MuGS: Multi-Baseline Generalizable Gaussian Splatting Reconstruction

Yaopeng Lou Liao Shen Tianqi Liu Jiaqi Li
Zihao Huang Huiqiang Sun Zhiguo Cao†
School of AIA, Huazhong University of Science and Technology

Abstract

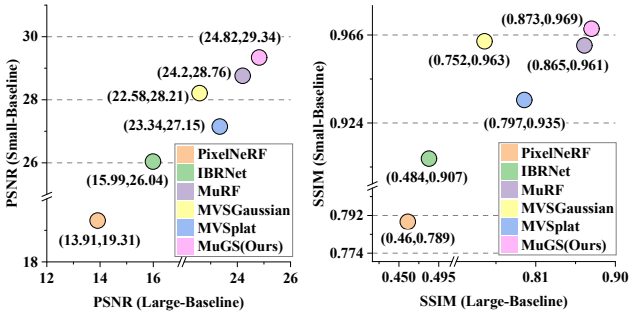
We present Multi-Baseline Gaussian Splatting (MuGS), a generalized feed-forward approach for novel view synthesis that effectively handles diverse baseline settings, including sparse input views with both small and large baselines. Specifically, we integrate features from Multi-View Stereo (MVS) and Monocular Depth Estimation (MDE) to enhance feature representations for generalizable reconstruction. Next, We propose a projection-and-sampling mechanism for deep depth fusion, which constructs a fine probability volume to guide the regression of the feature map. Furthermore, We introduce a reference-view loss to improve geometry and optimization efficiency. We leverage 3D Gaussian representations to accelerate training and inference time while enhancing rendering quality. MuGS achieves state-of-the-art performance across multiple baseline settings and diverse scenarios ranging from simple objects (DTU) to complex indoor and outdoor scenes (RealEstate10K). We also demonstrate promising zero-shot performance on the LLFF and Mip-NeRF 360 datasets. Code is available at https://github.com/EuclidLou/MuGS.

1. Introduction

Novel view synthesis (NVS) represents a fundamental and practical challenge in computer vision and graphics. Neural Radiance Fields (NeRF) [25], which encode scenes as implicit radiance fields, have demonstrated remarkable success. However, NeRF is computationally expensive, as it requires querying dense points for rendering. Recently, 3D Gaussian Splatting (3D-GS) [19] has emerged as an efficient alternative, leveraging anisotropic 3D Gaussians to represent scenes explicitly. This approach facilitates real-time, high-quality rendering through a differentiable tile-based rasterizer. Despite these advances, 3D-GS requires per-scene optimization, which remains time-consuming, limiting its practical applicability.

To tackle this issue, several generalizable methods [3, 6, 7, 24, 32, 38] achieve notable advancements in rendering high-quality novel views from unseen scenes. These methods accomplish this by introducing explicit geometry constraints and leveraging data-driven approaches rather than overfitting to a specific scene. Based on input view overlap, these methods can be categorized into two classes: small-baseline tasks, which handle images with large overlap, and large-baseline tasks, which operate on images having small overlap. However, existing methods tend to specialize in small-baseline or large-baseline settings, struggling to generalize across different baselines.

A key challenge in practical scenarios is that the baseline of input views are either small or large, which limits the generalization of baseline-specific methods. To address this issue, we propose the first 3D Gaussian splatting method designed for rendering novel views from sparse inputs across varying baselines, as shown in Fig. 1.



(a) MuGS achieves the best performance in both large- and small-baseline.

(b) MuGS can generalize across different baselines.

Figure 1. **MuGS supports multiple baseline settings**. MuGS is the first Gaussian-based method designed for different baselines. Our method outperforms the previous state-of-the-art methods.

†Corresponding author.

Generalizable Gaussian models typically extract feature volumes and depth probability volumes using multi-view stereo (MVS) techniques [15, 47], then regress depth and other Gaussian parameters. Accurate depth estimation is essential for retrieving reliable information, yet challenges arise with baseline mismatches. Small-baseline models tested on large-baseline datasets suffer from depth errors due to occlusions and insufficient overlap, leading to distortions. Conversely, large-baseline models evaluated on small-baseline datasets struggle with the lack of matching cues, resulting in inaccurate depth estimation. This inaccuracy causes inconsistent Gaussian placements and blurred rendered images.

Our insight is that accurate depth guidance can address common challenges in both large-baseline and small-baseline methods, while unifying them into a more generalized model. However, achieving precise depth under sparse, multi-baseline inputs is challenging due to three key obstacles: **First**, the preferred depth estimation strategy differs from the baseline. For instance, in a two-view setup, the smallest baseline corresponds to a binocular scenario, where the two views can be treated as adjacent video frames and processed with matching-based MVS techniques. The largest baseline, on the other hand, may imply no overlap between the two views, meaning depth information must rely on monocular depth methods. Resolving both types of problems within a single model is challenging. **Second**, unlike typical MVS tasks, we deal with sparse inputs where calculating the feature similarity is often infeasible. This limitation arises because at least two valid feature samples are required for variance at each candidate point. However, sparse inputs may lack sufficient overlap or be affected by occlusions, rendering the process ineffective. **Third**, for effective generalization, the model needs to store comprehensive prior knowledge to support inference across diverse baselines. Training on a dataset with a specific baseline while preserving multi-baseline adaptability is a nontrivial challenge, requiring careful optimization to prevent overfitting to a particular baseline configuration.

