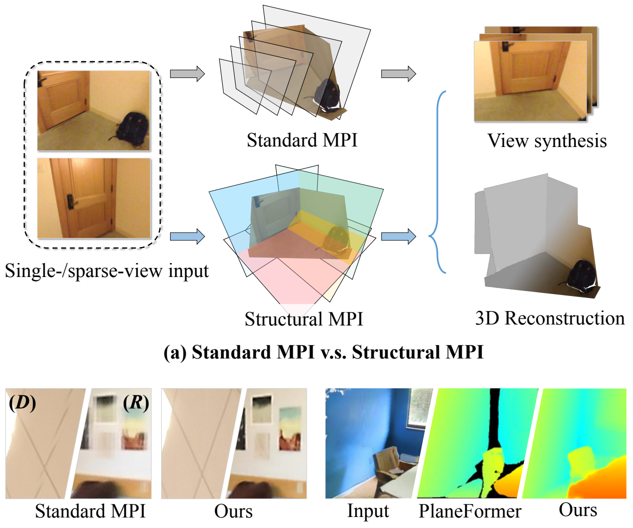
Structural Multiplane Image: Bridging Neural View Synthesis and 3D Reconstruction

Mingfang Zhang[1,2*], Jinglu Wang[2], Xiao Li[2], Yifei Huang[1], Yoichi Sato[1], Yan Lu[2]
[1]The University of Tokyo, [2]Microsoft Research Asia
{mfzhang,hyf,ysato}@iis.u-tokyo.ac.jp, {jinglwa,xili11,yanlu}@microsoft.com

Abstract

The Multiplane Image (MPI), containing a set of fronto-parallel RGBα layers, is an effective and efficient representation for view synthesis from sparse inputs. Yet, its fixed structure limits the performance, especially for surfaces imaged at oblique angles. We introduce the Structural MPI (S-MPI), where the plane structure approximates 3D scenes concisely. Conveying RGBα contexts with geometrically-faithful structures, the S-MPI directly bridges view synthesis and 3D reconstruction. It can not only overcome the critical limitations of MPI, i.e., discretization artifacts from sloped surfaces and abuse of redundant layers, and can also acquire planar 3D reconstruction. Despite the intuition and demand of applying S-MPI, great challenges are introduced, e.g., high-fidelity approximation for both RGBα layers and plane poses, multi-view consistency, non-planar regions modeling, and efficient rendering with intersected planes. Accordingly, we propose a transformer-based network based on a segmentation model [4]. It predicts compact and expressive S-MPI layers with their corresponding masks, poses, and RGBα contexts. Non-planar regions are inclusively handled as a special case in our unified framework. Multi-view consistency is ensured by sharing global proxy embeddings, which encode plane-level features covering the complete 3D scenes with aligned coordinates. Intensive experiments show that our method outperforms both previous state-of-the-art MPI-based view synthesis methods and planar reconstruction methods.



(a) Standard MPI v.s. Structural MPI

(b) View-synthesis result

(c) Reconstruction result

Figure 1. We propose the Structural Multiplane Image (S-MPI) representation to bridge the tasks of neural view synthesis and 3D reconstruction. It consists of a set of posed RGBα images with geometries approximating the 3D scene. The scene-adaptive S-MPI overcomes the critical limitations of standard MPI [20], e.g., *discretization artifacts (D)* and *repeated textures (R)*, and achieves a better depth map compared with the previous planar reconstruction method, PlaneFormer [1].

1. Introduction

Novel view synthesis [30, 51] aims to generate new images from specifically transformed viewpoints given one or multiple images. It finds wide applications in augmented or mixed reality for immersive user experiences.

The advance of neural networks mostly drives the recent progress. NeRF-based methods [28, 30] achieve impressive results but are limited in rendering speed and generalizability. The multiplane image (MPI) representation [42, 50] shows superior abilities in these two aspects, especially given extremely sparse inputs. Specifically, neural networks are utilized to construct MPI layers, containing a set of fronto-parallel RGBα planes regularly sampled in a reference view frustum. Then, novel views are rendered in real-time through simple homography transformation and integral over the MPI layers. Unlike NeRF models, MPI models do not need another training for a new scene.

Nevertheless, standard MPI has underlying limitations. 1) It is sensitive to discretization due to slanted surfaces in scenes. As all layered planes are parallel to the source image plane, slanted surfaces will be distributed to multiple MPI layers causing discretization artifacts in novel views,

[*] This work was done when Mingfang Zhang was an intern at MSRA.

as shown in Fig. 1 (b). Increasing the number of layers can improve the representation capability [38] but also increase memory and computation costs. 2) It easily introduces redundancy. It tends to distribute duplicated textures into different layers to mimic the lighting field [21], which can introduce artifacts with repeated textures as shown in Fig. 1 (b). The essential reason causing the above issues is that the MPI construction is dependent on source views but neglects the explicit 3D geometry of the scenes. Intuitively, we raise a question: Is it possible to construct MPIs adaptive to the scenes, considering both depths and orientations?

Slanted planes are smartly utilized in 3D reconstruction to save stereo matching costs [11] or represent the scene compactly [12, 17], especially for man-made environments. Recent advanced neural networks reach end-to-end planar reconstruction from images by formulating it as instance segmentation [25, 26], where planes are directly detected and segmented from the input image. However, non-planar regions are often not well-modeled or neglected [1, 25, 40], resulting in holes or discontinuities in depth maps, as shown in Fig. 1 (c).

In this paper, we aim to bridge the neural view synthesis and the planar 3D reconstruction, that is, to construct MPIs adaptive to 3D scenes with planar reconstruction and achieve high-fidelity view synthesis. The novel representation we propose is called **Structural MPI** (S-MPI), which is fully flexible in both orientations and depth offsets to approximate the scene geometry, as shown in Fig. 1 (a). Although our motivation is straightforward, there are great challenges in the construction of an S-MPI. (1) The network not only needs to predict RGBα values but also the planar approximation of the scene. (2) It is difficult to correspond the plane projections across views since they may cover different regions and present different appearances. Recent plane reconstruction works [1, 18] build matches of planes after they are detected in each view independently, which may increase costs and accumulate errors. (3) Non-planar regions are challenging to model even with free planes. Previous plane estimation methods [1,40,46] cannot simultaneously handle planar and non-planar regions well. (4) In the rendering process, as an S-MPI contains planes intersecting with each other, an efficient rendering pipeline needs to be designed so that the rendering advantages of MPI can be inherited.

