

# Multiplane Prior Guided Few-Shot Aerial Scene Rendering

Zihan Gao

Licheng Jiao \*

Lingling Li

Xu Liu

Fang Liu

Puhsa Chen

Yuwei Guo

School of Artificial Intelligence, Xidian University

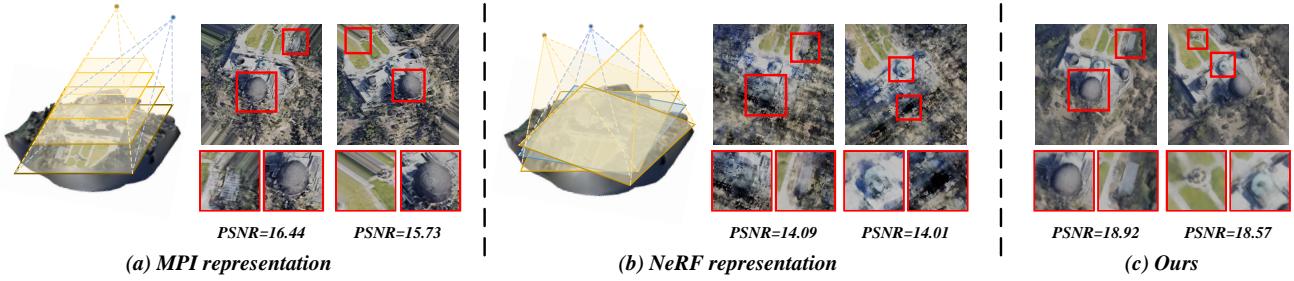


Figure 1. **Visualization of failure modes in NeRF and MPI.** (a) MPI models scenes only in each single camera frustums and performs homography warping to render novel views. Insufficient sampling leads to incorrect depth and thus results in an overlapping ghosting effect. Large camera movement leads to cropped corners. However, high-frequency details seem to be successfully preserved. (b) NeRF models scenes in a continuous volumetric manner. If only sparse views with large camera movement are provided, some parts of the scene may be sampled very little or even never. Insufficient sampling leads to collapsed details and unexpected floaters. (c) Our approach combines the capabilities of NeRF with the perspective-friendly nature of MPI in aerial scenes to achieve photorealistic novel view renderings.

## Abstract

*Neural Radiance Fields (NeRF) have been successfully applied in various aerial scenes, yet they face challenges with sparse views due to limited supervision. The acquisition of dense aerial views is often prohibitive, as unmanned aerial vehicles (UAVs) may encounter constraints in perspective range and energy constraints. In this work, we introduce Multiplane Prior guided NeRF (MPNeRF), a novel approach tailored for few-shot aerial scene rendering—marking a pioneering effort in this domain. Our key insight is that the intrinsic geometric regularities specific to aerial imagery could be leveraged to enhance NeRF in sparse aerial scenes. By investigating NeRF’s and Multiplane Image (MPI)’s behavior, we propose to guide the training process of NeRF with a Multiplane Prior. The proposed Multiplane Prior draws upon MPI’s benefits and incorporates advanced image comprehension through a SwinV2 Transformer, pre-trained via*

*SimMIM. Our extensive experiments demonstrate that MPN-eRF outperforms existing state-of-the-art methods applied in non-aerial contexts, by tripling the performance in SSIM and LPIPS even with three views available. We hope our work offers insights into the development of NeRF-based applications in aerial scenes with limited data.*

## 1. Introduction

Neural Radiance Field (NeRF) [38] has succeeded in rendering high-fidelity novel views and many 3D applications by modeling 3D scenes as a continuous implicit function. In contrast to indoor or synthetic scenes capturing simple objects using cell phones, aerial images provide a unique bird’s-eye view and overview of a landscape. Based on NeRF, many applications in aerial scenes have been developed, such as navigation, urban planning, data augmentation, autonomous vehicles, and environmental mapping [1, 10, 25, 34, 37, 45, 65].

In many real-world scenarios, unmanned aerial vehicles (UAVs) encounter constraints such as limited perspectives, energy limitations, or adverse weather conditions, which restrict their ability to acquire dense observational data. While NeRF forms a foundational technology for many aerial applications, it is prone to overfitting on sparse training views [17, 21, 39]. This limitation of NeRF becomes particularly salient in the context of aerial imagery. Alleviating this prob-

\*Corresponding author. This work was supported in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the Joint Funds of the National Natural Science Foundation of China (U22B2054), the National Natural Science Foundation of China (62076192, 61902298, 61573267, 61906150, and 62276199), the 111 Project, the Program for Cheung Kong Scholars and Innovative Research Team in University (IRT\_15R53), the ST Innovation Project from the Chinese Ministry of Education, the Key Research and Development Program in Shaanxi Province of China(2019ZDLGY03-06), the China Postdoctoral fund(2022T150506).

lem could save resources and may benefit numerous applications.

As the first to explore few-shot NeRF for aerial imagery, we stand at the forefront of this emerging field. The landscape of few-shot neural rendering to date has been predominantly shaped by its application to indoor and synthetic scenes. Transfer learning based methods [4, 51, 52, 62] aim to pre-train the model on a large number of scenes. Yet, these approaches necessitate extensive datasets for pre-training. This is not only resource-heavy but also impractical to fulfill for varied aerial scenarios. Another line of works [13, 17, 19, 21, 39, 43, 59, 61] seeks to impose regularization on NeRF by exploring the universal attributes of 3D scenes like local continuity and semantic consistency. Yet, in situations where available data is significantly sparse relative to the complexity of the scene, these methods might struggle to maintain their effectiveness. And the last is depth-prior-based methods [11, 50] gain additional supervision from the scene’s depth. These methods can be problematic in aerial images due to the frequent occurrence of ambiguous depth cues and the high cost of obtaining accurate depth maps. Despite their efficacy in controlled environments, these methods fall short of addressing the unique complexities of aerial scenes, leaving a gap that our work aims to fill. We therefore ask: **Can we harness the intrinsic geometric regularities specific to aerial imagery to broaden the capabilities of NeRF under sparse data conditions, thereby easing the data collection constraints?**

To answer this question, we turn to earlier works on 2.5D representations such as Multiplane Image (MPI) [27, 48, 55, 64]. MPI typically operates by extracting multiple RGB and density planes from a single image input by an encoder-decoder style MPI generator to represent scene geometry within the camera’s frustum. Although NeRF provides a continuous representation of a scene, MPI offers discrete, frustum-confined layers that can be particularly advantageous in the context of aerial imagery. This is due to UAVs frequently capturing images from overhead perspectives that align well with MPI’s planar representation. Additionally, the encoder-decoder architecture of the MPI generator can exploit the inductive biases inherent in advanced convolutional and self-attention-based image processing compared to the simple multi-layer perceptron (MLP) of NeRF, thus enhancing the rendering of local and global scene details. However, while MPIS present certain benefits in terms of their adaptability to aerial perspectives, their partial scene recovery and limitation to individual frustums pose challenges in creating a comprehensive 3D understanding.

In this work, we present Multiplane Prior guided NeRF (MPNeRF), a novel method for enhancing NeRF models in few-shot aerial scene rendering. We guide NeRF’s learning process by using a multiplane prior—a concept drawn from the strengths of MPI and refined with cutting-edge im-

age understanding from a Swin TransformerV2 pre-trained with SimMIM [32]. This approach unites the capabilities of NeRF with the perspective-friendly nature of MPI, tailored for the unique vantage points of aerial scenes. Concretely, our approach updates the NeRF branch using pseudo labels generated from the MPI branch. As training proceeds, NeRF can effectively pick up finer details from the MPI branch and the advantage of the MPI branch is implicitly distilled into NeRF. This strategy implicitly folds a multiplane prior to NeRF, boosting its performance in handling sparse aerial imagery data. Our contributions can be summarized as follows:

1. We introduce Multiplane Prior guided NeRF (MPNeRF), a novel framework that synergistically combines NeRF and MPIS for enhanced few-shot neural rendering in aerial scenes. To the best of our knowledge, this is the first method specially designed for this task.
2. Through an investigation, we pinpoint and analyze the typical failure modes of NeRF and MPI in aerial scenes. We devise a simple yet effective learning strategy that guides the training process of NeRF by learning a multiplane prior, effectively circumventing NeRF’s typical pitfalls in sparse aerial scenes.
3. We compare MPNeRF against a suite of state-of-the-art non-aerial scene methods, rigorously testing its adaptability and performance in aerial scenarios. Our experiments demonstrate MPNeRF’s superior performance, showcasing its significant leap over methods previously confined to non-aerial contexts.

## 2. Related Work

### 2.1. Scene Representations for View Synthesis.

Earlier works on light fields [9, 15, 26] achieve view synthesis by interpolating nearby views given a dense set of input images. Later works utilize explicit mesh [6, 28, 30, 49], or volumetric [16, 20, 23, 42, 44, 54] representation to represent the scene. More recently, layered representations have gained attention due to their efficiency in modeling occluded content. One such layered representation is the MPI [27, 48, 55, 64]. An MPI consists of multiple planes of RGB and  $\alpha$  values at fixed depths. Given an input image, an encoder-decoder network typically generates the MPI within the camera frustum. This MPI is then homography warped to the target camera position and integrated over the planes to produce novel views. It’s important to note that the generated MPI only models the geometry within each camera frustum at given depths, and the complete 3D scene is not fully recovered.

Recently, NeRF [38] has shown significant potential in novel view synthesis. NeRF works by modeling the scene with a continuous function of 3D coordinates and viewing directions to output the corresponding RGB and volume density values. Following NeRF, many methods have been proposed. mip-NeRF [3] introduces a more robust representa-

tion that uses a cone tracing technique and samples the cone with multivariate Gaussian. NeRF-W [36] and Ha-NeRF[7] have extended the applicability of NeRF to in-the-wild photo collections through object decomposition and hallucination techniques. For large-scale scenes, BungeeNeRF [56] proposes a multiscale progressive learning method to reconstruct cities from satellite imagery, while Block-NeRF [46] leverages individual NeRFs for each component of the scene to achieve large-scale scene rendering. ShadowNeRF [12] and Sat-NeRF [35] address the issue of strong noncorrelation between satellite images taken at different times by modeling solar light and transient objects. However, these NeRF-based methods are still limited by the need for densely sampled views of the scene.

