VGGS: VGGT-guided Gaussian Splatting for Efficient and Faithful Sparse-View Surface Reconstruction

Peng Xiang[1], **Liang Han**[1], **Hui Zhang**[1†], **Yu-Shen Liu**[1†], **Zhizhong Han**[2]
[1]School of Software, Tsinghua University, Beijing, China
[2]Department of Computer Science, Wayne State University, Detroit, USA
xiangp23@mails.tsinghua.edu.cn han123@mails.tsinghua.edu.cn huizhang@tsinghua.edu.cn
liuyushen@tsinghua.edu.cn h312h@wayne.edu

Abstract

Reconstructing a faithful geometric surface from sparse images remains a fundamental challenge in 3D computer vision. While recent methods have achieved remarkable progress, they still struggle to recover reliable geometry due to the lack of multi-view geometric cues, particularly in non-overlapping regions. To address this issue, we introduce VGGS, a Gaussian Splatting (GS) method that exploits multi-view geometric priors from VGGT for efficient and high-fidelity sparse-view surface reconstruction. Our primary contribution is an anchor-calibrated depth estimation scheme, which yields accurate depth maps. The insight is to align the VGGT depth prior to the underlying surface with a sparse set of multi-view consistent anchors, then infer depth for unreliable regions by relative depth estimation. Furthermore, to mitigate misalignment in complex scenes, we propose a relative depth consistency loss that penalizes the rendered depth if its relative depth relationship in local regions is inconsistent to the multi-view prior. Extensive experiments on widely-used benchmarks show that VGGS surpasses state-of-the-art methods in both accuracy and efficiency, delivering 4–7× faster optimization while reducing memory consumption compared to previous GS-based approaches.

Code — https://github.com/AllenXiangX/VGGS

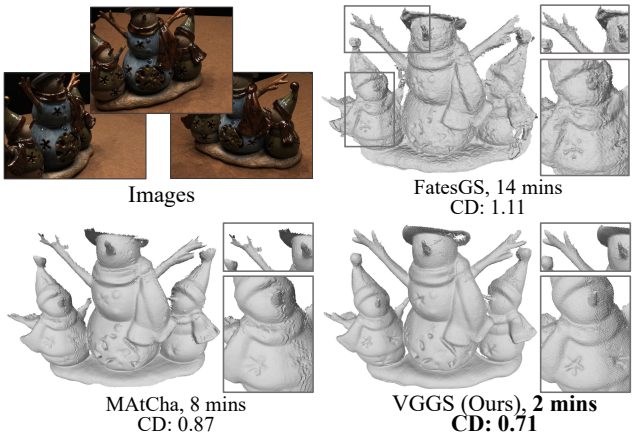


Figure 1: Surface reconstruction from 3-view images (small-overlap) of DTU scan 69. Compared with state-of-the-art FatesGS (Huang et al. 2025) and MAtCha (Guédon et al. 2025), VGGS manages to optimize superior geometry within only 2 minutes, achieving faithful and efficient surface reconstruction.