To address these challenges, we propose to build an S-MPI with an end-to-end transformer-based model for both planar geometry approximation and view synthesis, in which planar and non-planar regions are processed jointly. We follow the idea [25] of formulating plane detection as instance segmentation and leverage the segmentation network [4]. Our S-MPI transformer uniformly takes planar and non-planar regions as two structure classes and predicts their representative attributes, which are for reconstruction (structure class, plane pose and plane mask) and view synthesis (RGBα image). We term each instance with such attributes as a *proxy*. Note that non-planar layers are inclusively handled as fronto-parallel planes with adaptive depth offsets and the total number of the predicted proxy instances is adaptive to the scene.

Our model can manipulate both single-view and multi-view input. It aims to generate a set of proxy embeddings in the full extent of the scene, covering all planar and non-planar regions aligned in a global coordinate frame. For multi-view input, the proxy embeddings progressively evolve to cover larger regions and refine plane poses as the number of views increases. In this way, the predicted proxy instances are directly aligned, which avoids the sophisticated matching in two-stage methods [1, 18]. The global proxy embeddings are effectively learned with the ensembled supervision from all local view projections. Our model achieves state-of-the-art performance for single-view view synthesis (10% ↑ PSNR) and planar reconstruction (20% ↑ recall) in datasets [6, 36] of man-made scenes and also achieve encouraging results for multi-view input compared to NeRF-based methods with high costs.

In summary, our main contributions are as follows:

- We introduce the Structural MPI representation, consisting of geometrically-faithful RGBα images to the 3D scene, for both neural view synthesis and 3D reconstruction.
- We propose an end-to-end network to construct S-MPI, where planar and non-planar regions are uniformly handled with high-fidelity approximations for both geometries and light filed.
- Our model ensures multi-view consistency of planes by introducing the global proxy embeddings comprehensively encoding the full 3D scene, and they effectively evolve with the ensembled supervision from all views.

2. Related Works

View synthesis with explicit representations. Various methods are proposed for novel view synthesis based on different representations, such as point cloud [45] and mesh [16]. Layered representations have been the subject of people's interest. Layered Depth Image (LDI) [8, 34, 35, 43] uses several layers of depth maps and associated color values to represent a scene. Multiplane Image (MPI) [50] is a popular variant of LDI, in which layer depths are fixed and an alpha channel is introduced. [42] proposes a model to construct an MPI from single-view images and [10, 29] work with densely sampled multi-view image input. [38] increases the number of layers to enhance MPI's capacity, while [13, 21] claims that adding planes will make the over-parameterized problem worse and the network tends to repeat the content over multiple planes because of depth uncertainty. [13, 27] propose to solve the problem by placing planes at adaptive depths. We propose the Structural MPI that overcomes the MPI's drawbacks like discretization artifacts from sloped surfaces and abuse of redundant layers.

View synthesis with implicit representations. Implicit representations are popular for view synthesis recently, such as NeRF [28, 30, 48], which encodes 3D objects and scenes in the weights of an MLP. A recent method [23] leverages a collection of planar experts in NeRF, but it relies on pre-acquired point clouds. Although NeRF-based methods can achieve promising view synthesis results, they are limited in rendering speed and generalizability. People have tried to develop various solutions [2, 7, 19, 32, 39, 41, 47] and our method has natural advantages, especially with single-view input and no finetuning session. We can achieve comparable results as NeRF with sparse input views while rendering an image significantly faster than NeRF methods.

Planar 3D Approximation. MPI uses a set of fronto-parallel planes to model the scene while planes in a scene have various orientations. Piece-wise planar depth map reconstruction has been a traditional research topic. [12, 37] reconstruct 3D points and perform plane-fitting and recent learning-based methods directly detect and reconstruct 3D planes. [24, 25] generates plane segments and 3D plane parameters for the ScanNet dataset [6] and proposes a detection-based framework. [40, 49] learns an embedding for each pixel and groups them to generate plane instances. [1, 18, 46] solve the problem with multi-view images and conduct plane matching and fusion to generate corresponding instances. Our method generates plane reconstructions with correspondence directly without a second-stage matching and predicts RGBα contents on each plane.

3. Structural Multiplane Image

In this section, we introduce the structural multiplane image representation by first elaborating on the geometry formulation based on standard MPI formulation, and then detailing the process of rendering.

3.1. Geometry Formulation

MPI preliminaries. The standard MPI consists of a collection of N planes parallel to the image plane, $\mathcal{P}_c = \{(C_i, A_i, d_i)\}, i = 1, ..., N$, where $C_i \in \mathbb{R}^{H\times W\times 3}$ and $A_i \in \mathbb{R}^{H\times W\times 1}$ denote RGB and alpha transparency maps of size $H \times W$, and d_i denotes the plane offset to the optical center $\mathbf{o}$.

S-MPI formulation. Differently, our S-MPI contains a set of RGBα images on N_p plane layers with their geometries faithful to the scene, i.e., $\mathcal{P}_s = \{(C_i, A_i, \pi_i)\}_{i=1}^{N_p}$. The number of planes N_p is adaptive. The plane geometry $\pi = (\mathbf{n}, d)$ is represented by a normal vector $\mathbf{n}$ and an offset scalar d. Note that each 3D point $\mathbf{x}$ on the plane satisfies:

$$\mathbf{n} \cdot \mathbf{x} = d. \tag{1}$$

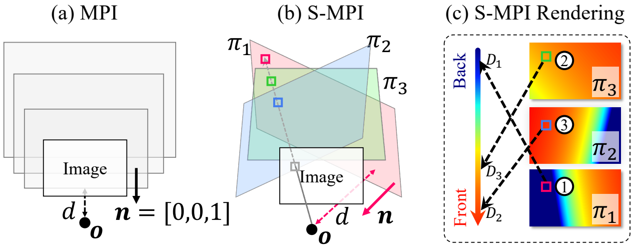


Figure 2. **S-MPI formulation and rendering.** Different from standard MPI with fronto-parallel image planes (a), S-MPI contains a set of adaptively posed planes (b). Each plane is represented by its normal $\mathbf{n}$ and offset d to the optical center $\mathbf{o}$. (c) The rendering order of the noted pixel is given by [①, ②, ③], which follows the depth descending order, $[D_1, D_3, D_2]$. The depth value D_i is obtained by backprojecting the pixel to plane π_i.