## 2.2. NeRF with Sparse Input

Many approaches have been developed to train a NeRF from sparse input in different directions. One straightforward direction [4, 31, 47, 51, 52, 62] is to learn the general appearance of a scene or object from a large number of data. PixelNeRF [62] adopts a CNN feature extractor to condition each input coordinate with image features. MVSNeRF [4] uses 3D CNN to process cost volume acquired by image warping. These methods often require a large number of multi-view images to be pre-trained on, which is sometimes hard to acquire in aerial imagery. Some other techniques [13, 21, 24, 39, 43] find it is more data-efficient to regularize NeRF with common properties of the 3D geometry. InfoNeRF [21] regularizes NeRF by putting a sparsity constraint on the density of each ray. RegNeRF [39] regularizes NeRF by local smoothness. Other works [11, 17, 50, 53] aim to take advantage of supervision from other sources, such as depth or appearance. DS-NeRF [11] supervises the geometry with sparse point cloud generated with structure from motion. DietNeRF [17] regularize NeRF by ensuring perceptual consistency within different views. ManifoldNeRF [19] builds upon DietNeRF and takes into account viewpoint-dependent perceptual consistency to refine supervision in unknown viewpoints. However, we noticed none of these methods is designed for aerial scenes and thus left a gap our work aims to fill.

## 3. Method

Our objective is to train a standard NeRF model to create highly realistic novel views of an aerial scene from a limited number of captured perspectives. To address the challenges of training NeRF with sparse aerial views, we introduce a novel training approach that leverages a Multiplane Prior. The proposed Multiplane Prior harnesses the strengths of MPI and is enriched by advanced image understanding capabilities derived from a SwinV2 Transformer pre-trained using SimMIM [57]. An overview of our approach is presented in Figure 2.

In Sec. 3.1, we briefly review the background related to our method. Sec. 3.2 provides an investigation into NeRF and MPI’s behavior in sparse aerial scenes. Sec. 3.3 describe our overall framework.

### 3.1. Preliminaries

**Neural Radiance Field.** Given a 3D coordinate  $\mathbf{x} = (x, y, z)$  and a 2D viewing direction  $\mathbf{d} = (\theta, \varphi)$ , NeRF aims to model the scene by solving a continuous function  $f(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$  using multi-layer perceptron (MLP) network, where  $\mathbf{c}$  and  $\sigma$  represent the emissive color and volume density at the given coordinate. NeRF cast rays  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  from the camera origin  $\mathbf{o}$  along the direction  $\mathbf{d}$  to pass through a pixel. NeRF then samples  $M$  points along this ray and computes its color by volume rendering:

$$\begin{aligned}\hat{\mathbf{C}}(\mathbf{r}) &= \sum_{i=1}^M T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i, \\ T_i &= \exp \left( - \sum_{j=1}^{i-1} \sigma_j \delta_j \right),\end{aligned}\quad (1)$$

where  $\mathbf{c}_i$  and  $\sigma_i$  are the color and volume density of i-th sample along the ray and  $\delta_i$  is the distance between adjacent samples.  $\hat{\mathbf{C}}(\mathbf{r})$  denotes the final color of that pixel rendered by NeRF. In NeRF, a dense 3D scene is recovered implicitly in the form of neural network weights.

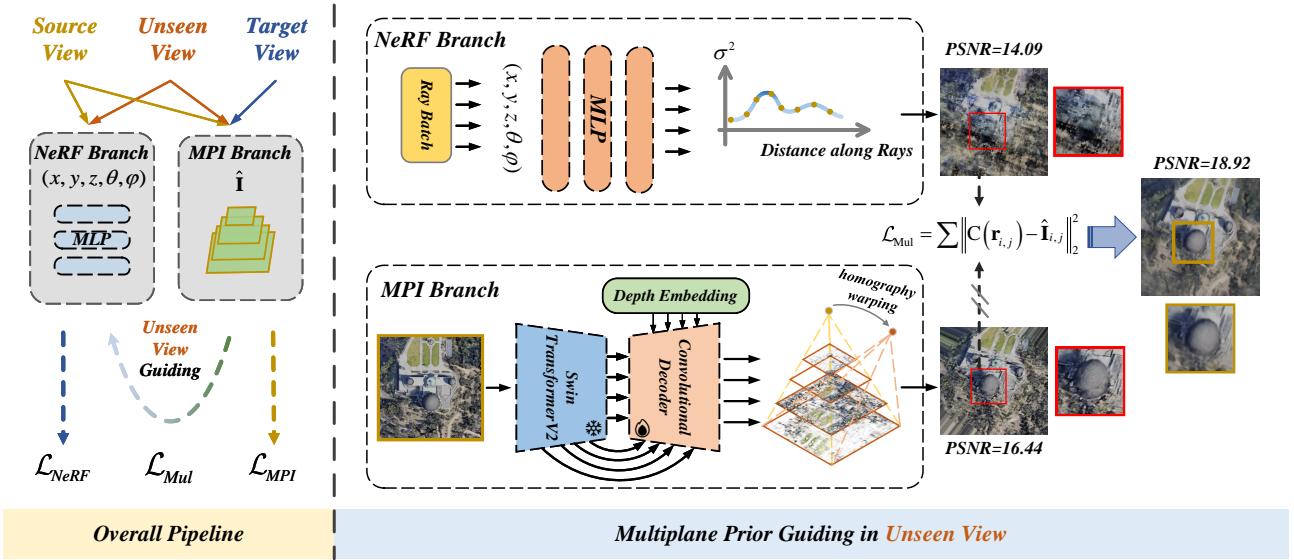
**Multiplane Image.** Multi-plane Image (MPI) represents the scene by dividing the 3D space into a collection of planes with RGB and density values in one camera frustum. In training, each batch consists of a pair of images  $\mathbf{I}_s, \mathbf{I}_t \in \mathbb{R}^{H \times W \times 3}$  with corresponding camera intrinsic  $\mathbf{K}_s, \mathbf{K}_t \in \mathbb{R}^{3 \times 3}$  and relative pose  $\mathbf{P}_{s2t} = [\mathbf{R}_{s2t} \in \mathbb{R}^{3 \times 3}, \mathbf{t}_{s2t} \in \mathbb{R}^3]$  denoted as  $\{(\mathbf{I}_s, \mathbf{K}_s), (\mathbf{I}_t, \mathbf{K}_t), \mathbf{P}\}$ , subscripts  $s$  and  $t$  represent source and target viewpoint respectively. Depth for each plane is sampled  $\{z = z_k | k = 1, 2, 3, \dots, N\}$  uniformly according to the scene bounds. An encoder-decoder based MPI-generator denoted as  $G_{\text{MPI}}$  is adopted to generate multiple planes of RGB and density at discrete depth as:

$$\{(\mathbf{c}_k, \sigma_k) | k = 1, 2, 3, \dots, N\} = G_{\text{MPI}}(\mathbf{I}_s). \quad (2)$$

Here the subscript  $k$  denotes the k-th plane. The rendering of MPI is performed in two steps: First, establish the correspondence between the pixel coordinates in the source and target plane through homography warping as:

$$[u_t, v_t, 1]^T = \mathbf{K}_t \left( \mathbf{R}_{s2t} - \frac{\mathbf{t}_{s2t} \mathbf{n}^T}{z_k} \right) \mathbf{K}_s^{-1} [u_s, v_s, 1]^T. \quad (3)$$

Second, similar to NeRF, apply differentiable rendering to



**Figure 2. Overall pipeline for training Multiplane Prior guided NeRF (MPNeRF).** Our novel MPNeRF architecture integrates a standard NeRF branch with an MPI branch, informed by a pre-trained SwinV2 Transformer. This design introduces a multiplane prior to guide the NeRF training, addressing the common challenges of rendering with sparse aerial data. The process begins by sampling three distinct views: a source and target view for training with known ground truth, and an unseen view from a novel viewpoint. The NeRF model is then refined using pseudo labels produced by the MPI branch, which are especially crucial for synthesizing views from previously unseen angles, as shown in the pipeline.

get target view 2D images  $\hat{\mathbf{I}}_t$  :

$$\hat{\mathbf{I}}_t = \sum_{k=1}^N T_k (1 - \exp(-\sigma_k \delta_k)) \mathbf{c}_k, \quad (4)$$

$$T_i = \exp \left( - \sum_{l=1}^{k-1} \sigma_l \delta_l \right).$$

### 3.2. A Closer Look at The Behavior of NeRF & MPI

To better understand the behavior of NeRF and MPI, we conducted an investigation into their failure modes. In Figure 1, we visualize the rendering results of NeRF and MPI when encountering large camera movements. Our findings reveal that NeRF often produces blurry renderings, while MPI tends to exhibit overlapping ghosting effects and cropped corners.

Recall that NeRF represents the whole 3D scene continuously by encoding the volume density and color into an 8-layer MLP’s weights. In other words, NeRF utilizes a learning-based approach by forcing correct rendering from every angle of the scene with multi-view consistency. Such a model is highly compact when supervised with sufficient training views. When the supervised angle is limited, areas covered less (as in the non-overlapped camera frustum in Figure 1 (b)) are uncontrolled and may exhibit high-density values [29, 39, 61], leading to blurry or even collapsed results. Considering the structured nature of 3D aerial scenes, we recognize two key factors of aerial scenes: *aligned perspectives with predominant planarity*, and *consistent geomet-*

*ric appearance*. First, the typical flight paths of UAVs over these scenes predominantly capture landscapes aligned with the XY planes, providing a unique geometrical consistency. Second, objects in aerial scenes contain common visual characteristics, offering additional cues for scene interpretation and analysis.

