Generalizable Novel-View Synthesis using a Stereo Camera

Haechan Lee[1,*] Wonjoon Jin[2,*] Seung-Hwan Baek[1,2] Sunghyun Cho[1,2]

POSTECH [1]GSAI & [2]CSE

{gocks8, jinwj1996, shwbaek, s.cho}@postech.ac.kr

Abstract

In this paper, we propose the first generalizable view synthesis approach that specifically targets multi-view stereo-camera images. Since recent stereo matching has demonstrated accurate geometry prediction, we introduce stereo matching into novel-view synthesis for high-quality geometry reconstruction. To this end, this paper proposes a novel framework, dubbed StereoNeRF, which integrates stereo matching into a NeRF-based generalizable view synthesis approach. StereoNeRF is equipped with three key components to effectively exploit stereo matching in novel-view synthesis: a stereo feature extractor, a depth-guided plane-sweeping, and a stereo depth loss. Moreover, we propose the StereoNVS dataset, the first multi-view dataset of stereo-camera images, encompassing a wide variety of both real and synthetic scenes. Our experimental results demonstrate that StereoNeRF surpasses previous approaches in generalizable view synthesis.

1. Introduction

Novel-view synthesis is a long-standing ill-posed problem in computer vision and graphics, which is inherently challenging due to the necessity of predicting both the geometry and texture from images of a target scene. Recently, Neural Radiance Fields (NeRF) [24] have achieved photorealistic results by jointly optimizing geometry and radiance fields with a coordinate-based network. However, the need for per-scene optimization, which has to learn representations for each target scene individually, restricts its applicability, as it requires additional training time for such optimization.

Recent approaches [2, 4, 16, 21, 30, 32, 33, 38] have addressed this issue of synthesizing novel-view images on-the-fly for unseen scenes without per-scene optimization. Early studies [4, 33, 38] utilize an image encoder to train a generic view interpolation function, enabling the estimation of NeRF parameters from unseen input images in a feed-forward manner. However, this single feed-forward manner for estimating geometry and color exacerbates the ill-posedness, resulting in low-quality geometries and rendering results. For better geometry reasoning, MVSNeRF [2] and GeoNeRF [16] leverage the multi-view stereo (MVS) approach to handle occlusions in a 3D scene. Nonetheless, they struggle with inaccurate geometry prediction and limited synthesis accuracy, as shown in the synthesis result of GeoNeRF (Fig. 1 (a)).

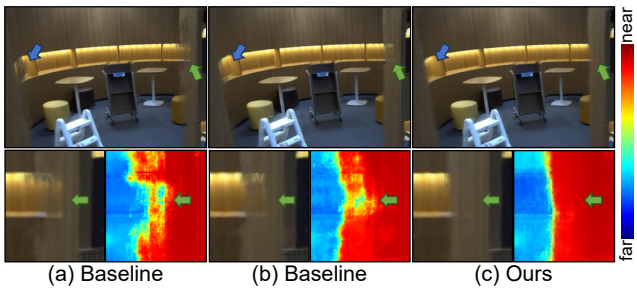

Figure 1. Novel-view synthesis results of a baseline method [16] and ours. The baseline shows degraded performances, even trained using stereo-camera images (b). In contrast, fully exploiting stereo-camera images, our method shows superior results (c).

To tackle this challenge, we propose the first generalizable NeRF approach that leverages *stereo-camera* images, which are easily accessible thanks to the ubiquity of stereo cameras in most mobile devices. Recent advance in learning-based stereo estimation has demonstrated accurate geometry prediction, often even outperforming learning-based MVS methods as shown in Fig. 2. The superior performance of stereo estimation can be attributed to several key factors. First, unlike MVS that assumes arbitrary number of inputs with arbitrary camera positions and orientations, stereo matching assumes two stereo inputs with a fixed baseline. This constraint allows a more optimal network architecture that can effectively find dense correspondences between input images, and significantly eases the training difficulty. Moreover, larger scale of stereo-matching datasets [12, 23, 27] compared to MVS datasets has facilitated remarkable generalization capabilities in stereo estimation network. Therefore, we aim to harness this accurate geometric information from stereo images to alleviate the ill-posedness of generalizable view synthesis.

*Equal contribution.

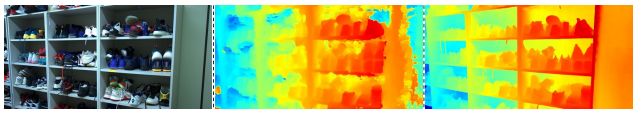

Figure 2. Usefulness of exploiting binocular stereo. Comparison on depth estimation between a learning-based MVS method [26] and a learning-based binocular stereo method [35].

However, since previous methods do not explicitly consider stereoscopic inputs, they cannot leverage the aforementioned benefits, resulting in degenerated performances shown in Fig. 1 (b) despite using stereo-camera images.

This paper proposes StereoNeRF, a novel generalizable view synthesis framework leveraging stereo images. StereoNeRF integrates stereo matching into NeRF-based generalizable view synthesis approach, where the stereo matching provides vital geometric information. To this end, we first introduce a stereo feature extractor, which extracts geometry-aware features by correlating horizontal epipolar lines within stereo images. In addition, the stereo feature extractor takes advantage of stereo-correlated features from an off-the-shelf stereo estimation network, which can transfer rich geometric knowledge to our model. Furthermore, with the reliable depth estimated from the stereo estimation network, we aggregate multi-view features through a depth-guided plane-sweeping technique. This technique ensures correspondence matching around the geometry in cost volume construction. We also present a stereo depth loss utilizing the estimated stereo depth. These additional geometric cues from the stereo matching effectively mitigate the ill-posedness in generalizable view synthesis. Notably, our framework leveraging stereo images surpasses the previous approaches [16, 21], which rely on extra depth information.