Then, each proxy of the S-MPI is represented as $(C_i, A_i, \mathbf{n}_i, d_i)$.

Non-planar region. There are often some non-planar regions not suitable to fit large planes. We consider them not necessary to be forced to fit into small fragmented planes, which can increase fitting errors and rendering costs. Luckily, the fronto-parallel MPI is a special case of S-MPI where all planes are with the same normal as the image plane, i.e., $\mathbf{n}_z = (0, 0, 1)$. We utilize the inclusive property to simply distribute non-planar regions into nearby fronto-parallel planes as standard MPI. Yet, the offsets d can be adaptive to the depth distribution of the scenes. We then get the S-MPI representation for non-planar regions as $\{(C_i, A_i, \mathbf{n}_z, d_i)\}_{i=1}^{N_n}$.

Therefore, our S-MPI contains hybrid structures, i.e., geometrically faithful planes approximating scenes' planar regions and depth-adaptive fronto-parallel planes for non-planar regions. Despite the differences in structures, the two types of multiplane images can be unified in our S-MPI formulation. The final S-MPI for the complete scene is extended as $\mathcal{P}_s = \{(C_i, A_i, \mathbf{n}_i, d_i)\}_{i=1}^{N_p+N_n}$, where N_n elements are with a fixed normal $\mathbf{n}_z$.

3.2. Rendering Formulation

S-MPI rendering. Unlike the standard MPI having a global back-to-front rendering order of planes for each pixel due to the fronto-parallel planes, our S-MPI has different rendering orders for pixels because planes could intersect with each other, as shown in Fig. 2. Each pixel has its own rendering order. First, we calculate depth values D for each pixel on each plane. Let $\mathbf{K}$ denote the camera intrinsic. We backproject the 2D pixel $\mathbf{q} = (u, v, 1)$ to the 3D plane $\pi = (\mathbf{n}, d)$, and then get the equation $\mathbf{n} \cdot (D\mathbf{K}^{-1}\mathbf{q}) = d$ from Eq. (1). Thus, the depth value takes the form:

$$D = \frac{d}{\boldsymbol{n} \cdot \boldsymbol{K}^{-1}\boldsymbol{q}}. \tag{2}$$

In the second step, we rearrange the RGBα images with the depth order for each pixel. The S-MPIs for each pixel

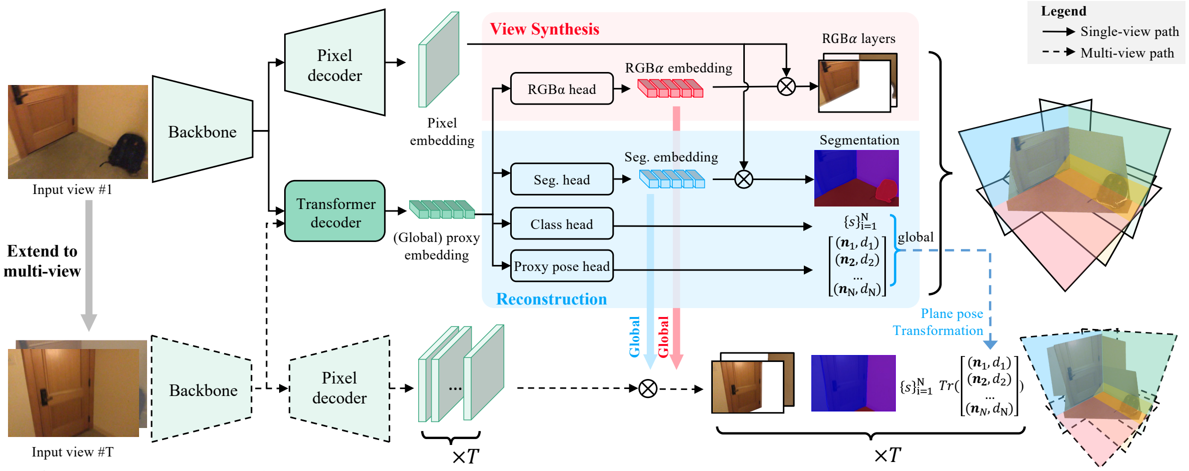


Figure 3. **Structural MPI Transformer.** We use a backbone to extract features for T ($= 1$ or > 1) images of different views and a pixel decoder extracts per-pixel embeddings. Then, a transformer decoder attends to multi-scale image features and produces N (global) proxy embeddings. The proxy embeddings generate global class predictions and global plane poses which are then transformed to different view coordinate frames, ensuring plane pose alignment. Also, RGBα embeddings and segmentation embeddings are generated globally. They are incorporated with pixel embeddings to generate $N \times T$ segmentation masks and $N \times T$ RGBα layers with dynamic convolution layers.

