

GSNeRF: Generalizable Semantic Neural Radiance Fields with Enhanced 3D Scene Understanding

Zi-Ting Chou^{*1}, Sheng-Yu Huang^{*1}, I-Jieh Liu¹, Yu-Chiang Frank Wang^{1,2},

¹ Graduate Institute of Communication Engineering, National Taiwan University ² NVIDIA, Taiwan
 {r11942101, f08942095, r11942087}@ntu.edu.tw frankwang@nvidia.com

Abstract

Utilizing multi-view inputs to synthesize novel-view images, Neural Radiance Fields (NeRF) have emerged as a popular research topic in 3D vision. In this work, we introduce a Generalizable Semantic Neural Radiance Fields (GSNeRF), which uniquely takes image semantics into the synthesis process so that both novel view image and the associated semantic maps can be produced for unseen scenes. Our GSNeRF is composed of two stages: Semantic Geo-Reasoning and Depth-Guided Visual rendering. The former is able to observe multi-view image inputs to extract semantic and geometry features from a scene. Guided by the resulting image geometry information, the latter performs both image and semantic rendering with improved performances. Our experiments not only confirm that GSNeRF performs favorably against prior works on both novel-view image and semantic segmentation synthesis but the effectiveness of our sampling strategy for visual rendering is further verified.

1. Introduction

3D scene understanding plays a pivotal role in many vision-related tasks, including 3D reconstruction and 3D reasoning. The former domain focuses on low-level understanding, including developing efficient models that interpret and reconstruct observations from various data sources, such as RGB, RGBD, and physical sensors. In contrast, the latter emphasizes high-level understanding, like encoding, segmentation, and common sense knowledge learning. For real-world applications like robotic navigation or augmented reality, both these facets - reconstruction and reasoning - are crucial to enhance interactions with the real world [10]. Among the challenges in the reconstruction domain, novel view synthesis has consistently been a challenging objective. Being able to generate a novel view image based on a few existing views shows that a model com-

prehends the scene’s geometry and possesses a foundational understanding akin to reconstruction.

The Neural Radiance Field (NeRF) [1, 4, 8, 19–22, 28, 35] has recently risen as an exciting research area, providing a novel way to tackle the task of novel view synthesis. By encoding the density and emitted radiance at each spatial location, NeRF is able to compress a scene into a learnable model given several images and the corresponding camera poses of the scene. By incorporating a volumetric rendering skill, images of unseen camera views can be generated with convincing quality. However, NeRF primarily focuses on reconstructing the color information of novel views and understanding the associated high-level semantic information (e.g., semantic segmentation or object detection) remains a significant challenge.

Recent works [9, 15, 25, 30, 31, 37] aim to amplify the high-level scene understanding ability of NeRF by integrating NeRF with semantic segmentation. By sharing the information between the semantic object class and their corresponding appearances, these two tasks are capable of benefiting from each other’s insights [17]. Nevertheless, these semantic understanding methods stick closely to the original NeRF paradigm, focusing on exploiting the model to represent a specific scene. As mentioned in [17], such a strategy requires additional annotations of semantic segmentation maps when applied to a new scene, limiting real-world applicability and generalizability.

To tackle this problem, generalizable NeRFs [3, 14, 18, 27, 32, 33, 36] have emerged as a promising solution. They adopt an on-the-fly approach for building a neural radiance field conditioned on extracted features from input images of different scenes rather than encoding the scene representation directly into the model. For instance, PixelNeRF [36] introduces the idea of conditioning a NeRF model with multi-view images across multiple scenes. This innovation boosts NeRF’s ability to generalize to unseen scenes and avoids retraining for each individual scene. Yet, it leaves uncertainties in its application to semantic understanding.

As a pioneer in integrating a generalizable NeRF with semantic segmentation capabilities, Semantic-Ray (S-

^{*}Equal Contribution

Method	generalized NeRF	generalized segmentation	sampling strategy	efficient sampling
Semantic NeRF [37]	-	-	hierarchical	✓
Panoptic Neural Fields [15]	-	-	uniform	-
Panoptic Lifting [25]	-	-	uniform	-
DM-NeRF [31]	-	-	hierarchical	✓
NeSF [30]	-	✓	hierarchical	✓
Semantic Ray [17]	✓	✓	hierarchical	✓
Ours	✓	✓	depth-guided	✓

Table 1. **Comparisons of semantic-aware NeRFs (i.e., NeRFs, which jointly perform novel view synthesis and semantic segmentation).** Note that NeSF only performs generalizable segmentation, yet its view synthesis remains restricted to the scene-specific training scheme. And, Semantic Ray requires ground truth depth as additional inputs for both training and inference.

Ray) [17] made notable advancements by introducing a generalizable semantic field concept. By incorporating a cross-projection attention module, S-Ray aggregates information of multi-view images from all points sampled along a ray and predicts RGB values and semantic labels. Although this approach appears to capture the necessary contextual information for semantic segmentation, aggregating details of all points introduces noisy features that could interfere with segmentation tasks, as segmentation is mainly concerned with the first object class that the ray encounters. Additionally, their hierarchical sampling approach further increases additional computation and model complexity in view synthesis. Thus, a properly designed rendering strategy for a generalizable semantic field would be desirable.

