mu$ is varied depending on the variance of the normalized pixel value $E$ of the view:

$$\mu(E) = \frac{1}{\text{Var}[E]}$$

(4.5)

4.4.3 CRF Estimation Error

To obtain the estimated CRF from our method after training, we sample 1000 evenly-space logarithmic radiance values from -5 to 3, and get their predicted LDR colors from our tone-mapper MLP. Then, we calculate the root mean square error (RMSE) between the predicted LDR colors and the ground-truths obtained from the ground-truth CRF. We calculate the CRF RMSE in 8-bit color domain, where the lowest possible RMSE is 0 and the highest possible RMSE is 255.
In this chapter, we cover the evaluation results of our experiments. We split this chapter into two sections. Section 5.1 shows the evaluation results on the HDR-NeRF dataset, and section 5.2 shows the evaluation results on the HDR-Plenoxels dataset. In both sections, we show, quantitatively and qualitatively, how our method outperforms other baselines in LDR rendering while producing HDR views with accurate CRF estimation.

5.1 HDR-NeRF Real Dataset

The quantitative results of LDR novel view synthesis in Tab. 5.1 show that our method outperforms HDR-NeRF on all scenes. The metric values are consistent with a qualitative comparison in Figure 5.1, where our LDR views show sharper details than those from HDR-NeRF. On average, our method achieves higher metrics than HDR-NeRF (PSNR: 37.34 versus 35.34, SSIM: 0.966 versus 0.956, and LPIPS: 0.041 versus 0.068). Note that HDR-NeRF uses the exposure times given in the EXIF metadata, while our method learns the exposure settings from scratch. The difference in rendering quality suggests that the EXIF metadata may not be accurate, motivating our use of learned exposure parameters even when the metadata is available.

Our method also obtains higher metrics than Ours (MLP) (PSNR: 37.34 versus 36.28, SSIM: 0.966 versus 0.958, and LPIPS: 0.041 versus 0.065). Additional LDR renderings in Fig. 5.2 show that our method can render high-quality LDR views that match the exposures of the ground-truth views. By sampling within the range of learned
Table 5.1: Results of quantitative evaluation on the real scenes from HDR-NeRF. Values are the average of the metrics for the test data of each scene. We color code the cells as best and second best.

<table>
<thead>
<tr>
<th>Type Method</th>
<th>Computer</th>
<th>Flower</th>
<th>LuckyCat</th>
<th>Box</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
<td>LPIPS↓</td>
<td>PSNR↑</td>
</tr>
<tr>
<td>HDR-NeRF</td>
<td>35.17</td>
<td>0.945</td>
<td>0.093</td>
<td>34.98</td>
</tr>
<tr>
<td>Ours (MLP)</td>
<td>35.47</td>
<td>0.944</td>
<td>0.094</td>
<td>36.91</td>
</tr>
<tr>
<td>Ours</td>
<td>36.34</td>
<td>0.950</td>
<td>0.061</td>
<td>37.21</td>
</tr>
</tbody>
</table>

Figure 5.1: Comparison of LDR renderings from the computer scene from the HDR-NeRF real dataset. The results from our method have sharper details and more closely resemble the ground truth.

exposure parameters, our method can also render LDR views under different exposures that are not included in the training set (novel views under novel exposures), as shown in Fig. 5.3.

Our tone-mapped HDR in Fig. 5.4 show that views reveal the over-exposed and under-exposed regions of the scene. Estimated CRFs in Fig. 5.5 show that our method’s CRF estimate is close to the ground-truth.

Built upon optimized primitives from Instant-NGP[32] and NerfAcc[22], our method is also much faster than HDR-NeRF. To reach 30 PSNR, HDR-NeRF takes 52 minutes, while our method takes 35 seconds, which is 90× faster (Fig. 5.11a).
Figure 5.2: Ground-truth LDR views of varying exposure and our rendered LDR views on HDR-NeRF real scene. Our PSNR values are on the bottom right of the error maps.

Figure 5.3: By sampling within the range of the learned exposure parameter vectors, our method can render LDR novel views under different exposure times (a-f) that are not included in the training set.
Figure 5.4: Ground-truth LDR views (top row) and our tone-mapped HDR views (bottom row) from the HDR-NeRF dataset. Our tone-mapped HDR views reveal under-exposed and over-exposed regions of the scenes.

Figure 5.5: Comparisons between our estimated CRFs and the ground-truth CRF calibrated using Debevec and Malik method [9] on the HDR-NeRF real dataset.
5.2 HDR-Plenoxels Real and Synthetic Datasets.

