SAMPLING: Scene-adaptive Hierarchical Multiplane Images Representation for Novel View Synthesis from a Single Image

Xiaoyu Zhou[1] Zhiwei Lin[1] Xiaojun Shan[1] * Yongtao Wang[1] † Deqing Sun[2] Ming-Hsuan Yang[2,3]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Google Research [3]University of California, Merced

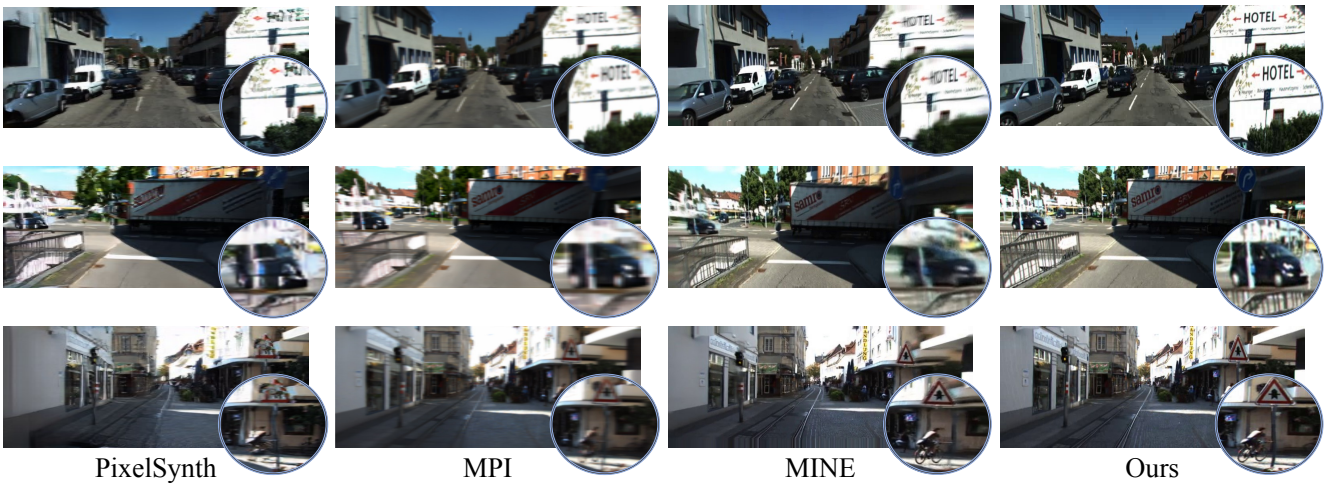

Figure 1: **Novel view synthesis result comparisons.** Given a single image captured in an outdoor scene, our method synthesizes novel views with fewer visual artifacts, geometric deformities, and blurs. Notably, our method models favorable intricate details, such as tiny objects, symbols, and traffic signs, resulting in more photo-realistic views.

Abstract

Recent novel view synthesis methods obtain promising results for relatively small scenes, e.g.*, indoor environments and scenes with a few objects, but tend to fail for unbounded outdoor scenes with a single image as input. In this paper, we introduce SAMPLING, a* ***S****cene-****a****daptive Hierarchical* ***M****ulti****pl****ane* ***I****mages Representation for* ***N****ovel View Synthesis from a Sin****g****le* ***I****mage based on improved multiplane images (MPI). Observing that depth distribution varies significantly for unbounded outdoor scenes, we employ an adaptive-bins strategy for MPI to arrange planes in accordance with each scene image. To represent intricate geometry and multi-scale details, we further introduce a hierarchical refinement branch, which results in high-quality synthesized novel views. Our method demonstrates considerable performance gains in synthesizing large-scale unbounded outdoor scenes using a single image on the KITTI dataset and generalizes well to the unseen Tanks and Temples dataset. The code and models will be made public.*

1. Introduction

Taking a photo and using it to synthesize photo-realistic images at novel views is an important task with a wide range of applications, such as generating realistic data for training AI models (*e.g.*, autonomous driving perception and robot simulation). This task is challenging as it requires a precise understanding of 3D geometry, reasoning about occlusions, and rendering high-quality, spatially consistent novel views. It becomes even more difficult for large-scale unbounded outdoor scenes, which contain complex geometric conditions, various objects, and diverse depth distributions corresponding to different scenes.

Recently, Neural Radiance Field (NeRF) [1] based methods have gained much attention by synthesizing photo-realistic images with dense multi-view inputs. By leveraging Multi-layer Perceptron (MLP) layers, NeRF implicitly models a specific scene via RGB values and volume occupancy density. However, NeRF-based methods are

*Work is done during the internship at Peking University.
†Corresponding author.

primarily applicable for rendering bounded objects or interiors, which are impeded by the stringent requirement for the dense views captured from different angles, precise corresponding camera poses, and unobstructed conditions [2, 3, 4]. Furthermore, these methods rely on per-scene fitting and cannot easily generalize to unseen scenes. Several methods [5, 6, 7, 8] try to utilize multi-modal data, *e.g.*, LiDAR scans and point clouds, to complicate the synthesis of novel views in large scenes. However, additional modalities are difficult to obtain and have greater memory consumption and computational costs. Besides, similar to NeRF, these multi-modal methods require multiple input views with large overlaps and need to be trained per scene.