$\mathbf{q}$ are sorted in the depth descending (back-to-front) order, $\boldsymbol{\sigma}(\mathbf{q}) = [\sigma_1^{\mathbf{q}}, \sigma_2^{\mathbf{q}}, ..., \sigma_N^{\mathbf{q}}]$, where $N = N_p + N_n$ is the total number of planar and non-planar proxies and $\sigma_i^{\mathbf{q}}$ is the order index used in rendering for pixel $\mathbf{q}$ on the i-th proxy. Then, we obtain the rearranged RGBα images $[(C_1', A_1'), (C_2', A_2'), ..., (C_N', A_N')]$, where the RGB$\alpha$ values of each pixel are given by $C_i'(\mathbf{q}) = C_{\sigma_i^{\mathbf{q}}}(\mathbf{q})$ and $A_i'(\mathbf{q}) = A_{\sigma_i^{\mathbf{q}}}(\mathbf{q})$. Finally, we apply the standard alpha composition [50] with the new order to render the image I:

$$I = \sum_{i=1}^{N} (C_i' A_i' \prod_{j=i+1}^{N} (1 - A_j')). \tag{3}$$

Similarly, a smooth depth map can be rendered by leveraging alpha to blend depth maps softly. First, we use Eq. (2) to get a depth map $\mathcal{D}_i$ for each plane. Then, we obtain the rearranged depth maps $[\mathcal{D}_1', \mathcal{D}_2', ..., \mathcal{D}_N']$ with the new order $\boldsymbol{\sigma}(\mathbf{q})$. Finally, the rendered depth map $\mathcal{D}_{\mathcal{I}}$ can be produced by $\mathcal{D}_{\mathcal{I}} = \sum_{i=1}^{N} (\mathcal{D}_i' A_i' \prod_{j=i+1}^{N} (1 - A_i'))$.

Rendering novel views. To render a novel view image with S-MPI, we first transform the plane parameters from the source view. Given a transformation matrix $\boldsymbol{H} \in \mathbb{R}^{4\times4}$, a 3D point $\mathbf{x}'$ in the source view can be transferred to $\boldsymbol{H}\mathbf{x}'$. By Eq. (1), We get the plane parameters in the target viewpoint as $\pi_{\mathbf{t}} = (\boldsymbol{H}^{-1})^T \pi_{\mathbf{s}}$. Then, we warp each plane of RGBα image to target views by sampling from the source view with inverse homography [14]:

$$[u_s, v_s, 1]^T = \boldsymbol{K}(\boldsymbol{R} - \frac{\boldsymbol{t}\boldsymbol{n}^T}{d})\boldsymbol{K}^{-1}[u_t, v_t, 1]^T, \tag{4}$$

where $\boldsymbol{R}, \boldsymbol{t}$ are rotation and translation decomposed from $\boldsymbol{H}$. Given transformed plane parameters and RGBα images in the novel view, our rendering process can be applied to get the target RGB image.

4. Structural Multiplane Image Transformer

Given T views of images with known camera poses, our goal is to construct the S-MPI representation $\mathcal{P}_s$ for novel view synthesis and planar reconstruction. Most MPI construction methods [20,50] assume that a pre-defined number of paralleled planes are preset. We propose a transformer-based model to predict S-MPI with an appropriate number of posed planes that faithfully approximate the scene, as well as RGBα layers for further view synthesis. Our method is inspired by previous neural planar reconstruction methods [25,40] that formulate plane detection as instance segmentation. Instead of neglecting non-planar regions or treating non-planar regions as a single layer, we differentiate non-planar instances according to their depth range and unify planar and nonplanar detection in the same pipeline.

Specifically, our model aims to predict proxies $\{(s_i, \mathbf{n}_i, d_i, M_i, C_i, A_i)\}_{i=1}^{N_p+N_n}$, where $s_i \in \{\text{planar}, \text{nonplanar}\}$ denotes the structure class, M_i is the visible projected mask of the plane in the current view. For multi-view input, we predict plane parameters $(\mathbf{n}_i, d_i)$ in a global coordinate frame for alignment.

4.1. Single-view Network

Network design. Our model is built upon a universal image segmentation network [4, 5], which contains a pixel branch for pixel-level decoding and an instance branch representing instance-level embeddings. We fully utilize the two-branch architecture and propose a new one as illustrated in Fig. 3. First, the backbone extracts visual features and the pixel decoder upsamples features and generates high-resolution per-pixel embeddings. The transformer decoder operates on image features to generate N proxy

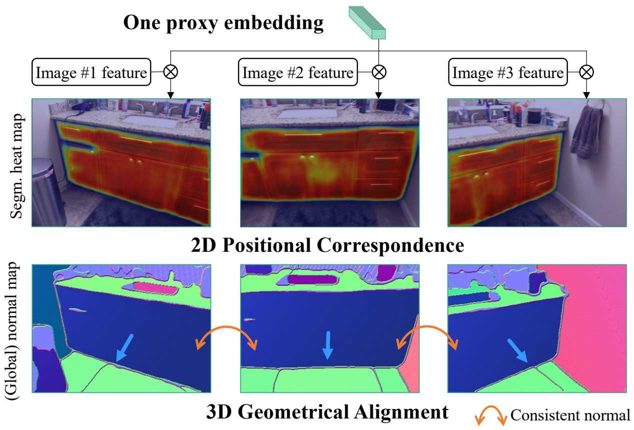


Figure 4. Both 2D positional and 3D geometry information are encoded in our instance embedding. By dot production with image features, the corresponding instances in different views can be visualized. In the normal map, identical plane normals in the global space are painted with the same color.

embeddings at the instance level. Such proxy embeddings are used as queries for various downstream heads. Specifically, per-instance structure classification $\{s_i\}_{i=1}^N$ and plane parameter estimation $\{(\mathbf{n}_i, d_i)_{i=1}^N\}$ are generated from per-instance embeddings directly with linear layer heads. The segmentation and RGBα predictions are generated by incorporating per-instance embeddings and high-resolution pixel embeddings with dynamic convolution [4]. Finally, our S-MPI is created by combining predicted plane parameters and RGBα layers of $N_p + N_n$ high-confidence instances.