To address the challenges above, we propose the following solutions. **First**, we introduce a pre-trained monocular depth model [43] to assist MVS, as the former offers more robust and smooth depth features for sparse inputs, whereas the latter typically exhibits large errors in challenging areas, despite performing well in regions with sufficient context. **Second**, for each depth candidate in MVS, we compute both projected depth and sampled depth, which represent the spatial position and the expected depth of each point, respectively. A 3D U-net is then employed to calculate the consistency between the two depths. **Third**, the consistency information mentioned above is then used as a query in a lightweight attention network, refining the depth probability volume. By prioritizing depth candidates near the surface, the MLP network better utilizes features and colors sampled from each source view, ultimately reducing artifacts and improving rendering quality. Additionally, we propose a reference-view loss for contextual supervision to learn geometric correspondence more effectively.

Our contributions can be summarized as follows:

- We propose MuGS, the first multi-baseline generalizable Gaussian based method that integrates the features of multi-view stereo and monocular models.
- We introduce the projection-sampling depth consistency network to guide the fine-grained probability volume and enhance robustness for challenging sparse inputs.
- We propose a reference-view loss for contextual supervision to improve rendering quality.
- We demonstrate that our method outperforms existing approaches across different baseline datasets and achieves superior performance on zero-shot datasets.

2. Related Work

Multi-Baseline. The idea of "multi-baseline" originates from multi-baseline stereo depth estimation [14, 26, 44, 45], in which several stereo pairs with different baselines are employed to overcome matching ambiguities and enhance accuracy. Recently, MuRF [39] has made significant progress in extending the multi-baseline problem to the NVS task by leveraging a pre-trained multi-view feature encoder to construct target view frustum volume, along with an efficient CNN decoder. This approach can handle both large- and small-baseline problems, even with very sparse inputs. However, as this method relies solely on MVS principles to obtain a density volume, it faces challenges when there is insufficient overlap or occlusion in the views. In such cases, the density along the ray tends to disperse instead of concentrating around the true surface. Moreover, due to the NeRF-like volume rendering approach, noise feature sampled from incorrect depths also contributes to the final output, resulting in blurriness and artifacts. In contrast, our work addresses these challenges from a more fundamental perspective, specifically depth precision. By doing so, we propose a unified solution that effectively handles the shared challenges encountered by both large-baseline and small-baseline methods.

Multi-View Stereo (MVS) aims to recover 3D geometry from multiple views. Traditional MVS methods [12, 13, 29, 30] rely on handcrafted features and similarity metrics, limiting performance, while MVSNet [47] first introduces an end-to-end pipeline that constructs a cost volume to aggregate 2D data in a 3D geometry-aware manner. Following this cost volume-based pipeline, subsequent works make improvements from various aspects, *e.g.* higher memory efficiency with coarse-to-fine architectures [8, 15, 41] or recurrent plane sweeping [41, 48], optimized cost aggregation [34, 37], enhanced feature representations [10, 23],

and improved decoding strategy [27, 49]. As the cost volume encodes the consistency of multi-view features and naturally performs correspondence matching, many feed-forward Gaussian methods [6, 7, 24] follow this spirit to learn better geometry. However, they inherently suffer from the limitation of feature matching in challenging situations like insufficient overlap or occlusion.

Monocular Depth Estimation (MDE). Recently, there has been notable advancement in depth estimation from a single image [2, 18, 28, 42, 51], with current methods delivering impressively accurate edge-aligned results across a wide range of real-world data. However, monocular depth techniques still face challenges with scale ambiguities and are unable to generate depth predictions that are consistent across multiple views, which are essential for tasks like 3D reconstruction [50] and video depth estimation [35]. In this paper, we propose the concepts of projected depth and sampled depth to integrate depth information from both the cost volume and monocular depth. By leveraging the robustness of a pre-trained monocular depth model [43], our approach mitigates the limitations of feature matching-based methods. The very recent work DepthSplat [40] attempts to combine MVS and MDE, focusing on the mutual enhancement of depth estimation and large-baseline view synthesis through a feature-level concatenation. In contrast, our work deeply explores and models the relationship between the two depth cues across different view baselines, enabling view synthesis under varying baselines.

3. Preliminary

Adjusted Multi-View Stereo Pipeline adapts the traditional MVS approach for novel view synthesis (NVS). The process starts by defining multiple fronto-parallel planes in the target view. Then, the feature maps extracted from N input views are warped onto these planes using a differentiable homography, expressed as:

$$H_i(z) = K_t R_t \mathbb{C}(z) R_i^{-1} K_i^{-1}, \tag{1}$$

where $[K_t, R_t]$ and $[K_i, R_i]$ are the camera intrinsics and rotations for the target view and the source view I_i. With $\mathbb{I}$ the identity matrix, z the depth candidate, n the principal axis of the target view camera and $[t_t, t_i]$ the camera translations for the target view and the source view I_i, we can obtain the correction term $\mathbb{C}(z)$ by:

$$\mathbb{C}(z) = \mathbb{I} - \frac{(R_t^{-1}t_t - R_i^{-1}t_i)n^T R_t}{z}. \tag{2}$$

With warped multi-view features $\{f_i\}_{i=1}^N$, we can obtain the cost volume by calculating learnable pair-wise similarity [53]. It can be expressed as:

$$Sim = \sum_{j<k} w_{jk} * cos(f_j, f_k), \quad j,k \in \{1,2,...,N\}, \tag{3}$$

where w_{jk} are learned weights.