Furthermore, we propose the StereoNVS dataset, which is the first dataset for training and evaluation of novel-view synthesis using stereo-camera images. Our StereoNVS dataset provides real-world and synthetic stereo images. Our extensive evaluation on the StereoNVS dataset shows that stereo-pair inputs can effectively enhance the quality of novel-view synthesis, and shows that StereoNeRF outperforms previous generalizable novel-view synthesis approaches in terms of image and shape qualities.

Our contributions are as follows:

- We propose a generalized NeRF approach that leverages stereo-camera images for the first time to alleviate the ill-posedness in generalizable novel-view synthesis.
- We propose a novel framework, StereoNeRF, which exploits the benefits of stereo images by integrating stereo matching into generalizable view synthesis. To this end, we present a stereo feature extractor, a depth-guided plane-sweeping, and a stereo depth loss in our framework.
- We also present the StereoNVS dataset, the first dataset for training and evaluation of novel-view synthesis obtained by stereo cameras.

2. Related Work

2.1. Novel-View Synthesis

Novel-view synthesis aims to synthesize target-view images from reference-view images. Early approaches synthesize novel-view images by blending pixels from multiple input images [6, 13, 20]. Recent work adopting neural volume representations has shown remarkable novel-view synthesis results. Zhou et al. [43] propose multi-plane images (MPI) representation estimated from input images, but their methods produce valid novel-view images only for narrow ranges of camera poses. Mildenhall et al. [24] propose Neural Radiance Fields (NeRF) that can synthesize photo-realistic target-view images via neural implicit representations and volume rendering. Albeit its photo-realism, computation-heavy per-scene optimization is needed. Recent variants such as [9, 17, 25, 37] have remarkably reduced the optimization time, but they still require large memory and several minutes for training.

Generalizable View Synthesis. Recently, many novel-view synthesis approaches without per-scene optimization have been proposed. Several studies directly predict pixel colors by aggregating image features from aligned pixels, without 3D representations. Suhail et al. [30] and Varma et al. [32] adopt a transformer-based network [7] to compute features along the epipolar lines, which needs a large number of training images for high-quality view synthesis. Du et al. [8] propose a framework to synthesize target views from two images with small overlapped regions.

Another research direction predicts volumetric representation [24] from aggregated features from reference views and synthesizes images via volumetric rendering. PixelNeRF, SRF, and GRF [4, 31, 38] predict radiance fields from pixel-aligned features using an MLP. Among these methods, SRF [4] uses two sampled images from an image collection captured by a monocular camera as a stereo pair, but shows limitied synthesis quality. IBRNet [33] proposes to learn generic view interpolation functions, but suffers from artifacts for challenging scenes with complex geometries. MVSNeRF, GeoNeRF and NeuRay [2, 16, 21] utilize the MVS approach using cost volume for better occlusion handling in generalizable view synthesis.

However, all the aforementioned methods often fail to capture accurate geometry, particularly in textureless regions, leading to limited synthesis results. Unlike these methods, our framework exploits stereo-camera images with a fixed baseline to effectively capture the geometry of complex scenes as well as textureless regions for high-quality novel-view synthesis.

2.2. Stereo Matching and Multi-View Stereo

Geometry estimation is a long-standing problem in computer vision that has a variety of applications. Among them,

stereo matching is a task that takes rectified stereo images and computes stereo correspondence to estimate disparities [15, 27]. Recently, a huge number of data-driven approaches have been introduced with the emergence of a vast amount of stereo-matching datasets [12, 23, 27] and made significant progress [1, 23, 35, 39]. MVS approaches that use more than two views have been extensively studied as well [11, 19, 29]. Recently, learning-based MVS approaches have been proposed, e.g., Cheng et al. [3], Gu et al. [14], and Yang et al. [36] present efficient frameworks that cascade cost volumes in a coarse-to-fine manner to enable high-resolution depth estimation.

In our work, we have the best of both worlds in our generalizable view synthesis framework. We adopt the MVS approach to aggregate multi-view information for occlusion-aware geometry estimation, as done in GeoNeRF [16]. Furthermore, our framework also assumes structured stereo images as inputs, and integrates the two-view stereo matching into our framework for robust and accurate geometry estimation in generalizable novel-view synthesis.

3. StereoNeRF

For novel-view synthesis, our framework takes a set of rectified stereo-camera images and estimates neural radiance fields [24] from which novel-view images are rendered. StereoNeRF integrates a well-designed stereo-matching algorithm into the existing generalizable view synthesis approach [16]. In this section, we first provide an overall pipeline that utilizes an pre-trained stereo estimation network [35]. Then, we explain each step of our framework and the training process in detail, highlighting how to integrate stereo matching into our framework.