In this paper, we propose Generalizable Semantic Neural Radiance Fields (GSNeRF) to jointly tackle the problems of generalized novel view synthesis and semantic segmentation. Given a set of source images of a scene with corresponding camera views and a novel target camera view, our GSNeRF is able to derive visual features and depth map predictions of each source view, which are used to predict the depth map of the target view. With the depth map of the target view and properly derived visual features of source views, the novel view RGB image and semantic segmentation can be derived accordingly. Our GSNeRF consists of two key learning stages: *Semantic Geo-Reasoning* and *Depth-Guided Visual Rendering*. The former is to derive visual features and aggregate the predicted depth information of each source view to estimate the depth of the novel view, while the latter renders the RGB image and semantic segmentation map of the target view. We further design two distinct sampling strategies to minimize noisy features and enhance rendering efficiency. Through conducting extensive experiments on both real-world and synthetic datasets, we show that our method outperforms current state-of-the-art generalizable NeRF methods in novel view synthesis and semantic segmentation.

The key contributions of our work are as follows:

- We propose GSNeRF for jointly rendering novel view images and producing the associated semantic segmentation

mask on unseen scenes.

- The proposed Semantic Geo-Reasoning stage learns color, geometry, and semantic information of the input scene, introducing generalization ability of our GSNeRF.
- Based on the inferred geometry information, the introduced Depth-Guided Visual Rendering stage customizes two different sampling strategies according to the predicted target view depth map, so that image and semantic map rendering can be performed simultaneously.

2. Related Work

2.1. Neural Radiance Fields

The Neural Radiance Field [20] has emerged as a widely embraced implicit representation by encoding a 3D scene within a neural network. Nonetheless, as mentioned in [7], the original NeRF method demands extensive training time, ranging from hours to days, and relies on dozens of multi-view images as input. As a result, a large number of subsequent works [8, 21, 28, 35] are put forward to confront these issues. To shorten the training process, methods such as Instant NGP [21] and DVGO [28] opt for a balance between speed and memory, using hash encoding and voxel encoding to minimize the training time to minutes. On the other hand, to address the demand for multiple input views, some methods [7, 14, 23] introduce depth as additional supervision, a notable example being DS-NeRF [7]. DS-NeRF applies extra depth supervision to ensure that the NeRF accurately encodes a scene’s geometry, thus enabling the construction of neural radiance fields with fewer images. This method shows how incorporating depth information can enhance the quality of novel view synthesis while reducing the number of required source images.

2.2. Generalizable Novel View Synthesis

Since NeRF directly encodes information of a scene into a neural network, it is only applicable to the training scene of interest. In response to this limitation, some methods [3, 14, 16, 18, 27, 29, 32, 36] propose on-the-fly construction of NeRF, allowing trained NeRFs applied for syn-

thesizing novel views in unseen scenes. The concept is to extract features from each source view image and condition the NeRF renderer on these scene-specific features, typically through [36] 2D CNN or [3, 14] cost volume techniques that aggregate features across different images. While these methods realize generalizable novel view synthesis, extensions of such NeRFs to high-level scene understanding tasks remain an open challenge.

2.3. Multi-tasking NeRF

Recently, several methods [9, 15, 17, 25, 30, 31, 37] seek to enhance NeRF with higher-level understanding abilities. Semantic NeRF [37] stands as the first method to integrate semantic segmentation with NeRF by designing two projection heads for predicting both semantics and color simultaneously. Although it achieves satisfactory performance and robustness on semantic segmentation tasks, it requires re-training for each scene. This implies a need for not only RGB images but also corresponding semantic segmentation maps to represent each new scene, hindering their generalizability. NeSF [30] has made an attempt to improve this aspect by proposing a half-generalizable method that includes a generalizable semantic segmentation module. That is, after training an original NeRF of each scene, they extract a 3D voxel-like grid density from the neural radiance field and train a generalizable 3D UNet to do a semantic segmentation task. However, this learning scheme still requires additional time to retrain NeRFs for novel scenes.

S-Ray [17] advances further by proposing a generalized semantic neural field. Following the backbone of Neuray [18], they propose a cross-reprojection attention module. Given a ray emitted from the target viewpoint, they uniformly sample points along the ray, and hierarchically sample more points on the second forward pass according to the density predicted from the first pass. With the cross-reprojection module, they aggregate the contextual information on all sampled points. Despite achieving good results on both real-world and synthetic datasets, the sampling strategy is time-consuming. Also, aggregating all information along the entire ray can introduce noisy features that may hamper segmentation tasks, as segmentation primarily focuses on the first object class encountered by the ray.

Contrarily, our proposed GSNeRF aims to predict the depth map of the target view and utilize it to sample points along the ray more efficiently. This unique strategy enables us to sample points focusing on the first encountered object and filter out unrelated noisy points for semantic segmentation prediction. In Table 1, we compare the characteristics of recent semantic-aware NeRFs with GSNeRF.