As shown in Tab. 5.2 and Tab. 5.3, our method achieves higher metrics than HDR-Plenoxels (PSNR: 30.35 versus 28.73, SSIM: 0.904 versus 0.891, and LPIPS: 0.121 versus 0.294). Fig. 5.6 shows a qualitative comparison.

Table 5.2: Results of quantitative evaluation on the synthetic scenes from HDR-Plenoxels.

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Book</th>
<th>Classroom</th>
<th>Monk</th>
<th>Room</th>
<th>Kitchen</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDR-Plenoxels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR↑</td>
<td>27.49</td>
<td>0.817</td>
<td>0.292</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>29.87</td>
<td>0.908</td>
<td>0.284</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>31.53</td>
<td>0.350</td>
<td>0.156</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours (MLP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR↑</td>
<td>28.80</td>
<td>0.852</td>
<td>0.219</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>28.57</td>
<td>0.878</td>
<td>0.163</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>35.56</td>
<td>0.308</td>
<td>0.066</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR↑</td>
<td>26.58</td>
<td>0.801</td>
<td>0.262</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>30.04</td>
<td>0.920</td>
<td>0.101</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>35.13</td>
<td>0.328</td>
<td>0.052</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Results of quantitative evaluation on the real scenes from HDR-Plenoxels.

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Character</th>
<th>Desk</th>
<th>Plant</th>
<th>Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDR-Plenoxels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR↑</td>
<td>33.14</td>
<td>0.960</td>
<td>0.343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>33.77</td>
<td>0.959</td>
<td>0.087</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>32.15</td>
<td>0.817</td>
<td>0.166</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours (MLP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR↑</td>
<td>33.46</td>
<td>0.950</td>
<td>0.080</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>28.52</td>
<td>0.912</td>
<td>0.125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>31.63</td>
<td>0.943</td>
<td>0.070</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR↑</td>
<td>31.69</td>
<td>0.920</td>
<td>0.080</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>31.03</td>
<td>0.857</td>
<td>0.141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>30.98</td>
<td>0.937</td>
<td>0.096</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.6: Comparison of LDR renderings from the room scene from the HDR-Plenoxels synthetic dataset.

Our method also outperforms Ours (MLP) on these datasets (PSNR: 30.35 versus 30.28, SSIM: 0.904 versus 0.894 and LPIPS: 0.121 versus 0.141). Sample LDR renderings in Fig. 5.7 and Fig. 5.8 show our method’s ability to render LDR views of varying exposure and white balance.
Figure 5.7: Ground-truth LDR views and our rendered LDR views on HDR-Plenoxels synthetic scene. Our PSNR values are on the bottom right of the error maps.

Our HDR renderings are shown in Fig. 5.9. Estimated CRFs are in Fig. 5.10. To reach 30 PSNR, HDR-Plenoxels takes 12 minutes, while our method takes 47 seconds, which is $15 \times$ faster (Fig. 5.11b).
Figure 5.8: By sampling within the range of the learned exposure parameter vectors, our method can render LDR novel views under different exposures and white-balance (a-f) that are not included in the training set.

Figure 5.9: Ground truth LDR views (top row) and our tone-mapped HDR views (bottom row) from the HDR-Plenoxels dataset.
Figure 5.10: Comparisons between our estimated CRFs and the ground-truth CRF calibrated using Debevec and Malik method [9] on the HDR-Plenoxels real dataset.

Figure 5.11: Training time comparisons between our method and (a) HDR-NeRF and (b) HDR-Plenoxels. Our method shows the fastest training time among the three.
In this chapter, we cover three ablation studies that further demonstrate the technical contributions of our paper. Our first study provides quantitative and qualitative comparisons to highlight a significant improvement of CRF estimation and HDR rendering quality when enforcing the tone-mapper MLP to increase monotonically. Our second study in section 6.2 is a comparison showing how our implicit CRF representation outperforms the explicit CRF representation from HDR-Plenoxels [17] in both CRF estimation accuracy and rendering quality. Our last study in section 6.3 demonstrates how applying tone-mapping before volume-rendering improves the LDR rendering quality, with the cost of a slight increase in CRF estimation error by 2.86%.

