

Geometry-aware Reconstruction and Fusion-refined Rendering for Generalizable Neural Radiance Fields

Tianqi Liu Xinyi Ye Min Shi Zihao Huang Zhiyu Pan Zhan Peng Zhiguo Cao*
School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology, Wuhan 430074, China
{tq-liu, xinyiye, min-shi, zihao Huang, zhiyupan, peng-zhan, zgcao}@hust.edu.cn

Abstract

Generalizable NeRF aims to synthesize novel views for unseen scenes. Common practices involve constructing variance-based cost volumes for geometry reconstruction and encoding 3D descriptors for decoding novel views. However, existing methods show limited generalization ability in challenging conditions due to inaccurate geometry, sub-optimal descriptors, and decoding strategies. We address these issues point by point. First, we find the variance-based cost volume exhibits failure patterns as the features of pixels corresponding to the same point can be inconsistent across different views due to occlusions or reflections. We introduce an Adaptive Cost Aggregation (ACA) approach to amplify the contribution of consistent pixel pairs and suppress inconsistent ones. Unlike previous methods that solely fuse 2D features into descriptors, our approach introduces a Spatial-View Aggregator (SVA) to incorporate 3D context into descriptors through spatial and inter-view interaction. When decoding the descriptors, we observe the two existing decoding strategies excel in different areas, which are complementary. A Consistency-Aware Fusion (CAF) strategy is proposed to leverage the advantages of both. We incorporate the above ACA, SVA, and CAF into a coarse-to-fine framework, termed *Geometry-aware Reconstruction and Fusion-refined Rendering (GeFu)*. *GeFu* attains state-of-the-art performance across multiple datasets. Code is available at <https://github.com/TQTQliu/GeFu>.

1. Introduction

Novel view synthesis (NVS) aims to generate realistic images at novel viewpoints given a set of posed images. By encoding the density and radiance fields of scenes into implicit representations, Neural Radiance Field (NeRF) [21] has shown impressive performance in NVS. However, NeRF re-

*Corresponding author

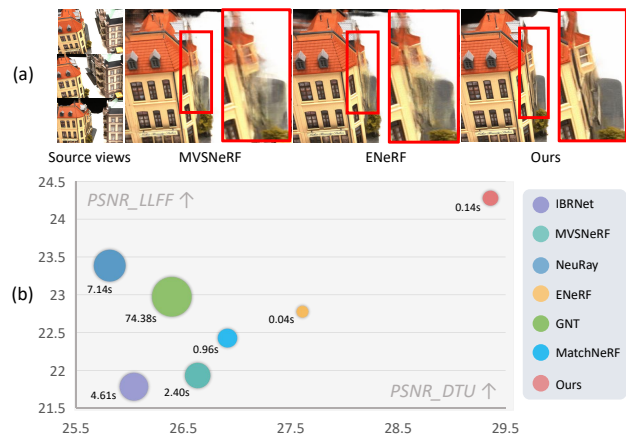


Figure 1. **Comparison with existing methods.** (a) With three input source views, our generalizable model synthesizes novel views with higher quality than existing methods [4, 16] in the severe occluded area. (b) Circle area represents inference time. The X-axis represents the PSNR on the DTU dataset [1] and the Y-axis represents the PSNR on the Real Forward-facing dataset [20]. Our method attains state-of-the-art performance.

quires a lengthy optimization with densely captured images for each scene, which limits its applications.

To address this issue, some recent methods [4, 6, 8, 13, 15, 16, 19, 26–28, 34, 40] generalize NeRFs to unseen scenes. Instead of overfitting the scene, they extract feature descriptors for 3D points in a scene-agnostic manner, which are then decoded for rendering novel views. Pioneer methods [28, 40] utilize 2D features warped from the source images. As this practice avoids explicit modeling of 3D geometric constraints, its generalization capability for geometry reasoning and view synthesis in new scenes is limited. Hence, subsequent methods [4, 16] introduce explicit geometry-aware cost volume from multi-view stereo (MVS) to model geometry at the novel view. Despite significant progress achieved by these methods, synthesis results remain unsatisfactory, especially in challenging areas, such

as the occluded area illustrated in Fig. 1 (a). The model struggles to accurately infer the geometry of these regions, as variance-based cost volumes cannot perceive occlusions.

Generalizable NeRFs’ pipeline consists of two phases: radiance field reconstruction and rendering. The reconstruction phase aims to recover scene geometry and encode 3D-aware features for rendering, *i.e.*, creating descriptors for 3D points. For geometry reasoning, similar to MVS, if the geometry (*i.e.*, depth) is accurate, features across various views are supposed to be similar with low variance. However, this variance-based cost metric is not universal, especially in occluded and reflective regions. Due to inconsistent features in these regions, equally considering the contributions of different views is unreasonable, leading to misleading variance values. Inspired by MVS methods [29, 31], we propose an Adaptive Cost Aggregation (ACA) module. ACA adaptively reweights the contributions of different views based on the similarity between source views and the novel view. Since the novel view is unavailable, we employ a coarse-to-fine framework and transfer the rendering view from the coarse stage into the fine stage to learn adaptive weights for the cost volume construction.

With the geometry derived from the cost volume, we can re-sample 3D points around the surface and encode descriptors for sampled points. For descriptors encoding, previous methods [6, 16, 28, 40] directly aggregate inter-view features into descriptors for subsequent rendering. These descriptors lack 3D context awareness, leading to discontinuities in the descriptor space. To this end, we design the Spatial-View Aggregator (SVA) to learn 3D context-aware descriptors. Specifically, we encode spatially context-aware and smooth features by aggregating 3D spatial information. Meanwhile, to preserve geometric details, we utilize smoothed features as queries to reassemble high-frequency information across views to create final descriptors.