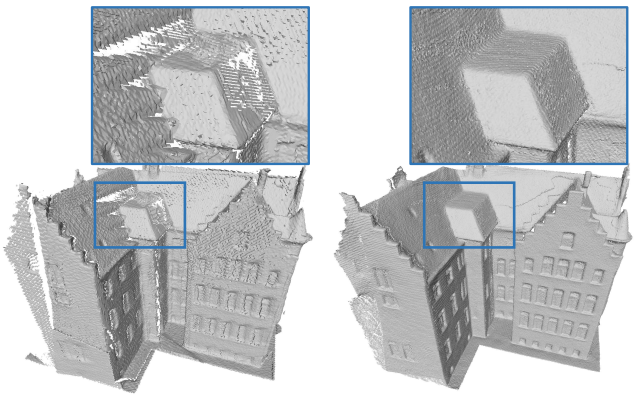
Introduction

Predicting faithful geometric surface from multi-view images remains an important and challenging task in 3D computer vision. In recent years, the rise of Neural Radiance Fields (NeRF) (Mildenhall et al. 2021; Yu et al. 2022) and 3D Gaussian Splatting (GS) (Kerbl et al. 2023) techniques has brought substantial progress to multi-view surface reconstruction. Particularly, 3DGS, benefiting from its explicit and unstructured 3D representation, manages to improve reconstruction quality while enabling real-time rendering, which significantly inspires follow-up work to achieve high-fidelity surface reconstruction. However, these methods typically demand dense inputs to infer accurate geometries; with limited input views, the scarcity of multi-view geometric cues makes it difficult to infer reliable geometries, especially in non-overlapping regions. An intuitive solution to this problem is to incorporate geometric priors from foundation models to compensate for the lack of geometric cues. Along this line, recent methods (Yu et al. 2022; Huang et al. 2025; Guédon et al. 2025) manage to improve reconstruction quality with monocular geometric priors. Nevertheless, due to the lack of multi-view geometric information of monocular priors, it's still difficult to achieve efficient and faithful sparse-view surface reconstruction.

Fortunately, the emergence of VGGT (Visual Geometry Grounded Transformer) (Wang et al. 2025) offers a new perspective for the above problem. Trained on large-scale multi-view datasets, this model can predict strong multi-view depth priors, exhibiting promising potential for sparse-view surface reconstruction. However, directly fusing VGGT depth maps for mesh extraction is suboptimal. First, although the VGGT depth prior is multi-view based, slight misalignment among different frames can severely degrade the surface quality, as shown in Figure 2. Second, the estimated camera parameters often deviate from the ground truth, which require additional coordinate transformation to

†Corresponding authors.

correct the inconsistency.



(a) VGGT (b) VGGS

Figure 2: (a) The surface of VGGT extracted by fusing multi-view depth maps with TSDF fusion, which suffers from artifacts and holes due to slight misalignment among different frames. (b) The surface obtained by VGGS.

To address these issues, we propose VGGS, a simple and efficient Gaussian Splatting method that exploits the multi-view geometric priors from VGGT. The core of VGGS is an anchor-calibrated depth estimation scheme that effectively aligns the depth prior with the underlying surface to obtain high-fidelity depth maps. Specifically, we first select pixels that are multi-view consistent as surface anchors. Then, we align the depth prior to the rendered depth at these anchors. After the alignment, VGGS derives accurate depth at each pixel in unreliable regions from its relative depth to the anchor pixels, as specified by the aligned depth priors. Meanwhile, since the estimated depth may exhibit inconsistency due to misalignment in complex scenes, we further introduce a relative depth consistency loss to penalize the rendered depth that is inconsistent to the multi-view depth prior. We achieved the state-of-the-art reconstruction accuracy under widely used benchmarks. Our main contributions can be summarized as follows.

- We introduce VGGS, a simple and effective method that manages to incorporate VGGT's multi-view geometric priors into 3DGS, achieving high-fidelity and efficient sparse-view surface reconstruction.
- We propose an anchor-calibrated depth estimation scheme that yields accurate depth estimation through anchor-guided alignment and relative depth estimation, together with a relative depth consistency loss that enforces overall depth consistency to the depth prior.
- We conduct extensive experiments on widely-used benchmarks and demonstrate that VGGS not only surpasses existing state-of-the-art methods in reconstruction accuracy, but also achieves a 4–7× speed-up in optimization and a substantial reduction in memory consumption over previous GS-based methods.

Related Work

Multi-View Surface Reconstruction

Multi-view surface reconstruction is a fundamental and challenging task in 3D computer vision, the emergence of Neural Radiance Fields (NeRF) (Mildenhall et al. 2021; Wang et al. 2021; Fu et al. 2022; Li et al. 2023) and 3D Gaussian Splatting (GS) (Kerbl et al. 2023; Liu et al. 2025; Zhang et al. 2025c) has significantly advanced this field. Particularly, the explicit and unstructured 3D representation of Gaussian Splatting enables numerous follow-up works to enhance both reconstruction quality and efficiency through diverse strategies (Yu, Sattler, and Geiger 2024; Guédon and Lepetit 2024; Zhang et al. 2025a; Gao et al. 2025). For example, 2DGS (Huang et al. 2024a) leverages planar Gaussian disk to optimize multi-view consistent geometry; similar works use gaussian surfels (Dai et al. 2024; Jiang et al. 2025) to realize precise reconstruction. PGSR (Chen et al. 2024a) adopts unbiased depth rendering with multi-view consistent constraints and achieves impressive performance. Despite the progress, these methods rely on dense views that contain abundant multi-view geometric cues (Li et al. 2025). Under sparse-view settings, the scarcity of multi-view geometric information leads to severe performance degradation.