3. Brief Review of Generalizable NeRFs

For the sake of completeness, we provide a brief review of generalizable NeRFs [3, 14, 18, 29, 36]. Given K multi-

view images $I_{1:K} = \{I_1, I_2, \dots, I_K\}$ with corresponding camera poses $\xi_{1:K} = \{\xi_1, \xi_2, \dots, \xi_K\}$ of a scene, and a target camera pose ξ_T , the goal is to synthesize the novel view image I_T . Rays are sampled across the scene from the target camera pose to render the novel view image. Along each ray, r , N_c points (coarse points) are sampled within the given near bound b_l and far bound b_u . Existing methods partition $[b_l, b_u]$ into N_c bins uniformly and draw one sample within each bin as defined by:

$$t_n^r \sim \mathcal{U}[b_l + \frac{n-1}{N_c}(b_u - b_l), b_l + \frac{n}{N_c}(b_u - b_l)], \quad (1)$$

where t_n^r is the n -th sample along r . With the corresponding camera position o^r and viewing directions ϕ^r , the n -th sampled point x_n^r can be derived as $x_n^r = o^r + t_n^r \phi^r$, with $x_n^r \in \mathbb{R}^3$. Several methods [3, 18, 29, 36] also use hierarchical sampling, conducting a second forward pass to sample N_f points (fine points) where the first N_c uniformly sampled points show higher predicted density.

With the above information, a generalized NeRF can be represented by a learnable continuous function A_θ , where θ denotes the learnable parameters. A_θ computes a volumetric density $\sigma_n^r \in \mathbb{R}^+$ and a view-dependent color representation $\mathbf{c}_n^r \in \mathbb{R}^3$ for each point x_n^r . That is,

$$\sigma_n^r, \mathbf{c}_n^r = A_\theta(x_n^r, \phi^r, f_{1:K}(w(x_n^r))), \quad (2)$$

where $f_{1:K}$ are features of $I_{1:K}$ and $w(\cdot)$ is the function that reprojects the 3D position x_n^r onto the image plane with known camera poses.

To this end, the color value $\mathbf{C}(r)$ of the pixel corresponding to the ray r is formulated as:

$$\mathbf{C}(r) = \sum_{n=1}^{N_c} T_n^r \alpha_n^r \mathbf{c}_n^r. \quad (3)$$

$$T_n^r = \exp(-\sum_{i=1}^{n-1} \sigma_i^r \delta_i^r), \quad \alpha_n^r = 1 - \exp(-\sigma_n^r \delta_n^r), \quad (4)$$

where T_n^r is the accumulated transmittance and $\delta_n^r = \|x_{n+1}^r - x_n^r\|_2$ represents the distance between every two adjacent points along r .

As is the case with other NeRF methods, a generalized NeRF also optimizes the model based on the rendering loss between the rendered image and the GT image:

$$\mathcal{L}_{image} = \sum_{r \in R} \|\mathbf{C}(r) - \hat{\mathbf{C}}(r)\|_2^2, \quad (5)$$

where $\hat{\mathbf{C}}(r)$ represents the GT color of the renderer image corresponding to a ray r .

Our method is built on top of generalized NeRF methods but with a more efficient sampling strategy that can deal with both semantic segmentation and image rendering.

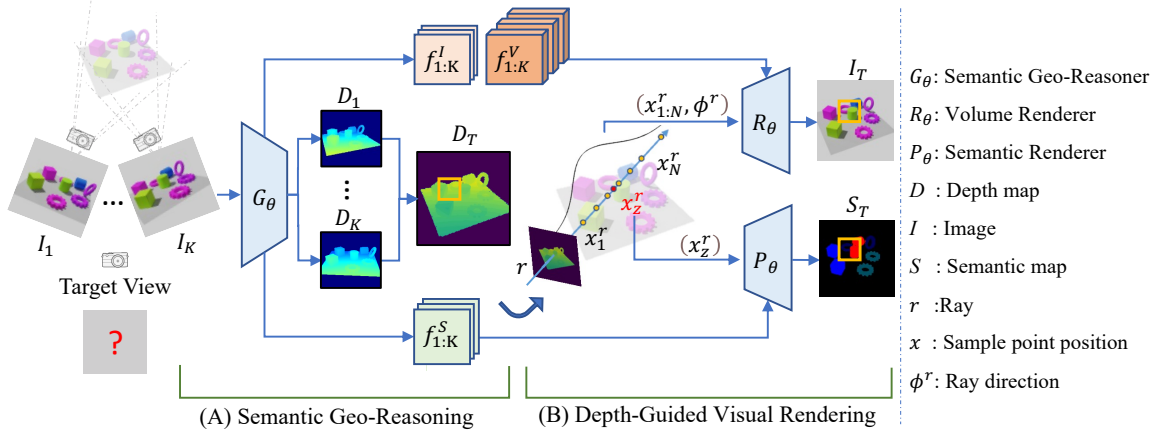


Figure 1. **Overview of GSNeRF**: including *Semantic Geo-Reasoning* and *Depth-Guided Visual Rendering*. Given K multi-view image $I_{1:K}$ of a scene, the Semantic Geo-Reasoner predicts the depth map $D_{1:K}$ for each source image, which is aggregated to estimate the target view depth map D_T . With D_T as key geometric guidance, we design Depth-Guided Visual Rendering to render target view image I_T and semantic segmentation S_T , by Volume Renderer R_θ and Semantic Renderer P_θ , respectively.

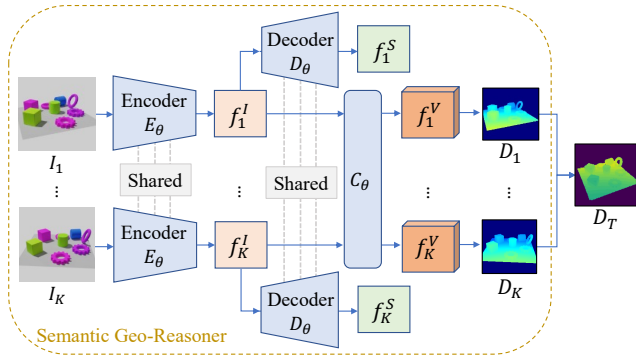


Figure 2. **Semantic Geo-Reasoning**. The Semantic Geo-Reasoner contains a shared Encoder E_θ , a shared Decoder D_θ , and a cost-volume aggregator C_θ , producing image and volume features ($f_{1:K}^I$ and $f_{1:K}^V$), with the associated depth maps $D_{1:K}$. The depth map D_T of the target view is estimated from $D_{1:K}$.