With the scene geometry and point-wise descriptors, the subsequent rendering phase aims to decode descriptors into volume density and radiance for rendering a novel view. For the radiance prediction, [16, 28] predict blending weights to combine color values from source views, while [4, 6, 40] directly regress from features. However, the analysis of these two approaches has not been conducted in existing works. In this paper, we observe that the blending approach performs better in most areas (Fig. 2 (a)), as the color values from source views provide referential factors. However, as shown in Fig. 2 (b)&(c), in challenging areas such as reflections and boundaries, the regression approach produces superior results with fewer artifacts, while the blending approach leads to suboptimal rendering due to unreliable referential factors. To unify the advantages of both strategies, we propose to separately predict two intermediate views using two approaches and design a weighted structure named Consistency-Aware Fusion (CAF) to dynamically fuse them

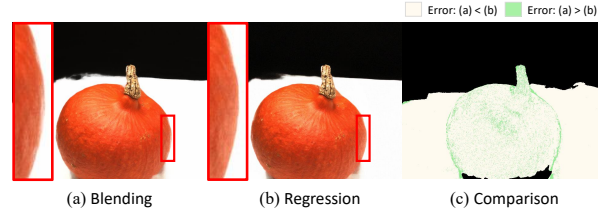


Figure 2. **Comparison of two rendering strategies.** (a) The view obtained using the blending approach that combines color values from source views. (b) The view obtained using the regression approach that directly regresses color values from features. (c) Accuracy comparison between two rendering strategies. The former strategy performs better in the white regions, while worse in the green ones.

into the final view. The fusing weights are learned by checking multi-view consistency, following an underlying principle that if color values are close to the ground truth, the multi-view features corresponding to the correct depth are supposed to be similar.

By embedding the above ACA, SVA, and CAF into a coarse-to-fine framework, we propose GeFu. To demonstrate the effectiveness, we evaluate GeFu on the widely-used DTU [1], Real Forward-facing [20], and NeRF Synthetic [21] datasets. Extensive experiments show that GeFu outperforms other generalizable NeRFs by large margins without scene-specific fine-tuning as shown in Fig. 1 (a)&(b). After per-scene fine-tuning, GeFu also outperforms other generalizable NeRFs and achieves performance comparable to or even better than NeRF [21]. Additionally, GeFu is capable of generating reasonable depth maps, surpassing other generalization methods.

Our main contributions can be summarized as follows:

- We propose ACA to improve geometry estimation and SVA to encode 3D context-aware descriptors for geometry-aware reconstruction.
- We conduct an analysis of two existing color decoding strategies and propose CAF to unify their advantages for fusion-refined rendering.
- GeFu achieves state-of-the-art performance across multiple datasets, showing superior generalization.

2. Related Work

Multi-View Stereo. Given multiple calibrated images, multi-view stereo (MVS) aims to reconstruct a dense 3D representation. Traditional MVS methods [10, 11, 24, 25] primarily rely on hand-crafted features and similarity metrics, which limits their performance, especially in challenging regions such as weak-texture and repetitive areas. Powered by the impressive representation of neural networks, MVSNet [37] first proposes an end-to-end cost volume-based pipeline, which quickly becomes the

mainstream in the MVS community. Following works explore the potential capacity of this pipeline from various aspects. *e.g.*, reducing memory consumption with recurrent approaches [31, 35, 38] or coarse-to-fine paradigms [7, 12, 36, 41], enhancing feature representations [9, 18] and modeling output formats [23, 39]. Another important line is to optimize the cost aggregation [29, 31] by adaptively weighting contributions from various views. In this paper, following the spirit, we introduce the adaptive cost aggregation tailored for the NVS task to mitigate the issue of inconsistent features caused by reflections and occlusions.

Generalizable NeRF. With implicit continuous representation and differentiable volume rendering, NeRF [21] achieves photo-realistic view synthesis. However, NeRF and its downstream expansion works [2, 3, 5, 22, 32, 33] require an expensive per-scene optimization process. To address this issue, some generalizable NeRF methods have been proposed, following a reconstruction-and-rendering pipeline. In the reconstruction phase, each sampled point is assigned a feature descriptor. Specifically, according to the descriptors, generalizable NeRF methods can be categorized into the following types: appearance descriptors [40], aggregated multi-view descriptors [16, 26, 28], cost volume interpolated descriptors [4, 16, 19], and correspondence matching descriptors [6]. Despite different forms, these descriptors only aggregate inter-view information or are interpolated from the low-resolution cost volume, lacking the ability to effectively perceive 3D spatial context. To remedy the issue, we utilize a proposed aggregator to facilitate the interaction of spatial information. In the rendering phase, volume density is obtained by decoding descriptors. For radiance, [16, 28] predict blending weights to combine color from source views, while [4, 6, 19, 26, 40] directly regress features. In this paper, we observe that these two strategies benefit different regions and thus propose a unified structure to integrate their advantages.