In contrast, the Multiplane Images (MPI) representation [9] has shown promising results in synthesizing scenes from sparse views, using a set of parallel semi-transparent planes to approximate the light field. The MPI representation is particularly effective at understanding complex scenes with challenging occlusions [10]. However, prior MPI-based approaches place planes at fixed depths with equal intervals, have limitations in modeling irregular geometry, such as texture details, and do not perform well in unbounded outdoor scenes, as shown in Figure 1. For complicated geographic features and differentiated depth ranges, the uniform static MPI [11, 9, 12] are often over-parameterized for large areas of space, yet under-parameterized for the occupied scenes. In addition, using single-scale scene representation in MPI also limits the quality of the synthesized images in large-scale scenes, leading to apparent artifacts and blurs.

In this paper, we introduce SAMPLING, a scene-adaptive hierarchical representation for novel view synthesis from a single image based on improved MPI. Instead of generating multiplanes with a static uniform strategy, we design the Adaptive-bins Generation strategy to adaptively distribute the planes according to each input image. This strategy enables a more efficient representation to better fit various unbounded outdoor scenes. Additionally, we propose a Hierarchical Refinement Branch that utilizes multi-scale information from large scenes, incorporating both global geometries and high-frequency details into the MPI representation. This branch enhances the quality of intermediate scene representations, resulting in more complete and high-quality synthesized images. Our method achieves high-quality view synthesis results on challenging outdoor scenes, such as urban scenes, and shows a well cross-scene generalization, enabling a more versatile scene representation. Our main contributions are:

- We present a novel scene-adaptive representation for synthesizing new views from a single image. Our approach is based on learnable bins for MPI, enabling the learning of a more efficient and effective unbounded scene representation from a single view.
- We develop a hierarchical refinement method for 3D representations of outdoor scenes. We show that representing scenes with hierarchical information can synthesize new images with favorable details.
- Our method achieves new state-of-the-art performance in outdoor view synthesis from a single image. Experimental results also show our method generalizes well for both outdoor and indoor scenes.

2. Related Work

2.1. Novel View Synthesis

Novel view synthesis (NVS) aims to render unseen viewpoints of the scene from the observed images. Recently, numerous deep models have been introduced to represent 3D objects or scenes and synthesize images in novel views. Some methods exploit generative models for image generation and completion [13, 14, 15, 16], while others exploit explicit or implicit 3D scene representations [1, 17, 18, 11] derived from input images and synthesize new viewpoints through differentiable rendering.

The recent methods based on neural rending fields (NeRF) [19, 20, 21] have achieved state-of-the-art results for implicit neural 3D scene representation. Given a set of posed images, a NeRF method maps the 3D position and direction to a density and radiance by the multilayer perceptron (MLP), followed by differentiable volume rendering to synthesize the images. Typically, the original NeRF [1] model is trained per scene and requires dense inputs with accurate camera poses. To make NeRF more practical, the NeRF in the wild method [4] requires only unstructured collections of in-the-wild photographs. In [22, 23, 24], Lin *et al*. train NeRF from imperfect (or even unknown) camera poses. Moreover, Yu *et al*. [25] introduce PixelNerf, which predicts a continuous neural scene representation conditioned on one or a few input images. Other approaches explore the possibilities of NeRF in more application scenarios, such as dynamic scenes [19, 26, 27], controlled editing [28, 29], and interior scene [30, 31]. However, if the input is sparse (or even a single image), or the scene is large and complicated (*e.g.*, street view), the novel views synthesized by NeRF-based methods will be of low quality and contain artifacts. Furthermore, existing works for novel view synthesis need to be trained per scene, lacking general representation for scene understanding.

In contrast to NeRF, Multiplane Images (MPI) methods can synthesize novel views from fewer images, due to the properties of explicitly modeling scenes with sparse inputs. Using a stack of RGB-α layers at various depths, the MPI representation mimics the light field in a differentiable manner. In recent years, significant advances have been made in MPI for novel view synthesis. For instance, Zhou *et al*. [11] use MPI for realistic rendering of novel views with a stereo

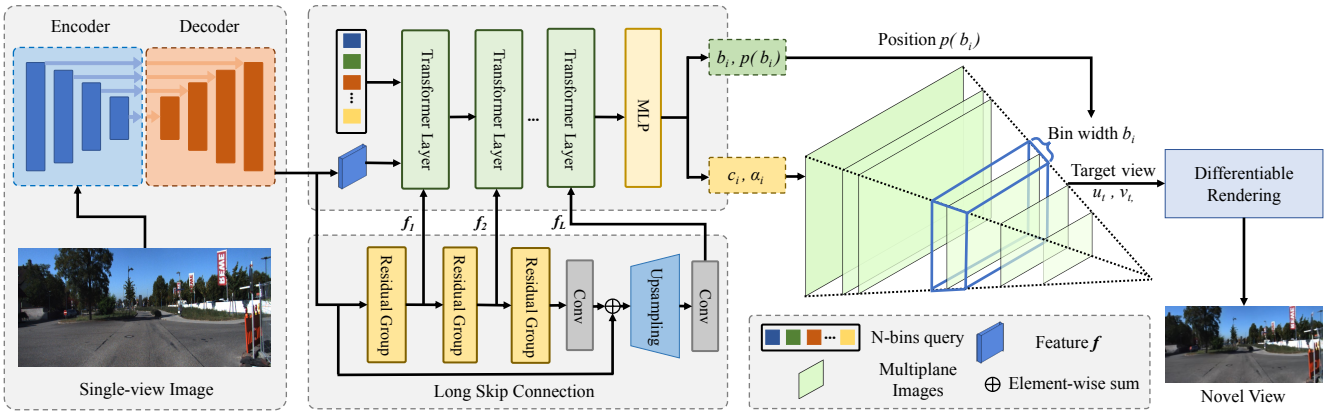

Figure 2: **An overall pipeline of our proposed method for novel view synthesis from a single image.** Given a single-view image as input, we first employ an encoder-decoder network combined with skip connections to extract features. The features are then fed into the Adaptive-bins MPI generation module along with an N-bins query, which calculates the adaptive positions of the MPI. Simultaneously, the Hierarchical Refinement Branch extracts hierarchical residual features with a set of Residual Groups and passes them to Transformer Layers. The MPI position $p(b_i)$ and representation (c_i, α_i) are then predicted by an MLP head to synthesize the novel views through the differentiable rendering.