4. Method

4.1. Problem Formulation and Model Overview

We first define the setting and notations used in this paper. Given a scene with K multi-view images $I_{1:K}$ and their corresponding camera poses $\xi_{1:K}$ along with a target camera pose as input, we aim to achieve novel-view synthesis and semantic segmentation simultaneously. During training, we synthesize both the novel view image I_T and the semantic segmentation map S_T from ξ_T , given ground truth (GT) image \hat{I}_T and semantic segmentation map \hat{S}_T of the target view ξ_T . Once the model completes training on a given set of training scenes, it is able to directly generalize to unseen scenes without finetuning.

To achieve this goal, we propose Generalizable Seman-

tic Neural Radiance Fields (GSNeRF), as illustrated in Figure 1. GSNeRF consists of two key learning stages: Semantic Geo-Reasoning and Depth-Guided Visual Rendering. For Semantic Geo-Reasoning, we employ a Semantic Geo-Reasoner G_θ on each input source image $I_k \in \mathbb{R}^{H \times W \times 3}$ to create 2D feature $f_k^I \in \mathbb{R}^{H \times W \times d}$, semantic feature $f_k^S \in \mathbb{R}^{H \times W \times d}$, 3D volume feature $f_k^V \in \mathbb{R}^{H \times W \times L \times d}$, and depth predictions $D_k \in \mathbb{R}^{H \times W}$, where H and W represent the height and width of the image, L is the depth sample number defined in [34] to determine the volume size for the 3D feature, and d is the feature dimension. The source view depth maps are integrated to estimate the depth for the target view D_T . As for Depth-Guided Visual Rendering, we conduct a unique sampling strategy based on D_T to minimize noisy features and enhance rendering efficiency. The sampled points, along with the previously extracted feature, are respectively input into the volume renderer R_θ and semantic renderer P_θ , synthesizing the target view’s image and semantic segmentation map. In the following subsection, we will explain the Semantic Geo-Reasoning and Depth-guided Visual Rendering process within our GSNeRF framework.

4.2. Generalizable Semantic NeRF

4.2.1 Semantic Geo-Reasoning

Given K multi-view source images $I_{1:K}$ of a scene, our Semantic Geo-Reasoner G_θ in Figure 1 is designed to achieve two goals. First, geometric clues are extracted from G_θ , comprising 3D Volume features $f_{1:K}^V$ and depth maps $D_{1:K}$ for each source image. In addition, 2D image features $f_{1:K}^I$ and semantic features $f_{1:K}^S$ for each source image can be jointly produced. Secondly, our G_θ is learned to predict the

target view depth map D_T , which is realized by deterministic aggregation and reprojection of all source view depth predictions $D_{1:K}$. We now detail this learning stage.

As depicted in Figure 2, our Semantic Geo-Reasoner consists of a shared 2D CNN encoder E_θ that extracts $f_{1:K}^I$ from every image $I_{1:K}$. To construct geometry-aware features, a cost volume aggregator C_θ [6, 12, 13, 34] is applied to gather 2D features $f_{1:K}^I$ across different images. As a result, C_θ produces 3D feature $f_{1:K}^V$, which are ultimately used to predict the depth map $D_{1:K}$ for every source view image. To obtain semantic-aware features for semantic prediction, a shared semantic feature decoder D_θ is additionally applied to derive 2D semantic features $f_{1:K}^S$ from $f_{1:K}^I$.

Since $D_{1:K}$ is a vital intermediate element that assists in estimating D_T , which allows the subsequent depth-guided rendering process, we strive to ensure the precise prediction of $D_{1:K}$, whether or not the ground truth depth images are accessible. To achieve this, we integrate ideas from DSNeRF [7] and GeoNeRF [14] to supervise our G_θ with either a depth-supervised loss or a self-supervised depth loss. If the ground-truth depth maps of $D_{1:K}$ are available, our depth-supervised loss \mathcal{L}_D is given as:

$$\mathcal{L}_D = \frac{1}{K} \left(\sum_{k=1}^K \|D_k - \hat{D}_k\|_{s1} \right), \quad (6)$$

Here, \hat{D}_k represents the GT depth map of view k , $\|\cdot\|_{s1}$ represents smooth L1 loss [11]. If \hat{D}_k is unavailable, we use a self-supervised depth loss \mathcal{L}_{ssl} to regularize our depth predictions by considering cross-view depth consistency between all source views. Following the self-supervised depth estimation approach of RC-MVSNet [2], \mathcal{L}_{ssl} is defined as:

$$\mathcal{L}_{ssl} = \lambda_1 \mathcal{L}_{RC} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{Smooth}, \quad (7)$$

where \mathcal{L}_{RC} , \mathcal{L}_{SSIM} , and \mathcal{L}_{Smooth} represent reconstruction, structure similarity, and depth smoothness losses [2]. The hyper-parameters λ_1 , λ_2 , and λ_3 are also selected following [2] (see supplementary material for details).