To address the above problem, many approaches attempt to improve sparse-view surface reconstruction, which can be divided into two categories: generalizable and scene-specific. Generalizable models (Long et al. 2022; Ren et al. 2023; Han et al. 2025) are trained on large-scale datasets to learn multi-view geometric information. After pre-training, they can deliver rapid feed-forward inference for new scenes in seconds(Charatan et al. 2024; Xu et al. 2025; Chen et al. 2024b). However, the pre-training is often time-consuming, and the reconstruction performance can drop severely when test scenes are significantly different from the training set. On the other hand, scene-specific methods directly infer per-scene geometry from sparse images without requiring costly pre-training (Huang et al. 2024b; Yu et al. 2022; Younes, Ouasfi, and Boukhayma 2024), allowing flexible adaptation to new scenes. To improve reconstruction quality, scene-specific methods typically devise efficient strategies that fully leverage the geometric cures from geometric priors. Our VGGS also follow this line of work, and we propose novel strategy to achieve better geometry inference.

Relation to Geometric Priors

Geometric priors (Yin et al. 2023; Yang et al. 2024a) are extensively explored for sparse-view surface reconstruction. For example, MonoSDF (Yu et al. 2022) leverages geometric cues in monocular normal and depth priors (Xiang et al. 2025; Wang et al. 2022; Zhang et al. 2025b) for neural implicit surface reconstruction. More recently, FatesGS (Huang et al. 2025) improves reconstruction by distilling local depth ranking information from monocular depth prior. MAtCha (Guédon et al. 2025) improves both novel view synthesis and surface reconstruction by fully utilizing monocular depth prior (Yang et al. 2024b). Although existing works have achieved promising progress with monocular priors, they remain inherently limited. First, the monocular

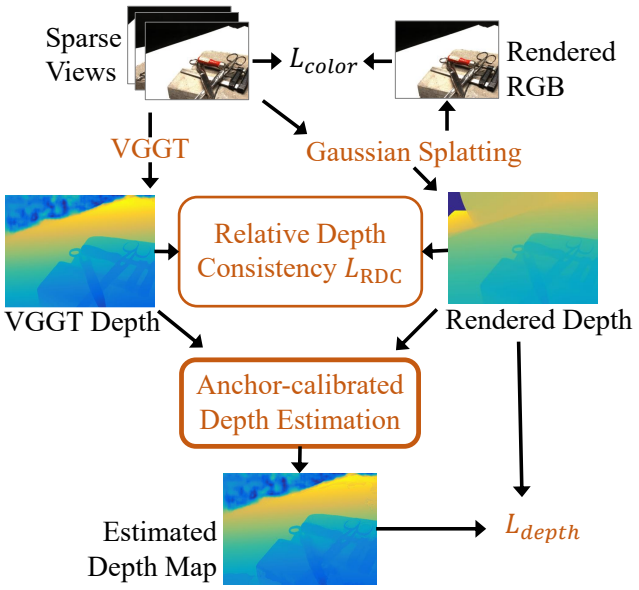


Figure 3: The overview of VGGS, which consists of anchor-calibrated depth estimation and a Relative Depth Consistency (RDC) loss.

prior cannot provide adequate multi-view geometric clues. Second, the widely used depth priors usually suffer from ambiguous scale and distortion. Additionally, some methods (Wu et al. 2025; Younes, Ouasfi, and Boukhayma 2024) attempt to leverage multi-view stereo (MVS) priors to enhance reconstruction quality, yet such priors struggle in challenging settings where overlap among views is small. Different from previous methods, we adopt sparse-view surface reconstruction by exploring VGGT (Wang et al. 2025), a powerful 3D foundation model trained on large-scale multi-view datasets and can predict multiple promising 3D attributes, including multi-view depth priors. Compared to foundation models with geometric priors, our method can provide more accurate geometry inference on a specific scene, achieving better generalization ability.