Overall pipeline. Fig. 3 shows an overview of our pipeline. Our framework utilizes N pairs of stereo images of a target scene $\{(I_L^n, I_R^n)\}_{n=1}^N$ for novel-view synthesis. In the first step, the pre-trained stereo estimation network takes the n-th image pair (I_L^n, I_R^n), and outputs stereo depths $(d_{s,L}^n, d_{s,R}^n)$ and stereo-correlated features (t_L^n, t_R^n). In the second step, a stereo feature extractor takes the stereo image pair and the stereo-correlated features, and outputs stereo image features (f_L^n, f_R^n), explained in Sec. 3.1. The third step aggregates stereo image features from all viewpoints $\{(f_L^n, f_R^n)\}_{n=1}^N$ to build 3D feature volumes $\{(\phi_L^n, \phi_R^n)\}_{n=1}^N$ using an MVS network (Sec. 3.2). For constructing feature volumes that faithfully reflect the 3D geometry of a target scene, the third step adopts a depth-guided plane-sweeping which utilizes the estimated stereo depths $(d_{s,L}^n, d_{s,R}^n)$. Finally, a neural renderer predicts radiance fields from the stereo image features $\{(f_L^n, f_R^n)\}_{n=1}^N$ and the feature volumes $\{(\phi_L^n, \phi_R^n)\}_{n=1}^N$ from all viewpoints, and novel-view images are synthesized from the radiance fields (Sec. 3.3).

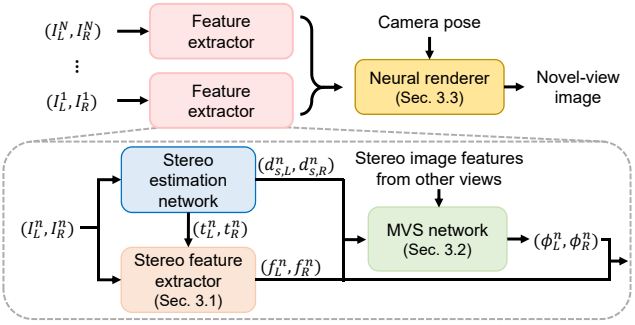

Figure 3. Overview of StereoNeRF. StereoNeRF consists of a shared feature extractor and a neural renderer. We design the feature extractor with a stereo estimation network, a stereo feature extractor, and a MVS network. StereoNeRF takes N pairs of stereo images and a camera pose as inputs, and synthesizes a novel-view image of the camera pose.

3.1. Feature Extraction from Stereo Image Pairs

Unlike previous approaches, which extract features of input images respectively, our stereo feature extractor takes rectified stereo-camera images and computes image feature maps by exploiting the epipolar geometry between them. To this end, our stereo feature extractor consists of three parts: CNN encoders, stereo attention modules (SAM) [5], and CNN decoders (Fig. 4). Moreover, we propose integrating the stereo-correlated features (t_L^n, t_R^n) from the pre-trained stereo estimation network into the SAM (green line in Fig. 4). This integration inherits geometric cues from the stereo estimation network. In the following, we provide the detailed description of the stereo feature extractor.

First, the weight-shared CNN encoders project each of I_L^n and I_R^n into the feature space. Then, we adopt the stereo attention module (SAM) [5] to fuse two feature maps on the horizontal epipolar lines. To this end, SAM estimates stereo correspondences between two feature maps and explicitly adds stereo-correspondent features, as shown in Fig. 5. In addition, we extend the SAM by aggregating the stereo-correlated features (t_L^n, t_R^n) in this feature fusion process. Note that the stereo-correlated features are computed from the pre-trained stereo estimation network, which provides vital geometric cues. These stereo-correlated features enhance stereo correspondences between two feature maps, leading to higher-quality fused feature maps.

Specific explanation of the aforementioned feature fusion of the SAM is as follows (Fig. 5). For the left feature, a warping matrix $\mathcal{W}_{R\to L}$ is built by multiplying a query matrix and a key matrix followed by a softmax layer. The query matrix is computed from the left feature and the stereo-correlated features, and the key matrix is computed from the right feature and the stereo-correlated features. Then, a value matrix, computed from the right feature and the stereo-correlated feature, is warped along the epipolar lines by multiplying the warping matrix $\mathcal{W}_{R\to L}$, then added to

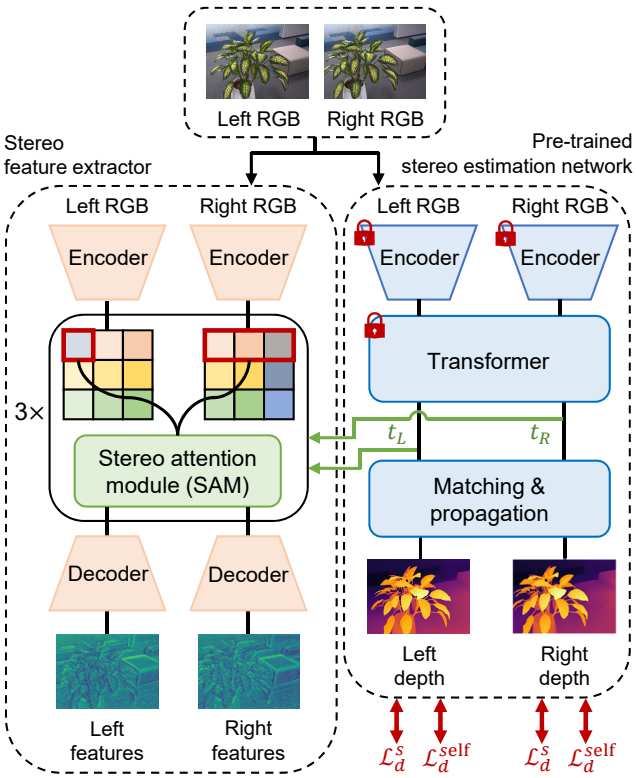

Figure 4. Stereo feature extractor of StereoNeRF, where an pre-trained stereo estimation network is incorporated.

the left feature. The same feature fusion process is also performed for the right feature.

For effective fusion of information from both stereo images and stereo-correlated features, the stereo feature extractor has three sequentially stacked SAM. Finally, the weight-shared CNN decoders take each of the fused features from the SAM, and generate stereo image features (f_L^n, f_R^n), which will be used for building feature volumes.