To this end, the target view depth map D_T is calculated using $D_{1:K}$ through a deterministic algorithm. In simple terms, we directly project the pixels of each depth map into 3D space and then reproject them back to the target view. More details about the self-supervised depth loss and corresponding pseudo-code of target view depth estimation are available in the supplementary material.

4.2.2 Depth-Guided Visual Rendering

We now discuss how we perform volume and semantic rendering with the guidance of the predicted D_T . That is, given a target view ξ_T , a ray is sampled for each pixel on the image plane to render the corresponding RGB value for I_T

and the semantic segmentation output for S_T via the Volume Renderer R_θ and the Semantic Renderer P_θ .

Volume Rendering For *volume rendering* of I_T , existing semantic-NeRF approaches [17, 30, 37] follow conventional NeRFs [20] by applying either uniform sampling or hierarchical sampling along each ray r , as depicted in Table 1. However, both strategies require dense sampling to locate the surface of objects in the given scene, limiting the sampling efficiency. To promote the efficiency of our sampling process, we conduct a depth-guided sampling strategy by focusing on sampling points near the depth values of D_T . More specifically, this is achieved by modifying the sampling strategy in Eq. 1 as follows:

$$t_n^r \sim \mathcal{G}(z, v^2), \quad v = \frac{\min(\|z - b_u\|, \|z - b_l\|)}{3}, \quad (8)$$

where t_n^r is the n -th sample along the ray r , and \mathcal{G} represents Gaussian (normal) distribution. Note that z and v represent the mean and standard deviation of the distribution, respectively, while the value of z is determined as the predicted depth for the pixel p that emits the ray r (i.e., $z = D_T(p)$). By these sampling point values, we derive N sampled points $x_{1:N}^r$ (i.e., $x_n^r = o^r + t_n^r \phi^r$) closer to the surface of the first encountered object by r , avoiding considering noisy points that are far from the object surface. Please see the supplementary material for further details.

We feed all these points alongside the 2D feature $f_{1:K}^I$ and 3D volume feature $f_{1:K}^V$ as additional conditions into the volume renderer R_θ to predict the novel view image of the scene. Specifically, we reproject the point x_n^r onto each source view to get aggregated features from these condition features. Thus, we write the aggregated feature $f_{n,k} \in \mathbb{R}^{2d}$ of the n -th sampled point x_n^r from the k -th view as:

$$f_{n,k} = [f_k^I(w(x_n^r)); f_k^V(x_n^r)], \forall k \in [1 : K] \quad (9)$$

where $w(\cdot)$ is the reprojection function used in Eq. 2 and $[\cdot; \cdot]$ denotes concatenation along feature dimension. We further define the global feature $f_{n,0}$ as guidance across all K source views as:

$$f_{n,0} = [\text{mean}(\{f_{n,k}\}_{k=1}^K); \text{var}(\{f_{n,k}\}_{k=1}^K)]. \quad (10)$$

Following GeoNeRF [14] methodology, the volume renderer R_θ (with learnable parameters θ) predicts per point density and radiance based on the feature $f_n = \{f_{n,k}\}_{k=0}^K \in \mathbb{R}^{(K+1) \times 2d}$ mentioned above:

$$\sigma_n^r, \mathbf{c}_n^r = R_\theta(x_n^r, \phi^r, f_n, M_n), \quad (11)$$

where $M_n = \{M_{n,k}\}_{k=0}^K$ denotes the mask used to mask out unrelated features of a view. If the sampled point x_n^r lies behind the predicted depth map D_k or if the reprojected

Generalized method	GT Depth	ScanNet [5]					Replica [26]				
	Train / Test	mIoU	acc. / class acc.	PSNR↑	SSIM↑	LPIPS↓	mIoU	acc. / class acc.	PSNR↑	SSIM↑	LPIPS↓
Neuray [18] + semhead	✓ / ✓	52.09	67.81 / 61.98	25.01	83.07	31.63	44.37	79.93 / 54.25	26.21	87.37	30.93
GeoNeRF [14] + semhead	✓ /	53.78	76.18 / 61.90	32.55	90.88	12.69	45.12	81.67 / 52.36	28.70	88.94	20.42
S-Ray [17]	✓ / ✓	55.53	77.79 / 60.92	25.19	83.66	30.98	45.30	80.48 / 53.72	26.38	88.13	30.04
GSNeRF (Ours)	✓ /	58.30	79.79 / 65.93	31.33	90.73	12.53	51.52	83.41 / 61.29	31.16	92.44	12.54
MVSNerf [3] + semhead		43.06	66.90 / 53.63	24.14	80.36	34.63	30.21	69.35 / 39.75	23.68	84.37	28.08
GeoNeRF [14] + semhead		45.11	67.12 / 53.44	30.75	88.27	16.48	40.35	74.63 / 49.18	29.92	91.14	17.60
GNT [32] + semhead		43.49	62.06 / 51.84	24.39	82.37	28.36	38.14	71.44 / 47.46	24.56	87.31	20.97
Neuray [18] + semhead		46.09	66.39 / 53.79	25.24	84.39	31.33	40.91	76.23 / 50.15	27.80	89.55	23.68
S-Ray [17]		47.69	64.90 / 54.47	25.13	84.18	30.44	43.27	77.63 / 52.85	26.77	88.54	22.81
GSNeRF (Ours)		52.21	74.71 / 60.14	31.49	90.39	13.87	51.23	83.06 / 61.10	31.71	92.89	12.93

Table 2. **Quantitative results on ScanNet & Replica.** Note that methods in the first four rows take GT depth as inputs or training supervision, while the methods in the last six rows do not observe GT depth during training/testing.

location is outside the source image, $M_{n,k}$ is set to 0. However, the mask for the global feature $M_{n,0}$ is always set to 1.