In contrast to NeRF, MPI models the scene within each camera frustum and decomposes it into an explicit set of discrete 2D planes at fixed depths. This mirrors the overhead views and planar surfaces commonly found in aerial scenes. Also, the convolution-based or self-attention-based MPI-generator is inherently suited to carry prior knowledge of the scene. However, with insufficient supervision provided, the MPI for different camera frustums may not be properly calibrated. As a consequence, we observe the occurrence of overlapping ghosting effects in rendered unseen views. Additionally, when there is substantial camera movement, the corners of the target views may be excluded from the source views, resulting in invalid renderings. However, MPI is successful in preserving high-frequency details in the rendered image. We attribute this capability to the power of CNNs and the implicit encoding of prior knowledge in the MPI generator.

Due to the distinct failure mode of MPI and its favorable properties, we explore a strategy to enhance NeRF in a sparse aerial context.

### 3.3. Guiding NeRF with a Multiplane Prior

Based on the investigation of the different properties shown in NeRF and MPI when encountering sparse input. We turn to the task of few-shot aerial scene rendering and propose a simple yet effective strategy that treats the MPI as a bridge to convey information that is hard to learn by the traditional NeRF pipeline.

We formulate the proposed MPNeRF with a NeRF branch, and an MPI branch denoted as  $\mathbf{G}_{\theta_1}$  and  $\mathbf{G}_{\theta_2}$ . Given a batch contains images from source and target viewpoints alongside the corresponding camera parameters. To train the NeRF branch, we cast rays for the source viewpoints pixels using the camera parameters following [38]. The MSE loss is adopted to supervise the NeRF branch with the ground truth color  $C(\mathbf{r})$ :

$$\mathcal{L}_{\text{NeRF}} = \sum_{\mathbf{r} \in B} \left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2, \quad (5)$$

where,  $B$  is the set of input rays during training. For the MPI branch, the encoder-decoder style MPI generator takes in images from the source view and outputs the corresponding MPI representation. In order to incorporate prior knowledge, we adopt a frozen Swin Transformer V2 model pre-trained with SimMIM [32] as a feature extractor to extract multi-scale features from aerial images. These features are fused to generate the final MPI representation. The loss function to optimize the MPI branch contains three components: L1 loss to match the synthesized target image  $\hat{\mathbf{I}}_t$  to ground truth  $\mathbf{I}_t$  at a pixel level, SSIM loss to encourage structure consistency, and LPIPS loss [63] for perceptual consistency.

$$\begin{aligned} \mathcal{L}_{L1} &= \left\| \hat{\mathbf{I}}_t - \mathbf{I}_t \right\|_1, \\ \mathcal{L}_{\text{SSIM}} &= 1 - \text{SSIM}(\hat{\mathbf{I}}_t, \mathbf{I}_t), \\ \mathcal{L}_{\text{LPIPS}} &= \left\| \phi(\hat{\mathbf{I}}_t) - \phi(\mathbf{I}_t) \right\|_1. \end{aligned} \quad (6)$$

The overall loss function to optimize the MPI branch would then be a sum of these three losses:

$$\mathcal{L}_{\text{MPI}} = \mathcal{L}_{L1} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{LPIPS}}. \quad (7)$$

These conventional loss functions train both branches to give predictions based on training views. Based on the investigation in Sec. 3.2, we aim to guide the training process of the NeRF with a multiplane prior learned by the MPI branch. An intuitive choice is sampling a random number of pixels from an unseen view and matching the predicted color of two branches with an MSE loss.

$$\mathcal{L}_{\text{Mul}} = \sum \left\| \hat{C}(\mathbf{r}_{i,j}) - \hat{\mathbf{I}}_{i,j} \right\|_2^2. \quad (8)$$

$i, j$  denotes the sampled pixel coordinates in the unseen viewpoint. In practice, this intuitive choice works surprisingly

well and the advantage of MPI is implicitly learned by NeRF. We give a further analysis of our design choice of  $\mathcal{L}_{\text{Mul}}$  in Sec. 4.5.

To summarize, the final objective functions of the NeRF branch  $\mathbf{G}_{\theta_1}$  and MPI branch  $\mathbf{G}_{\theta_2}$  are given as follows:

$$\begin{aligned} \mathcal{L}_{\mathbf{G}_{\theta_1}} &= \mathcal{L}_{\text{NeRF}} + \lambda \mathcal{L}_{\text{Mul}}, \\ \mathcal{L}_{\mathbf{G}_{\theta_2}} &= \mathcal{L}_{\text{MPI}}. \end{aligned} \quad (9)$$

In this way, the training experience of the MPI branch serves as a multiplane prior that guides the training process of NeRF. Even if the learned MPI is not entirely accurate, the NeRF branch benefits from this multiplane prior and thus avoids collapse during training.

## 4. Experiment

### 4.1. Implementation Details

Our method is implemented using PyTorch, and all experiments are conducted on a GeForce RTX 3090 GPU. For the NeRF branch, we use the original NeRF in [38]. During training, we randomly sample unseen views following the strategy proposed by [17]. The batch size is set to 1024 pixel rays in both source and unseen views. For each ray, we perform 64 coarse sampling and 32 fine sampling along the ray. For the MPI branch, we sample 16 layers of planes for each viewpoint. The optimization of the two branches of MPNeRF is performed using the Adam optimizer [22] with a learning rate of  $5 \times 10^{-4}$ . The hyperparameter  $\lambda$  in Eq. 9 is set to 1.

### 4.2. Datasets and Evaluation Metrics

The main experiments are conducted on 16 scenes collected by LEVIR-NVS [55]. These scenes contain various scenarios in common aerial imagery, including mountains, buildings, colleges, etc. 3 and 5 views are used for training and the rest for testing. Additional experiments and discussions can be found in the Appendix. In line with previous studies of few-shot neural rendering [17, 21, 39, 61], we report PSNR, SSIM and LPIPS [63].

### 4.3. Baseline Methods

We compare MPNeRF against various state-of-the-art methods including NeRF [38], Mip-NeRF [3], InfoNeRF [21], DietNeRF [17], PixelNeRF [62], RegNeRF [39] and FreeNeRF [61]. Among these methods, NeRF and Mip-NeRF are designed for dense view training, we mainly explore the performance gain achieved by MPNeRF. PixelNeRF aims to learn a generalized NeRF representation for all scenes and is pre-trained on the DTU dataset [18]. Since a large domain gap might exist when applied in aerial imagery, we report PixelNeRF's results with and without additional fine-tuning per scene. Other methods are designed for few-shot

Metrics	Methods	Building	Church	College	Mountain	Mountain	Observation	Building	Town	Stadium	Town	Mountain	Town	Factory	Park	School	Downtown	Mean
PSNR	NeRF[38]	13.64	12.06	14.45	18.60	18.40	14.56	14.10	14.98	15.02	13.18	20.88	14.11	14.28	15.44	14.68	13.67	15.13
	Mip-NeRF[3]	12.19	10.57	12.39	17.26	16.93	13.06	12.53	13.11	14.22	11.89	20.11	12.25	13.25	13.95	12.59	11.67	13.62
	InfoNeRF[21]	13.09	11.53	14.42	16.43	16.43	13.68	13.99	15.00	14.87	12.85	17.26	12.89	14.87	15.54	13.83	12.59	14.33
	DietNeRF[17]	13.44	12.20	14.86	19.34	18.67	15.27	13.73	15.78	16.68	14.21	20.66	14.73	16.73	16.55	14.97	13.61	15.71
	PixelNeRF[62]	6.07	6.26	7.68	12.09	10.24	6.77	6.19	7.35	5.74	6.02	12.47	6.89	4.96	6.53	4.52	5.88	7.23
	PixelNeRF ft[62]	12.44	11.76	11.74	17.42	17.15	14.44	11.18	15.86	19.65	16.09	23.99	15.24	13.02	15.70	13.86	13.86	15.21
	RegNeRF[39]	12.07	10.79	12.60	16.39	17.36	13.04	11.92	12.94	13.49	11.59	19.37	12.21	12.66	13.98	12.71	11.21	13.40
	FreeNeRF[61]	13.56	11.01	13.93	20.03	19.74	15.29	13.00	16.29	15.47	13.21	21.59	13.15	18.91	18.23	13.35	11.87	15.54
	Ours	18.81	17.93	20.71	25.50	24.92	19.56	18.64	21.59	22.08	21.20	28.57	21.41	22.61	23.57	20.71	19.73	21.72
SSIM	NeRF[38]	0.16	0.12	0.17	0.30	0.24	0.16	0.15	0.14	0.20	0.17	0.34	0.22	0.22	0.20	0.23	0.22	0.20
	Mip-NeRF[3]	0.12	0.09	0.16	0.30	0.24	0.14	0.14	0.13	0.21	0.15	0.35	0.17	0.24	0.17	0.21	0.12	0.18
	InfoNeRF[21]	0.15	0.12	0.17	0.26	0.21	0.14	0.15	0.15	0.22	0.15	0.28	0.17	0.30	0.24	0.21	0.11	0.19
	DietNeRF[17]	0.16	0.15	0.23	0.34	0.24	0.21	0.17	0.18	0.28	0.21	0.35	0.26	0.36	0.26	0.28	0.19	0.24
	PixelNeRF[62]	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.01
	PixelNeRF ft[62]	0.14	0.20	0.16	0.29	0.30	0.29	0.16	0.30	0.51	0.47	0.48	0.34	0.25	0.26	0.29	0.25	0.29
	RegNeRF[39]	0.12	0.11	0.16	0.28	0.27	0.14	0.13	0.13	0.20	0.13	0.34	0.16	0.20	0.16	0.21	0.11	0.18
	FreeNeRF[61]	0.24	0.11	0.21	0.37	0.33	0.26	0.17	0.28	0.27	0.23	0.38	0.21	0.51	0.42	0.24	0.14	0.27
	Ours	0.73	0.72	0.79	0.82	0.81	0.73	0.71	0.81	0.80	0.84	0.89	0.84	0.86	0.85	0.79	0.76	0.80
LPIPS	NeRF[38]	0.59	0.62	0.60	0.56	0.58	0.58	0.61	0.59	0.60	0.59	0.53	0.59	0.53	0.57	0.55	0.66	0.58
	Mip-NeRF[3]	0.64	0.66	0.66	0.60	0.64	0.64	0.62	0.62	0.61	0.63	0.60	0.64	0.56	0.62	0.63	0.65	0.63
	InfoNeRF[21]	0.60	0.60	0.60	0.57	0.59	0.59	0.61	0.68	0.58	0.60	0.57	0.61	0.53	0.55	0.60	0.62	0.59
	DietNeRF[17]	0.59	0.61	0.59	0.56	0.59	0.56	0.61	0.58	0.52	0.56	0.56	0.58	0.48	0.54	0.57	0.59	0.57
	PixelNeRF[62]	0.74	0.73	0.75	0.74	0.74	0.74	0.74	0.73	0.74	0.74	0.72	0.73	0.74	0.74	0.74	0.74	0.74
	PixelNeRF ft[62]	0.70	0.61	0.72	0.59	0.59	0.58	0.67	0.59	0.48	0.52	0.56	0.61	0.67	0.65	0.67	0.58	0.61
	RegNeRF[39]	0.65	0.65	0.67	0.61	0.62	0.63	0.63	0.62	0.63	0.64	0.59	0.65	0.58	0.62	0.64	0.66	0.63
	FreeNeRF[61]	0.61	0.67	0.64	0.55	0.58	0.58	0.63	0.58	0.61	0.61	0.57	0.64	0.48	0.54	0.61	0.65	0.60
	Ours	0.21	0.24	0.18	0.20	0.18	0.25	0.24	0.18	0.20	0.16	0.12	0.17	0.12	0.14	0.20	0.19	0.19