3.2. Depth-guided Feature Volume Construction

Once stereo image features are obtained, we aggregate these features from all viewpoints to create cost volumes. To this end, we adopt a plane-sweeping-based approach, which computes multi-view correspondences among these features. The plane-sweeping-based approach first defines a depth range within pre-computed near and far depths obtained from COLMAP [28]. Then, it hypothesizes depth planes in the entire depth range. On these depth planes, multi-view stereo image features are aggregated via plane-sweeping to build cost volumes. Then, these cost volumes are processed to yield feature volumes, which will be used later in predicting neural radiance fields.

However, this plane-sweeping-based approach often struggles with accurate geometry estimation. Since this approach searches the entire depth range of a scene, correspondence matching is inaccurate and prone to error, espe-

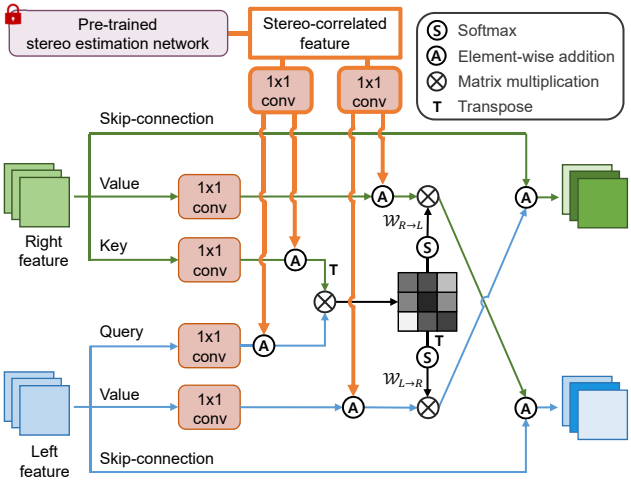

Figure 5. Stereo attention module used in the stereo feature extractor. We exploit the rich features from the pre-trained stereo estimation module.

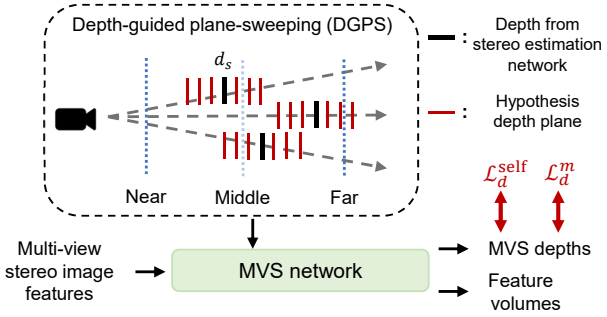

Figure 6. Feature volume construction using DGPS.

cially in textureless regions. To tackle this issue, we introduce a depth-guided plane-sweeping (DGPS), which constructs cost volumes around the stereo depth predicted from the stereo estimation network. Unlike the previous approach using the entire depth range, DGPS introduces a dynamic search range that varies based on the estimated stereo depth. This technique significantly constrains the search space and reduces outliers in correspondence matching.

Fig. 6 depicts the feature volume construction using DGPS. Instead of searching the entire depth range, we hypothesize depth planes around the stereo depths $(d_{s,L}^n, d_{s,R}^n)$. Then, through DGPS, we aggregate the stereo image features from all viewpoints $\{(f_L^n, f_R^n)\}_{n=1}^N$ on these depth planes to build cost volumes. These cost volumes are further processed via the MVS network to obtain feature volumes (ϕ_L^n, ϕ_R^n) and depth maps $(d_{s,L}^n, d_{s,R}^n)$. We repeat this process to obtain feature volumes and depth maps for every viewpoint. The resulting multi-view feature volumes and depth maps are denoted as $\{(\phi_L^L, \phi_R^L)\}_{n=1}^N$ and $\{(d_{m,L}^n, d_{m,R}^n)\}_{n=1}^N$, respectively. Refer to Sec. B.3 in the supplementary document for more details about our feature volume construction.

3.3. Rendering Novel Views

Once feature volumes are constructed, novel-view images can be rendered via neural rendering. For rendering novel-view images, our framework adopts the rendering procedure of GeoNeRF [16]. In the following, we briefly describe the rendering procedure. Given a target viewpoint, we cast a ray for each pixel and sample points along the ray in the 3D space. Then, for each sampled point, we sample features from all the feature volumes $\{(\phi_L^n, \phi_R^n)\}_{n=1}^N$. We also sample features from all the image features $\{(f_L^n, f_R^n)\}_{n=1}^N$ by projecting the sampled point onto input images. The sampled volume and image features are aggregated by a neural renderer network, and the color and density are estimated. Finally, by integrating the estimated color values and depth values of sampled points with their densities along each ray, we obtain the color $\hat{c}$ and the depth d_r at each pixel of a novel view, respectively.

3.4. Training StereoNeRF

We train StereoNeRF with ground-truth images and pseudo-ground-truth depths for high-quality novel-view synthesis. The pseudo-ground-truth depths d_{gt} are obtained from the pre-trained stereo estimation network [35]. Our training loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_d, \tag{1}$$

where $\mathcal{L}_c$ is a color loss and $\mathcal{L}_d$ is a depth loss. The color loss is defined as a mean-squared-error (MSE) between the rendered colors $\hat{c}$ and ground-truth colors [24].