To this end, we apply the same volume rendering as in Eq. 3, predicting the target view Image I_T . We also apply \mathcal{L}_{image} as in Eq. 5 to supervise the rendered image quality.

Semantic Rendering *Semantic rendering of S_T* , on the other hand, is different from the above volume rendering process. Specifically, image rendering considers the color of both the surface and nearby objects because of possible semi-transparent or transparent objects. Conversely, semantic segmentation focuses only on the first object a ray touches. Therefore, for each ray r , we sample only **one** surface point x_z^r estimated by the target depth map for each ray r to render its semantic value (*i.e.*, $x_z^r = o^r + z\phi^r$, where z is the same depth value as mentioned in Eq. 8). More specifically, with the aggregated semantic feature $f_z = \{f_k^S(w(x_z^r))\}_{k=1}^K \in \mathbb{R}^{K \times d}$, our semantic renderer P_θ with learnable parameters θ predicts the semantic logits $\mathbf{S}(r)$ of ray r as follows:

$$\mathbf{S}(r) = P_\theta(x_z^r, f_z', M), \quad (12)$$

where

$$f_z' = f_z + \{f_0\} \in \mathbb{R}^{(K+1) \times d}. \quad (13)$$

Note that we follow volume rendering in Eq. 10 to define a global semantic feature f_0 as guidance across all views:

$$f_0 = [\text{mean}(f_z); \text{var}(f_z)]. \quad (14)$$

As for the mask $M = \{M_k\}_{k=0}^K$ in Eqn. 12, it is similar to the one used in Eq. 11, where M_k is to mask out the unrelated features of a view if the sample point x_z^r is behind the corresponding depth value of D_k .

Eventually, we utilize a semantic loss to supervise the novel view semantic prediction. Here, we use a multi-class

cross-entropy loss to reduce the semantic prediction error:

$$\mathcal{L}_{sem} = \sum_{r \in R} (\mathbf{S}(r) \log \hat{\mathbf{S}}(r)), \quad (15)$$

where $\hat{\mathbf{S}}(r)$ represent the ground truth segmentation score along the ray r .

4.3. Training and Inference

4.3.1 Training

At the training stage, the total loss is calculated by combining the losses that are mentioned above. The aggregated loss function for our model, when ground-truth depth is available, is given by:

$$\mathcal{L} = \mathcal{L}_{image} + \mathcal{L}_D + \lambda \mathcal{L}_{sem}. \quad (16)$$

If the GT depth is *not* available, our model is trained by:

$$\mathcal{L} = \mathcal{L}_{image} + \mathcal{L}_{ssl} + \lambda \mathcal{L}_{sem}, \quad (17)$$

where the coefficient λ is set to 0.5 for simplicity.

In essence, this optimization process enables our model to estimate the depth map of the target camera view accurately. Utilizing this depth information, the model is capable of efficiently sampling points to render new view images and create corresponding semantic segmentation maps.

4.3.2 Inference

At the inference stage, our GSNeRF is able to generalize to unseen scenes without the need for retraining from scratch. Due to the design of the conditioning renderer on scene-specific features f^I , f^V , and f^S , our model is able to build a semantic neural radiance field on the fly. Once we finish training our model on training scenes, we are able to directly infer novel view images and semantic segmentation maps on any scene given multi-view images and corresponding camera poses of that scene.

