

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 AIA, Huazhong University of Science and Technology

{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) [25] has shown impressive performance in NVS. However, NeRF requires a lengthy optimization with densely captured images

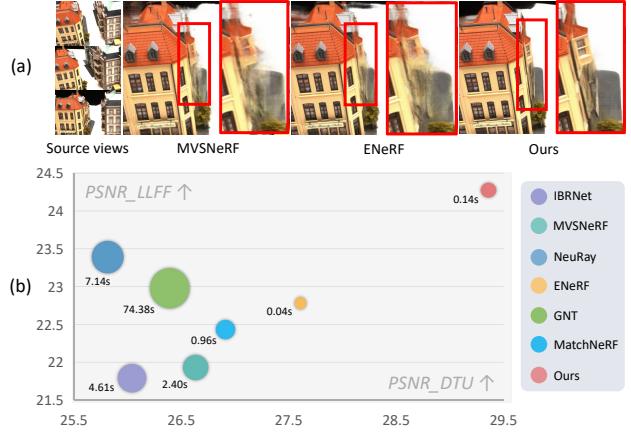


Figure 1. **Comparison with existing methods.** (a) With three input source views, our generalizable model synthesizes novel views with higher quality than existing methods [5, 20] in the severe occluded area. (b) Circle area represents inference time. The X-axis represents the PSNR on the DTU dataset [1] and the Y-axis represents the PSNR on the Real Forward-facing dataset [24]. Our method attains state-of-the-art performance.

for each scene, which limits its applications.

To address this issue, some recent methods [5, 7, 9, 15, 19, 20, 23, 30–32, 38, 44] 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 [32, 44] 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 [5, 20] 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,

*Corresponding author

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 [33, 35], 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 [7, 20, 32, 44] 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, [20, 32] predict blending weights to combine color values from source views, while [5, 7, 44] 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 into the final view. The fusing weights are learned by checking multi-view consistency, following an underlying princi-

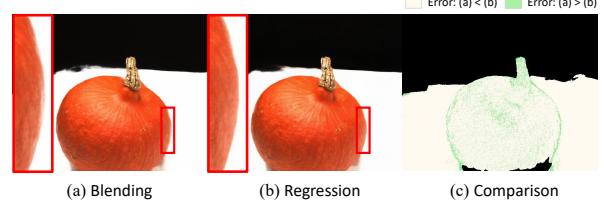


Figure 2. Comparison of two rendering strategies. (a) The view obtained using the blending approach that combines color values from source views. (b) The view obtained using the regression approach that directly regresses color values from features. (c) Accuracy comparison between two rendering strategies.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

ple 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 [24], and NeRF Synthetic [25] 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 [25]. 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 [11, 12, 28, 29] 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 [41] 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 [35, 39, 42] or coarse-to-fine paradigms [8, 14, 40, 45], enhancing feature representations [10, 22] and modeling output formats [27, 43]. Another important line is to optimize the cost aggregation [33, 35] 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 [25] achieves photo-realistic view synthesis. However, NeRF and its downstream expansion works [2, 3, 6, 26, 36, 37] 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 [44], aggregated multi-view descriptors [20, 30, 32], cost volume interpolated descriptors [5, 20, 23], and correspondence matching descriptors [7]. 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, [20, 32] predict blending weights to combine color from source views, while [5, 7, 23, 30, 44] 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 [25]. 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 [20, 32] 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 [5, 7, 44] 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:

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

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

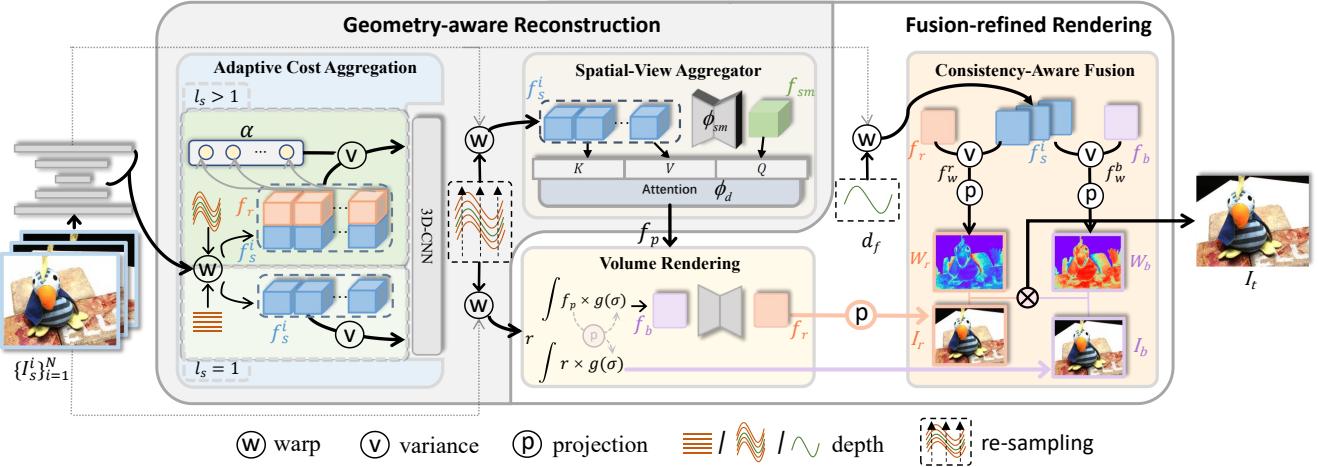


Figure 3. **The overview of GeFu.** In the reconstruction phase, we first infer the geometry from the constructed cost volume, and the geometry guides us to further re-sample 3D points around the surface. For each sampled point, the warped features from source images are aggregated and then fed into 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.

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 [5, 20] 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 [35], 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 [20, 32] 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 details. 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 [20], 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 [46] and ssim loss [34] 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 [5], 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 [24] dataset and 8 scenes from the NeRF Synthetic [25] 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 [34], and LPIPS [46] metrics.

Baselines. We compare our methods with state-of-the-art generalizable NeRF methods [5, 7, 20, 23, 30, 32, 44]. For generalization with three views and per-scene optimization,

Method	Settings	DTU [1]			Real Forward-facing [24]			NeRF Synthetic [25]		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
PixelNeRF [44]	3-view	19.31	0.789	0.382	11.24	0.486	0.671	7.39	0.658	0.411
IBRNet [32]		26.04	0.917	0.191	21.79	0.786	0.279	22.44	0.874	0.195
MVSNeRF [5]		26.63	0.931	0.168	21.93	0.795	0.252	23.62	0.897	0.176
NeuRay [23]		25.81	0.868	0.160	23.39	0.744	0.217	24.58	0.892	0.163
ENeRF † [20]		27.61	0.956	0.091	22.78	0.808	0.209	26.65	0.947	0.072
ENeRF [20]		27.61	0.957	0.089	23.63	0.843	0.182	26.17	0.943	0.085
GNF [30]		26.39	0.923	0.156	22.98	0.761	0.221	25.80	0.905	0.104
MatchNeRF [7]		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 [5]	2-view	24.03	0.914	0.192	20.22	0.763	0.287	20.56	0.856	0.243
NeuRay [23]		24.51	0.825	0.203	22.73	0.720	0.236	22.42	0.865	0.228
ENeRF [20]		25.48	0.942	0.107	22.78	0.821	0.191	24.83	0.931	0.117
GNF [30]		24.32	0.903	0.201	20.91	0.683	0.293	23.47	0.877	0.151
MatchNeRF [7]		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. ENeRF † represents results borrowed from the original paper. The comparison methods are organized based on the year of publication.

Method	DTU [1]			Real Forward-facing [24]			NeRF Synthetic [25]		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
NeRF $_{10.2h}$ [25]	27.01	0.902	0.263	25.97	0.870	0.236	30.63	0.962	0.093
IBRNet $_{ft-1.0h}$ [32]	31.35	0.956	0.131	24.88	0.861	0.189	25.62	0.939	0.111
MVSNeRF $_{ft-15min}$ [5]	28.51	0.933	0.179	25.45	0.877	0.192	27.07	0.931	0.168
NeuRay $_{ft-1.0h}$ [23]	26.96	0.847	0.174	24.34	0.781	0.205	25.91	0.896	0.115
ENeRF $_{ft-1.0h}$ [20]	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 [5, 7, 20] and borrow the results of [5, 7, 25, 32, 44] from [5, 7]. We evaluate [20] 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 [23] and retrain [30] with the released code, and evaluate them under our test settings. For generalization with two views, we borrow the results of [5, 7] from [7]. For other baselines [20, 23, 30], we evaluate them using the released model [20, 23] or the retrained model [30].

Implementation Details. Following [20], 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 using the Adam [18] 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 [44], which applies appearance descriptors, has reasonable results on the DTU test set, but insufficient generalization on the other two datasets. Other methods [5, 7, 20, 23, 30, 32] 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



Figure 4. Qualitative comparison of rendering quality with state-of-the-art methods [5, 7, 20] under generalization and three input views settings.