Method

Overview and Motivation

The overview pipeline of VGGS is shown in Figure 3, which leverages VGGT's depth prior D_p to supervise rendered depth D_r from sparse-view images $\mathcal{I} = \{I_i | i \in 1, 2, ..., N\}$, with known poses $\mathcal{T} = \{T_i | i \in 1, 2, ..., N\}$. The core of VGGS is an anchor-calibrated depth estimation scheme that estimates accurate depth map $\hat{D}_e$ by distillation multi-view prior from VGGT. Our insight is to align the depth prior D_p to the underlying surface by a sparse set of carefully selected anchor pixels, based on which we estimate depth for unreliable regions by relative depth estimation. Next, the rendered depth D_r is constrained by the estimated depth map $\hat{D}_e$ through depth loss L_{depth}. Meanwhile, as misalignment may occur in complex scenes where it's difficult to select reliable anchors, we further propose a relative depth consistency loss L_{RDC} to enforce local and global depth consistency of rendered depth D_r according to the depth prior D_p.

Anchor Selection and Surface Alignment

The anchor selection and surface alignment are illustrated in Figure 4 (a) and (b), where we use 2D example for clarity. Although VGGT is trained on large-scale multi-view datasets and the depth prior contains abundant multi-view information, this prior still deviates from the true surface by unknown shift and scale factors, as illustrated in Figure 4 (a). Previous methods (Eigen, Puhrsch, and Fergus 2014; Ranftl et al. 2020) like MonoSDF (Yu et al. 2022) propose to calculate the scale and shift factors using least-squares criterion to align monocular depth to the rendered depth. However, MonoSDF solves the scale and shift using pixels specified by a random batch of rays, which preserves overall alignment yet may degrade the accuracy of pixels that were originally precise. Unlike previous methods, VGGS aims to achieve accurate and robust alignment using anchors that lie on the underlying surface. We assume that with multi-view images, though sparse, surface anchors can be identified according to their cross-view consistency. Therefore, we leverage the multi-view photometric consistency from PGSR (Chen et al. 2024a) to select reliable anchor pixels. Specifically, given the multi-view consistency map C_m and the depth confidence map C_d from VGGT, we first select the top-K_1 pixels with the highest consistency scores,

$$\begin{aligned} \mathcal{P}_{\text{topK}_1} &= \left\{ \mathbf{p}_{(i)} \;\middle|\; 1 \le i \le \mathrm{K}_1, C_d(\mathbf{p}_{(i)}) > \tau \right\}, \\ \text{with } & C_m(\mathbf{p}_{(1)}) \ge C_m(\mathbf{p}_{(2)}) \ge \cdots \ge C_m(\mathbf{p}_{(\mathrm{K}_1)}), \end{aligned} \tag{1}$$

where τ is the threshold that excludes pixels with low depth confidence. Although the selected top-K_1 pixels exhibit high consistency, we observe that these pixels tend to cluster in local regions, which can lead to ambiguous scale factor and cause severe alignment failure. To mitigate this problem, we use furthest point sampling (FPS) (Qi et al. 2017) to choose K_f pixels from the next top-K_2 pixels by,

$$\begin{aligned} \mathcal{P}_{\text{FPS}} &= \text{FPS}(\mathcal{P}_{\text{topK}_2}, \mathrm{K}_f), \\ \mathcal{P}_{\text{topK}_2} &= \left\{ \mathbf{p}_{(i)} \;\middle|\; \mathrm{K}_1 + 1 \le i \le \mathrm{K}_1 + \mathrm{K}_2 \right\}, \end{aligned} \tag{2}$$

Note that both K_1 and K_2 are small integers, with $\mathrm{K}_1 = \mathrm{K}_2 = 15$ in our paper. Since too many anchors would inevitably compromise the accuracy of high-consistency pixels, the utilization of FPS aims to keep anchors uniformly distributed over reliable regions while avoiding excessive anchor number. Next, we can obtain the final anchor pixels from,