3. Preliminaries

Learning-based MVS. Given a target image and N source images, MVS aims to recover the geometry, such as the depth map of the target image. The key idea of MVS is to construct the cost volume from multi-view inputs, aggregating 2D information into 3D geometry-aware representation. Specifically, each voxel-aligned feature vector f_c of cost volume can be computed as:

$$f_c = \Omega(f_t, f_s^1, \dots, f_s^N), \quad (1)$$

where f_t and f_s^i represent the target feature vector and the warped source feature vector, respectively. And Ω denotes a consistency metric, such as variance. The underlying principle is that if a sampled depth is close to the actual depth, the multi-view features of the sampled point are supposed

to be similar, which naturally performs multi-view correspondence matching and geometry reasoning, facilitating the generalization to unseen scenes.

Generalizable NeRF. In generalizable NeRFs, each sampled point is assigned a geometry-aware feature descriptor f_p , as $(x, d, f_p) \rightarrow (\sigma, r)$, where x and d represent the coordinate and view direction used by NeRF [21]. σ and r denote the volume density and radiance for the sampled point, respectively. Specifically, the volume density σ can be obtained from the descriptor via $\sigma = \text{MLP}(x, f_p)$. For the radiance r , one approach [16, 28] is to predict blending weights to combine color values from source views, as:

$$r = \sum_{i=1}^N \frac{\exp(w_i)c_i}{\sum_{j=1}^N \exp(w_j)}, \text{ where } w_i = \text{MLP}(x, d, f_p, f_s^i), \quad (2)$$

where $\{c_i\}_{i=1}^N$ are color values from source views. These colors provide referential factors, facilitating convergence and better performance in most areas. However, in occluded and reflective regions, these colors introduce misleading bias to the combination, resulting in distorted colors. Besides, for pixels on object boundaries, it is hard to accurately locate reference points for blending, resulting in oscillated colors between the foreground and background. Another approach [4, 6, 40] is to directly regress the radiance from the feature per $r = \text{MLP}(x, d, f_p)$. As fewer inductive biases are imposed on the output space, the model can learn to predict fewer artifacts in challenging areas (Fig. 2). With the volume density and radiance of sampled points, the color values c of each pixel can be computed by volume rendering, given by:

$$c = \sum_k \tau_k (1 - \exp(-\sigma_k)) r_k, \text{ where } \tau_k = \exp(-\sum_{j=1}^{k-1} \sigma_j), \quad (3)$$

where τ represents the volume transmittance.

4. Method

Given a set of source views $\{I_s^i\}_{i=1}^N$, NVS aims to generate a target view at a novel camera pose. As illustrated in Fig. 3, our method consists of an NVS pipeline wrapped in a coarse-to-fine framework. In the pipeline, we first employ a feature pyramid network [17] to extract multi-scale features from the source views. Then, we propose an Adaptive Cost Aggregation module (Sec. 4.1) to construct a cost volume, which is further processed by a 3D-CNN to infer the geometry. Guided by the estimated geometry, we re-sample 3D points around the surface and apply the Spatial-View Aggregator (Sec. 4.2) to encode 3D-aware feature descriptors f_p for sampled points. Finally, we decode f_p into two intermediate views using two strategies introduced in Sec. 3 and fuse them into the final target view through Consistency-Aware Fusion (Sec. 4.3). This pipeline is iteratively called. Initially, the low-resolution target view is generated and the

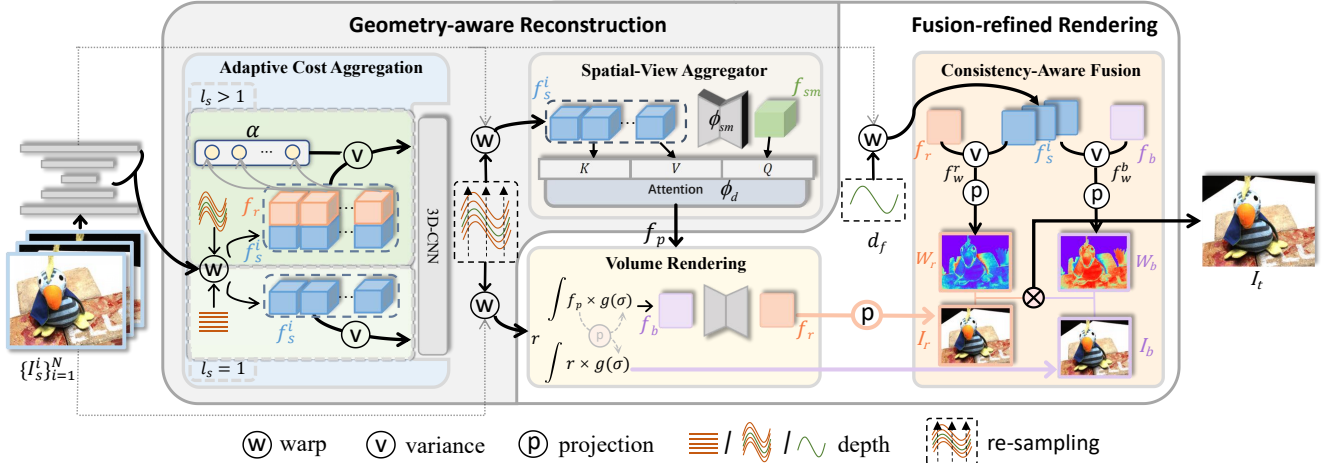


Figure 3. **The overview of GeFu.** In the reconstruction phase, we first infer the geometry from the constructed cost volume, and the geometry guides us to further re-sample 3D points around the surface. For each sampled point, the warped features from source images are aggregated and then fed into our proposed Spatial-View Aggregator (SVA) to learn spatial and inter-view context-aware descriptors f_p . In the rendering phase, we apply two decoding strategies to obtain two intermediate views and fuse them into the final target view in an adaptive way, termed Consistency-Aware Fusion (CAF). Our pipeline adopts a coarse-to-fine architecture, the geometry from the coarse stage ($l_s = 1$) guides the sampling at the fine stage ($l_s > 1$), and the features from the coarse stage are transferred to the fine stage for ACA to improve geometry estimation. Our network is trained end-to-end using only RGB images.

rough scene geometry is captured. Then, in the subsequent refining stage, high-resolution results with fine-grained geometry are obtained.