Loss function. To train our model, we jointly optimize view synthesis and planar estimation. For view synthesis, we render our predicted S-MPI in a novel view and employ $\mathcal{L}_{rgb}$, a combination of RGB L1 loss and SSIM loss between the rendered image and a ground truth image. For the planar estimation, we employ $\mathcal{L}_{ce}$, a cross-entropy loss to distinguish planar and non-planar regions, $\mathcal{L}_{seg}$, a combination of focal loss [22] and dice loss [31] for segmentation, and $\mathcal{L}_{pln}$, an L1 loss for plane parameters estimation. Finally, we generate a depth map prediction by alpha compositing depth maps of each estimated plane by Eq. (2) and Eq. (3) and employ $\mathcal{L}_{depth}$, a scale-invariant loss [9] between the predicted and a ground truth depth map. Our total loss is:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{view\_syn} + \mathcal{L}_{reconstruction} \\ &= \mathcal{L}_{rgb} + (\alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{seg} + \delta\mathcal{L}_{pln} + \mathcal{L}_{depth}). \end{aligned} \tag{5}$$

4.2. Multi-view Network

Multi-view input can enlarge the range of view synthesis, yet it also introduces new challenges. As to processing each view independently, single-view methods generating planar reconstructions are hardly aligned in the 3D space. Our goal is to deliver multi-view consistent planar reconstruction for better view synthesis.

Network design. Inspired by the application of transformer in video instance segmentation [3], as the proxy embeddings represent instance-level information of proxies in the current view in Sec. 4.1, we extend them to the global scene extent. Thus, the proxy embeddings across views are shared, representing all proxies in a global space. As is shown in Fig. 3, given T images, we first generate T pixel embeddings with the shared backbone and pixel decoder. Then, the global proxy embeddings are utilized to generate proxies in the same way as the single-view network with the multi-view alignment.

Multi-view alignment. As illustrated in Fig. 3, the global proxy embedding is shared across T views, as well as its outputs from the four heads. Thus, the predicted structure classes $\{s\}_{i=1}^N$ and poses $[(\mathbf{n}_1, d_1), (\mathbf{n}_2, d_2), ..., (\mathbf{n}_N, d_N)]$ of all planes are the same in T views. We consider these shared plane parameters to be a set of global plane poses in the full extent of the scene. Given camera poses, we then transform the global plane poses to the T views so that corresponding instances in each view are naturally aligned geometrically. Also, RGBα embeddings and segmentation embeddings are shared globally. By using these global proxy-level embeddings to query T pixel embeddings, each global proxy decodes out its corresponding segment mask and RGBα layer in each of the T views. Then, we employ the supervision in Sec. 4.1 for T views. In this way, the global proxy embeddings are learned with the ensembled supervision from all views with proxy-wise consistency, which can progressively cover regions to the full extent of scenes and refine predictions to obey the alignment. As illustrated in Fig. 4, the class activation maps and globally aligned normal maps from an example proxy embedding appear to be consistent.

Image merging. The final image generation is by merging generated images from multiple views. For each input view, we render an image in the target view and save the alpha weights used in the alpha composition process. Then, we follow [29] to fuse the T rendered images, where the alpha weights are used as confidence maps to blend images. Areas with low confidence in one view that are occluded or beyond the canvas will be overlaid by content with high confidence in other views.

5. Experiments

We demonstrate the effectiveness of our S-MPI by showing its state-of-the-art performance on both view synthesis and reconstruction from single-view and multi-view settings. Then we ablate the design of S-MPI confirming that the improvements stem from specific components.

Datasets. We use the image-based dataset, NYUv2 [36], for single-view reconstruction. The video-based dataset ScanNet [6] is adopted for single-view and multi-view reconstruction and view synthesis.

Table 1. Single-view depth estimation results on NYUv2 [36] and ScanNet [6]. The target view depth maps are acquired by novel view synthesis methods with source view images input. All the methods are trained on ScanNet and our method achieves higher accuracy than planar reconstruction methods [24, 40] and MPI-based methods [20, 42].

	NYUv2					ScanNet (src view)			ScanNet (tgt view)		
	MPI [42]	MINE [20]	PlaneTR [40]	PlaneRCNN [24]	Ours	MPI	MINE	Ours	MPI	MINE	Ours
rel↓	0.241	0.215	0.265	0.164	**0.155**	0.149	0.128	**0.082**	0.155	0.147	**0.101**
log↓	0.097	0.094	0.084	0.074	**0.065**	0.056	0.051	**0.030**	0.074	0.069	**0.053**
rmse↓	0.775	0.754	0.686	0.644	**0.530**	0.308	0.288	**0.214**	0.320	0.303	**0.251**
a1↑	0.665	0.683	0.734	0.753	**0.779**	0.858	0.914	**0.964**	0.842	0.887	**0.946**
a2↑	0.874	0.894	0.913	0.931	**0.956**	0.965	0.970	**0.985**	0.943	0.959	**0.979**
a3↑	0.950	0.952	0.961	0.982	**0.990**	0.984	0.989	**0.996**	0.976	0.978	**0.982**

Table 2. Single-view view synthesis results on ScanNet [6]. n is the number of frames between source views and target views. The numbers (32,64) for MPI and MINE are pre-set numbers of MPI layers.

	LPIPS↓			SSIM↑			PSNR↑		
	$n < 15$	$n < 30$	$n < 45$	$n < 15$	$n < 30$	$n < 45$	$n < 15$	$n < 30$	$n < 45$
MPI-32 [42]	0.2632	0.3365	0.4331	0.8081	0.7567	0.7328	22.737	20.405	19.358
MINE-32 [20]	0.2498	0.3064	0.3502	0.8220	0.7859	0.7487	22.514	20.803	19.517
MINE-64 [20]	0.2412	0.2845	0.3454	0.8296	0.7936	0.7561	22.862	21.058	19.705
Ours	**0.1853**	**0.1914**	**0.1984**	**0.8412**	**0.8285**	**0.8218**	**25.022**	**23.604**	**23.375**

5.1. Implementation

Data preparation. For the ScanNet dataset, we use the same ground truth labels of plane instance masks and plane parameters as [25]. For the non-planar regions, we simply sample S-MPIs according to the depth range of the scene uniformly and produce a set of masks of depth-wise segmentation. Then, we exclude planar masks as non-planar mask labels. It is worth noting that sophisticated depth division [13, 27] can be easily applied in our pipeline. We do not discuss the depth division for non-planar regions as it is not our main focus. The images are resized to 256×384 for training and evaluation.

Implementation details. We use ResNet50 [15] as the backbone. The Adam Optimizer is used with an initial learning rate of 0.0001 and a weight decay of 0.05 in our training. In Eq. (5), $\alpha = 2, \beta = 5, \delta = 5$. The model is trained on 4 NVIDIA-V100 GPUs for a total of 100k steps.