For novel view synthesis, it is essential not only to focus on depth but also to recover textures and colors. Thus, the cost volume is augmented by introducing multi-view features which are aggregated through a pooling network [22]. The augmented cost volume is regularized by CNNs to produce the target view frustum volume, from which we obtain the depth probability volume $\mathbb{V}^p$ and other rendering parameters via MLPs. Unlike previous studies [24, 39], which directly use the depth probability volume for rendering, our approach improves both depth and texture recovery, enhancing the final rendering quality.

3D Gaussian Splatting uses anisotropic 3D Gaussians to explicitly represent a 3D scene. Each Gaussian is defined by:

$$G(X) = exp[-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)], \tag{4}$$

where Σ and μ denotes 3D covariance matrix and mean. The covariance matrix Σ is usually decomposed into a scaling matrix S and a rotation matrix R by $\Sigma = RSS^TR^T$, which allows for effective optimization since Σ holds physical meaning only when it is positive semidefinite.

To render a view from 3D Gaussians, the first step is to splat Gaussians from 3D space to a 2D plane, yielding 2D Gaussians, which covariance matrix is calculated by $\Sigma' = JW\Sigma W^T J^T$. The J is the Jacobian matrix which represents the affine approximation of the projective transformation, and the W is the view transformation matrix. Next, the color of each pixel can be rendered by alpha-blending:

$$C = \sum_j c_j \alpha_j \prod_{k=1}^{j-1}(1-\alpha_k), \tag{5}$$

where the color c_j at depth-wise position j is defined by spherical harmonics (SH) coefficients and the density α_j equals to the multiplication of 2D Gaussians and a learnable point-wise opacity.

4. Methodology

4.1. Overview

Given a set of input views $\{I_i\}_{i=1}^N$, our objective is to render target views through a feed-forward, generalized process without per-scene optimization. The overview of our proposed framework is depicted in Fig. 2. Our method consists of two primary branches: the MVS branch and the MDE branch. In the MVS branch, a multi-view feature encoder is applied to the input views, constructing the target view frustum volume to regress a coarse depth probability volume. Meanwhile, the MDE branch generates monocular feature maps and predicts monocular depth maps for input views. Subsequently, using a projection-and-sampling approach,

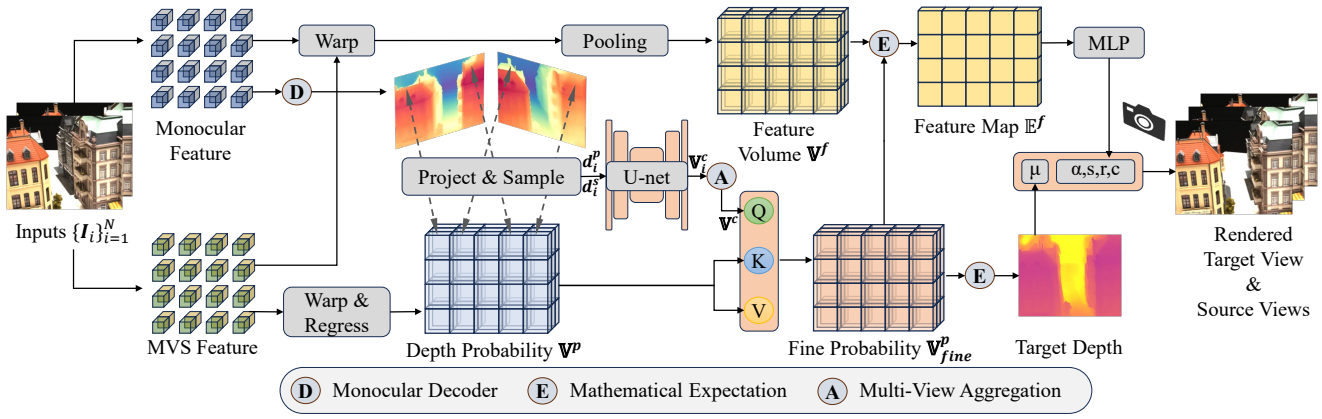


Figure 2. **Overview.** Given input images $\{I_i\}_{i=1}^N$, we first extract multi-view image features from both the monocular encoder and MVS's cross-view encoder. The MVS features are used to regress a target view depth probability volume, while monocular features are decoded into source view depth maps $\{\mathcal{D}_i\}_{i=1}^N$. By projecting the points of the depth probability volume to and sampling from the depth map $\mathcal{D}_i$, we obtain d_i^p and d_i^s, which are then fed into a U-net to query for a refined probability volume $\mathbb{V}^p_{fine}$. Besides, both features are concatenated to construct the feature volume. Next, we calculate the expected value of depth and feature using $\mathbb{V}^p_{fine}$, which produces the target depth and feature map. These are used to predict Gaussian parameters. Finally, the target view image and source reference views are rendered, which contribute to the total loss together.

monocular depth information from multiple views is integrated with the depth probability volume, refining the predicted target view depth. On the other hand, MVS features are enhances by monocular features to obtain the feature volume, which is then retrieved for the feature map by the refined probability volume, enabling the MLPs to regress Gaussian parameters while reducing noise. This pipeline is executed hierarchically, generating depth maps and rendered views in a coarse-to-fine manner.