4.1. Adaptive Cost Aggregation

The core process of geometry reasoning is to construct a cost volume that encodes the multi-view feature consistency. Previous works [4, 16] treat different views equally and employ the variance operator to construct a cost volume. However, due to potential occlusions and varying lighting conditions among different views, multi-view features of the same 3D point may exhibit notable disparities, thereby resulting in misleading variance values. Inspired by [31], we propose to adaptively weight the contribution of different views to the cost volume, termed Adaptive Cost Aggregation (ACA). ACA will suppress the cost contribution from features that are inconsistent with the novel views caused by reflections or occlusions, and enhance the contribution of better-matched pixel pairs. The voxel-aligned feature f_c of cost volume can be computed as:

$$f_c = \frac{1}{N} \sum_{i=1}^N (1 + \alpha(f_c^i)) \odot f_c^i, \text{ where } f_c^i = (f_s^i - f_t)^2, \quad (4)$$

where f_t and f_s^i denote the target feature vector and the warped feature vector of source image I_s^i , respectively. \odot denotes Hadamard multiplication and $\alpha(\cdot)$ represents the adaptive weight for each view. However, an important challenge is that the target view is available in MVS, but not for NVS. To remedy this, we adopt a coarse-to-fine framework

where a coarse novel view is first generated and serves as the target view in Eq. (4). Specifically, we obtain the coarse-stage feature f_b by accumulating descriptors f_p of sampled points along the ray, as:

$$f_b = \sum_k \tau_k (1 - \exp(-\sigma_k)) f_p^k. \quad (5)$$

The coarse-stage features are then fed into a 2D U-Net for spatial aggregation to obtain f_r . We replace f_t in Eq. (4) with the feature f_r to construct a robust cost volume, benefiting the geometry estimation.

4.2. Spatial-View Aggregator

With the estimated geometry, 3D points around objects' surfaces can be re-sampled, and the subsequent step is encoding descriptors for these sampled points. Existing methods [16, 28] aggregate inter-view features with a pooling network ρ to construct the descriptors $f_p = \rho(\{f_s^i\}_{i=1}^N)$. However, these descriptors only encode multi-view information and lack the awareness of 3D spatial context, leading to discontinuities in the descriptor space. To introduce 3D spatial information, a feasible approach is to interpolate the low-resolution regularized cost volume, but it lacks fine-grained details. To address these issues, we design a 3D-aware descriptors encoding approach. We first utilize a 3D U-Net termed ϕ_{sm} to aggregate 3D spatial context, as:

$$f_{sm} = \phi_{sm}(\rho(\{f_s^i\}_{i=1}^N)). \quad (6)$$

However, this may lead to the smoothing of 3D features and cause the loss of some high-frequency geometric de-

tails. Therefore, we use the smoothed features as queries to re-gather inter-view high-frequency details, as:

$$f_p = \phi_d(f_{sm}, \{f_s^i\}_{i=1}^N), \quad (7)$$

where ϕ_d is an attention module. f_{sm} is the input query q , and $\{f_s^i\}_{i=1}^N$ are the key sequences k and value sequences v . The sequence lengths of q , k , and v are 1, N , and N , respectively, which results in only a slight increase in computational costs.

4.3. Consistency-Aware Fusion

With the descriptor f_p , the volume density σ is first acquired through an MLP, and color values can be obtained using two decoding approaches discussed in Sec. 3. We observe that these two approaches exhibit advantages in different areas. To combine their strengths, we propose to predict two intermediate views using these two approaches separately, and then fuse them into the final target view.

Specifically, for the blending approach, the radiance r of each sampled point can be computed using Eq. (2). And then the pixel color c_b and feature f_b are obtained via the volume rendering manner per Eq. (3) and Eq. (5), respectively. For another regression approach, one practice is predicting radiance for sampled points and then accumulating it into pixel color. In contrast, to reduce computational costs, we accumulate point-wise descriptors into features and then decode them into pixel color. Specifically, we feed the accumulated feature into a 2D U-Net for spatial enhancement to obtain pixel feature f_r , followed by an MLP to yield the pixel color c_r .

Since these two approaches excel in different areas, instead of using a fixed operator, such as average, we propose dynamically fusing them via $c_t = w_b c_b + w_r c_r$, where w_b and w_r are predicted fusing weights. As the target view is unavailable, comparing the quality of c_b and c_r becomes a chicken-and-egg problem. A naive practice is to directly predict fusing weights from features, which is under-constrained (Sec. 5.5). In contrast, we propose to learn fusing weights by using the multi-view feature consistency as a hint. The underlying motivation is that if the predicted colors closely resemble the ground-truth colors, the corresponding features of the predicted view and source views under the correct depth are supposed to be similar.