Training. Before our main training, we add a bootstrap training phase for 50k steps, where we initialize the alpha prediction with the segmentation by turning off $\mathcal{L}_{rgb}$ and enforcing a segmentation loss on the alpha channel. To train the model with multi-view input, we set $T = 2$ and the trained model can be applied for $T \geq 1$. We cancel $\mathcal{L}_{pln}$ for non-planar regions in the second input view.

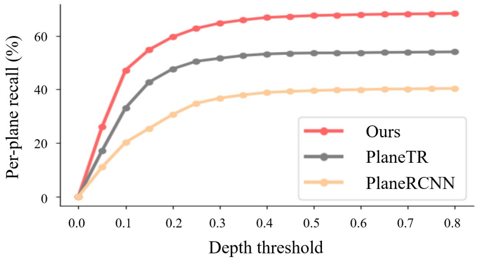


Figure 5. Single-view planar estimation results on ScanNet [6] by measuring the plane recall with a fixed Intersection over Union threshold 0.5 and a varying depth error threshold (from 0 to 0.8m).

5.2. Single-view Evaluation

Reconstruction. For planar reconstruction, we follow PlaneRCNN [24] and PlaneTR [40] to evaluate planar estimation metrics (per-plane recall) in Fig. 5, and segmentation metrics (Variation of Information, Rand Index, and Segmentation Covering) in Tab. 3. For depth estimation, we compare with planar reconstruction methods [24, 40] and MPI-based view synthesis methods [20, 42] in Tab. 1. All the methods are trained on ScanNet [6] and evaluated on NYUv2 [36] and ScanNet.

View synthesis. We compare with the standard MPI-based methods MPI [42] and MINE [20] in Tab. 2. All the methods are trained on ScanNet [6] with losses including $\mathcal{L}_{rgb}$ and $\mathcal{L}_{depth}$ in Eq. (5) given GT images and depth maps of source and target views. Note that we add the $\mathcal{L}_{depth}$ when training standard MPI methods for a fair comparison.

The results show that our method achieves better reconstruction and view synthesis performance. The reason is that our method benefits from proxy-level plane parameter prediction with transformer and alpha blended depth map generation in both planar and non-planar regions. Also, because of our adaptive planar structure and accurate scene reconstruction, our method outperforms standard MPI.

5.3. Multi-view Evaluation

Reconstruction. We compare with PlaneMVS [26] on plane detection and depth reconstruction accuracy in Tab. 7. Plane detection is measured by Average Precision with IoU 0.5 and a varying depth error (from 0.2 to 0.9m). Both

Table 3. Single-view planar segmentation results on ScanNet [6]. “VI”, “RI” and “SC” denote Variation of Information, Rand Index, and Segmentation Covering, respectively.

	VI↓	RI↑	SC↑
PlaneNet [25]	2.142	0.797	0.692
PlaneRCNN [24]	1.809	0.880	0.810
PlaneTR [40]	0.898	0.924	0.811
Ours	**0.791**	**0.947**	**0.851**

Table 4. Ablation study on multi-view consistency (MVC) and query number (the maximum number of RGBα layers).

# query	MVC	LPIPS↓	SSIM↑	PSNR↑
25	✓	0.199	0.803	22.97
50	✓	0.197	0.805	23.06
100		0.208	0.749	22.65
100	✓	**0.197**	**0.813**	**23.11**

Table 5. Ablation study on the difference (# frames gap) of input image pairs. Plane recalls with depth (0.1m,0.6m) and normal (5°,30°) error are evaluated.

# gap	$\text{RC}^{0.1m}$	$\text{RC}^{0.6m}$	RC^{5°	RC^{30°
0	52.61	76.43	**48.02**	74.76
20	**52.95**	**77.16**	47.17	**75.19**
40	46.43	75.47	43.02	73.33

Table 6. Ablation study on the percentage of plane area in the whole image.

planar area	LPIPS↓	SSIM↑	PSNR↑
0%˜40%	0.208	0.795	22.97
40%˜60%	0.196	0.807	23.61
60%˜80%	0.185	0.812	24.03
80%˜100%	**0.179**	**0.841**	**24.52**
0%*	0.212	0.796	21.10

Table 7. Multi-view reconstruction results on ScanNet [6]. We evaluate depth accuracy and planar detection results in comparison with PlaneMVS [26]. Both methods take image pairs as input. (Translation: 0.05∼0.15m)

	Pln-MVS	Ours		Pln-MVS	Ours
rel↓	0.088	**0.079**	$\text{AP}^{0.2m}$↑	0.456	**0.579**
rmse↓	**0.186**	0.205	$\text{AP}^{0.4m}$↑	0.540	**0.649**
a1↑	0.926	**0.946**	$\text{AP}^{0.6m}$↑	0.559	**0.704**
a2↑	0.988	**0.989**	$\text{AP}^{0.9m}$↑	0.562	**0.716**

Table 8. Quantitative view synthesis results with multi-view inputs on ScanNet. "(x)" indicates using x neighboring images of the test image in training for NeRF based methods, and x input views in inference for MINE [20] and ours. "'TPS" is short for training per scene. The bold number indicate the best performance and the blue underlying ones indicate the best performance without TPS.

	TPS	LPIPS↓	SSIM↑	PSNR↑	FPS
NeRF [30] (18)	✓	0.398	0.670	19.03	0.1
DS-NeRF [7] (18)	✓	0.344	0.713	20.85	-
NerfingMVS [44] (18)	✓	0.502	0.626	16.29	-
DP-NeRF (18) [33]	✓	0.294	**0.737**	**20.96**	0.1
DP-NeRF [33] (2)	✓	0.324	0.712	20.49	0.1
MINE [20] (2)		0.359	0.635	16.79	**2.5K**
Ours (2)		**0.267**	0.703	19.93	1.3K

methods are trained with paired images input from ScanNet. We prepare the test data following PlaneMVS consisting of image pairs with a translation from 0.05 to 0.15m. The results show the advantages of our global proxy embedding which incorporates multi-view information contributing to a global geometrical representation.