Table 1. **Quantitative comparison with different baseline methods in 3 views.** Our MPNeRF achieves the best results compared to prior arts for few-shot neural rendering in indoor and synthetic scenes. The best, second-best, and third-best entries are marked in ■, ■, and ■, respectively.

Metrics	Methods	Building	Church	College	Mountain	Mountain	Observation	Building	Town	Stadium	Town	Mountain	Town	Factory	Park	School	Downtown	Mean	
PSNR	NeRF[38]	13.44	12.64	15.62	19.91	19.39	15.79	15.11	16.69	16.44	14.90	22.06	13.44	13.76	15.25	13.67	13.31	15.71	
	Mip-NeRF[3]	12.42	10.94	12.91	18.13	17.24	14.14	12.74	13.65	15.08	12.40	20.79	13.21	14.33	14.78	13.08	11.94	14.24	
	InfoNeRF[21]	13.31	12.30	15.32	18.82	18.63	15.90	14.77	15.94	15.92	13.20	21.57	15.17	16.05	15.90	14.87	13.22	15.68	
	DietNeRF[17]	13.82	13.01	16.35	20.35	19.67	16.13	15.45	16.84	17.31	15.03	22.49	16.30	17.86	17.59	15.66	14.77	16.79	
	PixelNeRF [62]	6.07	6.31	7.81	12.03	10.22	6.84	6.29	7.41	5.80	6.03	12.45	6.85	5.04	6.59	4.58	5.92	7.27	
	PixelNeRF ft[62]	15.67	15.05	15.84	21.17	21.01	16.26	15.27	16.58	17.18	15.21	22.65	16.03	14.90	16.52	15.36	14.83	16.85	
	RegNeRF[39]	12.20	12.57	14.09	22.16	19.00	17.02	13.79	14.07	14.16	12.48	20.67	12.76	16.23	14.47	12.94	11.82	15.03	
	FreeNeRF[61]	16.83	16.66	19.54	21.97	21.44	18.28	17.93	19.81	16.63	17.90	23.95	20.37	21.42	17.90	18.25	15.77	19.04	
	Ours	20.50	19.56	23.08	26.02	24.88	21.29	20.99	21.92	23.07	21.57	29.00	22.19	22.58	23.59	21.72	20.50	22.65	
SSIM	NeRF[38]	0.16	0.12	0.17	0.30	0.24	0.16	0.15	0.14	0.20	0.17	0.34	0.22	0.22	0.20	0.23	0.22	0.20	
	Mip-NeRF[3]	0.12	0.09	0.15	0.30	0.23	0.17	0.14	0.13	0.21	0.14	0.36	0.17	0.25	0.18	0.21	0.12	0.19	
	InfoNeRF[21]	0.15	0.12	0.17	0.26	0.21	0.14	0.15	0.15	0.22	0.15	0.28	0.17	0.30	0.24	0.21	0.11	0.19	
	DietNeRF[17]	0.16	0.15	0.23	0.34	0.24	0.21	0.17	0.18	0.28	0.21	0.35	0.26	0.36	0.26	0.28	0.19	0.24	
	PixelNeRF[62]	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.01	
	PixelNeRF(ft)[62]	0.14	0.20	0.16	0.29	0.30	0.29	0.16	0.30	0.51	0.47	0.48	0.34	0.25	0.26	0.29	0.25	0.29	
	RegNeRF[39]	0.12	0.24	0.26	0.54	0.36	0.40	0.25	0.17	0.20	0.15	0.35	0.17	0.44	0.17	0.20	0.11	0.26	
	FreeNeRF[61]	0.47	0.44	0.44	0.44	0.39	0.35	0.41	0.47	0.33	0.48	0.44	0.54	0.65	0.41	0.46	0.37	0.44	
	Ours	0.81	0.80	0.87	0.87	0.83	0.81	0.81	0.84	0.85	0.87	0.90	0.87	0.87	0.86	0.83	0.81	0.84	
LPIPS	NeRF[38]	0.59	0.62	0.60	0.56	0.58	0.58	0.61	0.59	0.60	0.59	0.53	0.59	0.53	0.57	0.55	0.66	0.58	
	InfoNeRF[21]	0.60	0.60	0.60	0.57	0.59	0.59	0.61	0.61	0.68	0.58	0.60	0.57	0.61	0.53	0.55	0.60	0.62	0.59
	Mip-NeRF[3]	0.63	0.65	0.65	0.59	0.63	0.60	0.62	0.61	0.60	0.62	0.58	0.63	0.55	0.60	0.62	0.64	0.61	
	DietNeRF[17]	0.59	0.61	0.59	0.56	0.59	0.56	0.61	0.58	0.52	0.56	0.56	0.58	0.48	0.54	0.57	0.59	0.57	
	PixelNeRF[62]	0.74	0.73	0.75	0.74	0.74	0.74	0.74	0.73	0.74	0.74	0.72	0.73	0.74	0.74	0.74	0.74	0.74	
	PixelNeRF(ft)[62]	0.70	0.61	0.72	0.59	0.59	0.58	0.67	0.59	0.48	0.52	0.56	0.61	0.67	0.65	0.67	0.58	0.61	
	RegNeRF[39]	0.64	0.56	0.58	0.43	0.55	0.47	0.58	0.58	0.62	0.61	0.57	0.63	0.46	0.60	0.62	0.65	0.57	
	FreeNeRF[61]	0.52	0.50	0.52	0.52	0.55	0.53	0.53	0.50	0.56	0.48	0.55	0.47	0.43	0.55	0.51	0.56	0.52	
	Ours	0.17	0.20	0.14	0.18	0.18	0.21	0.19	0.16	0.17	0.15	0.12	0.15	0.10	0.12	0.17	0.14	0.16	

Table 2. **Quantitative comparison with different baseline methods in 5 views.** Our MPNeRF achieves the best results compared to prior arts for few-shot neural rendering in indoor and synthetic scenes. The best, second-best, and third-best entries are marked in ■, ■, and ■, respectively.