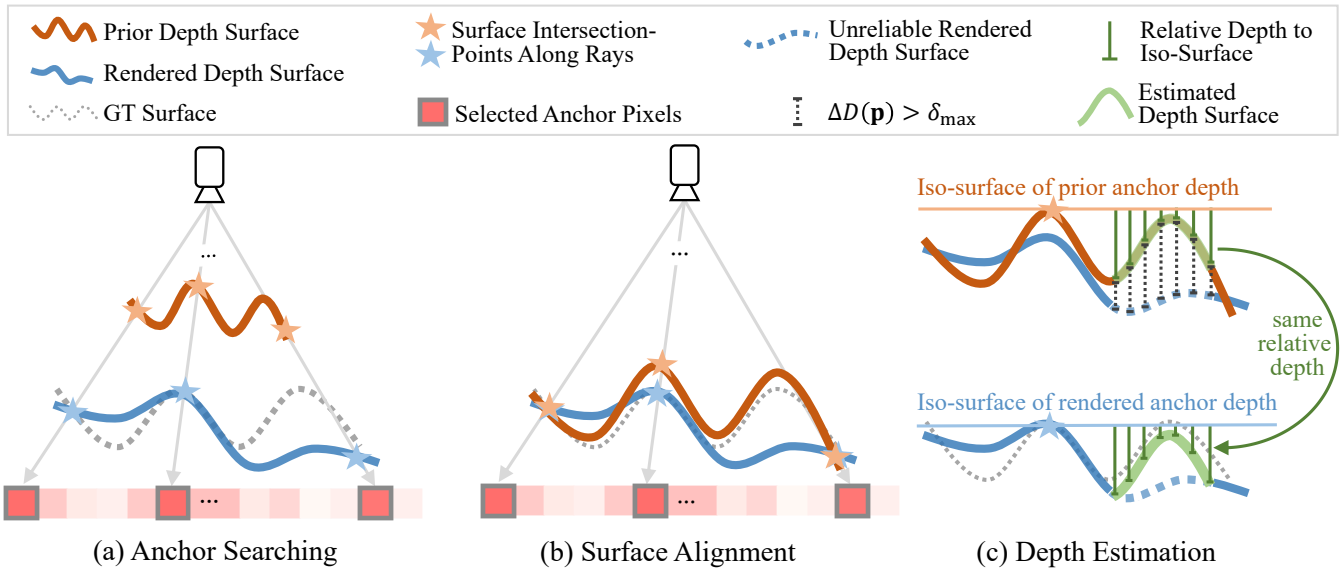
$$\mathcal{P}_{\text{anchor}} = \mathcal{P}_{\text{topK}_1} \cup \mathcal{P}_{\text{FPS}}. \tag{3}$$

With the selected anchors, we can conduct robust and accurate surface alignment by solving scale w and shift q factors with least-squares criterion,

$$(w, q) = \arg\min_{w,q} \sum_{\mathbf{p} \in \mathcal{P}_{\text{anchor}}} \left(w D_p(\mathbf{p}) + q - D_r(\mathbf{p}) \right)^2, \tag{4}$$

where D_p and D_r are the VGGT depth prior and the rendered depth, respectively. Equation 4 can be rewritten as,

$$\mathbf{h}^{\text{opt}} = \arg\min_{\mathbf{h}} \sum_{\mathbf{p} \in \mathcal{P}_{\text{anchor}}} \left(\mathbf{d}_{\mathbf{p}}^{\mathrm{T}} \mathbf{h} - D_r(\mathbf{p}) \right)^2, \tag{5}$$



(a) Anchor Searching (b) Surface Alignment (c) Depth Estimation

Figure 4: Anchor-calibrated depth estimation in 2D example. (a) Anchor selection based on multi-view consistency, indicated by red squares, where deeper red denotes higher consistency and three pixels (highlighted in gray border) are selected as surface anchors. The blue and brown stars denote surface intersections along camera rays. (b) Surface alignment using the selected anchors, where prior surface is scaled and shifted by the anchors. (c) Depth estimation for unreliable regions (denoted as blue dashed curve) by relative depth to the iso-surface of anchor depth. Here only one anchor is showed for clarity.

where $\mathbf{h} = (w, q)^{\mathrm{T}}$, $\mathbf{d}_\mathbf{p} = (D_p(\mathbf{p}), 1)^{\mathrm{T}}$, and $\mathbf{h}$ can be solved with closed-form solution,

$$\mathbf{h} = \left(\sum_{\mathbf{p}} \mathbf{d}_\mathbf{p}\mathbf{d}_\mathbf{p}^T\right)^{-1}\left(\sum_{\mathbf{p}} \mathbf{d}_\mathbf{p} D_r(\mathbf{p})\right). \tag{6}$$