To alleviate the shape-radiance ambiguity [40] in scene reconstruction using neural radiance fields, we introduce the depth loss $\mathcal{L}_d$, which is defined as:

$$\mathcal{L}_d = \lambda_d^{self}\mathcal{L}_d^{self} + \lambda_d^{stereo}\mathcal{L}_d^{stereo}, \tag{2}$$

where $\mathcal{L}_d^{self}$ is a self-supervised depth loss from GeoNeRF [16] and $\mathcal{L}_d^{stereo}$ is our proposed stereo depth loss. $\mathcal{L}_d^{self}$ penalizes the depths estimated from both the stereo estimation network (d_s) and the MVS network (d_m) by comparing them to the depth rendered from NeRF (d_r).

Our framework estimates depths from the stereo estimation network (d_s), the MVS network (d_m), and the neural renderer network (d_r). The stereo depth loss $\mathcal{L}_d^{stereo}$ guides these networks to predict more accurate depths using the pseudo-ground-truth depths d_{gt}. This loss introduces a further quality improvement, especially in geometric details as will be shown in Sec. 5.3. $\mathcal{L}_d^{stereo}$ is defined as:

$$\mathcal{L}_d^{stereo} = \lambda_1\mathcal{L}_d^s + \lambda_2\mathcal{L}_d^m + \lambda_3\mathcal{L}_d^r, \tag{3}$$

where $\mathcal{L}_d^s$, $\mathcal{L}_d^m$, and $\mathcal{L}_d^r$ penalizes depths d_s, d_m, and d_r, respectively, by comparing them with d_{gt}. Note that d_{gt} is different from d_s. We obtain d_{gt} from the pre-trained stereo estimation network with frozen parameters. On the other hand, we obtain d_s from the stereo estimation network in our framework, which has trainable parameters. Refer to Sec. B.4 in the supplementary document for more details about our stereo depth loss.

Due to the different characteristics of datasets such as baseline length, d_s may have estimation error, which leads to an error in the depth-guided plane-sweeping. To tackle this, we partially train the matching and propagation network in the stereo estimation network (Fig. 4) using $\mathcal{L}_d^{self}$. $\mathcal{L}_d^{self}$ provides multi-view supervision to the stereo estimation network by leveraging d_r estimated from multi-view images. However, since d_r may also have estimation error, we further regularize the stereo estimation network using $\mathcal{L}_d^s$. This training scheme for the stereo estimation network ensures consistent depth estimation, while preserving the stereo matching capability. Sec. D.1 in the supplementary document further discusses the training scheme of the stereo estimation network.

4. StereoNVS Dataset

We propose the StereoNVS dataset, the first stereo-camera image dataset for training and evaluating novel-view synthesis using stereo-camera images. The StereoNVS dataset provides both real and synthetic datasets, each of which is dubbed StereoNVS-Real and StereoNVS-Synthetic. In the following, we present the details of each dataset.

4.1. StereoNVS-Real

StereoNVS-Real provides real-world stereo-camera images for the training and evaluation of novel-view synthesis. The dataset provides stereo-camera images of 53 static scenes, and around 25 stereo-image pairs per scene. The images are undistorted and stereo-rectified, and have a resolution of 1792×896. The camera parameters such as the camera poses are provided as well. In our experiments, we divide the dataset into 45 and 8 scenes as training and test sets.

To capture stereo images, we built a camera rig with two Basler machine vision cameras. We measured the camera parameters including the intrinsic and distortion parameters by camera calibration using multi-view images of a checkerboard [42]. Then, we collected stereo images from multiple viewpoints for various indoor and bounded scenes using our camera system. The captured images were then undistorted and stereo-rectified. Finally, we obtained the camera poses of the captured images using COLMAP [28].

4.2. StereoNVS-Synthetic

For quantitative evaluation of synthesized geometries, we also present the StereoNVS-Synthetic dataset, constructed by rendering synthetic 3D models using the 3D-Front dataset [10]. The images have a resolution of 864×448,