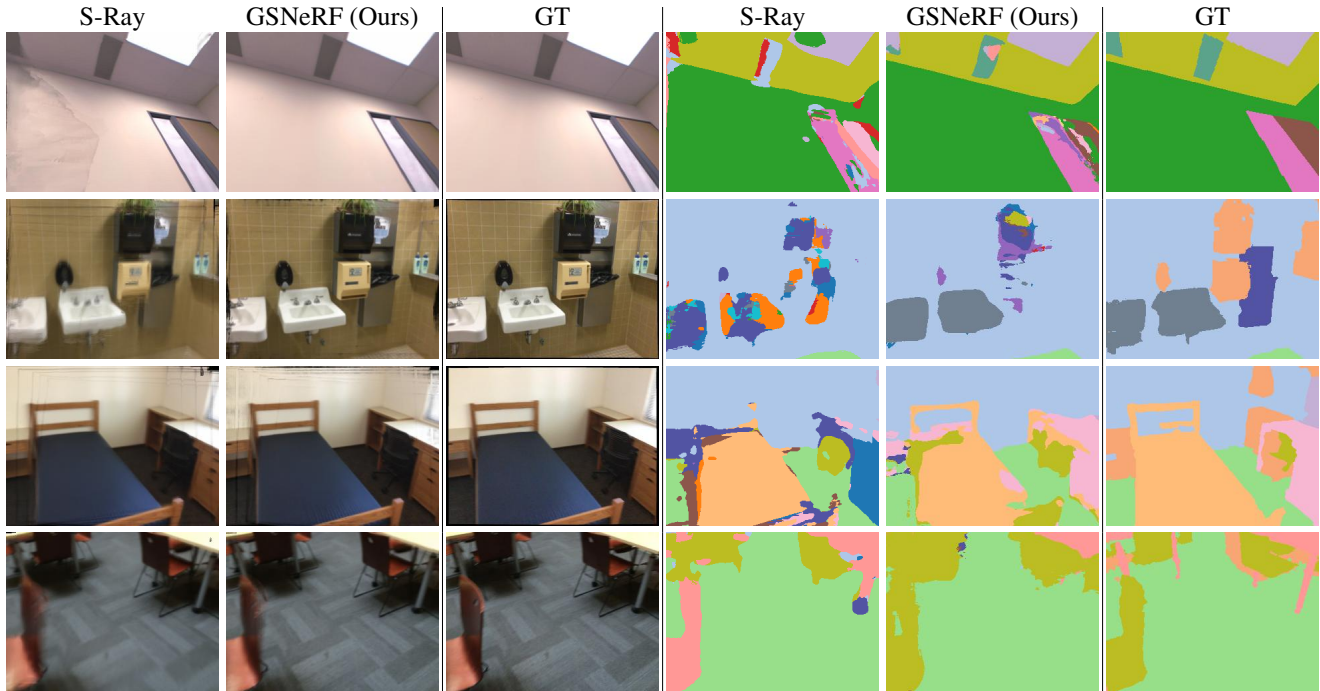


Figure 3. **Qualitative evaluation.** We compare the visual quality of the rendered novel view images (the first three columns) and semantic segmentation maps (the last three columns) with S-Ray [17].

5. Experiments

5.1. Datasets

To evaluate the effectiveness of our proposed method, we conduct experiments on both real-world and synthetic datasets. For real-world data, we use ScanNet [5], a large-scale, indoor RGB-D video dataset with over 2.5 million views from 1513 distinct scenes, including semantic annotations and camera poses. Following the setting presented in S-Ray [17], we train our model on 60 scenes and test its generalization on 10 new, unseen scenes. For synthetic data, we utilize Replica [26], a 3D reconstruction-based indoor dataset comprising 18 high-quality scenes with dense geometry, HDR textures, and semantic labels. Following the Semantic NeRF [37] supplied rendered data, we train our model on 6 distinct scenes across 12 video sequences and test on 2 novel scenes across 4 video sequences. Please see the supplementary material for details.

5.2. Results and analysis

5.2.1 Quantitative Results

Table 2 compares our method against several baselines, in which S-Ray [17] serves as a primary baseline since it is the first to address semantic segmentation with NeRF in a generalized setting. We further conduct experiments on MVS-NeRF [3], GeoNeRF [14], GNT [32] and NeuRay [18] with

the semantic head as described in [17]. It is worth noting that, even though S-Ray does not explicitly mention using a depth map of each source view image as input in their paper, their official implementation takes the depth maps $D_{1:K}$ as additional input along with $I_{1:K}$ during training and testing. Therefore, we follow their official implementation to reproduce S-Ray results.

From Table 2, it can be seen that our model generalizes well to unseen scenes. To show that our GSNeRF is also capable of producing impressive results without depth supervision, we conduct an additional experiment in which we only use the loss function in Eq. 17 to train our GSNeRF in Table 2. We also conduct experiments of S-Ray, NeuRay with only RGB images as input, and GeoNeRF without GT depth as supervision for a fair comparison. From Table 2, we see that even without access to GT depth and with the target depth information regularized in a self-supervised manner, our method still outperforms all the other baseline approaches. This suggests that our model can be trained without requiring depth map supervision, confirming the effectiveness and practicality of our proposed approach. We also note that, although GeoNeRF + semhead (with depth supervision) slightly outperforms our approach regarding PSNR and SSIM for the rendered RGB images on ScanNet, our method excels in all semantic segmentation metrics by approximately 5%. More comprehensive discussions are presented in the supplementary materials.

Model	I_T		S_T		PSNR	mIoU
	Depth-guided	renderer P_θ	Depth-guided			
A	-	-	-	-	28.48	37.88
B	✓	-	-	-	30.56	41.72
C	✓	✓	-	-	30.88	49.47
Ours	✓	✓	✓	✓	31.49	52.21

Table 3. **Ablation studies on GSNeRF.** We verify the effectiveness of using depth-guided rendering for rendering I_T and S_T . Note that, without P_θ , we simply take a pre-trained 2D segmentor on I_T to produce S_T .

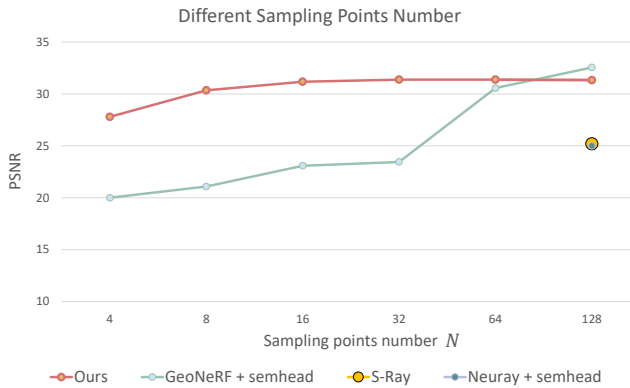


Figure 4. **Sampling efficiency on ScanNet.** Compared to SOTAs, our GSNeRF is able to achieve significantly improved rendering performance especially when the number of sampling points is small. Even with increasing numbers of sampling points, GSNeRF still performs favorably against existing models.