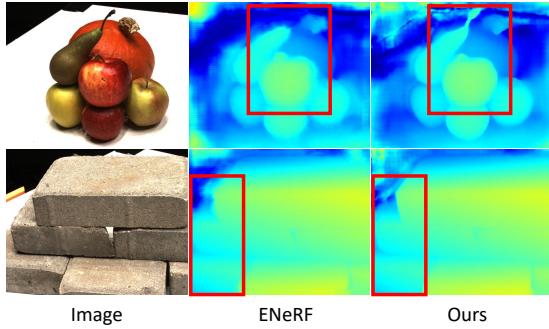


Figure 5. Qualitative comparison of depth maps with [20].

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 outperforms 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 [25] optimized

Method	Reference view			Novel view		
	Abs err ↓	Acc(2)↑	Acc(10)↑	Abs err ↓	Acc(2)↑	Acc(10)↑
MVSNet [41]	3.60	0.603	0.955	-	-	-
PixelNeRF [44]	49	0.037	0.176	47.8	0.039	0.187
IBRNet [32]	338	0.000	0.913	324	0.000	0.866
MVSNeRF [5]	4.60	0.746	0.913	7.00	0.717	0.866
ENeRF [20]	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.

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 results can be found in the supplementary material.

5.4. Depth Reconstruction Results

Following [5, 20], 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 [41] 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.

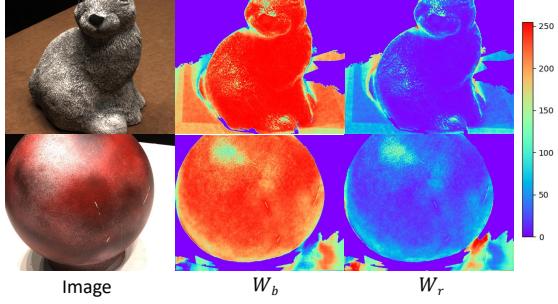


Figure 6. **Visualization of Fusion Weights.** W_b and W_r represent the weight maps of the blending approach and regression approach, respectively.

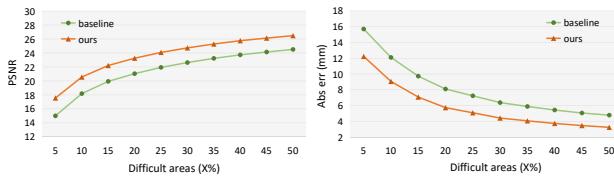


Figure 7. **Quantitative analysis of difficult areas.** The X-axis difficult areas ($X\%$) represents considering the area with the top $X\%$ of values in W_r as a difficult area. A smaller threshold X indicates a more challenging area.

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 baseline 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

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.

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 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 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](#), [11](#), [14](#), [15](#), [18](#)
- [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. Nerd: Neural reflectance decomposition from image collections. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 12684–12694, 2021. [3](#)
- [4] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: Unsupervised multi-view stereo with neural rendering. In *Proc. Eur. Conf. Comput. Vis.*, 2022. [13](#)
- [5] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, pages 14124–14133, 2021. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [11](#), [12](#), [15](#), [16](#), [17](#), [18](#), [19](#), [21](#), [22](#)
- [6] 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](#)
- [7] 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](#), [16](#), [17](#), [18](#), [19](#)
- [8] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Ernan 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](#)
- [9] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srdf): Learning view synthesis for sparse views of novel scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 7911–7920, 2021. [1](#)
- [10] 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](#)
- [11] 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](#)
- [12] 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](#)
- [13] Ross Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2015. [13](#)
- [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 *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2495–2504, 2020. [3](#), [13](#)
- [15] M. Johari, Y. Lepoittevin, and F. Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2022. [1](#)
- [16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [15](#)
- [17] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019. [13](#)
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#), [11](#)
- [19] 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](#)
- [20] 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](#), [11](#), [12](#), [15](#), [16](#), [17](#), [18](#), [19](#), [21](#), [22](#)
- [21] 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](#)
- [22] 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](#)
- [23] 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](#), [13](#), [15](#), [16](#), [21](#), [22](#)
- [24] 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](#), [11](#), [12](#), [13](#), [14](#), [15](#), [18](#)
- [25] 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](#), [11](#), [12](#), [13](#), [14](#), [15](#), [18](#), [21](#), [22](#)
- [26] 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](#)
- [27] 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
- [28] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4104–4113, 2016. 2
- [29] 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
- [30] 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, 15, 16, 21, 22
- [31] 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.
- [32] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, 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, 11, 15, 21, 22
- [33] 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, 14
- [34] 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, 11
- [35] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet adaptive aggregation recurrent multi-view stereo network. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6187–6196, 2021. 2, 3, 4, 14
- [36] 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
- [37] 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
- [38] 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
- [39] 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
- [40] 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
- [41] 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
- [42] 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
- [43] 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
- [44] 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, 15, 21, 22
- [45] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *Proc. Br. Mach. Vis. Conf.*, 2020. 3
- [46] 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, 11

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

Supplementary Material

7. Appendix

In the supplementary material, we present more details that are not included in the main text, including:

- More implementation and network details.
- Additional ablation experiments, including different numbers of views, feature discussions for the CAF module, intermediate view supervision, depth supervisions, and different implementations of ACA networks.
- More qualitative results, including novel view, depth maps, fusion weights, and error maps.
- Per-scene breakdown results.
- Limitations.

I. Implementation and Network Details

Implementation Details. Our generalizable model is trained on four RTX 3090 GPUs using the Adam [18] optimizer, with an initial learning rate of $5e - 4$. The learning rate is halved every 50k iterations. As shown in Fig. I, we present the metrics of the DTU test set [1] varying with the number of iterations. The model tends to converge after approximately 150k iterations, taking about 25 hours. It is worth noting that, with only 8k iterations, our model can achieve metrics of 27.68/0.961/0.080, surpassing the metrics of the SOTA method [20], which is 27.61/0.956/0.091 (PSNR/SSIM [34]/LPIPS [46]). Following ENeRF [20], during training, we select 2, 3, and 4 source views as inputs with probabilities of 0.1, 0.8, and 0.1. To save computational costs, the Consistency-aware Fusion (CAF) is exclusively employed in the fine stage, while in the coarse stage, a blending approach is used to synthesize a low-resolution target view, supervised by the ground-truth target view. During training, the final target view is generated by fusing two intermediate views as $I_t = W_b I_b + W_r I_r$. However, during evaluation for other datasets [24, 25], which have a different domain compared with DTU, increasing the weights of I_b can lead to slightly better performance. Therefore, we obtain the final view as $I_t = (I_b + (W_b I_b + W_r I_r))/2$. Our evaluation setup is consistent with ENeRF [20] and MVSNeRF [5]. The results of the DTU test set are evaluated using segmentation masks. The segmentation mask is defined based on the availability of ground-truth depth at each pixel. Since the marginal region of images is typically invisible to input images on the Real Forward-facing dataset [24], we evaluate the 80% area in the center of the images. Incidentally, all the inference time presented in Fig. 1 of the main text is measured at an

Input	Operation	Output
(C,D,H,W)	permute	(D,C,H,W)
(D,C,H,W)	cgr(C,C _h)	(D,C _h ,H,W)
(D,C _h ,H,W)	cgr(C _h ,C _h)	(D,C _h ,H,W)
(D,C _h ,H,W)	conv2d(C _h ,1)	(D,1,H,W)
(D,1,H,W)	sigmoid	(D,1,H,W)
(D,1,H,W)	permute	(1,D,H,W)

Table I. **The network architecture of Adaptive Cost Aggregation (ACA).** The cgr represents a block composed of conv2d, groupnorm, and relu. In our implementation, $C = 16$ and $C_h = 4$.

input image resolution of 512×640 . The number of our model’s parameters is 3.15M.

Network Details. Here, we will introduce the network details of the pooling network (Sec. 4.2) and ACA (Sec. 4.1) mentioned in the main text.

Pooling network. [20, 32] apply a pooling network ρ to aggregate multi-view features to obtain the descriptor via $f_p = \rho(\{f_s^i\}_{i=1}^N)$. The implementation details are as follows: initially, the mean μ and variance v of $\{f_s^i\}_{i=1}^N$ are computed. Subsequently, μ and v are concatenated with each f_s^i and an MLP is applied to generate a weight. The f_p is blended via a soft-argmax operator using obtained weights and multi-view features $(\{f_s^i\}_{i=1}^N)$.

ACA. Per the Eq. (4) in the main text, $\alpha(\cdot)$ represents the adaptive weight for each view, and the network architecture that learns these weights is shown in Table I.

Inference Speed. For an image with 512×640 resolution, the inference time of our method is 143ms. We decompose the inference time in Table II and results demonstrate that the inference time of our modules is only 71ms, with the remaining 72ms spent saving results.

II. Additional ablation experiments

Numbers of Views. As shown in Table III, we evaluate the performance of our trained generalization model and ENeRF [20] with different numbers of input views on the DTU test set [1]. With an increase in the number of input views, the performance improves as the model can leverage more multi-view information. In terms of both overall performance and the magnitude of performance improvement, our method outperforms ENeRF, indicating its superior capability in leveraging multi-view information for recon-

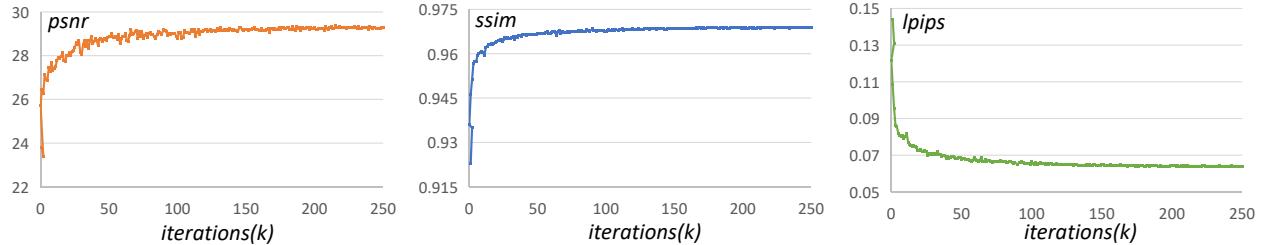


Figure I. The metrics of the DTU test set vary with the number of iterations.