image pair. In [9], an MPI-based method is developed to synthesize views directly from a single image input, leading to higher-quality results compared to traditional light fields. DeepView [32] further applies learned gradient descent to estimate multiplane images from sparse views, replacing the simple gradient descent update rule with a deep network. To improve the real-time performance, NeX [10] models view-dependent effects by performing basis expansion on the pixel representation. MINE [12] takes advantage of the MPI and NeRF, proposing a continuous depth MPI method for NVS and depth estimation. AdaMPI [33] designs two modules to improve MPI by adjusting plane depth and predicting depth-aware color, but it ignores multi-scale unbounded features in outdoor scenes. However, these approaches have limitations in modeling unbounded outdoor scenes with multi-scale information and complex geometry. They also fail to obtain detailed high-frequency information, leading to apparent artifacts, blurs, and defects when synthesizing images in large-scale scenes.

2.2. Large-Scale Neural Scene Rendering

Recent advances in neural rendering have exhibited considerable success in 3D object modeling and interior scene reconstruction. Nevertheless, current methods demonstrate suboptimal performance when applied to unbounded outdoor scenes. Numerous methods have been developed to address this issue. Block-NeRF [34] enables large-scale scene reconstruction by dividing large environments into multiple blocks and representing each block with an individual NeRF network. BungeeNeRF [35] introduces a progressive neural radiance field, which models diverse multi-scale scenes with varying views on multiple data sources. However, these methods can only model large outdoor driving scenes that are observed from dense input sensor views and precise camera poses. With high-speed shots, the outdoor driving scenes typically have very sparse viewpoints and limited view diversity. To tackle the above challenges, recent methods have explored multimodal fusion methods for neural rendering. Rematas *et al*. [36] extend NeRF to leverage asynchronously captured LiDAR data and to supervise the density of rays. Similarly, CLONeR [7] introduces the camera-LiDAR fusion to the outdoor driving scene, where LiDAR is used to supervise geometry learning. Li *et al*. [5, 37, 38] propose to synthesize photo-realistic scenes with the help of large-scale point clouds, using neural point-based rendering. However, current multimodal approaches take a two-stage synthesis strategy, that is, first pre-processing all multimodal data to reconstruct a rough 3D scene and then rendering a novel view image from the reconstructed 3D scene. Costly multimodal data collection, complex pre-processing, and per-scene training limit the efficiency and application of these methods. In contrast, we introduce a high-efficient representation method called SAMPLING. With only a single image as the input, our method can generate novel view images from end to end and produce more realistic images with fewer artifacts and deformities. Besides, our method does not necessitate per-scene optimization and thus reduces training costs.

3. Method

The overview architecture of SAMPLING is shown in Figure 2. Given a single image I, SAMPLING learns to generate the multiplane images (MPI) with discretized adaptive-bins and hierarchical feature refinement module. Synthesis images can then be rendered at novel viewing an-

gles from the generated MPI.

3.1. Adaptive-bins MPI Generation

We utilize MPI to explicitly represent the 3D geometry of the source view. MPI consists of N front-parallel RGB-α planes arranged at depths $d_0, \dots d_{N+1}$. Each plane i encodes an RGB color image c_i and an alpha map α_i.

Most existing works employ a uniform-fix MPI distribution strategy (*e.g.*, MPI [9] and MINE [12]), where planes are placed at fixed depths with equal intervals. However, depth distribution corresponding to different RGB inputs can vary dramatically, especially for outdoor scenes. Thus, we introduce an adaptive binning strategy for MPI generation. We discretize the depth interval into N bins, where the bin widths are adaptively obtained for each image, and distribute each plane of MPI according to the adaptive bins.

Specifically, we first extract the image feature f by sending a single-view image into an encoder-decoder network. The encoder-decoder network utilizes skip connections to produce the high-resolution image feature in a coarse-to-fine style. Then, we employ a transformer module to calculate the distribution of adaptive-bins MPI. The transformer module consists of several transformer layers, as shown in Figure 2. Similar to Adabins [39] and Binsformer [40], we randomly initialize N learnable bin queries f_b for depth prediction. Meanwhile, the feature f is viewed as the MPI query for RGB-α plane predictions in MPI. In each transformer layer, the MPI query f is sent to a Hierarchical Refinement Branch to produce the residual feature f_r. The residual feature f_r is viewed as values and keys to calculate the cross-attention with the concatenated queries f_b and f. Then, the updated concatenated queries f_b and f are subsequently sent to a self-attention layer and a feed forward layer, as shown in Figure 3. After that, a shared multi-layer perception head is performed over the N-bins query f_b and feature f to predict bin width $\widetilde{b}$ and generate (c_i, α_i) for each plane of MPI. We apply the softmax function to normalize the sum of widths $\widetilde{b}_i$ to 1 as follows:

$$\{b_i\}_{i=1}^{N} = Softmax(\{\widetilde{b}_i\}_{i=1}^{N}), \tag{1}$$

where b_i is the i^{th} normalized bin width. Finally, we calculate the adaptive depth location of each plane in MPI by:

$$p(b_i) = d_{near} + (d_{far} - d_{near})(\frac{b_i}{2} + \sum_{j=1}^{i-1} b_j), \tag{2}$$

where $p(b_i)$ is the position assigned to the i^{th} adaptive-bins MPI. d_{near} and d_{far} are the nearest and farthest distances of the planes in the frustum of the camera, respectively.