Specifically, we first obtain the final predicted depth d_f in a volume rendering-like way per $d_f = \sum_k \tau_k (1 - \exp(-\sigma_k)) d_k$, where d_k represents the depth of sampled point. With the depth d_f , we can obtain the warped features $\{f_s^i\}_{i=1}^N$ from source views, and the multi-view consistency can be computed using variance:

$$f_w^{[b,r]} = \text{var}(f_{[b,r]}, f_s^1, \dots, f_s^N). \quad (8)$$

We then feed the consistency $f_w^{[b,r]}$ into an MLP to obtain the fusing weight $w_{[b,r]}$, followed by a softmax operator for

normalization. The final target view can be represented in matrix form as $I_t = W_b I_b + W_r I_r$.

4.4. Loss Function

Our model is trained end-to-end only using the RGB image as supervision. Following [16], we use the mean squared error loss as:

$$L_{mse} = \frac{1}{N_p} \sum_{i=1}^{N_p} \|\hat{c}_i - c_i\|_2^2, \quad (9)$$

where N_p is the number of pixels and \hat{c}_i and c_i are the ground-truth and predicted pixel color, respectively. In addition, the perceptual loss [42] and ssim loss [30] can be applied, as:

$$\begin{aligned} L_{perc} &= \|h(\hat{I}) - h(I)\|, \\ L_{ssim} &= 1 - \text{ssim}(\hat{I}, I), \end{aligned} \quad (10)$$

where h is the perceptual function (a VGG16 network). \hat{I} and I are the ground-truth and predicted image patches, respectively. The loss at the k^{th} stage is as follows:

$$L^k = L_{mse} + \lambda_p L_{perc} + \lambda_s L_{ssim}, \quad (11)$$

where λ_p and λ_s refer to loss weights. The overall loss is:

$$L = \sum_{k=1}^{N_s} \lambda^k L^k \quad (12)$$

where N_s refers to the number of coarse-to-fine stages and λ^k represents the loss weight of the k^{th} stage.

5. Experiments

5.1. Settings

Datasets. Following MVSNeRF [4], we divide the DTU [1] dataset into 88 training scenes and 16 test scenes. We first train our generalizable model on the 88 training scenes of the DTU dataset and then evaluate the trained model on the 16 test scenes. To further demonstrate the generalization capability of our method, we also test the trained model (without any fine-tuning) on 8 scenes from the Real Forward-facing [20] dataset and 8 scenes from the NeRF Synthetic [21] dataset, both of which have significant differences in view distribution and scene content compared to the DTU dataset. The image resolutions of the DTU, the Real Forward-facing, and the NeRF Synthetic datasets are 512×640 , 640×960 , and 800×800 , respectively. The quality of synthesized novel views is measured by PSNR, SSIM [30], and LPIPS [42] metrics.

Baselines. We compare our methods with state-of-the-art generalizable NeRF methods [4, 6, 16, 19, 26, 28, 40]. For generalization with three views and per-scene optimization,

Method	Settings	DTU [1]			Real Forward-facing [20]			NeRF Synthetic [21]		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
PixelNeRF [40]	3-view	19.31	0.789	0.382	11.24	0.486	0.671	7.39	0.658	0.411
IBRNet [28]		26.04	0.917	0.191	21.79	0.786	0.279	22.44	0.874	0.195
MVSNeRF [4]		26.63	0.931	0.168	21.93	0.795	0.252	23.62	0.897	0.176
NeuRay [19]		25.81	0.868	0.160	23.39	0.744	0.217	24.58	0.892	0.163
ENeRF [16]		27.61	0.957	0.089	23.63	0.843	0.182	26.17	0.943	0.085
GNT [26]		26.39	0.923	0.156	22.98	0.761	0.221	25.80	0.905	0.104
MatchNeRF [6]		26.91	0.934	0.159	22.43	0.805	0.244	23.20	0.897	0.164
Ours	29.36	0.969	0.064	24.28	0.863	0.162	26.99	0.952	0.070	
MVSNeRF [4]	2-view	24.03	0.914	0.192	20.22	0.763	0.287	20.56	0.856	0.243
NeuRay [19]		24.51	0.825	0.203	22.73	0.720	0.236	22.42	0.865	0.228
ENeRF [16]		25.48	0.942	0.107	22.78	0.821	0.191	24.83	0.931	0.117
GNT [26]		24.32	0.903	0.201	20.91	0.683	0.293	23.47	0.877	0.151
MatchNeRF [6]		25.03	0.919	0.181	20.59	0.775	0.276	20.57	0.864	0.200
Ours		26.98	0.955	0.081	23.39	0.839	0.176	25.30	0.939	0.082

Table 1. **Quantitative results under the generalization setting.** We show the average results of PSNRs, SSIMs, and LPIPSs on three datasets under two settings for the number of input views. The comparison methods are organized based on the year of publication.