Modules	coarse stage	fine stage
feature extractor		1.32
build rays	9.11	15.94
geometry	cost volume	7.48
	regularization	1.91
descriptor	pooling ρ	0.57
	SVA _{sm}	1.74
	SVA _d	0.73
view decoding	1.62	17.80
save in dictionary	13.27	58.31

Table II. Time overhead for each module (in milliseconds). The term “save in dictionary” refers to storing tensor results in dictionary form for subsequent evaluation of various metrics.

structuring scene geometry and rendering novel views. Additionally, we also present the performance of our method on the Real Forward-facing [24] and NeRF Synthetic [25] datasets under different numbers of input views, as shown in Table IV. The results demonstrate the same trend, indicating the capability of our model in leveraging multi-view information is generalizable.

Features for Intermediate Views. In Sec. 4.3 of the main text, f_b and f_r are utilized as the feature representations for the two intermediate views I_b and I_r , respectively. Subsequently, their consistency with source views is individually computed to learn fusion weights. The choice of using f_b as the feature for I_b is based on their similar volume rendering generation manners, while the selection of f_r as the feature for I_r is driven by their direct projection relationship. Here, we will discuss different selection strategies for the features of intermediate views. An alternative approach for the features of I_b is to blend features from source views. Similar to Eq. (2) in the main text, the calculation of the features f_b for I_b is as follows:

$$f_b = \sum_{i=1}^N \frac{\exp(w_i) f_s^i}{\sum_{j=1}^N \exp(w_j)}, \quad (\text{I})$$

where $w_i = \text{MLP}(x, d, f_p, f_s^i)$, f_s^i is the feature of the source image I_s^i . x and d represent the coordinate and view direction, respectively. f_p is the descriptor for 3D point.

Another more intuitive alternative is to use a feature extractor to extract features for both intermediate views, as:

$$F_{[b,r]} = \phi_e I_{[b,r]}, \quad (\text{II})$$

where ϕ_e represents a feature extractor, instantiated as a 2D U-Net. f_b and f_r are the pixel-wise features of F_b and F_r , respectively. As shown in Table V, the strategy employed in the main text is slightly superior to the other two alternative strategies. For the first alternative Eq. (I), f_b is obtained by blending features from source views. f_b lacks 3D context awareness, leading to some information loss in the subsequently accumulated pixel features. For the second alternative Eq. (II), f_b and f_r are extracted from scratch at the RGB level. This practice is disadvantageous for the subsequent learning of 3D consistency weights, due to the lack of utilization of 3D information. Additionally, the introduction of a feature extractor also increases the burden on the model. However, the strategy in the main text maximally utilizes the obtained 3D-aware descriptors, while also having the smallest computational cost compared to the other two alternative approaches.

Intermediate View Supervision. In the main text, we only supervise the images fused through CAF. However, simultaneously supervising intermediate results is also a common practice, whose final result is 29.15/0.968/0.065. This result is slightly inferior to supervising only the fused view (29.36/0.969/0.064). Because each of the two intermediate views has its own advantages, supervising only the fused view allows the network to focus on the fusion process, leveraging the strengths of both. However, simultaneously supervising the intermediate views burdens the network, diminishing its attention to the fusion process. In theory, if both intermediate views are entirely correct, the final fused view will be accurate regardless of the fusion process. The network prioritizes predicting two accurate intermediate views, which is a more challenging task.

Depth Supervision. A critical factor in the model’s synthesis of high-quality views is its perception of the scene geometry. MVS-based generalizable NeRF methods [5, 20,

Views	PSNR↑	SSIM↑	LPIPS↓	Abs err↓	Acc(2)↑	Acc(10)↑
2	26.98 / 25.48	0.955 / 0.942	0.081 / 0.107	3.86 / 5.53	0.835 / 0.756	0.942 / 0.107
3	29.36 / 27.61	0.969 / 0.957	0.064 / 0.089	2.83 / 4.60	0.879 / 0.792	0.961 / 0.917
4	29.77 / 27.73	0.971 / 0.959	0.062 / 0.089	2.73 / 4.26	0.880 / 0.804	0.961 / 0.929
5	29.91 / 27.54	0.971 / 0.958	0.062 / 0.091	2.69 / 4.29	0.882 / 0.800	0.961 / 0.928

Table III. **The performance of our method and ENeRF with different numbers of input views on the DTU test set.** Each item represents (Ours/ENeRF’s). “Abs err” denotes the average absolute error and “Acc(X)” means the percentage of pixels with an error less than X mm.

Views	Real Forward-facing [24]			NeRF Synthetic [25]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
2	23.39	0.839	0.176	25.30	0.939	0.082
3	24.28	0.863	0.162	26.99	0.952	0.070
4	24.91	0.876	0.157	27.31	0.953	0.069

Table IV. **The performance of our method with varying numbers of input views on the Real Forward-facing and NeRF Synthetic datasets.**

	PSNR↑	SSIM↑	LPIPS↓
No.1	29.36	0.969	0.064
No.2	29.24	0.969	0.065
No.3	29.08	0.968	0.066

Table V. **Different strategies for the features of intermediate views.** No.1 represents the strategy in the main text. No.2 represents the strategy using Eq. (I). No.3 represents the strategy using Eq. (II).

Depth	PSNR↑	SSIM↑	LPIPS↓	Abs err↓	Acc(2)↑	Acc(10)↑
Self-supervision	29.21	0.968	0.064	3.21	0.873	0.957
Supervision	29.31	0.969	0.064	2.95	0.875	0.957
None	29.36	0.969	0.064	2.83	0.879	0.961

Table VI. **The comparison of different depth supervision signals.** The self-supervision represents using unsupervised depth loss and the supervision represents using ground-truth depth for supervision. The term “None” refers to training without any depth supervision signals.

[23], including our method, aim to improve the quality of view synthesis by enhancing the geometry prediction. By only supervising RGB images, excellent geometry predictions can be achieved (Sec. 5.4 in the main text). Here, we will discuss the impact of incorporating depth supervision signals on the model. We introduce supervision signals for depth in two ways: one through self-supervised and the other through supervision using ground-truth depth.

Following [4, 17], the unsupervised depth loss is:

$$L_d = \beta_1 L_{PC} + \beta_2 L_{SSIM} + \beta_3 L_{Smooth}, \quad (\text{III})$$

where L_{PC} represents the photometric consistency loss.

L_{SSIM} and L_{Smooth} are the structured similarity loss and depth smoothness loss, respectively. β_1 , β_2 , and β_3 are set to 12, 6, and 0.18 in our implementation, respectively. Refer to [4, 17] for more details. L_d is used to supervise the final depth, *i.e.*, d_f (Sec. 4.3 in the main text). Since the DTU dataset provides ground-truth depth, another approach is to utilize the ground-truth depth for supervision. The depth loss is as follows:

$$L_d = \xi(d_f, d_{gt}) \quad (\text{IV})$$

where d_f and d_{gt} represent the final predicted and ground-truth depth, respectively. ξ denotes a loss function. Following [14], ξ is instantiated as the Smooth L1 loss [13]. The quantitative results are presented in the Table VI. The performance of the three strategies in the table is comparable, indicating that supervising only the RGB images is sufficient, and there is no need for additional introduction of depth supervision signals.

More Comprehensive Depth Analysis. As shown in Fig. 3 in the main text, our pipeline first infers the geometry from the cost volume, re-samples 3D points around objects’ surfaces, and finally encodes 3D descriptors for rendering. We can obtain two depths: one inferred from the cost volume and the other obtained through volume rendering, which is the final depth. Here, we report the depth obtained from the cost volume and the final depth as shown in Table VII. Compared to the baseline, our method performs better on both depth metrics. Thanks to Adaptive Cost Aggregation (ACA), the depth obtained from the cost volume has been significantly improved. Based on this, as the Spatial-View Aggregator (SVA) encodes 3D-aware descriptors, the final depth has also been further improved. In addition, the well-designed decoding approach, *i.e.*, CAF, greatly facilitates the depth prediction of the model (Sec. 5.5 in the main text).