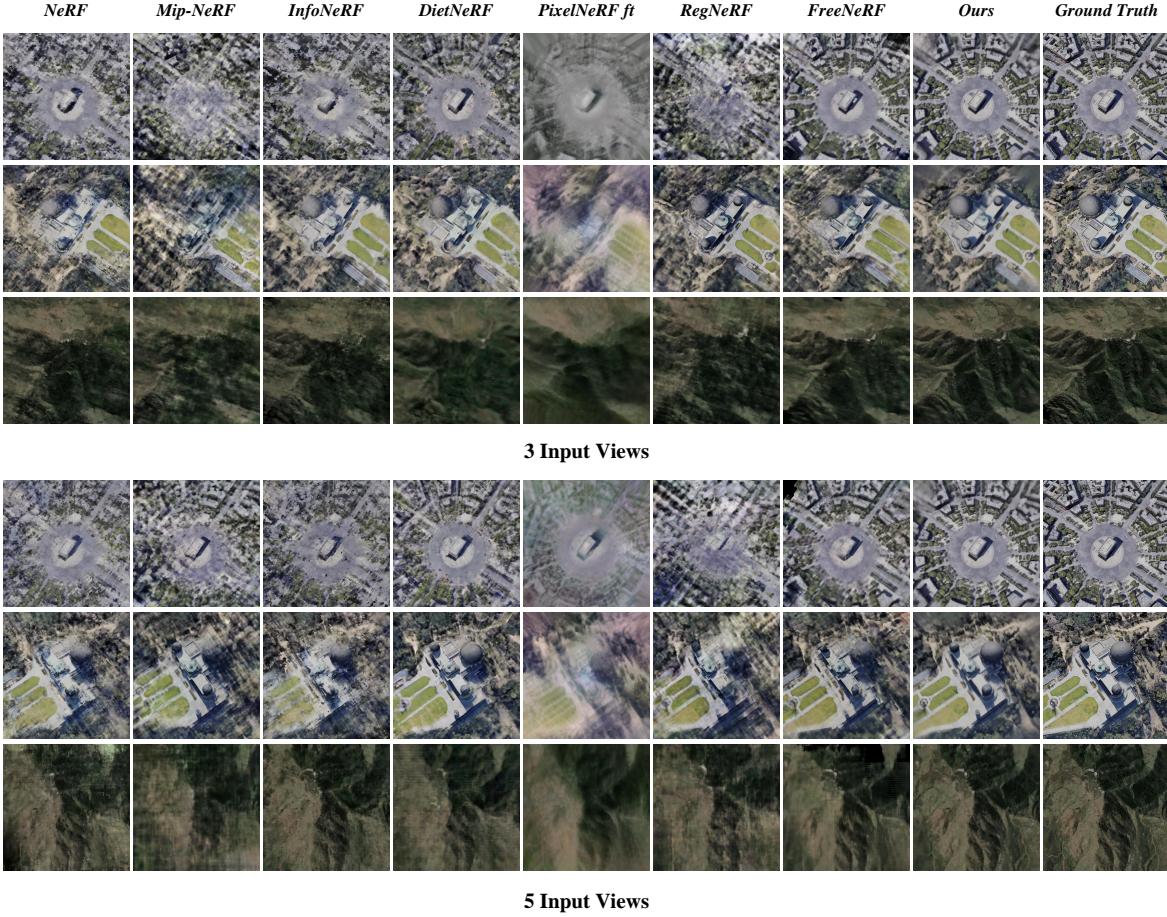


Figure 3. Visual comparisons on 3 selected scenes with 3 and 5 views. MPNeRF achieves photo-realistic quality in different scenes compared with ground-truth images on novel views.

neural rendering, and we conduct comparative experiments to investigate their performance when encountering aerial imagery.

#### 4.4. Comparative Results Analysis

Table 1 and Table 2 report the performance of MPNeRF and baseline methods in the 3-view and 5-view settings. Additionally, a qualitative comparison can be observed in Fig. 3. A very significant improvement can be found in all three metrics and rendering fidelity. The results demonstrate that PixelNeRF tends to produce blurry renderings, which we attribute to the poor localization of the CNN features. InfoNeRF and RegNeRF use local smoothness and sparsity to regularize NeRF explicitly. However, in scenarios with substantially limited information compared to the scene complexity, the performance of these methods could be compromised. DietNeRF implicitly distills the prior knowledge encoded in CLIP [41] and achieves better results. FreeNeRF investigates the frequency in NeRF training. By progressively learning each frequency component, FreeNeRF

has demonstrated remarkable effectiveness. Nonetheless, the progressive frequency regularization leads to relatively flat results, favoring PSNR but not metrics that consider local structures such as SSIM and LPIPS.

In fact, NeRF’s representation makes recovering 3D scenes from sparse inputs ill-posed. MPNeRF acquires superior results by the guiding of a multiplane prior to gaining a stronger understanding of local structures and semantics. In the more challenging scenes, such as Building in Figure 3, MPNeRF successfully avoids collapse during training.

#### 4.5. Ablation Studies and Further Analyses

**Ablation Analysis.** We ablate the proposed multiplane prior to our method, and the results are shown in Table 3. Intuitively, it seems better to use MPI as a guide after fully training it, we first construct experiments where a two-stage training strategy is employed. We then assess the impact of SwinV2’s pre-trained weights on performance by removing them. Next, we evaluate the contribution of multi-scale features by disconnecting the skip connections in the MPI

generator. Finally, we integrate the multiplane directly within NeRF’s sampling space, omitting the separate MPI branch, to examine the inductive biases’ influence on performance. Employing MPI concurrent branch during training leads to slight improvements. We believe this is because the MPI’s training experience itself carries information. The exclusion of SwinV2’s pre-trained weights declines performance, affirming the value of the encoded prior knowledge. Similarly, omitting the multi-scale feature connection diminishes the fidelity of the rendered images. Most significantly, the absence of the MPI generator results in a marked decrease in all metrics. This suggests that uncalibrated MPI generated by self-attention and convolution is important to avoid degenerate solutions. Collectively, these findings demonstrate that each element of the proposed multiplane prior is crucial for the superior performance of MPNeRF.

Methods	PSNR	SSIM	LPIPS
Ours	21.72	0.80	0.19
training MPI beforehand	21.49	0.76	0.20
w/o pre-trained weights	20.46	0.75	0.24
w/o multi-scale feature	20.05	0.71	0.25
w/o MPI generator	16.38	0.42	0.52
Baseline NeRF	15.13	0.20	0.58

Table 3. Ablation analysis on the proposed Multiplane Prior.

**Further Analyses on the design choice of the  $\mathcal{L}_{Mul}$ .** One intuitive thought of designing  $\mathcal{L}_{Mul}$  is that geometry recovered by the MPI branch may provide more information than color alone. So we design two experiments, one is to match the expected depth of both branches as an auxiliary depth loss, and another is to model density on each ray as a distribution [21] and minimize the KL divergence. Another intuitive thought is that the choice of  $\mathcal{L}_{Mul}$  should reflect the local or nonlocal relationships within the pixels. Therefore, we adopt the recently proposed S3IM loss [58] to measure this relationship.

Design Choice	PSNR	SSIM	LPIPS
w/t depth matching	21.34	0.75	0.20
w/t ray matching	21.11	0.71	0.21
w/t relation matching	21.19	0.70	0.21
Ours	21.72	0.80	0.19

Table 4. Design choice of the  $\mathcal{L}_{Mul}$

However, as shown in Table 4, the result suggests that these intuitive designs worsen the results. The first two design involves direct supervision of the depth generated by the MPI generator. The last involves capturing the non-local relationships between the predictions of the NeRF and the MPI branch. Since the learned MPI is not entirely accurate, we believe the noise within pseudo-labels may compromise performance with these enhanced supervisions applied.

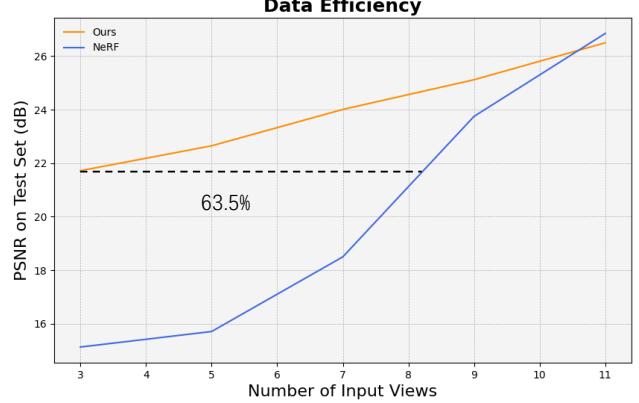


Figure 4. We investigate the data efficiency achieved by our method. Our method requires up to 63.5% training images to achieve a similar performance compared to a vanilla NeRF model.

**Impact of Different Pre-trained Models.** We perform a comparison study on three pre-trained vision transformers, i.e., CLIP [41], DINOV2 [40], and SimMIM [57]. We adopt the base model in our experiment. As shown in Table 5, all of these methods provide comparable results. The results show that the Swin Transformer pre-trained via SimMIM [57] outperforms others. We believe this can be attributed to the rich global and local details learned by SimMIM and the hierarchical structure of the Swin Transformer.

Pre-trained Image Encoder	PSNR	SSIM	LPIPS
CLIP [41]	20.15	0.70	0.25
DINOV2 [40]	20.01	0.68	0.25
SimMIM (Ours) [57]	21.72	0.80	0.19

Table 5. Impact of different pre-trained models.

**Data Efficiency** Since we aim to improve the capability of NeRF in aerial scenes when only sparse views are available, we investigate how much data MPNeRF can save to achieve similar rendering results compared to the original NeRF that requires dense view supervision. As shown in Fig. 4, the results show that our method requires up to 63.5% training images. This may help save energy and establish resource-efficient applications for UAVs based on NeRF.

## 5. Limitations and Conclusion

In this work, we introduce Multiplane Prior guided NeRF (MPNeRF), the first approach designed for few-shot aerial scene rendering. Through the guiding of the multiplane prior, MPNeRF effectively overcomes the typical pitfalls in sparse aerial scenes. We hope our work can provide insight into future NeRF-based applications in aerial scenes. However, further exploration of the guiding strategy design is needed. In particular, incorporating uncertainty prediction mechanisms or implementing grid-based representations holds promise for future research directions.