### 6.1 Monotonic Constraint on the Tone Mapper

We validated the necessity of enforcing the monotonic constraint (MC) on the tone-mapping function in an ablation study in Tab. 6.1. In experiments without MC, we remove the positive weight constraint for our tone-mapping function \( g \) to be monotonically increasing. Compared to our proposed method (Row 5, Tab. 6.1), our method without MC (Row 1, Tab. 6.1) achieves higher LDR rendering accuracy on the HDR-NeRF dataset, but fails to accurately model the ground truth CRF, resulting in a higher average CRF RMSE. Specifically, the qualitative comparison in Fig. 6.1a shows that the red channel of the estimated CRF is non-monotonic, resulting in an extreme red tint to the HDR tone-mapping images.
6.2 Explicit Versus Implicit CRF Representation

We tested replacing our monotonic tone-mapper with the explicit CRF representation from HDR-Plenoxels [17]. Our method (Row 5, Tab. 6.1) shows higher accuracy in both LDR rendering and CRF estimation than the version with explicit CRF (Row 2, Tab. 6.1). When using an explicit CRF representation, the estimated CRFs of the blue channel had a bias, which caused a blue hue in the HDR tone-mapped views (Fig. 6.2). When using our proposed implicit CRF representation, the CRFs matched more closely with the ground-truth ones, resulting in tone-mapped HDR views where the colors are more balanced.

6.3 Tone Mapping Before Versus After Volume Rendering

Our method and HDR-NeRF [16] both apply tone mapping to the color at each point before volume rendering (pre-integration), while HDR-Plenoxels [17] performs tone mapping to the integrated color after volume rendering (post-integration). While tone mapping post-integration better matches the actual image formation process [37], in our experiments we found that tone mapping pre-integration achieved higher accuracy in LDR rendering and CRF estimation. Specifically, for the implicit CRF representation, our proposed version using pre-integration tone mapping (Row 5, Tab. 6.1) shows higher LDR rendering accuracy and comparable CRF RMSE to the version using post-integration (Row 4, Tab. 6.1). For the explicit CRF representation, the version of pre-integration tone mapping (Row 2, Tab. 6.1) performs better in all metrics compared to the equivalent post-integration version (Row 3, Tab. 6.1). These findings may provide interesting avenues for further research into the relationship between neural rendering and CRF modeling. Qualitative comparisons are shown in Fig. 6.2.
Table 6.1: Quantitative results of LDR rendering accuracy and CRF estimation error of our ablation studies. Although, on average, our method achieves second-best in LDR rendering accuracy on the HDR-NeRF real dataset, our HDR-tonemapped views (Fig. 6.1b) look better as the CRFs are modeled more accurately.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>CRF RMSE↓</th>
<th>HDR–NeRF real PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
<th>HDR–Plenoxels real PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without MC</td>
<td>21.391</td>
<td>38.22</td>
<td>0.970</td>
<td>0.036</td>
<td>29.65</td>
<td>0.916</td>
<td>0.111</td>
</tr>
<tr>
<td>Explicit CRF</td>
<td>11.210</td>
<td>36.48</td>
<td>0.959</td>
<td>0.054</td>
<td>26.57</td>
<td>0.894</td>
<td>0.149</td>
</tr>
<tr>
<td>Explicit, post-integ. CRF</td>
<td>14.183</td>
<td>36.68</td>
<td>0.962</td>
<td>0.051</td>
<td>25.99</td>
<td>0.888</td>
<td>0.155</td>
</tr>
<tr>
<td>Post-integ. CRF</td>
<td>9.176</td>
<td>36.88</td>
<td>0.966</td>
<td>0.042</td>
<td>27.29</td>
<td>0.897</td>
<td>0.136</td>
</tr>
<tr>
<td>Ours</td>
<td>9.437</td>
<td>37.34</td>
<td>0.966</td>
<td>0.041</td>
<td>29.94</td>
<td>0.913</td>
<td>0.110</td>
</tr>
</tbody>
</table>

Figure 6.1: Qualitative results on the computer scene with/without enforcing monotonic constraint on the tone mapper. The ground-truth CRF was calibrated using Debevec and Malik method [9].
Figure 6.2: Qualitative comparisons among different model configurations mentioned in our ablation studies.
Chapter 7

REAL WORLD APPLICATIONS

In this chapter, we show the applications of our methods in several real-world scenes. Section 7.1 and 7.2 demonstrate how robust our method is in learning HDR radiance fields in multiple camera shooting conditions, and Section 7.3 demonstrates how we can control the tonemapper MLP to simulate auto-exposure effects that are commonly seen in video games and filming.

7.1 Scenes of Varying White-Balance

We used our method to render novel views for the scene of varying white balance shown in Fig. 1.1. The LDR spiral views shown in Fig. 7.1 are consistent in exposure settings, without any artifacts as compared to NeRF [30].