Visualization of ACA. Previous approaches using variance struggle to encode efficient cost information for challenging areas, such as the occluded areas marked in the black box in Fig. II. Our proposed ACA module learns an

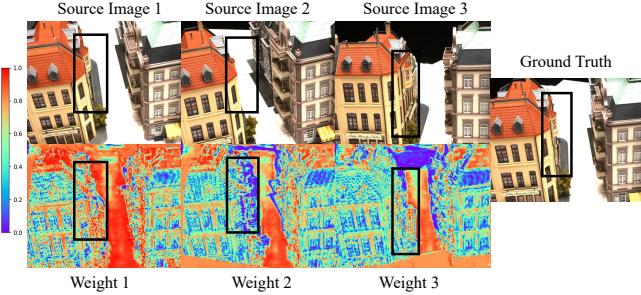


Figure II. **Visualization of ACA.**

adaptive weight for different views to encode accurate geometric information. As illustrated in Fig. II, the weights learned by ACA amplify the contribution of consistent pixel pairs, such as the visible areas in source image 1 and 3, while suppressing inconsistent ones, as shown in the occluded areas in the source image 2.

Different ACA networks. The primary challenge of applying ACA to the NVS task is the unavailability of the target view, which we addressed by adopting a coarse-to-fine framework. In the main text, the weight learning network utilized in ACA is illustrated in Table I, following the MVS method, *i.e.*, AA-RMVSNet [35]. Moreover, other networks can also be embedded into our coarse-to-fine framework to learn inter-view weights. Here, we adopt another MVS method, *i.e.*, MVSTER [33], to learn adaptive weights. The result on the DTU test set is 29.31/0.969/0.064 (PSNR/SSIM/LPIPS), which is comparable with the result obtained using [35]. In summary, our main contribution is to propose an approach for applying ACA to the NVS task, without specifying a particular network for learning weights.

Analysis of SVA. Previous approaches directly uses a pooling network to aggregate multi-view 2D features for encoding 3D descriptors, which are not spatially context-aware, leading to discontinuities in the decoded depth map and rendered view (see Fig. III (a)). To address this issue, convolutional networks can be used to impose spatial constraints on adjacent descriptors. However, due to the smooth nature of convolution, some high-frequency details may be lost. Since detailed information comes from the multi-view features, we employ a divide-and-conquer approach to aggregate descriptors. Firstly, we employ a 3D U-Net to aggregate spatial context and obtain smooth descriptors. Despite resolving the issue of discontinuities, an unsharpened object edge occurs (Fig. III (b)). Secondly, we propose using smoothed features as queries, with multi-view features serving as keys and values. Applying the attention mechanism allows us to gather high-frequency details adaptively.

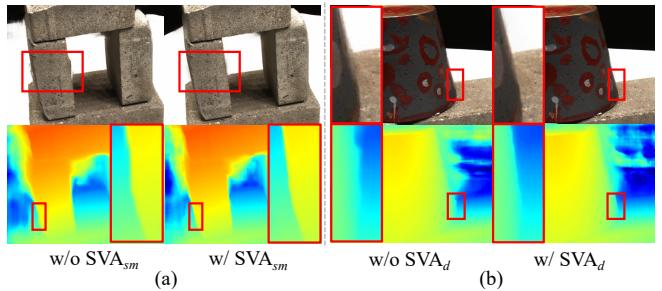


Figure III. **Visualization of SVA.** SVA_{sm} and SVA_d represent ϕ_{sm} and ϕ_d of SVA, respectively.

Method	Reference view			Novel view		
	Abs err ↓	Acc(2)↑	Acc(10)↑	Abs err ↓	Acc(2)↑	Acc(10)↑
Baseline-mvs	3.71	0.815	0.942	4.49	0.778	0.928
Baseline-final	3.58	0.842	0.944	4.32	0.800	0.928
Ours-mvs	2.79	0.836	0.965	3.15	0.816	0.958
Ours-final	2.47	0.900	0.971	2.83	0.879	0.961

Table VII. **More comprehensive depth metrics.** “-mvs” represents the depth obtained from the cost volume and “-final” represents the final depth obtained through volume rendering.

Approach	PSNR↑	SSIM↑	LPIPS↓
Regression	27.32	0.945	0.119
Blending	28.40	0.962	0.091
Overall	29.36	0.969	0.064

Table VIII. **Quantitative results for intermediate results.** Overall represents the fused views.

This practice results in continuities and sharp boundaries in both rendered views and depth maps.

III. More Qualitative Results

Qualitative Results under the Generalization Setting. As shown in Fig. IV, V, and VI, we present the qualitative comparison of rendering quality on the DTU [1], NeRF Synthetic [25], and Real Forward-facing [24] datasets, respectively. Our method can synthesize views with higher fidelity, especially in challenging areas. For example, in occluded regions and geometrically complex scenes, our method can reconstruct more details while exhibiting fewer artifacts at objects’ edges and in reflective areas.

Qualitative Results under the Per-scene Optimization Setting. Benefiting from the strong initialization of our generalizable model, excellent performance can be achieved within just a short fine-tuning period, such as 15 minutes. As shown in Fig. VII, we present the results after fine-tuning. After per-scene optimization, the model demonstrates enhanced capabilities in handling scene de-

tails, resulting in views with higher fidelity.

Qualitative Comparison of Depth Maps. As shown in Figs. VIII, IX, and X, we present the qualitative comparison of depth maps on the DTU [1], NeRF Synthetic [25], and Real Forward-facing [24] datasets, respectively. The depth maps generated by our method can maintain sharper object edges and preserve more details of scenes, which verifies the strong geometry reasoning capability of our method.

Fusion Weight Visualization. As shown in Fig. XI, we present the fusion weights of the Consistency-aware Fusion (CAF) module. The blending approach generally demonstrates higher confidence in most areas, while the regression approach shows higher confidence in challenging regions such as object boundaries and reflections.

Error Map Visualization. As shown in Fig. XII, we present the error maps obtained by two decoding approaches, as well as the error maps of the fused views. The blending approach tends to exhibit lower errors in most areas, while the regression approach may have lower errors in some regions with reflections and edges. In addition, we also present quantitative results, as shown in Table VIII. The views fused through Consistency-aware Fusion (CAF) integrate the advantages of both intermediate views, achieving a further improvement in quality.

IV. Per-scene Breakdown

As shown in Tables IX, X, XI, and XII, we present the per-scene breakdown results of three datasets (DTU [1], NeRF Synthetic [25], and Real Forward-facing [24]). These results align with the averaged results in the main text.

V. Limitations

Although our approach can achieve high performance for view synthesis, it still has the following limitations. 1) Like many other baselines [5, 32], our method is tailored specifically for static scenes and may not perform optimally when applied directly to dynamic scenes. 2) During per-scene optimization, the training speed and rendering speed of NeRF-based methods, including our method, are time-consuming. We will explore the potential of Gaussian Splatting [16] in generalizable NVS to address this issue in the future.

Scan	#1	#8	#21	#103	#114
Metric	PSNR \uparrow				
PixelNeRF [44]	21.64	23.70	16.04	16.76	18.40
IBRNet [32]	25.97	27.45	20.94	27.91	27.91
MVSNeRF [5]	26.96	27.43	21.55	29.25	27.99
NeuRay [23]	28.59	27.63	23.05	29.71	29.23
ENeRF [20]	28.85	29.05	22.53	30.51	28.86
GNT [30]	27.25	28.12	21.67	28.45	28.01
Ours	30.72	30.87	23.96	31.78	29.84
NeRF _{10.2h} [25]	26.62	28.33	23.24	30.40	26.47
IBRNet _{ft-1.0h} [32]	31.00	32.46	27.88	34.40	31.00
MVSNeRF _{ft-15min} [5]	28.05	28.88	24.87	32.23	28.47
NeuRay _{ft-1.0h} [23]	27.77	25.93	23.40	28.57	29.14
ENeRF _{ft-1.0h} [20]	30.10	30.50	22.46	31.42	29.87
Ours _{ft-15min}	31.54	31.41	24.07	32.97	30.52
Ours _{ft-1.0h}	31.58	31.61	24.07	33.09	30.53
Metric	SSIM \uparrow				
PixelNeRF [44]	0.827	0.829	0.691	0.836	0.763
IBRNet [32]	0.918	0.903	0.873	0.950	0.943
MVSNeRF [5]	0.937	0.922	0.890	0.962	0.949
NeuRay [23]	0.872	0.826	0.830	0.920	0.901
ENeRF [20]	0.958	0.955	0.916	0.968	0.961
GNT [30]	0.922	0.931	0.881	0.942	0.960
Ours	0.971	0.965	0.943	0.974	0.965
NeRF _{10.2h} [25]	0.902	0.876	0.874	0.944	0.913
IBRNet _{ft-1.0h} [32]	0.955	0.945	0.947	0.968	0.964
MVSNeRF _{ft-15min} [5]	0.934	0.900	0.922	0.964	0.945
NeuRay _{ft-1.0h} [23]	0.872	0.751	0.845	0.868	0.900
ENeRF _{ft-1.0h} [20]	0.966	0.959	0.924	0.971	0.965
Ours _{ft-15min}	0.973	0.967	0.945	0.976	0.969
Ours _{ft-1.0h}	0.973	0.967	0.945	0.976	0.969
Metric	LPIPS \downarrow				
PixelNeRF [44]	0.373	0.384	0.407	0.376	0.372
IBRNet [32]	0.190	0.252	0.179	0.195	0.136
MVSNeRF [5]	0.155	0.220	0.166	0.165	0.135
NeuRay [23]	0.157	0.201	0.156	0.140	0.128
ENeRF [20]	0.086	0.119	0.107	0.107	0.076
GNT [30]	0.143	0.210	0.171	0.149	0.139
Ours	0.061	0.088	0.068	0.085	0.065
NeRF _{10.2h} [25]	0.265	0.321	0.246	0.256	0.225
IBRNet _{ft-1.0h} [32]	0.129	0.170	0.104	0.156	0.099
MVSNeRF _{ft-15min} [5]	0.171	0.261	0.142	0.170	0.153
NeuRay _{ft-1.0h} [23]	0.155	0.272	0.142	0.177	0.125
ENeRF _{ft-1.0h} [20]	0.071	0.106	0.097	0.102	0.074
Ours _{ft-15min}	0.057	0.082	0.067	0.080	0.061
Ours _{ft-1.0h}	0.056	0.082	0.066	0.079	0.059

Table IX. Quantitative results of five sample scenes on the DTU test set.