View Synthesis. We follow the data settings in DP-NeRF [33] where they sample images sparsely in selected scenes in ScanNet and generate dense depth maps for additional supervision. The average view gap between two views is 68 frames which is much more challenging than our paired training data (20 frames). Our training scenes have no overlap with their test set, while NeRF-based methods [7, 30, 33, 44] are trained on the test scenes. For each test image, we select the nearest two views as our input. For a fair comparison, we additionally train DP-NeRF on the two nearest images and the results are noted with "(2)". To test MINE [20], we follow [29] to merge two MPI-generated images. The results in Tab. 8 show that our method outperforms MPI-based methods and achieves comparable results to NeRF-based methods, while no training in the target scene is performed. Qualitative comparisons are shown in Fig. 6. We also compare the rendering speed with NeRF based method in the last column in Tab. 8. Our method inherits the speed advantage of MPI-based methods. The speed of rendering depends on the number of plane layers, while this number is adaptive to different scenes in our method. The average number of planes in ScanNet is about 12.5. Our method is slower than MINE since we need additional plane intersection checking, but it is much faster than NeRF-based methods.

5.4. Ablation Study

We conduct the ablation study on ScanNet [6]. Our multi-view model is trained with input image pairs with a gap of < 20 frames and the view synthesis target image is < 30 frames away from the middle of two input images.

Multi-view consistency. Tab. 4 shows that with our global proxy embedding strategy, synthesized images are better aligned than processing multi-view images one by one.

Difference of input image pairs. Tab. 5 shows that our method benefits from the consistent multi-view geometry supervision and produces more accurate planar reconstruction than given two identical images (gap = 0). However, the performance drops inevitably if the view gap is too large.

Percentage of plane area. We divide our test data according to the planar area ratio. Tab. 6 shows that our method performs better in scenes with more planar coverage. When there are few planar regions, our method degrades to standard MPI with adaptive plane numbers and our performance is compatible with MINE [20]. To further validate it, we force N_p in Sec. 3.1 to be 0 in the groundtruth to regard all regions as non-planar (0%* in Tab. 6) to make comparison.

Query number. Tab. 4 shows a marginal performance

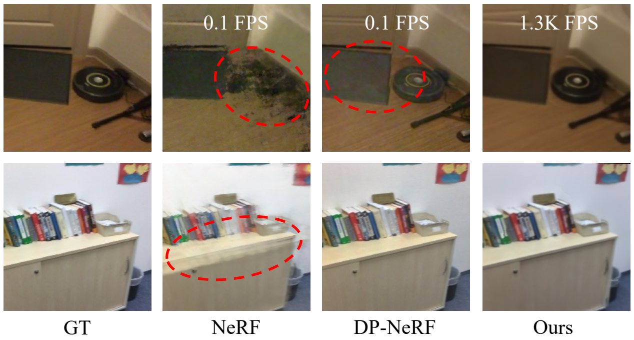
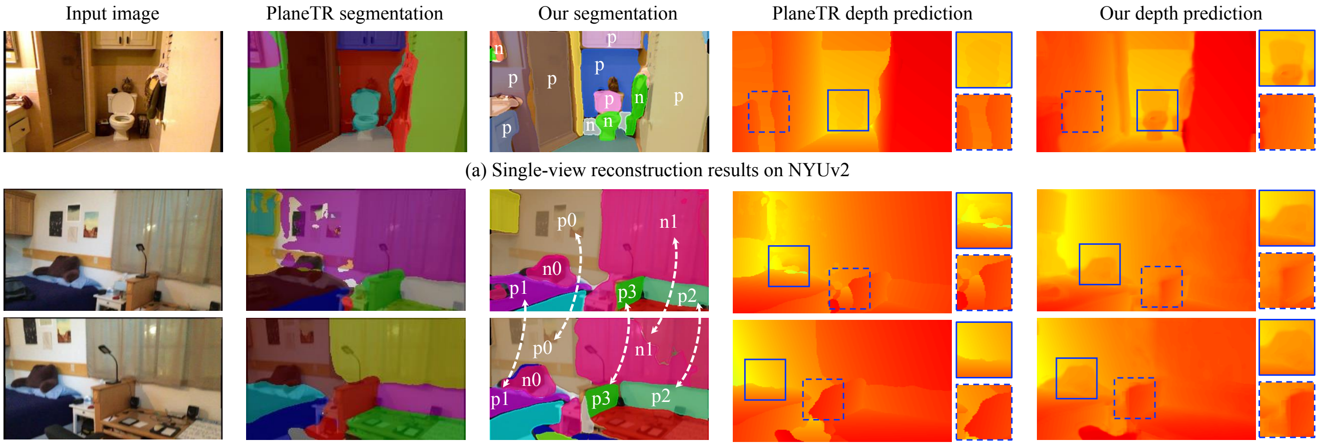


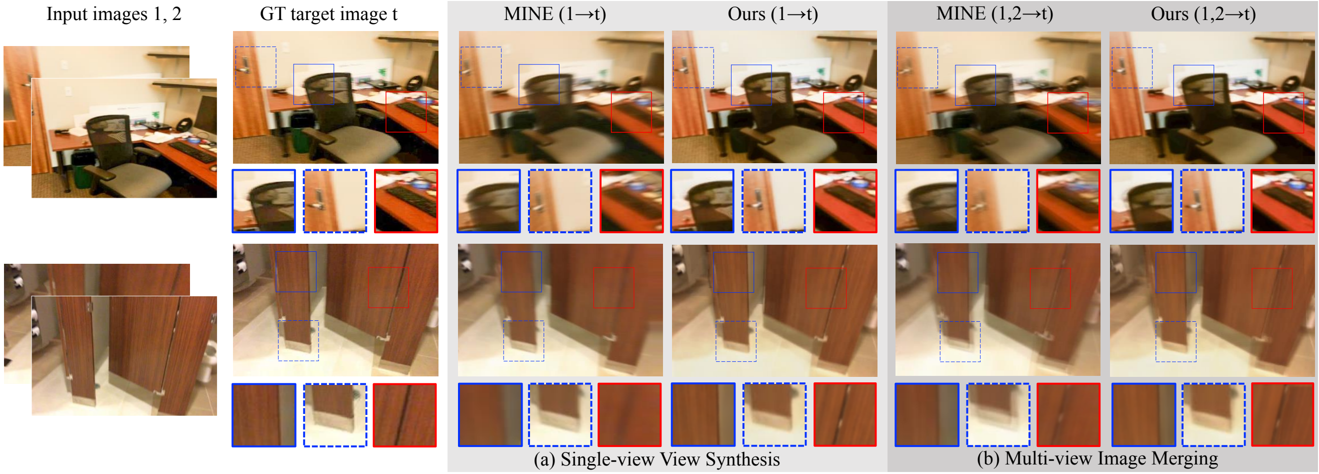
Figure 6. Compared to NeRF-based methods (NeRF [30], DP-NeRF [33]), we achieve better results with less noise and faster rendering speed, while ours do not need per-scene training.