3.2. Hierarchical Refinement Branch

Synthesizing novel views from a single image often faces difficulty in capturing multi-scale scene features, resulting in visually obvious holes and blurs. To address this issue, we propose a Hierarchical Refinement Branch to improve the feature with multi-scale information, which has proven effective in both 3D scene representation [41] and single image super-resolution [42, 43] tasks.

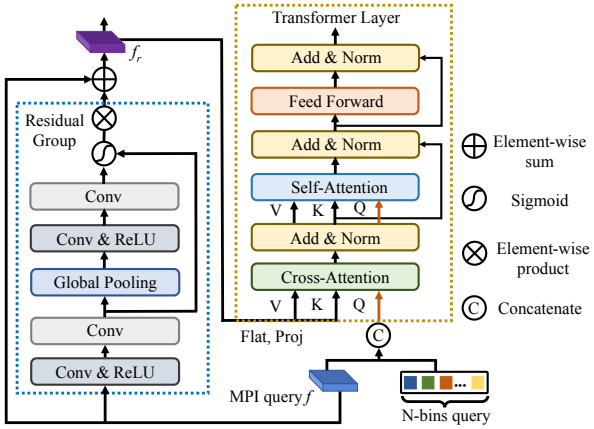

Figure 3: **Network details of the connection of Transformer Layer and Residual Group.** The combination of the two modules enables MPI representation to obtain both precise distribution and multi-scale detailed information.

Specifically, we employ a coarse-to-fine architecture, where the low-resolution planes enforce the smoothness in scenes, and high-resolution planes refine the geometry details. Given the shallow feature f from the encoder-decoder network, we employ a set of residual groups (RG) [44, 45] with upscale modules to extract hierarchical residual features $\{f_r | r = 1, 2, ..., L\}$, which can be formulated as:

$$f_r = H_{RG_r}(f_{r-1}), \tag{3}$$

where H_{RG_r} represents the r^{th} residual group, L is the number of the residual groups. RG aims to restore the high-frequency information and to extract rich edge and texture information of the outdoor scenes. The structural details of the RG are shown in Figure 3.

Besides, to stabilize the training process, we introduce a long skip connection, an additional upsampling block, and two convolution layers when calculating the last high-resolution residual feature f_L. Subsequently, we feed the output f_L into the last transformer layer, encouraging the generation of MPI to pay more attention to the informative details of scenes.

3.3. Differentiable Rendering in MPI

The synthesized MPI can be rendered in the target view by first warping each plane from the source viewpoint and then applying the composite operator to aggregate the warping results of each plane. The overall MPI rendering can be formulated as follows:

$$\hat{I}_t = O(W(C), W(A)), \tag{4}$$

where $\hat{I}_t$ denotes the synthesized image, W is the homography warping function, and O is the composite operator. $C = \{c_1, \ldots, c_N\}$ denotes the set of RGB channels and $A = \{\alpha_1, \ldots, \alpha_N\}$ is the corresponding alpha channel.

We first employ the homography warping operation for the i^{th} plane from the target to source view depending on the position $p(b_i)$ of each plane. Given the rotation matrix R, the translation matrix t from the target to source view, and the intrinsic matrix K_s and K_t for source and target views, we can generate the synthesized target-view image through W as follows:

$$[u_s, v_s, 1]^\top \sim K_s(R - \frac{tn^\top}{p(b_i)})(K_t)^{-1}[u_t, v_t, 1]^\top, \quad (5)$$

where $[u_s, v_s]$ and $[u_t, v_t]$ are coordinates in the source and target views, respectively. n is the norm vector of the i^{th} plane at the position $p(b_i)$. The MPI representation of the target view can be obtained by warping each layer from the source viewpoint to the desired target viewpoint using Eq. (5), finding the corresponding pixel for each pixel in the target frame. The MPI representation under the target view (c'_i, α'_i) can be defined as:

$$f(x) = \begin{cases} c'_i(u_t, v_t) = c_i(u_s, v_s), \\ \alpha'_i(u_t, v_t) = \alpha_i(u_s, v_s). \end{cases} \quad (6)$$

Finally, the synthesized target-view image can be then rendered via the compositing procedure [46] as follow:

$$\hat{I}_t = \sum_{i=1}^{N}(c'_i\alpha'_i \prod_{j=i+1}^{N}(1 - \alpha'_j)). \quad (7)$$

This rendering equation is completely differentiable, so our model can be trained from end-to-end.

3.4. Loss Function

Our overall loss combines an adaptive-bins loss to constrain the distribution of MPI according to each scene image and a synthesis loss to guide the network to synthesize images following the target views images.