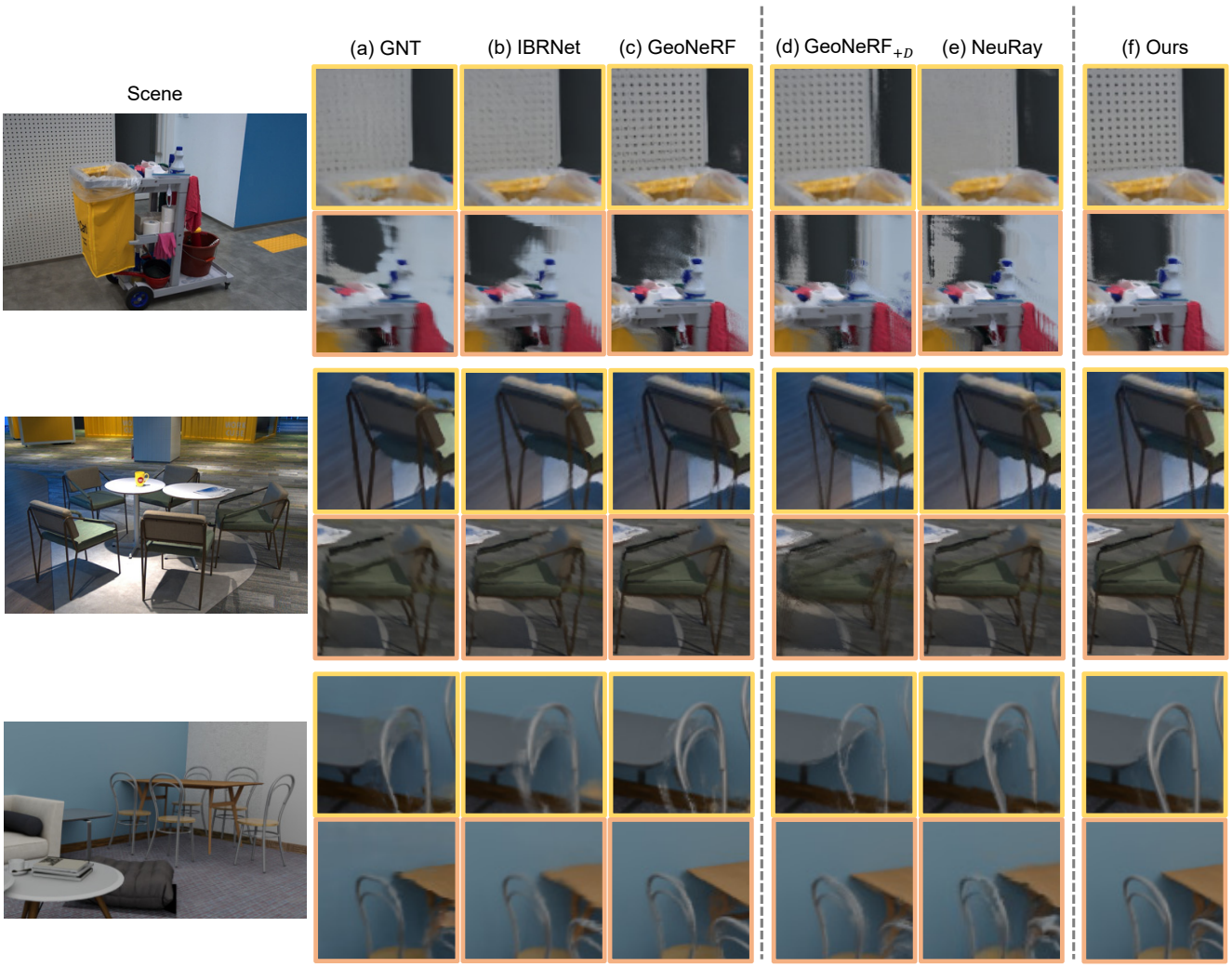

Figure 7. Qualitative comparison of novel-view synthesis on the StereoNVS dataset, showing rendering results for two real-world scenes (above) and one synthetic scene (below). All models are trained using stereo images. Our method outperforms the baseline methods [16, 21, 32, 33] on both real-world and synthetic scenes, especially in thin structures and textureless regions.

which is half of the real dataset. StereoNVS-Synthetic provides stereo-camera images of 50 scenes, and around 150 stereo image pairs per scene as well as the ground-truth camera parameters, camera poses, and depth maps. For more details about the StereoNVS dataset, refer to Sec. E in the supplementary document.

5. Experiments

We conduct extensive validation of our method on the StereoNVS dataset. In the following, we will provide implementation details of our method (Sec. 5.1), compare our method with other baselines (Sec. 5.2), and conduct a comprehensive analysis of our proposed components, which proves the effectiveness of our framework including the stereo feature extractor, the depth-guided plane-sweeping, and the stereo depth loss (Sec. 5.3).

Method	StereoNVS-Real				StereoNVS-Synthetic			
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ABS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ABS(↓)
SRF	21.12	0.6933	0.4164	2.9176	22.36	0.7162	0.4245	0.8125
IBRNet	26.12	0.8421	0.2095	0.7566	30.83	0.8889	0.1823	0.2628
GeoNeRF	28.01	0.8929	0.1460	0.5064	32.13	0.9179	0.1438	0.1577
GNT	26.08	0.8434	0.2285	1.0959	26.17	0.8406	0.2650	0.4512
GeoNeRF$_{+D}$	25.65	0.8172	0.2082	0.8693	32.85	0.9321	0.1171	0.0782
NeuRay	26.51	0.8538	0.1887	0.6147	32.26	0.9104	0.1478	0.1571
Ours	28.44	0.9000	0.1396	0.4868	33.45	0.9336	0.1203	0.1056

Table 1. Quantitative comparison between the baseline methods [4, 16, 21, 32, 33] and ours. All models are trained using stereo images. Our method shows superior performance on both StereoNVS-Real and StereoNVS-Synthetic datasets.

5.1. Implementation Details

We train our model on the training set of StereoNVS-Real, which has multi-view stereo image pairs of real scenes. Our model is trained for 250K iterations. For each iteration, we randomly select one scene and one target viewpoint of the scene. For both training and evaluation, image

features and feature volumes are extracted from three stereo image pairs (i.e., total six images) at three viewpoints nearest to the target viewpoint. During training, 512 rays are randomly selected for the training batch. We use the Adam optimizer [18] with learning rates of 0.0005 and the cosine annealing scheduling [22]. We employ UniMatch [35] for both the pre-trained stereo estimation network within the stereo feature extractor and the pseudo-ground-truth depth of the stereo depth loss. Refer to Sec. B in the supplementary document for additional implementation details.

5.2. Comparison

We compare our method with recent generalizable novel view synthesis methods: SRF [4], IBRNet [33], GeoNeRF [16], GNT [32] and NeuRay [21]. Like our method, we use three stereo-camera image pairs to synthesize each target-view image for all the baseline methods. While the previous methods do not explicitly assume stereo-camera images as their inputs, we also train them using stereo-camera images as they can handle stereo-camera images as independent inputs. All the baseline models are trained on the training set of StereoNVS-Real, as done for our method.