Method	DTU [1]			Real Forward-facing [20]			NeRF Synthetic [21]		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
NeRF _{10.2h} [21]	27.01	0.902	0.263	25.97	0.870	0.236	30.63	0.962	0.093
IBRNet _{ft-1.0h} [28]	31.35	0.956	0.131	24.88	0.861	0.189	25.62	0.939	0.111
MVSNeRF _{ft-15min} [4]	28.51	0.933	0.179	25.45	0.877	0.192	27.07	0.931	0.168
NeuRay _{ft-1.0h} [19]	26.96	0.847	0.174	24.34	0.781	0.205	25.91	0.896	0.115
ENeRF _{ft-1.0h} [16]	28.87	<u>0.957</u>	0.090	24.89	0.865	0.159	27.57	0.954	0.063
Ours _{ft-15min}	30.10	0.966	<u>0.069</u>	<u>26.62</u>	<u>0.903</u>	<u>0.110</u>	28.57	0.958	<u>0.060</u>
Ours _{ft-1.0h}	<u>30.18</u>	0.966	0.068	26.76	0.905	0.106	<u>28.81</u>	<u>0.960</u>	0.058

Table 2. **Quantitative results under the per-scene optimization setting.** The best result is in **bold**, and the second-best is underlined.

we follow the same setting as [4, 6, 16] and borrow the results of [4, 6, 21, 28, 40] from [4, 6]. We evaluate [16] using the official code and trained models. To keep consistent with the same settings for a fair comparison, such as the number of input views, dataset splitting, view selection, and image resolution, we use the released code and trained model of [19] and retrain [26] with the released code, and evaluate them under our test settings. For generalization with two views, we borrow the results of [4, 6] from [6]. For other baselines [16, 19, 26], we evaluate them using the released model [16, 19] or the retrained model [26].

Implementation Details. Following [16], the number of coarse-to-fine stages N_s is set to 2. In our coarse-to-fine framework, we sample 64 and 8 depth planes for the coarse-level and fine-level cost volumes, respectively. And we sample 8 and 2 points per ray for the coarse-level and fine-level view rendering, respectively. We set $\lambda_p = 0.1$ and $\lambda_s = 0.1$ in Eq. (11), while $\lambda^1 = 0.5$ and $\lambda^2 = 1$ in Eq. (12). We train our model on four RTX 3090 GPUs us-

ing the Adam [14] optimizer. Refer to the supplementary material for more implementation and network details.

5.2. Generalization Results

We report quantitative results on DTU, Real Forward-facing, and NeRF Synthetic datasets in Table 1 under generalization settings. PixelNeRF [40], which applies appearance descriptors, has reasonable results on the DTU test set, but insufficient generalization on the other two datasets. Other methods [4, 6, 16, 19, 26, 28] that model the scene geometry implicitly or explicitly by aggregating multi-view features can maintain relatively good generalization. Thanks to our proposed modules tailored for both the reconstruction and rendering phases, our method achieves significantly better generalizability. As shown in Fig. 4, the views produced by our method preserve more scene details and contain fewer artifacts. In challenging areas, such as occluded regions, object boundaries, and regions with complex geometry, our method significantly out-

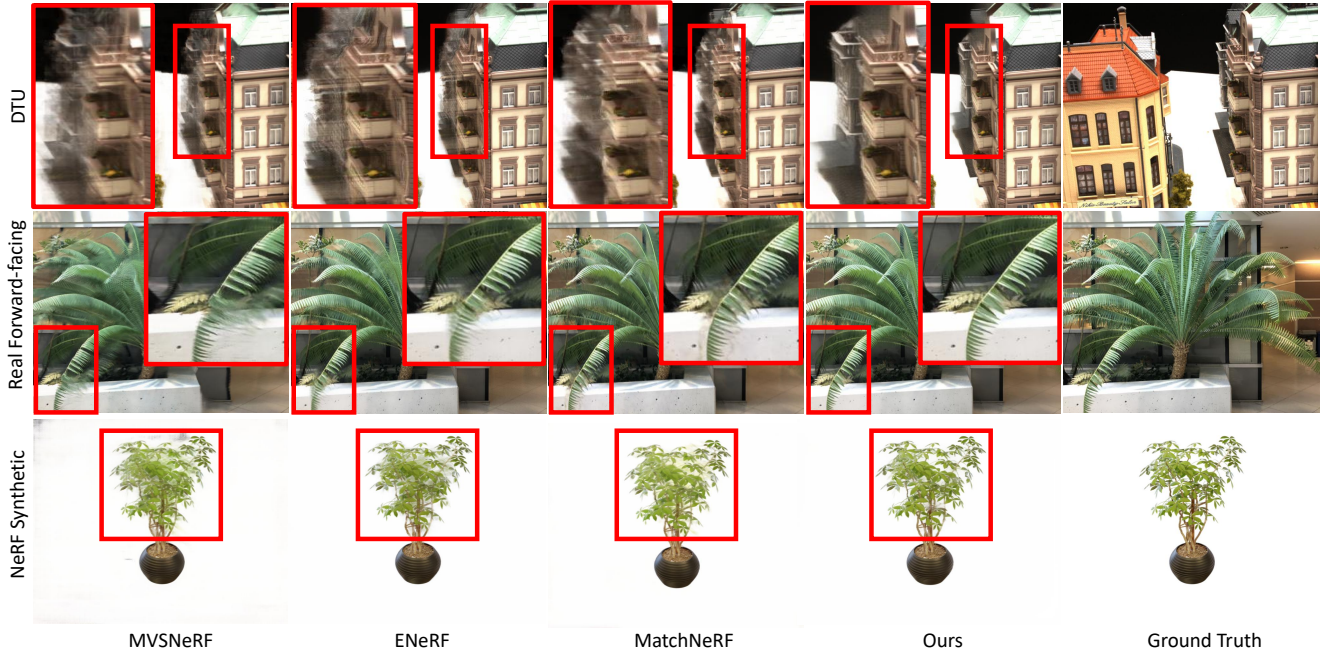


Figure 4. Qualitative comparison of rendering quality with state-of-the-art methods [4, 6, 16] under generalization and three input views settings.