4.2. MDE-based Depth Refining

Fusion Strategy. The most straightforward approach to fusing monocular depth estimation (MDE) and multi-view stereo (MVS) depth would be to merge both target view aligned monocular depth map and target view MVS depth probability volume together through a neural network [21, 46]. However, this method proves inadequate for our novel view synthesis task, as it faces a fundamental limitation: while MVS can estimate the depth map for the target view, the corresponding monocular depth map is inherently unavailable since the target view itself is what we aim to generate. Therefore, we propose a fusion strategy based on projection and sampling manner. Specifically, after obtaining the depth probability volume, we first estimate the monocular depth map for each input view, *i.e.*, $\{\mathcal{D}_i\}_{i=1}^N$. We then calculate projected and sampled depth information. Given the camera intrinsic K_i, rotation R_i and translation t_i of the input view I_i, each depth candidate point $\mathcal{P}$ on the fronto-parallel plane can be projected to input view I_i by:

$$\mathcal{P}_i * d_i^p = K_i(R_i\mathcal{P} + t_i), \tag{6}$$

where we can simultaneously obtain both the the projected depth d_i^p, which is the distance from the point to the camera plane, and the projection coordinates $\mathcal{P}_i$ in the camera coordinate system. The sampled depth d_i^s is obtained by performing grid sampling on the monocular depth map $\mathcal{D}_i$ according to the projection coordinates $\mathcal{P}_i$.

For candidate points near the object's surface, the projected depth and sampled depth exhibit a high degree of consistency, as both represent the same spatial location. In contrast, for candidates far from the object surface, the two depths become inconsistent. This property enables us to infer the authenticity of a candidate point based on the consistency between the two depths. To leverage this insight, we subsequently employ a four-layer 3D U-Net $\mathcal{U}$ for each view i to regress consistency cue $\mathbb{V}_i^c$ from the volume composed of d_i^p and d_i^s. Additionally, to mitigate the inherent scale ambiguity in monocular depth estimation, we introduce the depth ratio d_i^s/d_i^p as a third input channel. This normalization helps reduce discrepancies caused by varying depth scales across different views.

$$\mathbb{V}_i^c = \mathcal{U}(d_i^p, d_i^s, d_i^s/d_i^p). \tag{7}$$

Each consistency cue $\mathbb{V}_i^c$ indicates the consistency between the MVS's target view depth probability volume and the MDE's source view depth map $\mathcal{D}_i$. This operation is executed on every input view, yielding the set of the consistency cues $\{\mathbb{V}_i^c\}_{i=1}^N$.

Probability Refinement. With the multi-view consistency cues $\{\mathbb{V}_i^c\}_{i=1}^N$, we first aggregate them based on the visibility, which can be expressed as:

$$\mathbb{V}^c = \sum_{i=1}^N w_i \mathbb{V}_i^c, \tag{8}$$

where w_i are learnable weights and the outcome $\mathbb{V}^c$ serves as the overall consistency cue. This aggregation allows the

model to focus on those informative cues and discard the noisy ones caused by occlusion.

To refine the depth probability volume, we not only consider the consistency between projected and sampled depths, but also take into account the original MVS estimation results, since the latter cab be credible if sufficient context is given. Therefore, we use a lightweight attention network to integrate consistency cues with MVS results, which helps the network to balance the information. Specifically, we take consistency cue $\mathbb{V}^c$ as the query, the depth probability volume $\mathbb{V}^p$ as both the key and value, to conduct a depth-wise attention. The output result has the dimension of depth probability and is added with the residual of the original volume, which can be expressed as:

$$\mathbb{V}^p_{fine} = Attention(\mathbb{V}^c, \mathbb{V}^p, \mathbb{V}^p) + \mathbb{V}^p \tag{9}$$

4.3. Gaussian Parameter Prediction

Feature Enhancement. The import of a monocular model provides not only depth information but also well-encoded features, which carry informative inductive bias. To this end, we leverage the power of the monocular features to assist the prediction of Gaussian parameters. Specifically, both the monocular feature maps and the MVS feature maps are warped to the target view fronto-parallel planes according to Eq. (1) and concatenated. Subsequently, utilizing a pooling network [22], the features from different views are aggregated to construct the feature volume $\mathbb{V}^f$.

Gaussian Construction. With the refined depth probability volume, we can first compute the expected values of depths, yielding the target view depth map. The depths are then used to unproject each pixel to obtain the positions μ of the 3D Gaussians. Next, we regress from feature volume $\mathbb{V}^f$ to obtain the remaining parameters for each Gaussian, namely the color c, opacity α, scale s, and rotation r. Unlike NeRF-based methods that require the whole volume to be processed and prone to vaporific noise in the result, our method is constructed on a 2D feature map and focuses more on features describing the actual object's surface. Specifically, we compute the expectation of the feature volume along the depth channel using the refined depth probability, and the resulting features $\mathbb{E}^f$ are used as the input to the MLP ϕ to calculate:

$$\begin{aligned} &c = Sigmoid(\phi_c(\mathbb{E}^f)), \alpha = Sigmoid(\phi_\alpha(\mathbb{E}^f)), \\ &s = Softplus(\phi_s(\mathbb{E}^f)), r = Norm(\phi_r(\mathbb{E}^f)). \end{aligned} \tag{10}$$

With the set $\{\mu, s, r, \alpha, c\}$, pixel-aligned Gaussians are represented and alpha-blending can be performed to obtain the color of each pixel.