We evaluate both image and depth qualities on the StereoNVS-Real and the StereoNVS-Synthetic datasets. For StereoNVS-Real, we utilize pseudo-ground-truth depths obtained from COLMAP [28] to assess depth quality. On the other hand, for StereoNVS-Synthetic, we use rendered depth maps as ground truths for depth quality assessment. We employ PSNR, SSIM [34], and LPIPS [41] as metrics for image quality and absolute error (ABS) as a metric for depth quality.

Among the compared methods, for training and inference, NeuRay [21] requires depth maps and GeoNeRF [16] can use depth maps as additional inputs. We denote such variation of GeoNeRF as GeoNeRF$_{+D}$. For their training and inference, we used the pseudo-ground-truth depth maps estimated by UniMatch [35] as done for our method.

Qualitative comparison. Fig. 7 presents a qualitative comparison on novel view synthesis between our method and previous methods [16, 21, 32, 33] using the StereoNVS dataset. Previous methods estimate inaccurate geometry for scenes with thin structures or textureless regions, resulting in severe artifacts in the synthesized novel view images. In contrast, our method clearly outperforms the baseline methods in view synthesis results with significantly fewer artifacts even in textureless regions, thanks to more accurately estimated geometry. Moreover, our method shows better synthesis results compared to GeoNeRF$_{+D}$ [16] and NeuRay [21], even though they explicitly use depth maps in the inference time, demonstrating the robust utilization of depth maps in our framework.

	StereoNVS-Real				StereoNVS-Synthetic			
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ABS (↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ABS (↓)
Baseline (a)	28.05	0.8953	0.1372	0.5598	32.22	0.9172	0.1386	0.1679
+ Stereo setting (b)	28.01	0.8929	0.1460	0.5064	32.13	0.9179	0.1438	0.1577
+ SAM (c)	28.21	0.8967	0.1398	0.4935	31.90	0.9167	0.1386	0.1531
+ Correlated feature (d)	28.31	0.8988	0.1370	0.5057	32.35	0.9245	0.1282	0.1416
+ DGPS (e)	28.42	0.8997	0.1403	0.5098	33.09	0.9304	0.1230	0.1246
+ Stereo depth loss (f)	28.44	0.9000	0.1396	0.4868	33.45	0.9336	0.1203	0.1056

Table 2. Quantitative ablation study.

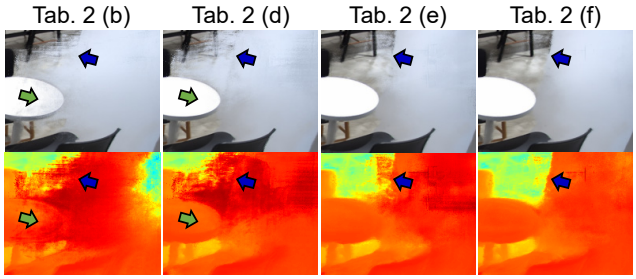

Figure 8. Qualitative ablation study. Our full method enables accurate synthesis of a novel-view image and a depth map.

Quantitative comparison. Tab. 1 shows that our method generally surpasses other baseline methods [4, 16, 21, 32, 33] in terms of image and depth qualities. GNT [32] exhibits degraded performances, likely due to its data-hungry transformer backbone. While performance differences with GeoNeRF are not substantial on the StereoNVS-Real dataset, our method shows the superior perceptual quality (Fig. 7) and significantly better performance on the StereoNVS-Synthetic dataset. While GeoNeRF$_{+D}$ [16] and NeuRay [21] achieve comparable performances on the StereoNVS-Synthetic dataset, they show considerably degenerated results on the StereoNVS-Real dataset. In contrast, our method demonstrates superior performance thanks to the robustness of our framework. Sec. D.1 in the supplementary document further discusses the sensitivity of GeoNeRF$_{+D}$, NeuRay, and our method to depth errors in real-world images.

5.3. Analysis and Discussion

5.3.1 Ablation Study

To assess the impact of our proposed methods, we conduct an ablation study starting with our baseline model [16], which is trained on three views in a monocular setting (i.e., three images). Although equipped with a stereo camera setting, the baseline model trained on three stereo views (i.e., six images) shows similar image qualities as shown in Tab. 2 (b). This result indicates that sophisticated methods are needed to exploit the invaluable information from stereo-camera images.

Tab. 2 shows that our model consistently demonstrates improvements in view synthesis results, as we introduce our proposed components. Our stereo feature extractor enables us to estimate better geometry, leading to fewer artifacts in novel view synthesis, as shown in Fig. 8 (d). The stereo attention module (c) and the stereo-correlated features (d)

help extract robust stereo image features, which are particularly effective for real scenes and synthetic scenes, respectively. Our DGPS is essential for better geometry estimation, leading to significantly improved results as shown in Fig. 8 (e), especially in textureless regions. This is further evident in Tab. 2 (e), with considerable performance gain on the StereoNVS-Synthetic dataset. Our final model shows the best performances (Tab. 2 (f)), with high-quality depth and view synthesis results (Fig. 8 (f)).

5.3.2 Effectiveness of Depth-Guided Plane-Sweeping

We conduct two experiments to demonstrate the effectiveness of our DGPS. In the first experiment, we compare our final model against a model without using DGPS. The "Final model" in Tab. 3 denotes our final model with DGPS. As shown in Tab. 3, the model without using DGPS shows degenerated results compared to our final model, highlighting the effectiveness of DGPS.