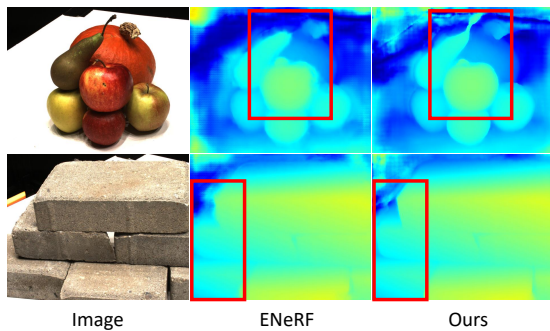


Figure 5. Qualitative comparison of depth maps with [16].

performs other methods. Refer to the supplementary material for more qualitative comparisons.

5.3. Per-scene Fine-tuning Results

The quantitative results after per-scene optimization are shown in Table 2 and we report the results of our method after 15 minutes and 1 hour of fine-tuning. Due to the excellent initialization provided by our generalization model, only a short period of fine-tuning is needed to achieve good results. Our results after 15 minutes of fine-tuning are comparable to or even superior to those of NeRF [21] optimized for substantially longer time (10.2 hours), and also outperform the results of other generalization methods after fine-tuning. With a longer fine-tuning duration, such as 1 hour, the rendering quality can be further improved. Qualitative

Method	Reference view			Novel view		
	Abs err ↓	Acc(2)↑	Acc(10)↑	Abs err ↓	Acc(2)↑	Acc(10)↑
MVSNet [37]	3.60	0.603	0.955	-	-	-
PixelNeRF [40]	49	0.037	0.176	47.8	0.039	0.187
IBRNet [28]	338	0.000	0.913	324	0.000	0.866
MVSNeRF [4]	4.60	0.746	0.913	7.00	0.717	0.866
ENeRF [16]	3.80	0.837	0.939	4.60	0.792	0.917
Ours	2.47	0.900	0.971	2.83	0.879	0.961

Table 3. Quantitative results of depth reconstruction on the DTU test set. MVSNet is trained with depth supervision while other methods are trained with only RGB image supervision. “Abs err” represents the average absolute error and “Acc(X)” means the percentage of pixels with an error less than X mm.

results can be found in the supplementary material.

5.4. Depth Reconstruction Results

Following [4, 16], we report the performance of depth reconstruction in Table 3. Our method can achieve higher depth accuracy than other methods, even including the MVS method MVSNet [37] that is trained with depth supervision. As shown in Fig. 5, the depth map produced by our method is more refined, such as sharper object edges.

5.5. Ablations and Analysis

Ablation studies. As shown in Table 4, we conduct ablation studies to investigate the contribution of each proposed module. Each individual component can benefit the base-

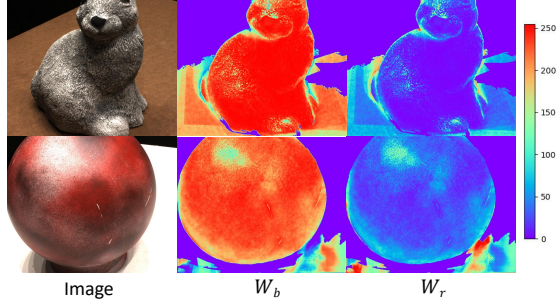


Figure 6. **Visualization of Fusion Weights.** W_b and W_r represent the weight maps of the blending approach and regression approach, respectively.

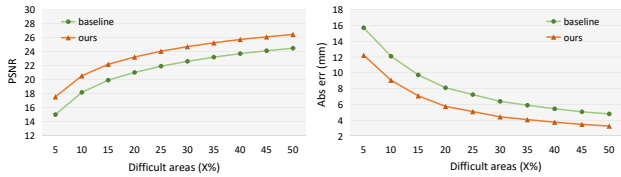


Figure 7. **Quantitative analysis of difficult areas.** The X-axis difficult areas ($X\%$) represents considering the area with the top $X\%$ of values in W_r as a difficult area. A smaller threshold X indicates a more challenging area.

line model in both view quality and depth accuracy, with CAF having the highest gain. An interesting phenomenon is that CAF greatly improves depth accuracy, indicating that a well-designed view decoding approach also facilitates depth prediction. Combining all components results in the greatest gain, with a 5.9% increase in PSNR and a 34.5% improvement in depth error compared to the baseline model.

CAF working mechanism. As shown in Fig. 6, we visualize the fusion weights of two decoding approaches. The regression approach exhibits higher confidence in challenging areas such as object edges and reflections, while the blending approach shows higher confidence in most other areas, which is consistent with the observation in Fig. 2. As shown in Fig. 7, we define challenging areas as those with high confidence in W_r and divide them by a series of thresholds. A smaller threshold X indicates a more difficult region. When $X = 5\%$, our method improves PSNR by 2.55db and “Abs err” by 3.47mm. When the threshold increases, such as $X = 50\%$, our method improves PSNR by 1.97db and “Abs err” by 1.54mm, which further demonstrates the superiority of our method in challenging areas.