4.4. Multi-View Gaussian Splatting

Due to the characteristic of target view pixel-aligned Gaussians, using only target view RGB supervision during training limits the ability to reflect the spatial information of the Gaussians in the results, ultimately hindering the achievement of accurate geometry. The explicit Gaussian representation allows us to quickly render not only the novel view but also source views without additionally constructing source view volumes. Therefore, we incorporate supervision from the source views to improve spatial accuracy. Specifically, after obtaining the parameter set $\{\mu, s, r, \alpha, c\}$, we input the camera parameters of both the source and target views, *i.e.* $\{[K_i, R_i]\}_{i=1}^N$ and $[K_t, R_t]$ into the Gaussian rasterizer to generate multiple rendered views in sequence for optimization rapidly.

4.5. Training Objective

Hierarchical Training. Our model is trained level-by-level in a coarse-to-fine manner. Specifically, the depth probability volume constructed by the previous level is transformed into a probability distribution function (PDF), which is then utilized to sample a smaller number of more accurate depth candidates for the subsequent level. This approach facilitates a more precise and memory-efficient training and rendering process.

Training Loss. Our model is trained solely under the supervision of RGB images. Different from the existing feed-forward 3D-GS methods [3, 6, 24], we introduce a novel approach by integrating reference views into the overall supervision as the reference loss $\mathcal{L}^{src}$. The inclusion of additional contextual views not only enhances geometric cues but also enriches texture information, thereby accelerating optimization and improving rendering quality. Specifically, for each layer k, the loss function $\mathcal{L}^k_{total}$ comprises both the target view loss and the source view loss. The former includes L1 loss, SSIM loss [36], and perceptual loss [54], while the latter consists of L1 loss computed for each individual source view, as demonstrated in Sec. 4.4. The overall loss can be computed by:

$$\mathcal{L}^k_{total} = \mathcal{L}_1^{target} + \mathcal{L}_{SSIM} + \mathcal{L}_{LPIPS} + \sum\nolimits_{i=1}^{N} \mathcal{L}_1^{src}{}_i. \tag{11}$$

5. Experiments

5.1. Settings

Datasets. To evaluate the cross-baseline performance of our method, we select two widely used datasets for training and testing: the object-centric dataset DTU [16], which can provide small-baseline inputs and the RealEstate10K [55] which can serve as the large-baseline dataset. To further evaluate the generalizable performance, we select the forward-facing dataset LLFF [25] and the large-baseline Mip-NeRF 360 dataset [1] to conduct zero-shot evaluation.

Baselines. We compare our method against several generalizable NeRF-based methods [4, 11, 17, 22, 31, 33, 52] as well as two typical 3D-GS methods MVSplat [6] and MVSGaussian [24] which are designed for large-baseline

Table 1. DTU small-baseline.

Method	3-view			2-view			Inference Time (s)	
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Encode↓	Render↓
PixelNeRF [52]	19.31	0.789	0.382	-	-	-	**0.005**	5.294
IBRNet [33]	26.04	0.907	0.191	-	-	-	0.016	4.592
MVSNeRF [4]	26.63	0.931	0.168	24.03	0.914	0.192	0.042	2.363
ENeRF [22]	27.61	0.957	0.089	25.48	0.942	0.107	0.019	0.032
MuRF [39]	28.76	0.961	0.077	27.02	0.949	0.088	0.142	1.122
PixelSplat [3]	-	-	-	14.01	0.662	0.389	0.102	2.3E-3
MVSplat [6]	27.15	0.935	0.121	25.02	0.915	0.126	0.040	3.9E-3
MVSGaussian [24]	28.21	0.963	0.076	25.78	0.947	0.095	0.021	2.4E-3
Ours	**29.34**	**0.969**	**0.075**	**27.56**	**0.958**	**0.084**	0.153	**2.1E-3**

Table 2. RealEstate10K large-baseline.

Method	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [52]	13.91	0.46	0.591
SRF [9]	15.40	0.486	0.604
GeoNeRF [17]	16.65	0.511	0.541
IBRNet [33]	15.99	0.484	0.532
GPNR [31]	18.55	0.748	0.459
AttnRend [11]	21.38	0.839	0.262
MuRF [39]	24.20	0.865	0.170
MVSGaussian [24]	22.58	0.752	0.206
MVSplat [6]	23.34	0.797	0.188
MuGS(Ours)	**24.82**	**0.873**	**0.153**

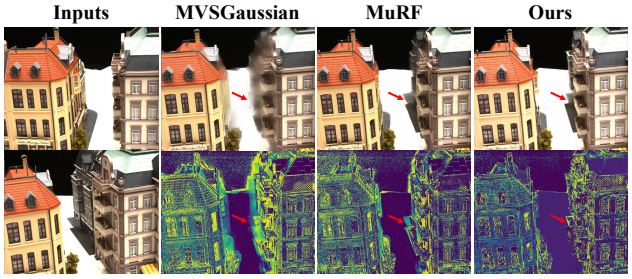


Figure 3. **2-view small-baseline results on the DTU [16] dataset**. Our method renders higher quality with fewer errors than other small-baseline methods.

inputs and small-baseline inputs, respectively. Besides, we compare our method with MuRF [39], the state-of-the-art NeRF-based multi-baseline method.