![Figure 7.1: LDR spiral views of the dresser scene between NeRF (top) and our method (bottom)](image-url)
7.2 Scenes of Varying Exposure and White-Balance

We captured two more sequences in real scenes outside of a laboratory setting, to assess the real-world applicability of the method. We chose scenes where the variability in light among different viewpoints is too wide to be captured using a single exposure setting. We capture the downtown scene (Fig. 7.2) with a digital camera using both exposure and white-balance bracketing. Our method works reliably on this scene, with an average LDR PSNR, SSIM, and LPIPS of 28.76, 0.816, and 0.199 respectively. We captured the window seat scene (Fig. 7.3) with a smartphone using auto-exposure, meaning that the exposure of each view is not systematically sampled. Our tone-mapped HDR view in Fig. 7.3 reveals the under-exposed and over-exposed regions, with an average LDR PSNR, SSIM, and LPIPS of 26.19, 0.835, and 0.163 respectively.

Figure 7.2: Results for downtown scene, captured with a DSLR with exposure and white-balance bracketing. Our HDR results show details in both the bright coffee logo and dark coffee tables.
Figure 7.3: Results for *window seat* scene, captured with a smartphone in auto-exposure mode. Our HDR results show the details of the bright paper and dark book cover.
7.3 Simulation of Auto-Exposure Effects

Auto Exposure (Eye Adaptation) simulates how the human eye adjusts to changes in brightness in real-time. When lighting conditions change (e.g., moving from a dimly lit room to a brightly lit outdoor area), our eyes adapt by altering the amount of light they let in. The camera can simulate the same effect by dynamically adjusting the exposure of an image to match its mid-tone.

As shown in previous chapters, we can change the exposure of the novel views by modifying the logarithmic exposure parameter vectors \( \ln \lambda \) passed into our tone-mapper MLP. To simulate the auto-exposure effects, we can dynamically alter the logarithmic exposure parameter vectors depending on the brightness of the current LDR view. Specifically, starting with the unit logarithmic exposure parameters of \([0,0,0]\) of view \(i\), the logarithmic exposure parameters of the upcoming view \(i + 1\) can be calculated by:

\[
\ln \lambda_{i+1} = \ln \lambda_i - (\alpha - \epsilon) \Delta s
\]

(7.1)

where \(\alpha\) is the ratio of over-exposed pixels in the LDR view \(i\). In 8-bit color domain, we consider a pixel to be over-exposed when its values exceed 240 for all three color channels. \(\epsilon\) is the threshold of the over-exposed pixels, which we set to 0.005 for the best viewing experience. \(\Delta s\) is the step size, which varies based on the range of learned logarithmic exposure parameters:

\[
\Delta s = \frac{|\max\{\ln \lambda_{\text{learned}}\} - \min\{\ln \lambda_{\text{learned}}\}|}{2}
\]

(7.2)

We simulate the auto-exposure effect when moving the camera around the downtown scene. As shown in Fig. 7.4, when the camera moves toward the over-exposed regions
such as the coffee logo, the exposure of the view is dynamically decreased to reveal the details of these regions.

Figure 7.4: Our method’s simulation of auto-exposure effects. When the camera traverses to over-exposed regions of the downtown scene, our method can dynamically decrease the exposure of the LDR view to reveal the details of these regions.
Instant HDR-NeRF has two main limitations: (1) it sometimes outputs HDR radiance map with extremely high radiance values in some regions, and (2) the scale of the learned exposure parameters is nondeterministic and varied among scenes. We discuss each limitation in the following sections below.

8.1 Extreme Radiance Values

In scenes with significant exposure variation, where certain regions are dramatically over-exposed, our method occasionally generates HDR radiance maps containing extremely high radiance values in those over-exposed areas. While our learned tone-mapper MLP successfully maps these values to LDR colors, our HDR tone-mapping operator $\mu$-law (Eq. 4.4) struggles to compress the wide range of radiance values into a visually pleasing HDR tone-mapped image. One potential workaround involves clipping out the extreme radiance values at the 99th percentile, which has shown promising results in our internal testing. Alternatively, introducing additional regularization to our network could help mitigate this issue. It’s important to note that theoretically, the original radiance of a scene can extend to $\infty$, so it is still reasonable for the HDR recovering method to behave similarly.
8.2 Learned Exposure Parameters with Nondeterministic Scale

Our method learns the exposure parameters for each color channel and for each image from scratch, and the scale and unit of our exposure do not necessarily match with those in a digital camera. Specifically, there is no guarantee that our learned exposures are in the unit of seconds like the exposure time in the EXIF metadata. Since we optimize the exposure parameters using stochastic gradient descent along with two implicit components (a radiance field and a tone-mapper), it is up to the optimizing process to determine the scale and unit of the learned exposure parameters. In contrast, the original HDR-NeRF method [16] fixes the scale of the exposure time using ground-truth metadata from EXIF. This allows it to render HDR views with arbitrary exposure time in the unit of seconds. Despite this difference, our method can still render novel views with arbitrary settings by sampling within the range of the learned exposures.
In this chapter, we propose potential extensions for our method, including: (1) modeling local tone-mapping, (2) adding support for photo collections in the wild, and (3) integrating our extension into Nerfstudio. We provide a brief overview of each possible extension in the following sections below.