In the second experiment, we compare our model using DGPS against a model using more depth planes for cost volume construction. The term "Base model" in Tab. 4 refers to our model that is solely equipped with the stereo feature extractor, excluding DGPS and the stereo depth loss. We train an additional model that uses approximately 1.5 times the number of depth planes, compared to the base model. Although this additional model utilizes more depth planes, it shows similar results to the base model on StereoNeRF-Synthetic, as reported in Tab. 4. This is because increasing the number of planes does not guarantee accurate correspondence matching, especially in textureless regions. In contrast, our DGPS guarantees correspondence matching across multi-view image features near the geometry, resulting in better image and depth qualities. Refer to Sec. C.2.1 in the supplementary document for more details.

5.3.3 Benefit of Stereo Estimation in Depth Loss

To show the effectiveness of using stereo estimation networks for depth supervision, we conduct an additional experiment as follows. First, we obtain two pseudo-GT depths: one from the pre-trained stereo network [35] (d_{gt}) as stated in Sec. 3.4 and the other from the state-of-the-art learning-based MVS network (d_{gt}^{mvs}) [26]. Then, we train two models with our methods using d_{gt} and d_{gt}^{mvs}, respectively. Then, we compare their synthesis results based on image and shape qualities.

As shown in Tab. 5, our model trained with d_{gt} surpasses the other model using d_{gt}^{mvs} on StereoNVS-Synthetic. Note that the MVS network takes more images than the stereo network, seven images and two images, respectively. These results show that the stereo network provides a reliable depth signal for high-quality view synthesis thanks to its generalization ability, which came from standardized inputs of stereo-camera images and large-scale stereo datasets.

	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ABS(↓)
Final model (Tab. 2 (f))	**33.45**	**0.9336**	**0.1203**	**0.1056**
- DGPS	32.30	0.9240	0.1290	0.1300

Table 3. Effectiveness of our depth-guided plane-sweeping (DGPS) compared to the baseline models without using DGPS.

	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ABS(↓)
Base model (Tab. 2 (d))	32.35	0.9245	0.1282	0.1416
+ more depth planes	32.33	0.9247	0.1299	0.1470
+ DGPS (Tab. 2 (e))	**33.09**	**0.9304**	**0.1230**	**0.1246**

Table 4. Efficiency of our depth-guided plane-sweeping (DGPS) compared to the baseline models using more depth planes.

	PSNR (↑)	SSIM (↑)	LPIPS (↓)	ABS (↓)
Ours w/ MVS depth (d_{gt}^{mvs})	32.73	0.9277	0.1253	0.1216
Ours w/ Stereo depth (d_{gt})	**33.45**	**0.9336**	**0.1203**	**0.1056**

Table 5. Effectiveness of using stereo depths as pseudo-ground truth for depth loss compared to using MVS depths.

6. Conclusion

This paper proposes StereoNeRF, a novel generalizable view synthesis framework leveraging stereo-camera images for high-quality novel-view synthesis. Due to the ill-posedness, previous methods struggle with accurate geometry estimation, which leads to severe artifacts in novel-view synthesis. Since the stereo matching provides vital information for accurate geometry reconstruction, our framework incorporates the stereo matching into NeRF-based generalizable view synthesis approach. To this end, we introduce a stereo feature extractor, a depth-guided plane-sweeping, and a stereo depth loss. We also present the StereoNVS dataset, the first stereo-camera image dataset for training and evaluating novel-view synthesis. Our extensive experiments show that StereoNeRF is effective in generalizable novel-view synthesis, particularly in scenes with complex structures or textureless regions.

Limitations and Future Work. Our method is not free from limitations. In sparse view settings, our method produces blurry images and inaccurate geometry, issues also present in other methods. This limitation arises from the insufficient information for novel viewpoints in the sparse view settings. Our future work will involve additional geometric or generative prior, along with stereo prior, to compensate for the lack of information in such settings.

Acknowledgement. We thank Woohyeok Kim and Hyeongmin Lee for their assistance in acquisition of the StereoNVS-Real dataset. This work was supported by the NRF grant (No.2023R1A2C200494611, 2022R1A6A1A03052954, RS-2023-00211658) and IITP grant (No.2019-0-01906, Artificial Intelligence Graduate School Program (POSTECH)) funded by the Korea government (MSIT) and Samsung Electronics Co., Ltd.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 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 *ICCV*, 2021. 1, 2

[3] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *CVPR*, pages 2524–2534, 2020. 3

[4] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *CVPR*. 1, 2, 6, 7

[5] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPR*, pages 1239–1248, 2022. 3

[6] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. *arXiv preprint arXiv:2304.08463*, 2023. 2

[9] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022. 2

[10] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 5

[11] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 3

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 1, 3

[13] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996. 2

[14] 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 *CVPR*, pages 2495–2504, 2020. 3

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[16] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *CVPR*, pages 18365–18375, 2022. 1, 2, 3, 5, 6, 7

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[19] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 3

[20] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 2

[21] 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 *CVPR*, pages 7824–7833, 2022. 1, 2, 6, 7

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[23] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 1, 3

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3, 5

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 2

[26] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *CVPR*, pages 8645–8654, 2022. 2, 8

[27] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002. 1, 3

[28] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 4, 5, 7

[29] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35:151–173, 1999. 3

[30] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *ECCV*, pages 156–174. Springer, 2022. 1, 2

[31] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *ICCV*, pages 15182–15192, 2021. 2

[32] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that nerf needs? In *ICLR*, 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 *CVPR*, pages 4690–4699, 2021. 1, 2, 6, 7

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[35] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 2, 3, 5, 7, 8

[36] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, pages 4877–4886, 2020. 3

[37] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, pages 5752–5761, 2021. 2

[38] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 1, 2

[39] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019. 3

[40] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 5

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[42] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000. 5

[43] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG*, 2018. 2