Implementation Details. Our implementation is mainly based on PyTorch and the 3D-GS rendering implemented in CUDA. We sample 64 isometric depth candidates for the coarse model and 16 for the fine model. All models are trained on 2 Nvidia A6000 GPUs with the Adam [20] optimizer. For the pre-trained monocular model, we use Depth Anything V2 [43].

5.2. Results

Small-Baseline on DTU. The DTU dataset [16] is a small-baseline dataset since the input images are object-centric and provide significant overlap between views. We follow the setting of MuRF [39], which takes the nearest 3 views around the target view to serve as source views. Additionally, we evaluate the 2-view scenario, which is more challenging since there is more occlusion and less context. We achieve more than 0.5dB PSNR improvement compared to the previous best methods in both 2-view and 3-view settings. Besides, our method provides higher inference speed compared to MuRF thanks to the Gaussian representation.

As shown in Fig. 3, this setting presents a significant challenge due to the large occlusion between the *buildings* and the limited availability of only two input views. Since the texture in the occluded regions is visible in only one of the input views, methods such as MVSGaussian [24], which heavily rely on MVS pipeline, struggle to predict accurate depth. This results in blurry rendered output images. While MuRF [39] achieves better quality than MVSGaussian, it still exhibits geometric inaccuracies, as the marked area is misplaced in Fig. 3. In contrast, our method demonstrates superior performance in both rendering quality and geometric accuracy, which we attribute this performance to our fusion strategy and multi-view training supervision.

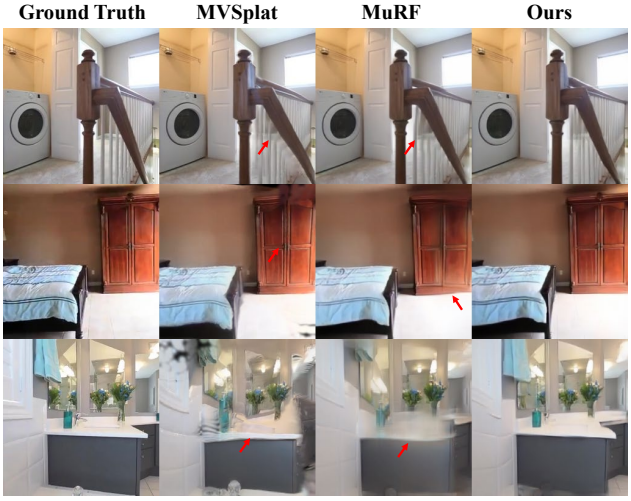


Figure 4. **2-view large-baseline results on the RealEstate10K dataset**. The images rendered by our method exhibit superior geometric accuracy and reduced distortion.

Large-Baseline on RealEstate10K. In this dataset [55], we follow the setting of AttnRend [11] and MuRF [39], where the 2 input views are selected from a video with a distance of 128 frames, and the target view to synthesis is an intermediate frame. This large-baseline setting provides relatively small overlap, which is challenging for methods like MVSGaussian [24], which relies heavily on multi-view feature matching, and monocular cues can be useful to inference geometry. As shown in Tab. 2, our method achieves more than 1dB PSNR improvement compared to the previous best large-baseline 3D-GS method MVSplat as well as more than 0.5dB PSNR improvement compared to the previous state-of-the-art multi-baseline method MuRF. The visual comparison in Fig. 4 indicates that our method produces clearer rendering results than MVSplat [6] and MuRF [39]. Meanwhile, the images generated by our method show both geometrical precision and reduced distortion.

Depth Accuracy on DTU. To evaluate the quality of recon-

Table 3. Depth evaluation results from 3-view inputs on DTU.

Method	Reference view			Novel view		
	Abs err↓	Acc(2)↑	Acc(10)↑	Abs err↓	Acc(2)↑	Acc(10)↑
MVSNet [47]	3.60	0.603	0.955	-	-	-
PixelNeRF [52]	49.0	0.037	0.176	47.8	0.039	0.187
IBRNet [33]	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 [22]	3.80	0.837	0.939	4.60	0.792	0.917
MuRF [39]	-	-	-	12.73	0.583	0.906
MVSGaussian [24]	**3.11**	0.866	0.956	3.66	0.838	0.945
MuGS(Ours)	3.23	**0.872**	**0.963**	**3.52**	**0.853**	**0.952**

Table 4. Zero-shot performance on DTU and Mip-NeRF 360 dataset.

Method	DTU			Mip-NeRF 360 Dataset		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AttnRend [11]	11.35	0.567	0.651	14.00	0.474	0.712
MVSplat [6]	13.94	0.473	0.385	-	-	-
MVSGaussian [24]	19.26	0.716	0.284	21.19	0.752	0.322
MuRF [39]	22.19	0.894	0.211	23.98	0.800	0.293
MuGS(Ours)	**22.43**	**0.916**	**0.202**	**24.25**	**0.845**	**0.256**

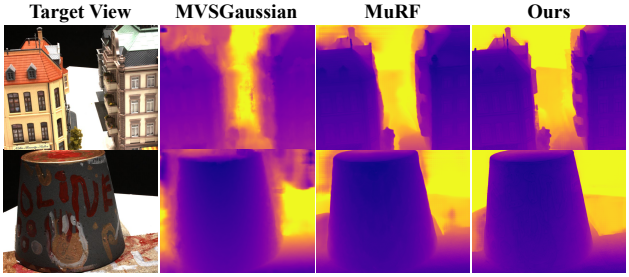


Figure 5. **2-view depth prediction on DTU**. Our method yields better detailed geometric information on novel views.