(a) Single-view reconstruction results on NYUv2

(b) Multi-view reconstruction results on ScanNet

Figure 7. **Planar reconstruction results on (single-view) NYUv2 [36] and (multi-view) ScanNet [6].** Compared with PlaneTR [40], our method can uniformly predict planar (p) and non-planar (n) instances and multi-view consistent segmentation and geometry prediction with matched instances.



(a) Single-view View Synthesis

(b) Multi-view Image Merging

Figure 8. **View synthesis results on ScanNet [6].** We show single-view results (1→t) and multi-view results (1,2→t) in comparison with MINE [20]. The multi-view results of MINE are generated by image merging following [29].

improvement as the query number increases in our transformer, as the numbers of planes are usually not large in man-made scenes.

6. Conclusion

In this paper, we introduce the Structural MPI (S-MPI) representation, consisting of geometrically-faithful RGBα images to the scene, for both neural view synthesis and 3D reconstruction. To construct the S-MPI, we propose an end-to-end model in which planar and non-planar regions are uniformly handled. For multi-view input, our model provides a direct global alignment scheme by generating global proxy embeddings at the full extent of the 3D scene for delivering aligned images for view synthesis.

Limitation and future work. Compared to MPI [50], although our Structural MPI achieves better reconstruction and view synthesis results, our method takes more time to construct the S-MPI according to the scene geometry and to render an image because of the intersecting planes. Yet, our rendering process still reaches real-time. For data preparation, we need ground-truth plane segmentation and poses which should be produced from depth map [25]. Finally, in an ideal situation, MPI has the capability to use multiple parallel layers to simulate non-Lambertian effects. However, it is hard for our current method to simulate them as we only use one proxy to simulate a surface. A simple extension can be performed to add more paralleled layers based on our posed planes. This also has advantages over the MPI since we have appropriate orientations to fit the light field. We will leave it for future work.

References

[1] Samir Agarwala, Linyi Jin, Chris Rockwell, and David F. Fouhey. Planeformers: From sparse view planes to 3d reconstruction. In *ECCV*, 2022. 1, 2, 3

[2] 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. 3

[3] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 5

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 1, 2, 4, 5

[5] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 4

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 3, 5, 6, 7, 8

[7] 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. 3, 7

[8] Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari. Peeking behind objects: Layered depth prediction from a single image. *Pattern Recognition Letters*, 125:333–340, 2019. 2

[9] 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*, 27, 2014. 5

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. 2

[11] David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[12] David Gallup, Jan-Michael Frahm, and Marc Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1418–1425. IEEE, 2010. 2, 3

[13] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH*, 2022. 2, 6

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021. 2

[17] Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. Structured indoor modeling. In *Proceedings of the IEEE international conference on computer vision*, pages 1323–1331, 2015. 2

[18] Linyi Jin, Shengyi Qian, Andrew Owens, and David F Fouhey. Planar surface reconstruction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12991–13000, 2021. 2, 3

[19] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3

[20] 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. 1, 4, 6, 7, 8

[21] Qinbo Li and Nima Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Trans. Graph.*, 39(6):229–1, 2020. 2

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[23] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. Neurmips: Neural mixture of planar experts for view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15702–15712, 2022. 3

[24] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019. 3, 6

[25] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018. 2, 3, 4, 6, 8

[26] Jiachen Liu, Pan Ji, Nitin Bansal, Changjiang Cai, Qingan Yan, Xiaolei Huang, and Yi Xu. Planemvs: 3d plane reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8665–8675, 2022. 2, 6, 7

[27] Diogo C Luvizon, Gustavo Sutter P Carvalho, Andreza A dos Santos, Jhonatas S Conceicao, Jose L Flores-Campana,

Luis GL Decker, Marcos R Souza, Helio Pedrini, Antonio Joia, and Otavio AB Penatti. Adaptive multiplane image generation from a single internet picture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2556–2565, 2021. 2, 6

[28] 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. 1, 3

[29] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 2, 5, 7, 8

[30] 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, 3, 7

[31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[33] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 7

[34] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, 1998. 2

[35] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 2

[36] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 2, 5, 6, 8

[37] Sudipta Sinha, Drew Steedly, and Rick Szeliski. Piecewise planar stereo for image-based rendering. In *2009 International Conference on Computer Vision*, pages 1881–1888, 2009. 3

[38] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 2

[39] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 3

[40] Bin Tan, Nan Xue, Song Bai, Tianfu Wu, and Gui-Song Xia. Planetr: Structure-guided transformers for 3d plane recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4186–4195, 2021. 2, 3, 4, 6, 8

[41] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 3

[42] 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. 1, 2, 6

[43] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 2

[44] 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. 7

[45] 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

[46] Yiming Xie, Matheus Gadelha, Fengting Yang, Xiaowei Zhou, and Huaizu Jiang. Planarrecon: Real-time 3d plane detection and reconstruction from posed monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6219–6228, 2022. 2, 3

[47] 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. *arXiv preprint arXiv:2204.00928*, 2022. 3

[48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: 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. 3

[49] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1029–1037, 2019. 3

[50] 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. 1, 2, 4, 8

[51] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 1