Fusion strategy. As shown in Table 5, we investigate the performance of different fusion strategies. Comparing No.1 and No.2, the result of using the blending approach alone is better than that of using the regression approach alone. For No.4, the DWF represents fusion weights derived directly from features of the two intermediate views, which

ACA	SVA _{sm}	SVA _d	CAF	PSNR ↑	Abs err ↓
✗	✗	✗	✗	27.73	4.32
✓	✗	✗	✗	28.35	3.85
✗	✓	✗	✗	28.30	3.53
✗	✗	✓	✗	27.89	3.94
✗	✗	✗	✓	28.85	3.36
✓	✓	✗	✗	28.53	3.54
✓	✓	✓	✗	28.64	3.06
✓	✓	✓	✓	29.36	2.83

Table 4. **Ablation studies on the DTU dataset.** We report the image quality (PSNR) and depth accuracy (Abs err) metrics without per-scene fine-tuning under three input views settings. SVA_{sm} and SVA_d represent ϕ_{sm} and ϕ_d of SVA, respectively.

	Blending	Regression	Fusion	PSNR ↑	SSIM ↑	LPIPS ↓
No.1	✓	✗	✗	27.73	0.956	0.088
No.2	✗	✓	✗	26.83	0.955	0.091
No.3	✓	✗	AE	27.89	0.959	0.081
No.4	✓	✓	DWF	28.10	0.963	0.075
No.5	✓	✓	CAF	28.85	0.966	0.070

Table 5. **Comparison of different fusion strategies.** The AE represents the refinement by an autoencoder. The DWF represents direct weighted fusion. The CAF is our proposed Consistency-Aware Fusion.

greatly degrades performance compared to our proposed way of checking the multi-view consistency (No.5). Our fusion approach utilizes the advantages of the two decoding approaches to refine the synthesized view. In No.3, we refine the synthesized view decoded in a single way, where the synthesized view is fed into an auto-encoder for refinement, which has limited improvement.

6. Conclusion

In this paper, we present a generalizable NeRF method capable of achieving high-fidelity view synthesis. Specifically, during the reconstruction phase, we propose Adaptive Cost Aggregation (ACA) to improve geometry estimation and Spatial-View Aggregator (SVA) to encode 3D context-aware descriptors. In the rendering phase, we conduct an analysis of two existing color decoding strategies and introduce the Consistency-Aware Fusion (CAF) module to unify their advantages to refine the synthesized view quality. We integrate these modules into a coarse-to-fine framework, termed GeFu. Extensive evaluations and ablations demonstrate the effectiveness of our proposed modules.

References

- [1] Henrik Aanaes, Rasmus Ramsbol Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.*, 120:153–168, 2016. [1](#), [2](#), [5](#), [6](#)
- [2] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. [3](#)
- [3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerf: Neural reflectance decomposition from image collections. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 12684–12694, 2021. [3](#)
- [4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnrf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 14124–14133, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [5] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM Trans. Graph.*, 41(1):1–26, 2022. [3](#)
- [6] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)
- [7] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2524–2534, 2020. [3](#)
- [8] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 7911–7920, 2021. [1](#)
- [9] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet global context-aware multi-view stereo network with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8585–8594, 2022. [3](#)
- [10] Pascal Fua and Yvan G Leclerc. Object-centered surface reconstruction combining multi-image stereo and shading. *Int. J. Comput. Vis.*, 16(ARTICLE):35–56, 1995. [2](#)
- [11] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 873–881, 2015. [2](#)
- [12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuoqun Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2495–2504, 2020. [3](#)
- [13] M. Johari, Y. Lepoittevin, and F. Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2022. [1](#)
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)
- [15] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2021. [1](#)
- [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 Conference Proceedings*, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [3](#)
- [18] Tianqi Liu, Xinyi Ye, Weiyue Zhao, Zhiyu Pan, Min Shi, and Zhiguo Cao. When epipolar constraint meets non-local operators in multi-view stereo. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 18088–18097, 2023. [3](#)
- [19] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2022. [1](#), [3](#), [5](#), [6](#)
- [20] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 38(4):1–14, 2019. [1](#), [2](#), [5](#), [6](#)
- [21] 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 *Proc. Eur. Conf. Comput. Vis.*, 2020. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)
- [22] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 5865–5874, 2021. [3](#)
- [23] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo a unified representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8645–8654, 2022. [3](#)
- [24] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4104–4113, 2016. [2](#)
- [25] Johannes L Schonberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 501–518. Springer, 2016. [2](#)
- [26] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that nerf needs? In *Proc. Int. Conf. Learn. Repr.*, 2023. [1](#), [3](#), [5](#), [6](#)
- [27] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 15182–15192, 2021.
- [28] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snaveley, and Thomas Funkhouser. Ibrnet:

- Learning multi-view image-based rendering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [29] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster epipolar transformer for efficient multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 573–591. Springer, 2022. [2](#), [3](#)
- [30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. [5](#)
- [31] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvnet adaptive aggregation recurrent multi-view stereo network. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6187–6196, 2021. [2](#), [3](#), [4](#)
- [32] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 9421–9431, 2021. [3](#)
- [33] Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. Neutex: Neural texture mapping for volumetric neural rendering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 7119–7128, 2021. [3](#)
- [34] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5438–5448, 2022. [1](#)
- [35] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *Proc. Eur. Conf. Comput. Vis.*, pages 674–689. Springer, 2020. [3](#)
- [36] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4877–4886, 2020. [3](#)
- [37] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet depth inference for unstructured multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 767–783, 2018. [2](#), [7](#)
- [38] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5525–5534, 2019. [3](#)
- [39] Xinyi Ye, Weiyue Zhao, Tianqi Liu, Zihao Huang, Zhiguo Cao, and Xin Li. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 17661–17670, 2023. [3](#)
- [40] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2021. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)
- [41] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *Proc. Br. Mach. Vis. Conf.*, 2020. [3](#)
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 586–595, 2018. [5](#)