Adaptive-bins loss. This loss term enforces that the distribution of MPI follows the ground truth value of the adaptive depth for each image:

$$L_{ada} = \sum_{x \in X} \min_{y \in p(b_i)} \|x - y\|^2 + \sum_{y \in p(b_i)} \min_{x \in X} \|x - y\|^2, \quad (8)$$

where $p(b_i)$ denotes the arranged depth of MPI and the set of all depth values in the ground truth image as X.

Synthesis loss. This loss aims at matching the synthesized target image with the ground truth by measuring the mean square error of RGB value and SSIM value [47]:

$$L_{syn} = \frac{1}{HW} \sum \left| \hat{I}_t - I_t \right| - SSIM(\hat{I}_t, I_t), \quad (9)$$

where $\hat{I}_t$ and I_t are the synthesized novel image and ground truth image with the same size of $H \times W$.

The total loss is given by:

$$L = \lambda_{ada} L_{ada} + L_{syn}, \quad (10)$$

where λ_{ada} is the parameter to balance the loss terms.

4. Experiments

We present quantitative and qualitative evaluations of our method on the KITTI [50] dataset and generalization performance on Tanks and Temples (T&T) [51], compared with prior view synthesis methods. To assess the quality of the synthesized novel views, we mainly focus on the evaluation metrics of Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [47], and Learned Perceptual Image Patch Similarity (LPIPS) [52]. All metrics are computed over all pixels.

4.1. Evaluating Quality

To demonstrate the efficacy of our method, we compare it to state-of-the-art methods for novel view synthesis. Following the settings of [9, 55], we train our model on the city subset of the raw KITTI dataset, randomly taking either the left or the right image as the source (the other being the target) at each training step. Following [12, 9], we evaluate the model on 4 test sequences of KITTI, cropping 5% from all sides of all images.

We compare our method with state-of-the-art approaches for NVS using different types of 3D representations, including the traditional NeRF-based method [4], Neural Point-based methods [38, 5], generative model-based methods [14, 48, 49], layer representation such as LDI-based method [18], and MPI-based methods [32, 9, 12]. Note that the traditional NeRF-based and Neural Point-based methods require per-scene optimization and pre-processing for exploiting additional supervision. Quantitative comparison results are presented in Table 1.

Compared with NeRF-based methods. We observe that our method outperforms NRW [4] by a large margin, although NRW introduces multiple supervision as well as paired poses to guarantee the training of the MLP. NPBG [38] and READ [5] are two Neural Point-based methods, exploiting the extra point clouds for supervision and synthesizing large-scale driving scenes with neural rendering. Our method achieves competitive results across all three metrics and improves the SSIM from 0.781 to 0.883 compared with the state-of-the-art methods on KITTI.

Compared with generative models. Based on generative models, SynSin [14] and PixelSynth [48] both utilize a high-resolution point cloud representation of learned features. 3D-Photo [49] presents a learning-based inpainting model combined with a Layered Depth Image, using depth

Table 1: **Overall comparison of SAMPLING with existing state-of-the-art approaches for novel view synthesis on the KITTI city dataset.** Note that ↑ denotes higher is better and ↓ means otherwise. The symbol † denotes the need for per-scene optimization and we use the average over all scenes as the final score. To ensure fairness, we follow the settings of MPI [9] and MINE [12] and show the results with $N = 64$.

Methods	Supervision	PSNR ↑	SSIM ↑	LPIPS ↓
NRW† [4]	RGB + Point Clouds + Depth	18.02	0.568	0.310
NPBG† [38]	RGB + Point Clouds	19.58	0.627	0.248
READ† [5]	RGB + Point Clouds	23.48	0.781	0.132
Synsin [14]	RGB + Point Clouds	16.70	0.520	-
PixelSynth [48]	RGB + Point Clouds + Depth	17.13	0.602	-
3D-Photo [49]	RGB + Depth + Edges	18.39	0.742	0.175
LSI [18]	RGB	16.52	0.572	-
Deepview [32]	RGB	17.28	0.716	0.196
MPI [9]	RGB	19.54	0.733	0.158
MINE [12]	RGB	21.65	0.818	0.117
SAMPLING (Ours)	RGB	**23.67**	**0.883**	**0.101**

Figure 4: **Qualitative comparison of novel view synthesis on the KITTI dataset.** Visualization results show our method generates better details compared to other single-view NVS methods, including PixelSynth [48], MPI [9], and MINE [12].

and linked depth edges as additional supervision. Although these methods perform well in indoor scenes, they struggle with complex unbounded outdoor scenes due to the absence of strict geometric constraints and multi-scale features.

Compared with layered representation methods. Similar to MPI, LSI [18] applies a layer-structured 3D representation of a scene from a single input image. Compared with LSI, our method boosts the results by 7.15% on PSRN. DeepView [32], MPI [9], and MINE [12] are MPI-based or MPI-NeRF methods for novel view synthesis. Notably, we improve the performance of MPI for outdoor scenes on all metrics, compared with these existing methods.

We also visually compare the view synthesis results in Figure 4. Our method produces more realistic images with high-quality details, more complete edge geometries, and fewer artifacts and distortions. For small objects (*e.g.*, pedestrians and traffic cones) and scene text (*e.g.*, traffic signs), our method also shows favorable synthesis perfor-

Table 2: **Generalization study on T&T.** We evaluate the generalization of our method on the Tanks and Temples (T&T) dataset that provides different scenes from KITTI.