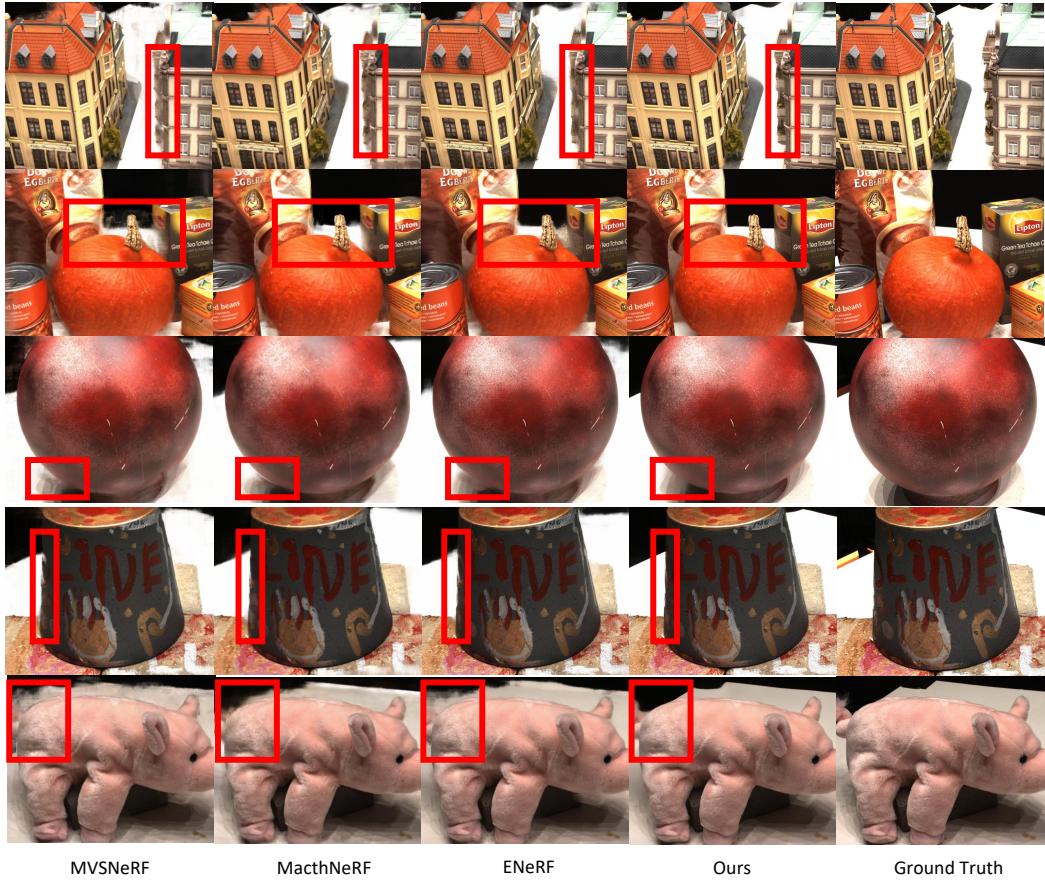


Figure IV. Qualitative comparison of rendering quality with state-of-the-art methods [5, 7, 20] on the DTU dataset under generalization and three views settings.

Scene	#30	#31	#34	#38	#40	#41	#45	#55	#63	#82	#110
Metric	PSNR ↑										
NeuRay [23]	21.10	23.35	24.46	26.01	24.16	27.17	23.75	27.66	23.36	23.84	29.90
ENeRF [20]	29.20	25.13	26.77	28.61	25.67	29.51	24.83	30.26	27.22	26.83	27.97
GNT [30]	27.13	23.54	25.10	27.67	24.48	28.10	24.54	28.86	26.36	26.09	26.93
Ours	30.94	26.95	28.21	29.87	28.62	31.24	26.01	32.46	29.24	29.78	29.30
Metric	SSIM ↑										
NeuRay [23]	0.916	0.851	0.767	0.800	0.812	0.872	0.878	0.870	0.927	0.919	0.927
ENeRF [20]	0.981	0.937	0.934	0.946	0.947	0.960	0.948	0.973	0.978	0.971	0.974
GNT [30]	0.954	0.907	0.880	0.921	0.893	0.908	0.918	0.934	0.938	0.949	0.930
Ours	0.986	0.956	0.954	0.961	0.966	0.972	0.963	0.983	0.984	0.980	0.980
Metric	LPIPS ↓										
NeuRay [23]	0.141	0.161	0.234	0.225	0.209	0.172	0.121	0.163	0.104	0.119	0.116
ENeRF [20]	0.052	0.108	0.117	0.118	0.120	0.091	0.077	0.069	0.048	0.066	0.069
GNT [30]	0.110	0.172	0.201	0.231	0.116	0.168	0.134	0.155	0.127	0.138	0.127
Ours	0.039	0.075	0.085	0.082	0.082	0.065	0.051	0.045	0.032	0.044	0.052

Table X. Quantitative results of other eleven scenes on the DTU test set.

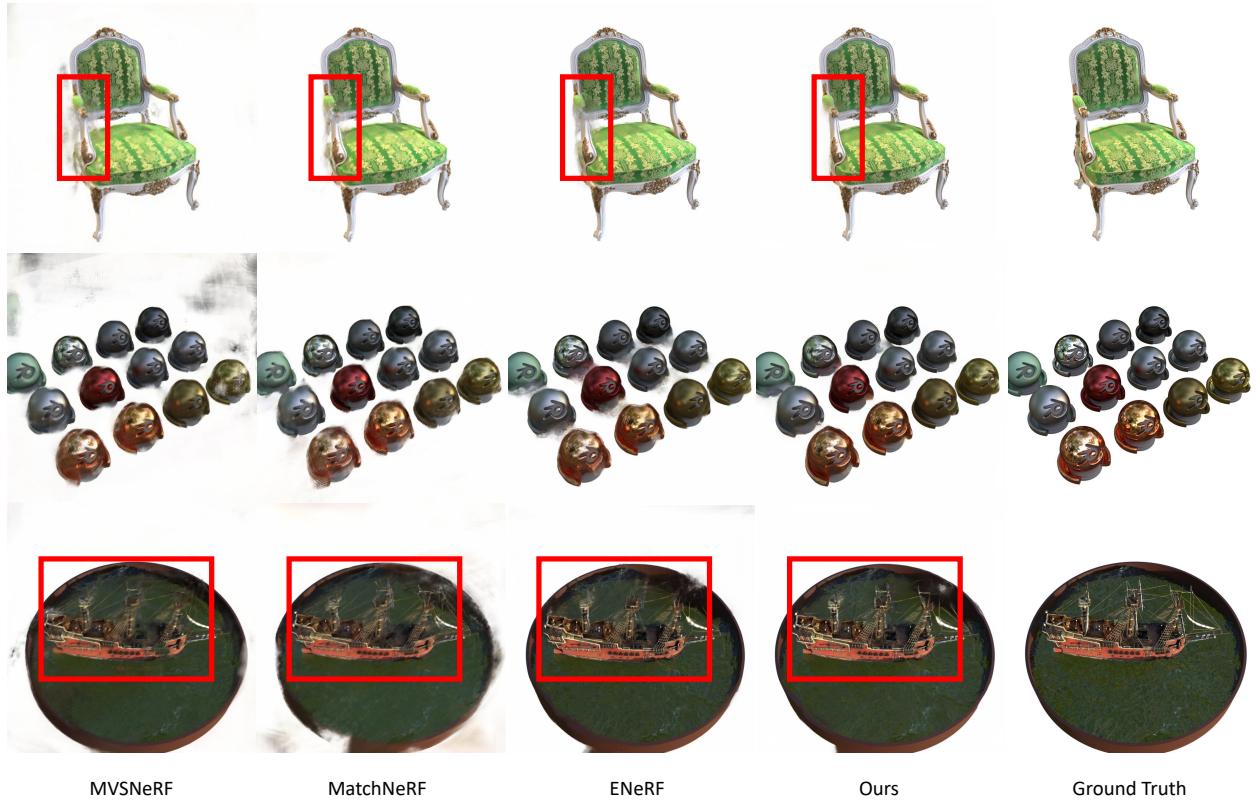


Figure V. Qualitative comparison of rendering quality with state-of-the-art methods [5, 7, 20] on the NeRF Synthetic dataset under generalization and three views settings.



Figure VI. Qualitative comparison of rendering quality with state-of-the-art methods [5, 7, 20] on the Real Forward-facing dataset under generalization and three views settings.