5.2.2 Qualitative Results

The qualitative results of our approach compared to S-Ray [17] are depicted in Fig. 3. This comparison illustrates that while the existing method can approximate the contours of objects, it is unable to capture the geometrical details, causing segmentation inconsistencies. In contrast, our GSNeRF accurately interprets scene geometry through our targeted view depth estimation design, allowing us to not only precisely segment objects but also render novel view images more realistically. This confirms that our depth-guided sampling technique effectively captures the correct geometry and that the sampled point features are appropriately aggregated from features of the Semantic Geo-Reasoner.

5.2.3 Ablation Studies

To further analyze the effectiveness of our designed modules, we conduct ablation studies on ScanNet, as shown in Table 3. Baseline model A employs our Semantic Geo-Reasoner G_θ and volume renderer R_θ to create the target view image I_T , using uniform sampling instead of depth-

guided sampling $x_{1:N}$. Also, a pre-trained 2D segmentor [24] is applied on the rendered image I_T to produce S_T without our semantic renderer P_θ . Model B enhances visual quality by replacing uniform sampling with depth-guided sampling. We further replace the 2D segmentor with our semantic renderer P_θ using uniform sampling for semantic rendering in model C, illustrating mutual benefits for both tasks. Our full model in the last row uses depth-guided sampling exclusively for surface points in the semantic renderer, achieving optimal results. This verifies the success of our proposed modules and sampling strategies.

5.2.4 Sampling Efficiency

Another benefit of our depth-guided sampling is that the performance of our GSNeRF is less sensitive to the number of sampled points along each ray. In Fig. 4, we conduct experiments of reducing the number of sampling points N , showing PSNR of the synthesized novel view images as evaluation. We find that by employing depth-guided sampling, our method accurately preserves visual details with less noise (compared to Neuray and S-Ray with 128 sampled points along each ray). It can be seen that GeoNeRF + semhead shows a catastrophic drop in PSNR with fewer points, while ours leveraging depth-guided sampling maintains satisfactory visual quality even with a significantly reduced number of sampled points (e.g., when sampling only 4 points along a ray, the PSNR remains near 28). This result further attests to the success of our design in estimating target depth and implementing the corresponding sampling strategy with better novel view synthesis performance.

6. Conclusion

In this paper, we proposed Generalizable Semantic Neural Raidance Fields (GSNeRF) to achieve generalized novel view synthesis and semantic segmentation. Our GSNeRF is trained to derive visual features and perform depth map prediction of each source view so that the depth map for a novel target view can be estimated. With such target-view depth information observed, the associated RGB image and semantic segmentation can be jointly produced via depth-guided rendering. In our experiments, we quantitatively and qualitatively confirm that our GSNeRF performs favorably against existing generalizable semantic-aware NeRF methods on both real-world and synthetic datasets.

Acknowledgement This work is supported in part by the National Science and Technology Council via grant NSCT112-2634-F-002-007 and the NTU Center of Data Intelligence: Technologies, Applications, and Systems via grant NTU-113L900902. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. [1](#)
- [2] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rcmvsnet: unsupervised multi-view stereo with neural rendering. In *European Conference on Computer Vision*, pages 665–680. Springer, 2022. [5](#)
- [3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnrnf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. [1](#), [2](#), [3](#), [6](#), [7](#)
- [4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. [1](#)
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. [6](#), [7](#)
- [6] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*, pages 1–8. Ieee, 2019. [5](#)
- [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. [2](#), [5](#)
- [8] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinlong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. [1](#), [2](#)
- [9] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. [1](#), [3](#)
- [10] Sourav Garg, Niko Sünderhauf, Feras Dayoub, Douglas Morrison, Akansel Cosgun, Gustavo Carneiro, Qi Wu, Tat-Jun Chin, Ian Reid, Stephen Gould, et al. Semantics for robotic mapping, perception and interaction: A survey. *Foundations and Trends® in Robotics*, 8(1–2):1–224, 2020. [1](#)
- [11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. [5](#)
- [12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. [5](#)
- [13] Baichuan Huang, Hongwei Yi, Can Huang, Yijia He, Jingbin Liu, and Xiao Liu. M3vsnet: Unsupervised multi-metric multi-view stereo network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3163–3167. IEEE, 2021. [5](#)
- [14] 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. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)
- [15] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. [1](#), [2](#), [3](#)
- [16] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. [2](#)
- [17] Fangfu Liu, Chubin Zhang, Yu Zheng, and Yueqi Duan. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In *CVPR*, pages 17386–17396, 2023. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [18] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. [1](#), [2](#), [3](#), [6](#), [7](#)
- [19] 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](#)
- [20] 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*, 65(1):99–106, 2021. [2](#), [5](#)
- [21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multi-resolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. [2](#)
- [22] Christian Reiser, Songyu Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. [1](#)
- [23] 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. [2](#)

- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 8
- [25] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kotschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 1, 2, 3
- [26] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6, 7
- [27] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pages 156–174. Springer, 2022. 1, 2
- [28] 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. 1, 2
- [29] 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. 2, 3
- [30] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 1, 2, 3, 5
- [31] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022. 1, 2, 3
- [32] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 1, 2, 6, 7
- [33] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 1
- [34] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 4, 5
- [35] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1, 2
- [36] 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. 1, 2, 3
- [37] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 1, 2, 3, 5, 7