## References

- [1] Brendan Alvey, Derek T Anderson, Andrew Buck, Matthew Deardorff, Grant Scott, and James M Keller. Simulated photo-realistic deep learning framework and workflows to accelerate computer vision and unmanned aerial vehicle research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2021. 1
- [2] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 4(11):e21, 2019. 14
- [3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2, 5, 6, 16
- [4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2, 3
- [5] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8924–8933, 2019. 14
- [6] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019. 2
- [7] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12952, 2022. 3
- [8] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 16
- [9] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. In *Computer Graphics Forum*, pages 305–314. Wiley Online Library, 2012. 2
- [10] Celso M de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. Next-generation deep learning based on simulators and synthetic data. *Trends in cognitive sciences*, 2022. 1
- [11] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2, 3
- [12] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. *Computer Vision and Pattern Recognition*, 2021. 3
- [13] Thibaud Ehret, Roger Marí, and Gabriele Facciolo. Regularization of nerfs using differential geometry. *arXiv preprint arXiv:2206.14938*, 2022. 2, 3
- [14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 16
- [15] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. *The Lumigraph*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2
- [16] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Learning a neural 3d texture space from 2d exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8356–8364, 2020. 2
- [17] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 1, 2, 3, 5, 6, 16
- [18] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanaes. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 5
- [19] Daiju Kanaoka, Motoharu Sonogashira, Hakaru Tamukoh, and Yasutomo Kawanishi. Manifoldnerf: View-dependent image feature supervision for few-shot neural radiance fields. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. BMVA*, 2023. 2, 3
- [20] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017. 2
- [21] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022. 1, 2, 3, 5, 6, 8, 16
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 17
- [23] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 2
- [24] Minseop Kwak, Jiuhn Song, and Seungryong Kim. Geconerf: Few-shot neural radiance fields via geometric consistency. *arXiv preprint arXiv:2301.10941*, 2023. 3
- [25] Obin Kwon, Jeongho Park, and Songhwai Oh. Renderable neural radiance map for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9108, 2023. 1
- [26] Marc Levoy and Pat Hanrahan. Light field rendering. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996. 2
- [27] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 2, 12, 16

- [28] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2
- [29] Zhemin Li, Hongxia Wang, and Deyu Meng. Regularize implicit neural representation by itself. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10280–10288, 2023. 4
- [30] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 2
- [31] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 3
- [32] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2, 5, 16
- [33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 17
- [34] Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4018–4025. IEEE, 2023. 1
- [35] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-NeRF: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1311–1321, 2022. 3
- [36] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3
- [37] Christopher Maxey, Jaehoon Choi, Hyungtae Lee, Dinesh Manocha, and Heesung Kwon. Uav-sim: Nerf-based synthetic data generation for uav-based perception. *arXiv preprint arXiv:2310.16255*, 2023. 1
- [38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 5, 6, 12, 14, 16
- [39] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Reg-nerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1, 2, 3, 4, 5, 6, 16
- [40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 8
- [42] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35:151–173, 1999. 2
- [43] Seunghyeon Seo, Yeonjin Chang, and Nojun Kwak. Flipnerf: Flipped reflection rays for few-shot novel view synthesis. *arXiv preprint arXiv:2306.17723*, 2023. 2, 3
- [44] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2
- [45] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 1
- [46] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 3, 14
- [47] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. 2020. 3
- [48] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 2, 12, 16
- [49] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 836–850. Springer, 2014. 2
- [50] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*, 2023. 2, 3
- [51] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 2, 3
- [52] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2, 3
- [53] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural

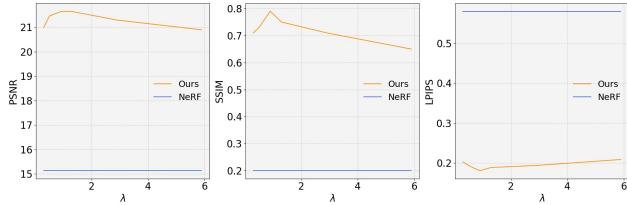
- radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 3
- [54] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 2
- [55] Yongchang Wu, Zhengxia Zou, and Zhenwei Shi. Remote sensing novel view synthesis with implicit multiplane representations. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2, 5, 15, 16
- [56] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering. 2021. 3
- [57] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3, 8, 16
- [58] Zeke Xie, Xindi Yang, Yujie Yang, Qi Sun, Yixiang Jiang, Haoran Wang, Yunfeng Cai, and Mingming Sun. S3im: Stochastic structural similarity and its unreasonable effectiveness for neural fields. In *International Conference on Computer Vision*, 2023. 8
- [59] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2
- [60] Lining Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8306, 2023. 14
- [61] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 2, 4, 5, 6, 12, 16
- [62] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelfnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 3, 5, 6, 16
- [63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5
- [64] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2
- [65] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1

In this supplement, we first conduct more experimental results and discussion to evaluate the robustness and efficiency of our proposed Multiplane Prior guided NeRF (MPNeRF). We also include more qualitative results to discuss the motivation and limitations of MPNeRF. Finally, we add more details of experimental settings and implementations.

## A. Additional Experiments and Analysis

**Robustness to Hyperparameters.** We have conducted a series of experiments to assess the sensitivity of our model to hyperparameters. Specifically, we focus on the hyperparameter  $\lambda$ , which plays a crucial role in balancing different components of our loss function. In Figure 5, we illustrate the impact of varying  $\lambda$  on the performance of the proposed MPNeRF and a standard NeRF [38] model.

As  $\lambda$  increases, we observe that the PSNR and SSIM metrics tend to plateau, suggesting that there is an optimal range for  $\lambda$  wherein the model achieves a balance between fidelity and perceptual quality. On the other hand, the LPIPS metric shows an initial decrease followed by a gradual increase, indicating a sweet spot where the model best captures the perceptual features of the aerial scenes. The trends exhibited by MPNeRF show its relative insensitivity to  $\lambda$  within a reasonable range, which underscores the robustness of our method. Notably, MPNeRF consistently outperforms the baseline NeRF model across all metrics, demonstrating the effectiveness of incorporating the multiplane prior to the rendering process.



**Figure 5. Hyperparameter Sensitivity Analysis.** Performance comparison of our method (MPNeRF) and a baseline NeRF model across different values of hyperparameter  $\lambda$ . The graphs show PSNR, SSIM, and LPIPS metrics. MPNeRF is robust to a wide  $\lambda$  choice.

**MPNeRF vs MPI-based Methods.** The present study of MPI mainly focuses on overcoming shortcomings such as “failure to represent continuous 3D space” in [27]. In contrast, our approach utilizes MPI as a bridge to convey complex information that a single NeRF struggles with. We construct comparison experiments under 3-view settings among the proposed MPNeRF, the original MPI [48], and MINE [27]. In Table 6, the original MPI achieves an 18.57 PSNR, 0.54 SSIM, and 0.45 LPIPS. While MINE performs better with 19.99 PSNR, 0.61 SSIM, and 0.40 LPIPS. Our MPNeRF outperforms these methods by a large margin. These MPI-based methods face inherent limitations like ghosting effects

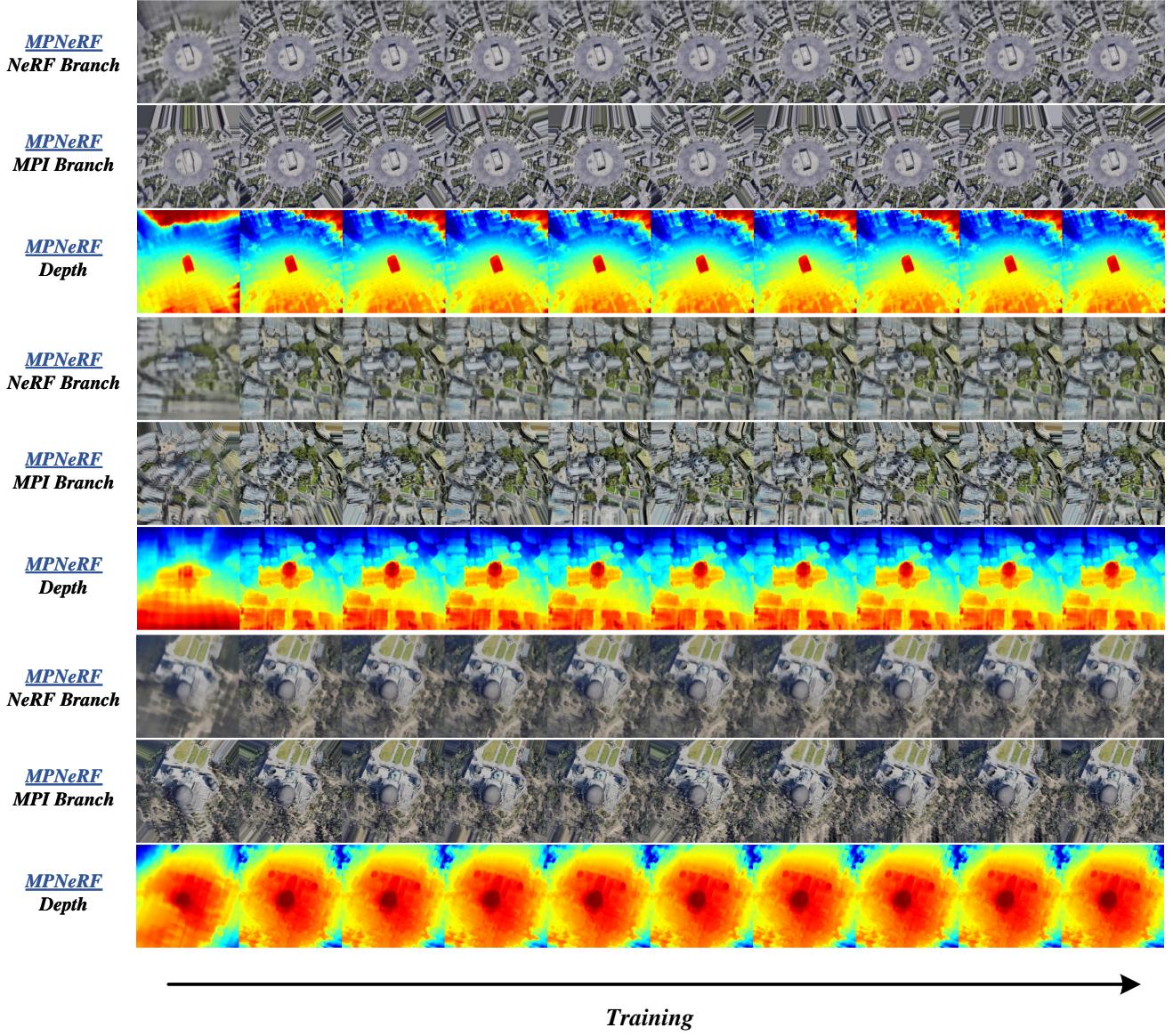
and cropped corners under sparse inputs and large camera movements.