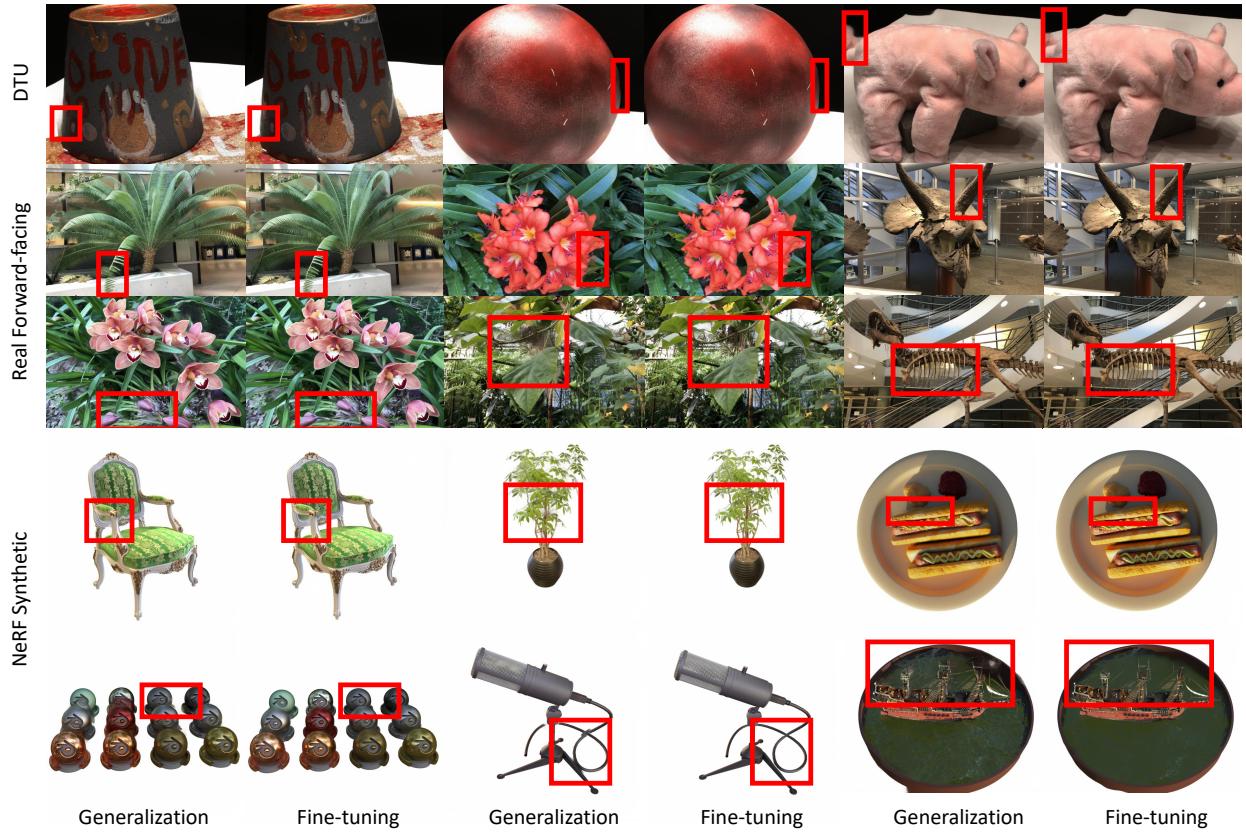


Figure VII. Qualitative comparison of results before and after fine-tuning on the DTU [1], Real Forward-facing [24], and NeRF Synthetic [25] datasets.

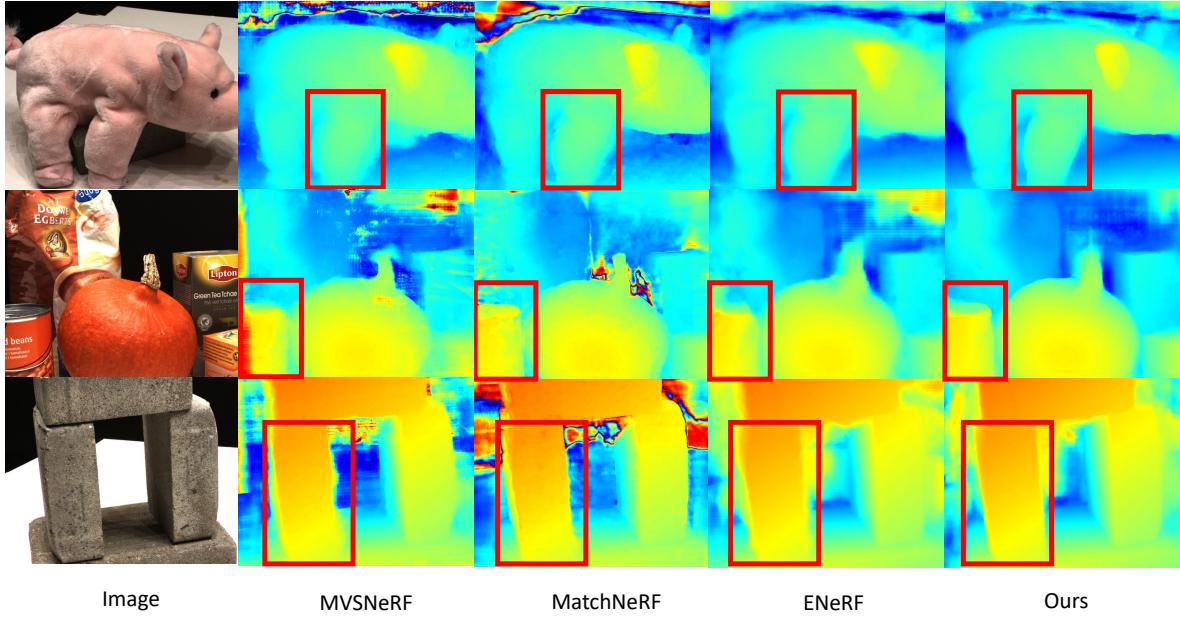


Figure VIII. Qualitative comparison of depth maps with state-of-the-art methods [5, 7, 20] on the DTU dataset.

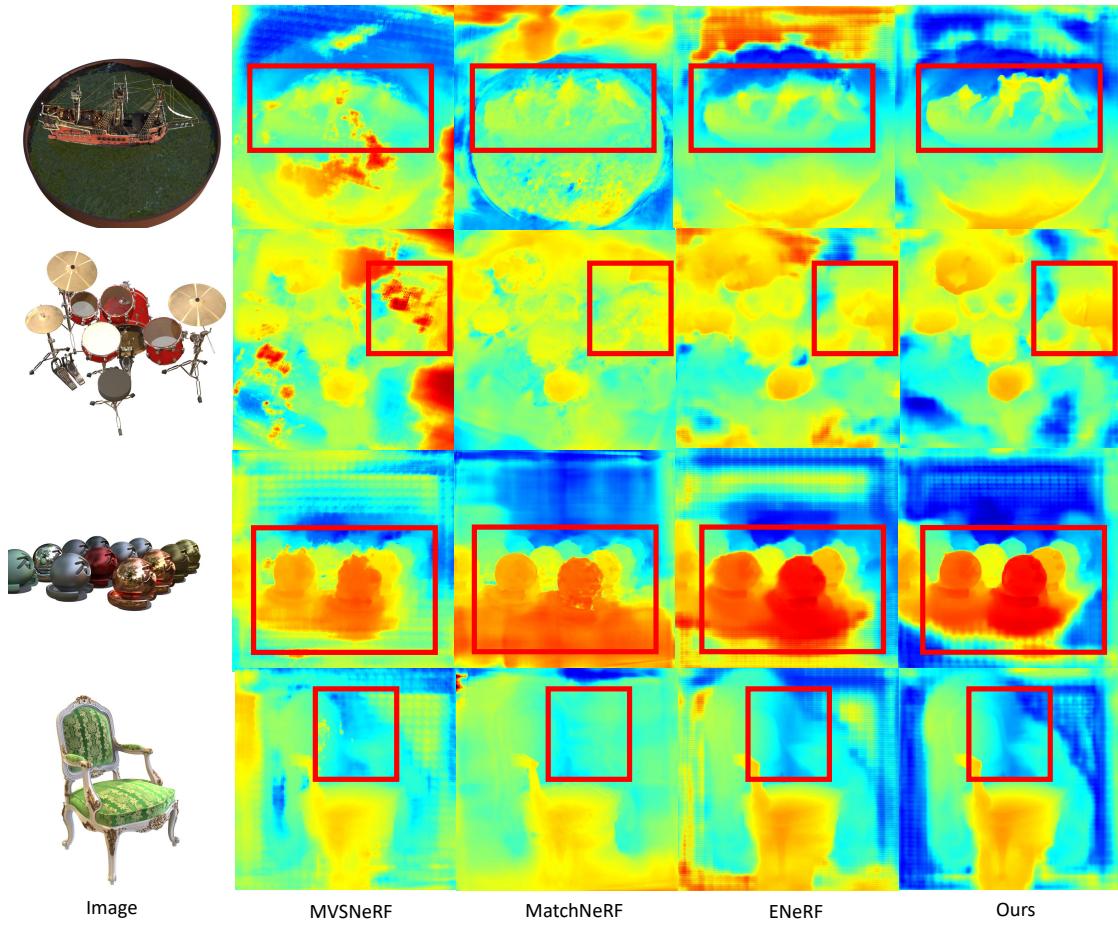


Figure IX. Qualitative comparison of depth maps with state-of-the-art methods [5, 7, 20] on the NeRF Synthetic dataset.