Methods	Training Set	T&T		
		PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [1]	T&T	22.14	0.676	-
NerfingMVS [30]		19.31	0.464	-
Monosdf [53]		21.48	0.689	-
ResNeRF [54]		23.39	0.795	-
3D-Photo [49]		**23.63**	0.848	0.136
MPI [9]	KITTI	18.62	0.614	0.260
MINE [12]		21.04	0.748	0.196
Ours		23.56	**0.852**	**0.125**

mance. The visualization confirms the effectiveness of our method in modeling the geometry and texture details of unbounded outdoor scenes.

We further show a qualitative comparison of disparity maps on the KITTI dataset in Figure 5. Similar to [9, 12], we use the models trained with KITTI to synthesize disparity maps from MPIs generated by our method and MINE [12]. We can observe that MINE [12] displays missing and distorted areas in depth maps, leading to unpleasant visual artifacts. In contrast, our method excels in adaptively aligning the depth of various outdoor scenes, promoting the synthesis of more precise geometric shapes and well-aligned boundaries of visible objects. The proposed hierarchical refinement branch also serves as guidance for generating smooth and refined disparity maps, as well as synthesized outputs. More results and video are available in the supplementary material.

4.2. Generalization

We further examine the generalization ability of our method using Tanks and Temples (T&T) dataset. Specifically, we train our model on the KITTI dataset and evaluate it on the advanced sets of T&T that contains indoor scenes.

We compare our method with existing methods for indoor scenes synthesis, such as NerfingMVS [30], Monosdf [53], and ResNeRF [54]. These methods employ explicit or implicit representation techniques to model a single scene with dense views as inputs. Note that these methods need to be trained separately for each scene, while our method can be trained in all scenes at once. We also compare our method with 3D-Photo [49] as well as MPI-based methods, including MPI [9] and MINE [12]. The quantitative results are presented in Table 2 and synthesis views of T&T are shown in Figure 6.

When evaluated on T&T, our method still maintains a high level of performance. Quantitative results demonstrate that our method outperforms implicit representation approaches (*e.g.*, Monosdf [53] and ResNeRF [54]), despite their use of multiple dense views as input. 3D-Photo [49] exploits a multi-layer representation for novel view synthesis and achieves state-of-the-art performance. Our method catches up with 3D-Photo [49] on PSNR and exceeds it in terms of SSIM and LPIPS. Due to the relatively tight geometric constraints and different depth distributions of the interior, it can be observed that MPI-based methods [9, 12] show a certain degree of decline on T&T, using the trained model on outdoor scenes (*e.g.*, KITTI). Nevertheless, SAMPLING still exhibits good performance with minimal degradation. This can potentially be attributed to the employing of adaptive-bins MPI, leading to the image-level scene representation adaption. Additionally, our proposed hierarchical refinement branch aids in obtaining the multi-scale details of scenes, enhancing the generalization capability.

Table 3: **Comparison of the proposed method with varying numbers of planes on the KITTI dataset.** N denotes the preset number of planes for MPI.

N	PSNR ↑	SSIM ↑	LPIPS ↓
8	20.02	0.812	0.139
16	22.18	0.856	0.114
32	23.53	0.879	0.105
64	23.67	0.883	0.101
128	**23.68**	**0.885**	**0.100**

Table 4: **Comparison of different strategies for MPI distribution.** Uniform-Fix and Log-Fix are two strategies for arranging MPI, both of which employ a static method for generating MPI and sampling.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓
Uniform-Fix MPI	21.98	0.837	0.118
Log-Fix MPI	22.53	0.862	0.112
Adaptive-bins MPI (Ours)	**23.67**	**0.883**	**0.101**

4.3. Ablation Studies

In this section, we conduct ablation experiments to analyze the effectiveness of each setting of our method, including the main components and hyperparameters. We evaluate our method on KITTI dataset in the following experiments.

Number of MPI planes. The performance of MPI representation is related to the number N of planes. To study the influence of the number of MPI, we train our network with various values of N and report results in Table 3. We can see consistent improvements with increasing N in our method. As the number of MPI increases, they can represent more complex scenes with a wider range of depth values. By contrast, sparse MPI (*e.g.*, 8 planes) settings can lead to inadequate scene representation for large-scale outdoor scenes with a wide depth range. We use 64 planes of MPI in our experiments, which achieves good performance and computation cost trade-off.