Methods	PSNR	SSIM	LPIPS
MPI [48]	18.57	0.54	0.45
MINE [27]	19.99	0.61	0.40
NeRF branch w/o $\mathcal{L}_{mul}$	15.13	0.20	0.58
MPI branch	20.32	0.57	0.34
NeRF branch w/t $\mathcal{L}_{mul}$	21.72	0.80	0.19

**Table 6. Comparation between the NeRF and MPI.** This table presents the evaluation of the MPI based method in previous studies, NeRF branch without multiplane loss ( $\mathcal{L}_{mul}$ ), the MPI branch independently, and the NeRF branch with  $\mathcal{L}_{mul}$  within our MPNeRF framework. The metrics of PSNR, SSIM, and LPIPS demonstrate the significant impact of the multiplane prior on the rendering performance in sparse aerial scenes.

**NeRF Branch vs MPI Branch.** In Table 6, we examine the performance impact of the NeRF and MPI branches within our proposed MPNeRF. Initially, the NeRF branch without the multiplane loss  $\mathcal{L}_{mul}$  (equals to a plain NeRF model) demonstrates a PSNR of 15.13, an SSIM of 0.20, and an LPIPS of 0.58. These values indicate a baseline level of performance where the NeRF branch struggles with sparse aerial views, as evidenced by the low PSNR and SSIM scores, along with a high LPIPS value which suggests a significant perceptual difference from the ground truth. In contrast, the MPI branch alone shows better across all metrics, with a PSNR of 20.32, an SSIM of 0.57, and a reduced LPIPS of 0.34. The MPI branch’s improved performance is likely due to its discrete depth-based representation that aligns better with the structured nature of aerial scenes, thus capturing the scene geometry more effectively. And the inductive bias of CNN and Transformer makes MPI generalize better. The most significant performance gains are observed when the NeRF branch is combined with the multiplane loss  $\mathcal{L}_{mul}$ , resulting in a PSNR of 21.72, an SSIM of 0.80, and an LPIPS of 0.19. The addition of  $\mathcal{L}_{mul}$  to the NeRF branch enhances its ability to recover details from sparse views, as reflected by the substantial improvements in PSNR, SSIM, and LPIPS. The proposed Multiplane Prior serves as a bridge to convey information that is hard to learn by the traditional NeRF pipeline. These results underscore the efficacy of incorporating multiplane priors into the NeRF framework for few-shot aerial scene rendering.

**Other Few-shot NeRF Methods Combined with Multiplane Prior.** It stands to reason that it’s worth evaluating other Few-shot NeRF methods combined with multiplane prior. We incorporated FreeNeRF’s [61] frequency regularization and evaluated it under a 3-view setting. This integration results in a marginal increase in the PSNR by 0.2db. We believe the MPI’s noisy predictions help reduce early



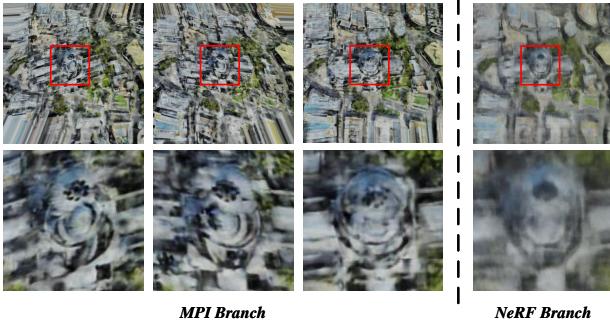
**Figure 6. Training Progression of MPNeRF.** The sequence shows comparative results from the NeRF and MPI branches at various training stages under 3 view settings. Left to right: early, mid, and late phases of training. The NeRF branch initially shows noisier reconstructions with indistinct depth estimations, while the MPI branch exhibits crop edge and overlapping ghosting effects. Over time, the NeRF branch, guided by the MPI-derived multiplane prior, progressively captures finer details and more accurate depth information, as reflected in the sharpening of depth map visualizations.

training overfitting in high-frequency details. This mechanism seems to parallel the underlying concept of FreeNeRF, potentially explaining the minimal improvement.

## B. Discussion and Future Works

**Why MPNeRF Works?** Despite the advantage of the MPI representation in aerial scenes, a simple question is: Why MPNeRF is kept away from the cropped edge and overlapping ghosting effect of the MPI? Avoiding the cropped

edge is simple, we sample rays from unseen views following the mask generated during homography warping. To better illustrate why the overlapping ghosting effect can not be learned by NeRF, we visualize the same target view rendered by the MPI branch in Figure. 7. With the source viewpoint varying, the overlapping ghosting effect in the rendered target view differs. Since MPI derived from different viewpoints does not share a common world space, these overlapping ghosting effects are not multi-view consistent across all views. Thus these effects violate the multi-view



**Figure 7. Detail Comparison between MPI and NeRF Branch Outputs.** The images on the left column represent the MPI branch’s output, displaying sharper details with overlapping ghosting effects in the highlighted regions. In contrast, the right column shows the NeRF branch’s output, where the same regions appear more blurred.

consistency assumption of NeRF [38]. With these noises provided as pseudo-supervision, the MLP optimized with gradient descent tends to give blurry rendering. The blurring signifies NeRF’s attempt to average out the incongruities across views. These pseudo-labels, while derived from an informed place, act as imperfect guides, introducing a trade-off that MPNeRF must navigate. On one hand, they provide a rich, albeit noisy, signal that captures the complexity of aerial scenes. On the other, they present a risk of polluting the training process with artifacts.

Although MPNeRF shows that a simple MSE loss can perform well, this delicate balance highlights the importance of a carefully crafted training regimen, one that can differentiate between useful signals and misleading noise. Our future work will delve into refining this balance, potentially through the development of more sophisticated noise-filtering mechanisms or through the implementation of more robust training strategies that can better leverage the nuanced information within these pseudo-labels. In doing so, we may further enhance MPNeRF’s rendering quality, pushing the boundaries of few-shot aerial scene rendering.

**Semantic Integration for Improved Scene Understanding.** Integrating semantic segmentation into the MPNeRF framework offers an exciting direction for enhancing scene understanding. By associating semantic labels with the MPI branch, MPNeRF may provide more contextually aware reconstructions and pave the way for applications in urban planning and navigation under limited data.

**Scene Editing.** An exciting avenue for future research is the possibility of editing NeRF-rendered scenes by directly manipulating the MPIS generated by the MPI branch. This could enable users to alter scene characteristics such as color, texture, or even geometric structure, through an intuitive interface. A potential direction is utilizing differentiable rendering techniques to backpropagate the desired edits from the

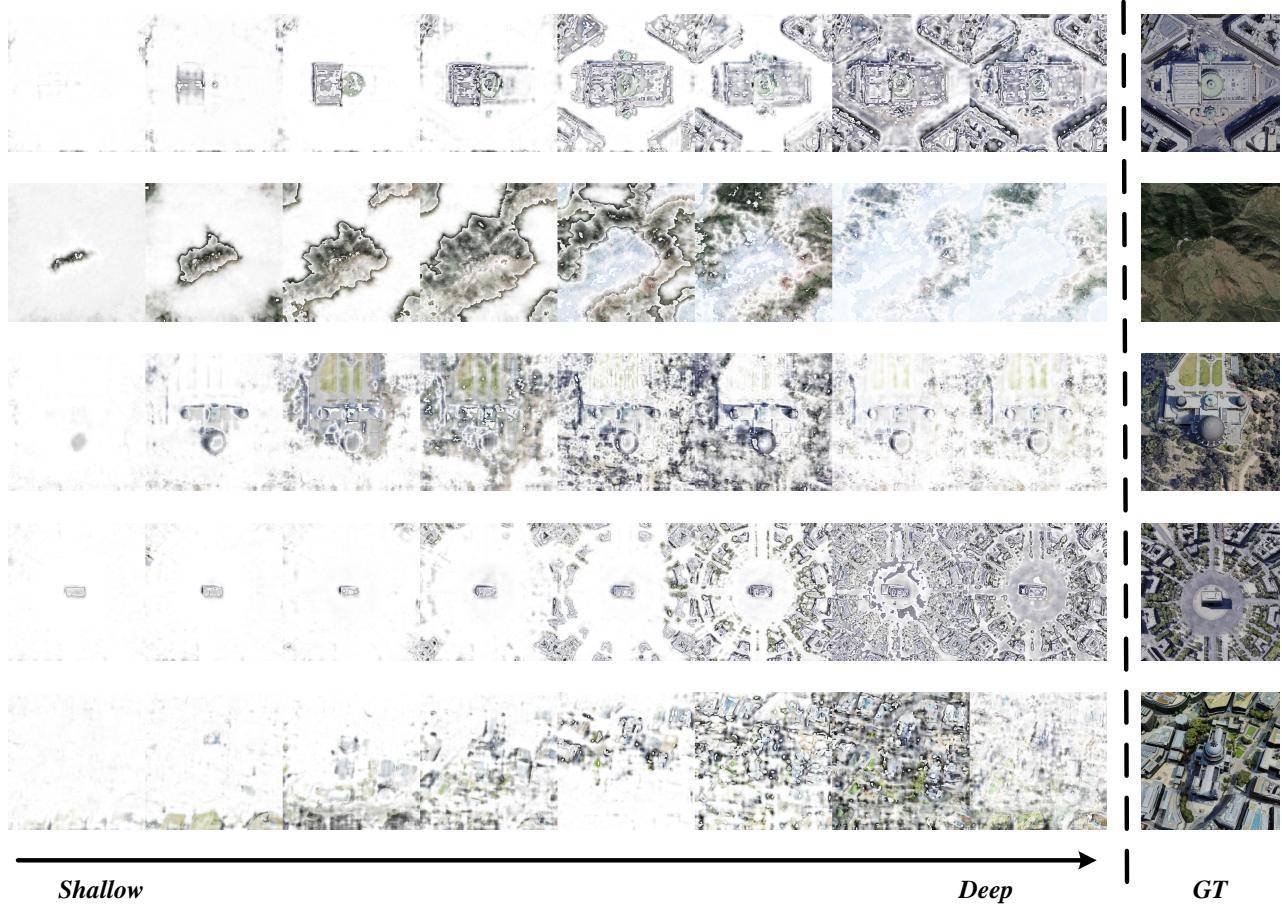
scene rendering back to the MPI and NeRF representations.