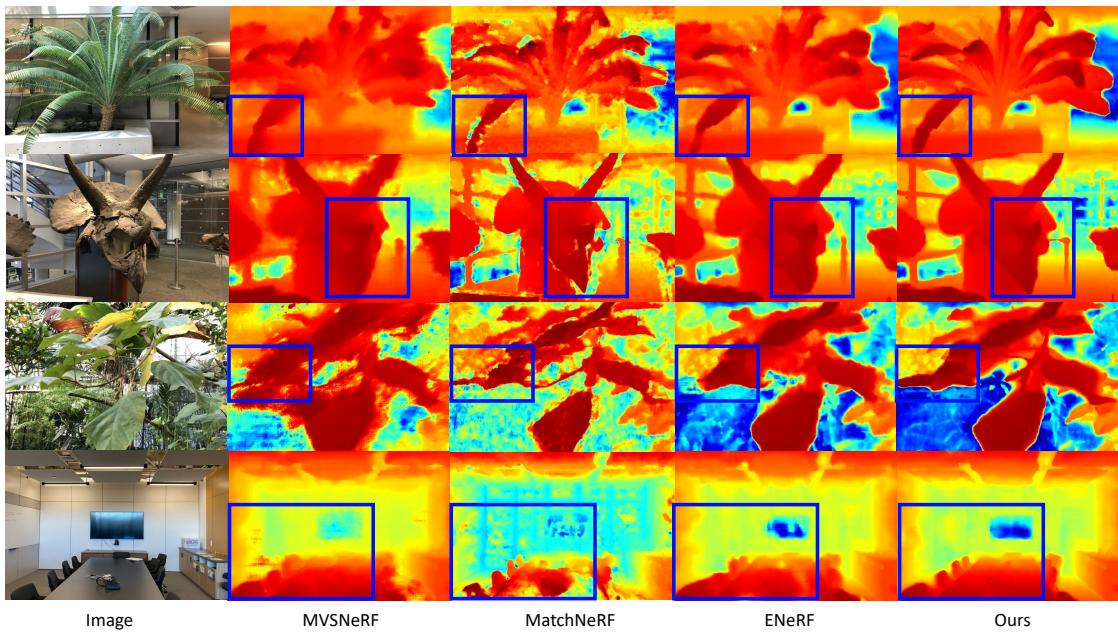


Figure X. Qualitative comparison of depth maps with state-of-the-art methods [5, 7, 20] on the Real Forward-facing dataset.

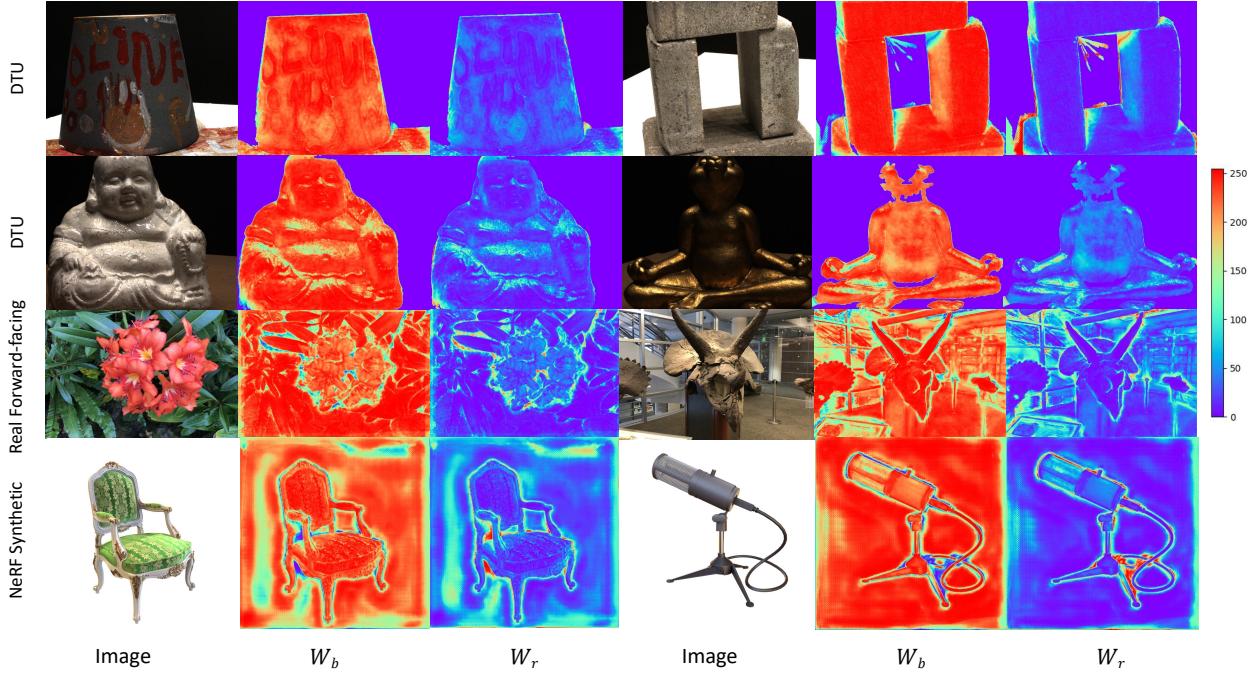


Figure XI. **Visualization of Fusion Weights.** W_b and W_r represent the weight maps of the blending approach and regression approach, respectively.

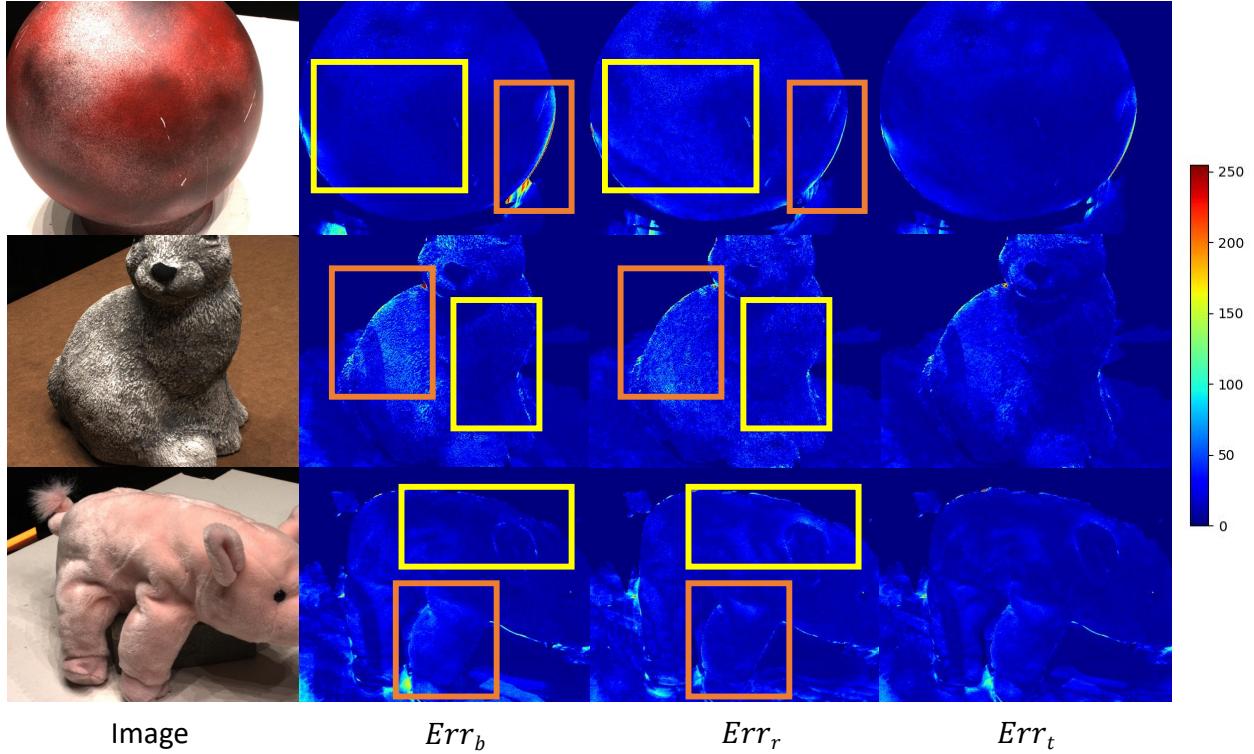


Figure XII. **Visualization of error maps.** Err_b and Err_r represent the error maps of the views obtained through the blending approach and the regression approach, respectively. Err_t is the error map of the final fused target view. The yellow boxes indicate that the blending approach outperforms the regression approach, while the orange boxes indicate regions where the regression approach outperforms the blending approach.