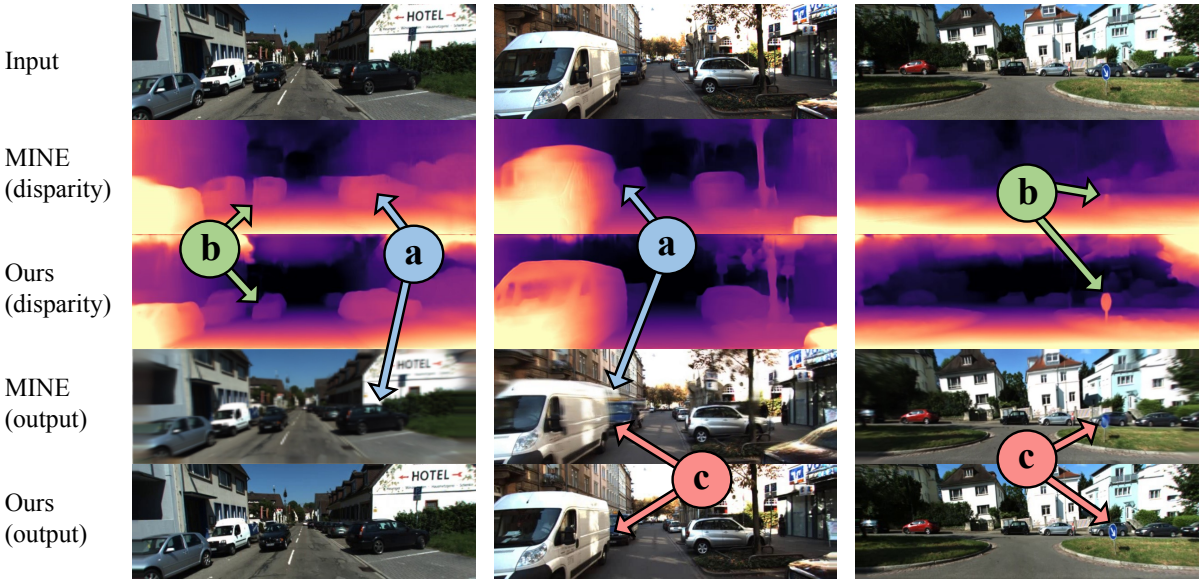

Figure 5: **Qualitative comparison of disparity map and novel view synthesis on the KITTI dataset.** (a) Disparity maps in [12] exhibit structural biases and missing objects, leading to unpleasant artifacts and distortions in the output. (b) The comparative disparity maps show that our method is capable of better recovering the spatial structure of complex scenes and intricate object boundaries. (c) Our method consistently delivers higher-quality and flawlessly disparity maps and outputs, even in challenging regions.

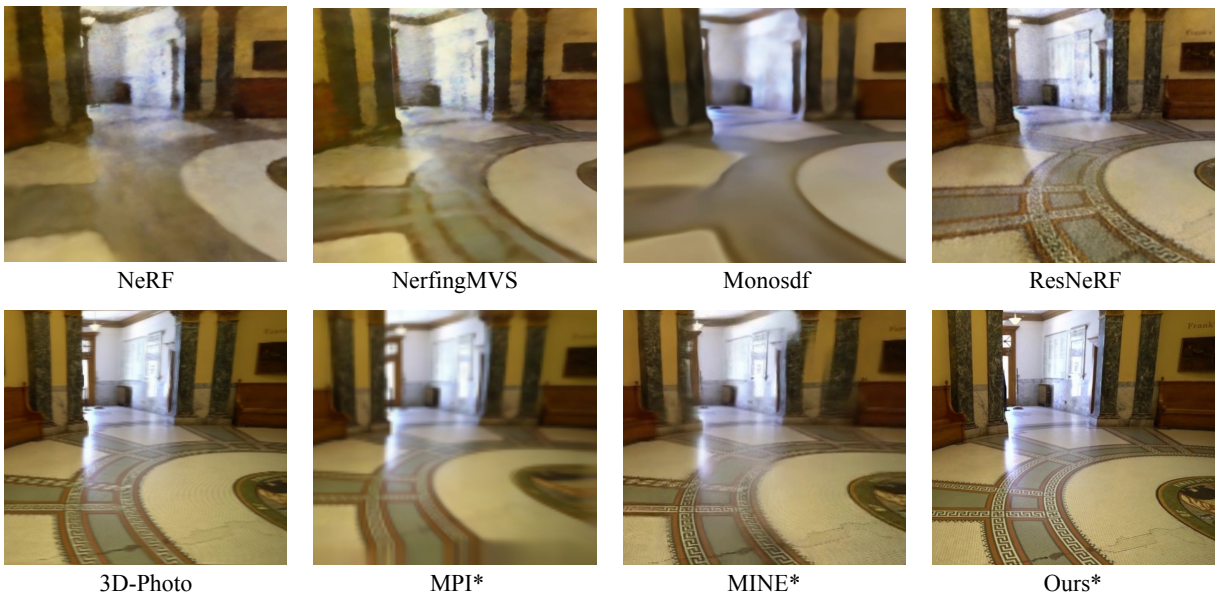

Figure 6: **The qualitative results of our method generalize to unseen dataset (T&T).** The symbol $*$ denotes the model is trained on KITTI and evaluated on T&T.

Type of MPI distribution. We examine the performance of three different strategies for MPI distribution and sampling, *i.e.*, Uniform-Fix, Log-Fix, and proposed Adaptive-bins MPI. In our experiment, we replace the Adaptive-bins MPI module with the uniform-fix or log-fix strategy without changing the other modules. Uniform-Fix MPI is a classical strategy employed by most MPI-based methods, such as MPI [9] and MINE [12]. It divides the depth range at fixed intervals and randomly samples on MPI. Log-Fix MPI introduces a priori for the depth distribution and distributes the MPI according to the depth range in a log scale. As shown in Table 4, compared with the two strategies, our

Table 5: **Ablation study on network design.** Ada-bins stands for the Adaptive-bins MPI module. HRB is the abbreviation for Hierarchical Refinement Branch.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Ada-bins	21.98	0.837	0.118
w/o HRB	22.87	0.869	0.109
w/o L_{ada}	22.25	0.858	0.113
Ours	**23.67**	**0.883**	**0.101**

Ground Truth Ours Ground Truth Ours

Figure 7: **Failure cases.** Due to the extremely narrow geometries (*e.g.*, street light pole) and inhomogeneous diffuse reflections, our method fails in modeling these areas and generates images distorted and misaligned.

Adaptive-bins MPI achieves optimal results by employing the adaptive bin distribution strategy per image, leading to a more efficient representation for unbounded outdoor scenes.