**Scalability.** Currently, scalability remains a potential limitation when our MPNeRF model is applied to larger scenes. The primary bottleneck arises from the inherent capacity constraints of NeRF models. They are typically optimized for smaller, more controlled environments and can struggle to maintain fidelity at the increased scale and complexity of larger scenes. As scenes expand in size, the NeRF’s neural network requires a corresponding increase in capacity to model the additional detail, which can lead to a significant escalation in computational and memory requirements [46, 60]. Furthermore, the encoder-decoder architecture employed within our MPI branch is not ideally suited for high-resolution imagery [2, 5]. It tends to consume substantial amounts of memory, especially when processing the finer details necessary for large-scale scene rendering. The memory footprint grows rapidly with the resolution of input images due to the quadratic increase in the number of pixels that need to be processed simultaneously.

## C. Additional Visualizations.

**Training Progression of MPNeRF.** Figure. 6 presents a detailed visual account of the training evolution within our MPNeRF, delineating the comparative outcomes from the NeRF and MPI branches across three distinct training phases. The left columns illustrate the initial stage where the NeRF branch outputs are notably noisier, and the depth maps lack precise definition, signifying the model’s initial struggle to interpret the sparse aerial views. These preliminary results are characterized by a lack of clarity and detail, with the depth maps displaying broad, undifferentiated regions of low confidence. As training progresses to the midpoint, displayed in the center columns, the MPI branch starts to assert its strengths. It delivers reconstructions with improved clarity and begins to better capture the geometric intricacies of the aerial scenes. This enhancement is evident in the depth maps, where we observe a transition from broad, undefined areas to more distinct regions of depth estimation, indicative of the MPI branch’s capability to delineate structural features more effectively at this stage. Reaching the later stages of training, shown in the right columns, the NeRF branch, now informed by the multiplane prior, shows significant advancement. It starts to match and, in certain aspects, surpasses the MPI branch’s performance by delivering images with greater detail fidelity. This is most apparent in the depth maps, where the once diffused and expansive high-confidence regions have now evolved into sharply defined areas, highlighting the network’s improved proficiency in depth perception.

The visualization of the depth maps is particularly telling; the sharpening of these maps directly correlates with the improved model’s depth estimations. The NeRF branch, leveraging the multiplane prior, demonstrates an enhanced ability to resolve the complex spatial relationships inherent in aerial



**Figure 8. Visualization of MPI Branch Depth Layers.** Sequential depth layers from the MPI branch reveal the aerial scene’s structure, evolving from translucent to opaque as we move from shallow to deep layers, culminating in the ground truth (GT) image for reference.

scenes, moving beyond the initial limitations evidenced in the early training outputs.

This sequential improvement underscores the efficacy of the MPNeRF training process, which effectively leverages the distinct advantages of both NeRF and MPI branches to progressively refine the model’s understanding of the scene, culminating in high-quality renderings from sparse inputs. The journey from noisy, indistinct initial attempts to clear, detailed final outputs exemplifies the potent potential of MPNeRF for aerial scene rendering.

**Different Layers of the MPI Branch.** To explore the geometry and appearance captured by the MPI branch, we visualize the color with transparency computed by the density of different MPI layers. Figure. 8 showcases a series of images that represent different layers of the MPI branch, each corresponding to a specific depth level within the aerial scene, as labeled from ‘Shallow’ to ‘Deep’. The images progress from the topmost layers, which capture high-elevation features like roofs, to the bottom layers, which reveal ground-level details. However, it is evident that the fidelity of the recon-

struction varies across depth layers. The initial layers, while capturing the broad layout, lack the finer details and the sharpness present in the ground truth (GT). The middle layers begin to show more structure and texture, indicating an intermediate range where the MPI branch most effectively captures the scene’s appearance. The deeper layers, while richer in detail, start to exhibit artifacts, such as blurring and possible misalignments, before converging towards the ground truth. This suggests that while the MPI branch of MPNeRF shows promise in reconstructing aerial scenes from limited data, it is still highly inaccurate and contains artifacts.

## D. More Implementation Details.

**Datasets and Metrics.** Our evaluation is conducted on a dataset that presents a rich tapestry of aerial landscapes, the LEVIR-NVS [55], comprising 16 diverse scenes that span mountains, urban centers, villages, and standalone architectural structures. Each scene in the dataset is represented by a collection of 21 multi-view images, each with a resolu-

Layer	Kernel Size	In-Channels	Out-Channels	Input	Activation
convdown1	1	768	512	encoder_layer4	ELU [8]
convdown2	3	512	256	convdown1	ELU
convup1_extra	3	256	256	convdown2	ELU
convup2_extra	1	256	768	convup1_extra	ELU
convup5	3	768 + 21	256	cat(convup2_extra, depth_embedding)	ELU
conv5	3	256 + 768 + 21	256	cat(convup5, encoder_layer3, depth_embedding)	ELU
convup4	3	256	128	conv5	ELU
conv4	3	128 + 384 + 21	128	cat(convup4, encoder_layer2, depth_embedding)	ELU
output4	3	128	4	conv4	Sigmoid (for RGB) and abs (for $\sigma$ )
convup3	3	128	64	conv4	ELU
conv3	3	64 + 192 + 21	64	cat(convup3, encoder_layer1, depth_embedding)	ELU
output3	3	64	4	conv3	Sigmoid (for RGB) and abs (for $\sigma$ )
convup2	3	64	32	conv3	ELU
conv2	3	32 + 96 + 21	32	cat(convup2, encoder_conv1, depth_embedding)	ELU
output2	3	32	4	conv2	Sigmoid (for RGB) and abs (for $\sigma$ )
convup1	3	32	16	conv2	ELU
conv1	3	16	16	convup1	ELU
output1	3	16	4	conv1	Sigmoid (for RGB) and abs (for $\sigma$ )

**Table 7. Decoder Architecture for the MPI Branch.** Each convup layer within our architecture is composed of a convolution layer, followed by batch normalization and the specified activation layer, as delineated in the table. This sequence is then succeeded by a  $2 \times$  nearest neighbor upsampling process. Conversely, the convdown blocks are structured beginning with a max pooling layer with a stride of 2, followed by a convolution layer, and culminating with an activation layer. This architecture choice follows previous research in MPI representations and depth estimation [14, 27, 55].

tion of  $512 \times 512$  pixels. This selection ensures a broad representation of scenarios that MPNeRF might encounter in real-world applications. The LEVIR-NVS dataset encapsulates a variety of pose transformations that mimic the dynamic nature of UAV flight patterns, including wrapping and swinging motions. These pose variations introduce realistic challenges in aerial photography, such as changes in viewpoint and scale, making the dataset a rigorous testing ground for our model. The inclusion of these complex transformations in the simulation process is crucial for assessing the robustness of MPNeRF’s performance in conditions that closely approximate actual aerial image capture.

In our experimental setup, we strategically select specific views for training to assess the capability of our model in both interpolation and extrapolation scenarios. For the three-view setting, we utilize view IDs: 0, 7, and 15. This selection is designed to provide a spread of perspectives that challenges the model to extrapolate the scene effectively. In the five-view setting, we expand our selection to include view IDs: 0, 7, 10, 15, and 20. This broader range tests the model’s interpolation skills and its ability to extrapolate scenes from more diverse viewpoints.

In our experiments, we employ three standard metrics. Peak Signal-to-Noise Ratio (PSNR) is used to measure the image reconstruction quality, calculated as the negative logarithm of the mean squared error between the predicted and ground truth images. Structural Similarity Index Measure (SSIM), obtained via the skimage<sup>1</sup> library, assesses image quality based on luminance, contrast, and structural infor-

mation. Learned Perceptual Image Patch Similarity (LPIPS), computed using a VGG-based model from the lpips<sup>2</sup> package, evaluates perceptual similarity, reflecting more human-centric assessments of image quality.

**Implementation of Baseline Methods.** We implement the baseline methods following their open-source code base. We adopt 64 coarse sampling and 32 fine sampling for the NeRF backbone of these methods. In particular, the RegNeRF [39] and FreeNeRF [61] are implemented based on Mip-NeRF [3], and others [17, 21, 38, 62] are based on a vanilla NeRF. All methods are trained for 30 epochs for each scene and the hyperparameters are strictly consistent across all experiments.

**Implementation of MPNeRF.** We implement MPNeRF based on the nerf-pl codebase<sup>3</sup>, which provides a PyTorch Lightning framework for efficiently operationalizing NeRF architectures. The settings of hyperparameters are strictly consistent with baseline methods. Our NeRF branch adheres closely to the original NeRF paper specifications, ensuring a faithful reproduction of the baseline model. We adopt 64 coarse sampling and 32 fine sampling for the NeRF branch. Inspired by previous works [27, 48, 55], our MPI branch is constructed following an encoder-decoder architecture MPI generator. The encoder is a strand SwinV2 Transformer [32] pretrained via SimMIM [57]. The encoder is kept frozen during training. A detailed description of our decoder architecture is presented in Table. 7. The MPI generator embeds depth hypotheses into the input features, which are then pro-

<sup>1</sup><https://scikit-image.org/>

<sup>2</sup><https://github.com/richzhang/PerceptualSimilarity>

<sup>3</sup><https://github.com/kwea123/nerfp1>

cessed through convolutional layers to output MPIS with RGB and density values, leveraging skip connections and multi-scale representations for detail enhancement.

For optimization, we utilize the Adam optimizer [22] with a learning rate of  $5 \times 10^{-4}$ , and a cosine learning rate decay scheduler [33]. Our model is trained on a single NVIDIA RTX 3090 GPU for 30 epochs, taking about 2.5 hours to converge. The batch size is set to 1024 rays per iteration for both seen and unseen views, allowing sufficient diversity of data points for gradient estimation while maintaining manageable memory requirements.