structed geometry, we select the DTU dataset as it provides the ground truth of depth. We use quantitative metrics, including the average absolute error "Abs err" [47] and the percentage of pixels with an error less than X mm, denoted as "Acc(X)" [24]. As shown in Tab. 3, our method recovers depth with higher accuracy than the previous best method MVSGaussian [24] in novel view while achieving close accuracy in the reference view. Compared to MuRF [39], which prioritizes rendering quality at the expense of geometry accuracy, our method effectively balances both aspects. Moreover, due to the volume rendering approach and implicit representation, MuRF cannot estimate the depth map for reference view since its volume is constructed on the target view, while our 3D-GS-based method can directly generate high-quality reference depth maps. The visual results of the 2-view setting are shown in Fig. 5. MVSGaussian fails to recover accurate depth due to insufficient context, while our method achieves better details than MuRF.

Generalization Performance. We also compare the generalization ability of the model trained on large-baseline or small-baseline datasets. In Tab. 5, we evaluate zero-shot performance on the LLFF dataset [25]. All models are trained on the DTU dataset. Our method achieves better scores in PSNR and SSIM, and close scores in LPIPS, compared with the small-baseline method MVSGaussian in a 3-view setting. Regarding the 2-view setting, our method outperforms others in all metrics, showing robustness in handling limited contextual information.

Table 5. Zero-shot performance on LLFF after trained in DTU.

Method	Settings	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [52]	3-view	11.24	0.486	0.671
IBRNet [33]		21.79	0.786	0.279
MVSNeRF [4]		21.93	0.795	0.252
ENeRF [22]		23.63	0.843	0.182
MatchNeRF [5]		22.43	0.805	0.244
MuRF [39]		23.67	0.860	0.206
MVSGaussian [24]		24.07	0.857	**0.164**
MuGS(Ours)		**24.21**	**0.872**	0.165
MVSNeRF [4]	2-view	20.22	0.763	0.287
ENeRF [22]		22.78	0.821	0.191
MatchNeRF [5]		20.59	0.775	0.276
MuRF [39]		22.82	0.846	0.208
MVSGaussian [24]		23.11	0.834	0.175
MuGS(Ours)		**23.33**	**0.855**	**0.169**

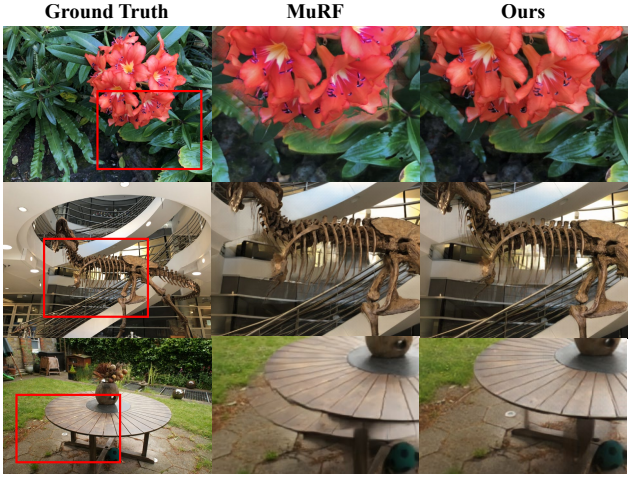


Figure 6. **Generalization performance**. The 1st and 2nd rows are from LLFF dataset with 3 input views, and the 3rd row is from the Mip-NeRF 360 Dataset with 2 input views. Fewer artifacts and blurry areas in our results than MuRF.

For models trained on RealEstate10K dataset [55], we evaluate them in both DTU [16] and Mip-NeRF 360 dataset [1], which is challenging since large-baseline training dataset provides limited supervision for the MVS pipeline due to insufficient overlap between views. The further challenge is that the small-baseline test dataset demands the model to recover precise geometry to obtain high-quality rendering results, which is inherently difficult for large-baseline methods like AttnRend [11] and MVSplat [6]. As shown in Tab. 4, multi-baseline methods, including MuRF [39] and ours, outperform the specific-baseline methods by a large margin on both datasets, while our method demonstrates even better results compared with MuRF. The visual results shown in Fig. 6 further indicate that our method yields sharper outputs with fewer artifacts than MuRF.

Figure 7. **Ablation of fusion strategy**. The 1st row is from small-baseline dataset DTU, and the 2nd row is from large-baseline dataset RealEstate10K. Our proposed fusion strategy works well with the area with occlusion or out of overlap.

Table 6. **Ablation study on each component of our method.**

Module	DTU (Small-Baseline)			RealEstate10K (Large-Baseline)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o feature enhancement	27.35	0.955	0.087	24.62	0.870	0.159
w/o depth refinement	27.28	0.954	0.087	24.56	0.869	0.161
w/o reference loss	27.52	0.957	0.085	24.67	0.871	0.156
Full model	**27.56**	**0.958**	**0.084**	**24.82**	**0.873**	**0.153**

5.3. Ablation and Analysis

As shown in Table 6 and Fig. 7, we conduct ablation experiments under both large-baseline and small-baseline settings to assess the effectiveness of our designs. The models are evaluated after training separately on DTU [16] and RealEstate10K [55].