Effectiveness of Each Module. We further investigate how each proposed module contributes to the final performance. We first verify the effectiveness of Adaptive-bins MPI by removing this module and exploiting uniform-fixed MPI followed by random sampling. We report the results in Table 5. The result shows that the Adaptive-bins MPI module plays a key role in modeling the entire outdoor scenes and generating more efficient representations. Then, we remove the hierarchical refinement branch, proving its usefulness in improving image quality and capturing texture details. Moreover, experimental results indicate that introducing the adaptive-bins loss function helps with better distributing planes according to each image. Qualitatively, our method achieves favorable results with the overall combination of each module.

4.4. Limitations & Failure Cases

Our method is based on MPI representation and, as a result, inherits certain limitations. When the synthesis view is significantly distant from the observation view, the generated images have relatively obvious visual distortions and artifacts. As with other MPI-based methods, the areas with strong diffuse light and slender geometric shapes may lead to distorted representation in planes as well as rendered output, as shown in Figure 7. Learning how to synthesize images in these hard cases could be a promising research topic.

5. Conclusion

In this paper, we present SAMPLING, an MPI-based novel view synthesis method from a single-view image for outdoor scenes. To address the difficulty of representing intricate geometries in unbounded outdoor scenes, we introduce Adaptive-bins MPI, which can adaptively distribute the planes of MPI in different depths for each scene image. Besides, we propose a Hierarchical Refinement Branch to fuse multi-scale information for better image detail generation. Experiment results show that our method enhances the efficiency and quality of MPI representation, especially in modeling complex geometries and high-frequency details. Our method achieves new state-of-the-art view synthesis results on the large-scale outdoor dataset. Furthermore, experimental results show that our method has a strong generalization ability on unseen scenes.

References

[1] 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. *Communications of the ACM*, 2021. 1, 2, 7

[2] 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 *CVPR*, 2021. 2

[3] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022. 2

[4] 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 *CVPR*, 2021. 2, 5, 6

[5] Zhuopeng Li, Lu Li, Zeyu Ma, Ping Zhang, Junbo Chen, and Jianke Zhu. Read: Large-scale neural scene rendering for autonomous driving. *arXiv preprint arXiv:2205.05509*, 2022. 2, 3, 5, 6

[6] Ruslan Rakhimov, Andrei-Timotei Ardelean, Victor Lempitsky, and Evgeny Burnaev. Npbg++: Accelerating neural point-based graphics. In *CVPR*, 2022. 2

[7] Alexandra Carlson, Manikandasriram Srinivasan Ramanagopal, Nathan Tseng, Matthew Johnson-Roberson, Ram Vasudevan, and Katherine A Skinner. Cloner: Camera-lidar fusion for occupancy grid-aided neural representations. *arXiv preprint arXiv:2209.01194*, 2022. 2, 3

[8] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023. 2

[9] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 2, 3, 4, 5, 6, 7, 8

[10] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time

view synthesis with neural basis expansion. In *CVPR*, 2021. 2, 3

[11] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2

[12] 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 *ICCV*, 2021. 2, 3, 4, 5, 6, 7, 8

[13] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[14] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 2, 5, 6

[15] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *CVPR*, 2022. 2

[16] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances In Neural Information Processing Systems*, 2022. 2

[17] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *ECCV*, 2022. 2

[18] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *ECCV*, 2018. 2, 5, 6

[19] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[20] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

[21] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *CVPR*, 2022. 2

[22] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *CVPR*, 2021. 2

[23] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[24] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *CVPR*, 2021. 2

[25] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[26] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in Neural Information Processing Systems*, 35:32653–32666, 2022. 2

[27] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2

[28] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *CVPR*, 2022. 2

[29] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, 2022. 2

[30] 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 *ICCV*, 2021. 2, 7

[31] Zheng Chen, Chen Wang, Yuan-Chen Guo, and Song-Hai Zhang. Structnerf: Neural radiance fields for indoor scenes with structural hints. *arXiv preprint arXiv:2209.05277*, 2022. 2

[32] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, 2019. 3, 5, 6

[33] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 3

[34] 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 *CVPR*, 2022. 3

[35] Yuanbo Xiangli, Linning 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. In *ECCV*, 2022. 3

[36] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 3

[37] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (TOG)*, 41(4):1–14, 2022. 3

[38] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020. 3, 5, 6

[39] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 4

[40] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 4

[41] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022. 4

[42] Bingchen Li, Xin Li, Yiting Lu, Sen Liu, Ruoyu Feng, and Zhibo Chen. Hst: Hierarchical swin transformer for compressed image super-resolution. In *ECCV*, 2023. 4

[43] Qing Cai, Yiming Qian, Jinxing Li, Jun Lv, Yee-Hong Yang, Feng Wu, and David Zhang. Hipa: Hierarchical patch transformer for single image super resolution. *arXiv preprint arXiv:2203.10247*, 2022. 4

[44] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 4

[45] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 4

[46] Thomas Porter and Tom Duff. Compositing digital images. In *SIGGRAPH Comput. Graph.*, 1984. 5

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[48] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*, 2021. 5, 6

[49] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020. 5, 6, 7

[50] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. 2013. 5

[51] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 2017. 5

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[53] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 7

[54] Yuting Xiao, Yiqun Zhao, Yanyu Xu, and Shenghua Gao. Resnerf: Geometry-guided residual neural radiance field for indoor scene novel view synthesis. *arXiv preprint arXiv:2211.16211*, 2022. 7

[55] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 2014. 5