Scene	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
Metric	PSNR \uparrow							
PixelNeRF [44]	7.18	8.15	6.61	6.80	7.74	7.61	7.71	7.30
IBRNet [32]	24.20	18.63	21.59	27.70	22.01	20.91	22.10	22.36
MVSNeRF [5]	23.35	20.71	21.98	28.44	23.18	20.05	22.62	23.35
NeuRay [23]	27.27	21.09	24.09	30.50	24.38	21.90	26.08	21.30
ENeRF [20]	28.29	21.71	23.83	34.20	24.97	24.01	26.62	25.73
GNT [30]	27.98	20.27	26.86	29.34	23.17	30.75	23.19	24.86
Ours	28.87	22.33	24.55	34.96	24.90	26.08	27.98	26.22
NeRF [25]	31.07	25.46	29.73	34.63	32.66	30.22	31.81	29.49
IBRNet _{ft-1.0h} [32]	28.18	21.93	25.01	31.48	25.34	24.27	27.29	21.48
MVSNeRF _{ft-15min} [5]	26.80	22.48	26.24	32.65	26.62	25.28	29.78	26.73
NeuRay _{ft-1.0h} [23]	27.37	21.69	23.45	32.26	26.87	23.03	28.12	24.49
ENeRF _{ft-1.0h} [20]	28.94	25.33	24.71	35.63	25.39	24.98	29.25	26.36
Ours _{ft-15min}	30.93	23.29	25.46	36.28	26.96	26.91	31.20	27.51
Ours _{ft-1.0h}	31.07	23.38	25.62	36.73	27.24	27.05	31.49	27.87
Metric	SSIM \uparrow							
PixelNeRF [44]	0.624	0.670	0.669	0.669	0.671	0.644	0.729	0.584
IBRNet [32]	0.888	0.836	0.881	0.923	0.874	0.872	0.927	0.794
MVSNeRF [5]	0.876	0.886	0.898	0.962	0.902	0.893	0.923	0.886
NeuRay [23]	0.912	0.856	0.901	0.953	0.899	0.881	0.952	0.779
ENeRF [20]	0.965	0.918	0.932	0.981	0.948	0.937	0.969	0.891
GNT [30]	0.935	0.891	0.941	0.940	0.897	0.974	0.791	0.874
Ours	0.971	0.931	0.939	0.983	0.956	0.953	0.980	0.899
NeRF [25]	0.971	0.943	0.969	0.980	0.975	0.968	0.981	0.908
IBRNet _{ft-1.0h} [32]	0.955	0.913	0.940	0.978	0.940	0.937	0.974	0.877
MVSNeRF _{ft-15min} [5]	0.934	0.898	0.944	0.971	0.924	0.927	0.970	0.879
NeuRay _{ft-1.0h} [23]	0.920	0.869	0.895	0.949	0.912	0.880	0.954	0.788
ENeRF _{ft-1.0h} [20]	0.971	0.960	0.939	0.985	0.949	0.947	0.985	0.893
Ours _{ft-15min}	0.978	0.936	0.946	0.987	0.959	0.958	0.987	0.909
Ours _{ft-1.0h}	0.979	0.938	0.947	0.988	0.963	0.960	0.989	0.912
Metric	LPIPS \downarrow							
PixelNeRF [44]	0.386	0.421	0.335	0.433	0.427	0.432	0.329	0.526
IBRNet [32]	0.144	0.241	0.159	0.175	0.202	0.164	0.103	0.369
MVSNeRF [5]	0.282	0.187	0.211	0.173	0.204	0.216	0.177	0.244
NeuRay [23]	0.146	0.211	0.184	0.113	0.126	0.165	0.104	0.256
ENeRF [20]	0.055	0.110	0.076	0.059	0.075	0.084	0.039	0.183
GNT [30]	0.065	0.116	0.063	0.095	0.112	0.025	0.243	0.115
Ours	0.035	0.089	0.064	0.060	0.064	0.054	0.021	0.175
NeRF [25]	0.055	0.101	0.047	0.089	0.054	0.105	0.033	0.263
IBRNet _{ft-1.0h} [32]	0.079	0.133	0.082	0.093	0.105	0.093	0.040	0.257
MVSNeRF _{ft-15min} [5]	0.129	0.197	0.171	0.094	0.176	0.167	0.117	0.294
NeuRay _{ft-1.0h} [23]	0.074	0.136	0.105	0.072	0.091	0.137	0.072	0.230
ENeRF _{ft-1.0h} [20]	0.030	0.045	0.071	0.028	0.070	0.059	0.017	0.183
Ours _{ft-15min}	0.024	0.080	0.059	0.028	0.052	0.044	0.015	0.181
Ours _{ft-1.0h}	0.023	0.076	0.058	0.026	0.050	0.043	0.012	0.179

Table XI. Quantitative results on the NeRF Synthetic dataset.

Scene	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	Trex
Metric	PSNR \uparrow							
PixelNeRF [44]	12.40	10.00	14.07	11.07	9.85	9.62	11.75	10.55
IBRNet [32]	20.83	22.38	27.67	22.06	18.75	15.29	27.26	20.06
MVSNeRF [5]	21.15	24.74	26.03	23.57	17.51	17.85	26.95	23.20
NeuRay [23]	21.17	26.29	27.98	23.91	19.51	18.81	28.92	20.55
ENeRF [20]	21.92	24.28	30.43	24.49	19.01	17.94	29.75	21.21
GNT [30]	22.21	23.56	29.16	22.80	19.18	17.43	29.35	20.15
Ours	22.53	25.73	30.54	25.41	19.46	18.76	29.79	22.05
NeRF _{ft-10.2h} [25]	23.87	26.84	31.37	25.96	21.21	19.81	33.54	25.19
IBRNet _{ft-1.0h} [32]	22.64	26.55	30.34	25.01	22.07	19.01	31.05	22.34
MVSNeRF _{ft-15min} [5]	23.10	27.23	30.43	26.35	21.54	20.51	30.12	24.32
NeuRay _{ft-1.0h} [23]	22.57	25.98	29.17	25.40	20.74	20.36	27.06	23.43
ENeRF _{ft-1.0h} [20]	22.08	27.74	29.58	25.50	21.26	19.50	30.07	23.39
Ours _{ft-15min}	23.67	27.89	31.63	27.47	22.41	20.63	33.69	25.53
Ours _{ft-1.0h}	23.82	28.09	31.63	27.66	22.59	20.80	33.97	25.54
Metric	SSIM \uparrow							
PixelNeRF [44]	0.531	0.433	0.674	0.516	0.268	0.317	0.691	0.458
IBRNet [32]	0.710	0.854	0.894	0.840	0.705	0.571	0.950	0.768
MVSNeRF [5]	0.638	0.888	0.872	0.868	0.667	0.657	0.951	0.868
NeuRay [23]	0.632	0.823	0.829	0.779	0.668	0.590	0.916	0.718
ENeRF [20]	0.774	0.893	0.948	0.905	0.744	0.681	0.971	0.826
GNT [30]	0.736	0.791	0.867	0.820	0.650	0.538	0.945	0.744
Ours	0.798	0.912	0.947	0.924	0.773	0.725	0.975	0.848
NeRF _{ft-10.2h} [25]	0.828	0.897	0.945	0.900	0.792	0.721	0.978	0.899
IBRNet _{ft-1.0h} [32]	0.774	0.909	0.937	0.904	0.843	0.705	0.972	0.842
MVSNeRF _{ft-15min} [5]	0.795	0.912	0.943	0.917	0.826	0.732	0.966	0.895
NeuRay _{ft-1.0h} [23]	0.687	0.807	0.854	0.822	0.714	0.657	0.909	0.799
ENeRF _{ft-1.0h} [20]	0.770	0.923	0.940	0.904	0.827	0.725	0.965	0.869
Ours _{ft-15min}	0.825	0.930	0.963	0.948	0.869	0.785	0.986	0.915
Ours _{ft-1.0h}	0.829	0.932	0.963	0.949	0.873	0.791	0.987	0.915
Metric	LPIPS \downarrow							
PixelNeRF [44]	0.650	0.708	0.608	0.705	0.695	0.721	0.611	0.667
IBRNet [32]	0.349	0.224	0.196	0.285	0.292	0.413	0.161	0.314
MVSNeRF [5]	0.238	0.196	0.208	0.237	0.313	0.274	0.172	0.184
NeuRay [23]	0.257	0.162	0.163	0.225	0.253	0.283	0.136	0.254
ENeRF [20]	0.224	0.164	0.092	0.161	0.216	0.289	0.120	0.192
GNT [30]	0.223	0.203	0.157	0.208	0.255	0.341	0.103	0.275
Ours	0.185	0.126	0.101	0.130	0.188	0.243	0.150	0.176
NeRF _{ft-10.2h} [25]	0.291	0.176	0.147	0.247	0.301	0.321	0.157	0.245
IBRNet _{ft-1.0h} [32]	0.266	0.146	0.133	0.190	0.180	0.286	0.089	0.222
MVSNeRF _{ft-15min} [5]	0.253	0.143	0.134	0.188	0.222	0.258	0.149	0.187
NeuRay _{ft-1.0h} [23]	0.229	0.173	0.162	0.209	0.243	0.257	0.160	0.208
ENeRF _{ft-1.0h} [20]	0.197	0.121	0.101	0.155	0.168	0.247	0.113	0.169
Ours _{ft-15min}	0.156	0.090	0.069	0.093	0.123	0.192	0.052	0.106
Ours _{ft-1.0h}	0.151	0.087	0.068	0.089	0.116	0.182	0.051	0.103

Table XII. Quantitative results on the Real Forward-facing dataset.