9.1 Local Tone Mapping

Our method currently models global (or spatially uniform) tone-mapping, where every pixel is mapped in the same way according to the estimated CRF, and independent of the surrounding pixels in the image. In contrast, local (or spatially varying) tone-mapping functions map each pixel differently according to the features extracted from surrounding pixels, such as average luminance. The two main benefits of using the local tone-mapping function are (1) enhancing local contrast and brightness, and (2) avoiding loss of fine details due to the high variance of radiance in different regions of the scene. With these advantages, phone manufacturers such as Google have introduced complex local tone-mapping operations in their phone’s camera processing pipeline[33, 14]. As a potential extension, we could explore even more sophisticated tone-mapping pipelines that handle heavily post-processed images with local tone-mapping functions. Such advancements would contribute to a more faithful HDR reconstruction of the scene.
9.2 HDR Novel View Synthesis from Photo Collections in the Wild

Our method primarily focuses on HDR reconstruction from photos captured by the same camera, where it is reasonable to model that camera with a single CRF. A potential extension of our work is to support HDR reconstruction from photo collections in the wild, where viewpoints can be from different cameras, each with a different hardware configuration and post-processing algorithms. One possible approach for this extension would be to have a separate tone-mapper MLP for each camera. This method, however, does not scale well with an increasing number of cameras in large photo collections. A possible alternative approach would be to extend the architecture design of the tone-mapper MLP so that it now takes in both the camera index and the logarithmic radiance value as the input, and produces the LDR color using the settings from the input camera index.

9.3 Integration to Nerfstudio

Nerfstudio [40] is an open-source Pytorch library that offers a suite of radiance field methods within a modular framework. It simplified the entire workflow with plug-and-play components, from data preparation to training and evaluation, across many different radiance field technologies ranging from vanilla NeRF [30], Instant-NGP [32], Mip-NeRF [1] to more recent, state-of-the-art methods like 3D Gaussian Splatting [20] and Zip-NeRF [3]. Our Instant HDR-NeRF method was designed with modularity in mind, particularly our tone-mapper MLP, which is a plug-and-play component that seamlessly integrates with NerfAcc’s Instant-NGP implementation. We believe that this modular component will simplify the incorporation of HDR support across all radiance field methods within Nerfstudio. Such integration would not only broaden the
accessibility and applicability of our work but also ensure that our HDR tone mapper remains compatible with future radiance field technologies adopted by Nerfstudio.
We present Instant HDR-NeRF, a novel method of learning HDR view synthesis from LDR views with unknown and varying exposure and white balance. The key to our method is to model the camera processing pipeline with two implicit components: (1) a radiance field built upon optimized primitives from Instant-NGP and Nerfacc to model the scene radiance, and (2) a monotonically-increasing tone-mapper MLP to model the CRF. Our method is capable of rendering novel HDR views without the need for ground-truth HDR views. Additionally, it can render LDR views with any exposure and white-balance, including those that align with the ground-truths. In terms of view synthesis quality and training time, our method surpasses previous methods. It successfully models the CRF calibrated by Debevec and Malik’s method [9], ensuring accurate HDR reconstruction.

The practical implications of our method are extensive. Primarily, it enables high-fidelity and exposure-consistent 3D reconstruction of real-world scenes with varying exposure and white balance, a common occurrence in casual capture. The tone-mapper MLP in our method can be easily controlled to simulate auto-exposure effects, making it useful in filming and video games. Furthermore, the HDR radiance maps produced by our method can be edited and tone-mapped as per user requirements.

Advancements in novel view synthesis technologies will continue to improve 3D reconstruction quality and speed. These advancements, if successfully integrated with our tone-mapper module, will become more accessible by accommodating a broader range of input conditions outside of experimental settings. We also hope that our
method, along with others [16, 17] will inspire further breakthroughs in modeling more components of the camera processing pipelines to achieve even more faithful reconstruction of the world around us.