Depth Refinement. Our method integrates the source view aligned MDE depth maps into the depth probability volume constructed by the MVS pipeline. To assess the contribution of this depth refinement process, we remove the project-and-sample step along with the subsequent 3D U-net, using the coarse probability volume directly for predicting the target depth map and feature map. As shown in Tab. 6, omitting depth refinement leads to a performance decline, which proves that the refined depth prediction indeed improves the final results. The visual comparison in Fig. 7 further indicates that our depth refinement is helpful to retrieve better geometry, particularly in occluded or low-overlap regions.

Feature Enhancement. We further explore the difference between the features enhanced by pre-trained monocular features or not. The quantitative results in Tab. 6 demonstrate that performance drops when only features encoded by the MVS encoder are used for rendering. This suggests that useful inductive bias introduced by monocular features is important for rendering quality. This can be verified in Fig. 7, as inaccurate color and texture artifacts emerge when the feature enhancement process is removed.

Training Loss. Our model is trained with an additional reference loss. To evaluate its effectiveness, we separately conduct the training with and without the reference loss under identical settings. As shown in Tab. 6, removing the reference loss leads to a decline in final performance, particularly in large-baseline scenarios. Moreover, the visual comparison in Fig. 8 indicates that the reference loss efficiently accelerates the optimization process, yielding more than 2dB PSNR improvement at specific iterations.

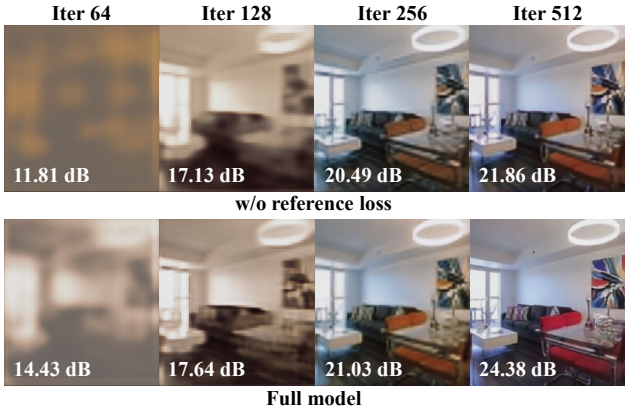


Figure 8. **Ablation of reference Loss**. This loss not only elevates the rendering quality, but also boosts the optimization process.

6. Conclusion

We present MuGS, a feed-forward, generalized 3D Gaussian Spatting approach for novel view synthesis that effectively generalizes across diverse baseline settings. Specifically, we leverage both Multi-View Stereo (MVS) and Monocular Depth Estimation (MDE) to infer depth and enhance the MVS feature with the powerful pre-trained MDE feature. To take advantage of the precision of MVS and the robustness of MDE, we propose a projection-and-sampling mechanism for depth fusion and refine the depth probability volume. To further introduce induction bias for better generalization, we introduce a novel loss function proposed to assist in better geometry and rendering quality. Experiments shows MuGS achieves better multi-baseline generalization as well as better zero-shot performance, proving the effectiveness of our method.

Limitations. As our method relies on MVS and MDE for depth estimation, it inherits limitations from both, such as decreased depth accuracy in areas with weak textures or specular reflections, resulting in degraded view quality.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 5, 7

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 1, 5, 6

[4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 5, 6, 7

[5] 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. 7

[6] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 1, 3, 5, 6, 7

[7] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *arXiv preprint arXiv:2411.04924*, 2024. 1, 3

[8] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2524–2534, 2020. 2

[9] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 6

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8585–8594, 2022. 2

[11] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4970–4980, 2023. 5, 6, 7

[12] Pascal Fua and Yvan G Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16(1):35–56, 1995. 2

[13] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2015. 2

[14] David Gallup, Jan-Michael Frahm, Philippos Mordohai, and Marc Pollefeys. Variable baseline/resolution stereo. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[15] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2

[16] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 5, 6, 7, 8

[17] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 5, 6

[18] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[20] Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*. San Diego, California;, 2015. 6

[21] Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21539–21548, 2023. 4

[22] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3, 5, 6, 7

[23] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18088–18097, 2023. 2

[24] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference*

on Computer Vision, pages 37–53. Springer, 2024. 1, 3, 5, 6, 7

[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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 5, 7

[26] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993. 2

[27] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8645–8654, 2022. 3

[28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[29] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[30] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 2

[31] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pages 156–174. Springer, 2022. 5, 6

[32] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 1

[33] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 5, 6, 7

[34] 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 *European Conference on Computer Vision*, pages 573–591. Springer, 2022. 2

[35] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9466–9476, 2023. 3

[36] 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

[37] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6187–6196, 2021. 2

[38] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European conference on computer vision*, pages 456–473. Springer, 2024. 1

[39] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: Multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20050, 2024. 2, 3, 6, 7

[40] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 3

[41] 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 *European conference on computer vision*, pages 674–689. Springer, 2020. 2

[42] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3

[43] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2, 3, 6

[44] Ruigang Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–I, 2003. 2

[45] Ruigang Yang, G. Welch, and G. Bishop. Real-time consensus-based scene reconstruction using commodity graphics hardware. In *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, pages 225–234, 2002. 2

[46] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8574–8584, 2022. 4

[47] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2, 7

[48] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2

[49] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17661–17670, 2023. 3

[50] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6480–6494, 2022. 3

[51] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 3

[52] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 5, 6, 7

[53] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214, 2023. 3

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[55] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 5, 6, 7, 8