Next, the aligned prior depth is obtained by,

$$\hat{D}_p = wD_p + q \tag{7}$$

Depth Estimation with VGGT

Although the aligned VGGT prior $\hat{D}_p$ is highly close to the true surface, minor deviations can still appear in non-anchor regions and even within the anchors in Figure 4 (b). Therefore, instead of directly supervising D_r with $\hat{D}_p$, we obtain more accurate depth map $\hat{D}_e$ through relative depth estimation (Figure 4 (c)). Intuitively, $\hat{D}_e$ should repair unreliable geometries while preserving reliable regions. To this end, we separate unreliable $\mathcal{P}_{\text{unrel}}$ and reliable regions $\mathcal{P}_{\text{rel}}$ based on the absolute depth difference $\Delta D = |\hat{D}_p - D_r|$, the maximal depth difference among anchors serves as threshold,

$$\begin{aligned} \delta_{\max} &= \max_{\mathbf{p}\in\mathcal{P}_{\text{anchor}}} \Delta D(\mathbf{p}), \\ \mathcal{P}_{\text{unrel}} &= \{\mathbf{p} | \Delta D(\mathbf{p}) > \delta_{\max}\}, \\ \mathcal{P}_{\text{rel}} &= \{\mathbf{p} | \Delta D(\mathbf{p}) \le \delta_{\max}\}, \end{aligned} \tag{8}$$

where the unreliable region $\mathcal{P}_{\text{unrel}}$ is denoted as dashed blue curves in Figure 4 (c). Next, given a selected anchor $\mathbf{p}_s \in \mathcal{P}_{\text{anchor}}$, we compute the depth for $\mathcal{P}_{\text{unrel}}$ through relative depth estimation (RDE),

$$\begin{aligned} \hat{D}_e(\mathbf{p}) &= \mathrm{RDE}(\mathbf{p}, \mathbf{p}_s) \\ &= D_r(\mathbf{p}_s) + (\hat{D}_p(\mathbf{p}) - \hat{D}_p(\mathbf{p}_s)), \mathbf{p} \in \mathcal{P}_{\text{unrel}}. \end{aligned} \tag{9}$$

The motivation behind Equation 9 is to propagate VGGT's multi-view geometries from anchor $\mathbf{p}_s$ to unreliable regions $\mathcal{P}_{\text{unrel}}$, which can significantly improve the quality of challenging areas, such as non-overlapping regions under sparse-view setting. Meanwhile, since the depths of anchors may not be perfectly accurate, especially in the early optimizing stage, the anchor pixel $\mathbf{p}_s$ used for relative depth estimation should also be carefully selected. To this end, we use the absolute depth difference as the indication of geometric accuracy, then select the top-K_3 anchors with minimal depth difference and average the estimated depths,

$$\begin{aligned} &\hat{\mathcal{D}}_e(\mathbf{p}) = \frac{1}{\mathrm{K}_3}\sum_{\mathbf{p}^{(i)}} \mathrm{RDE}(\mathbf{p}, \mathbf{p}^{(i)}), \mathbf{p} \in \mathcal{P}_{\text{unrel}} \\ &\text{with } \Delta D(\mathbf{p}^{(1)}) \le \Delta D(\mathbf{p}^{(2)}) \le \ldots \le \Delta D(\mathbf{p}^{(\mathrm{K}_3)}). \end{aligned} \tag{10}$$

Experimentally, we set $\mathrm{K}_3 = 4$ in the early stage of training for stable optimization. As the anchor depths become more close to the underlying surface during training, we gradually reduce K_3 to 1 for more precise depth estimation. As for the reliable regions $\mathcal{P}_{\text{rel}}$, we directly preserve their rendered depth. Finally, the estimated depth map $\hat{D}_e$ can be written as,

$$\hat{D}_e(\mathbf{p}) = \begin{cases} D_r(\mathbf{p}), & \mathbf{p} \in \mathcal{P}_{\text{rel}} \\ \frac{1}{\mathrm{K}_3}\sum_{\mathbf{p}^{(i)}} \mathrm{RDE}(\mathbf{p}, \mathbf{p}^{(i)}), & \mathbf{p} \in \mathcal{P}_{\text{unrel}} \end{cases}, \tag{11}$$

and the depth loss is given as,

$$L_{\text{depth}} = \mathrm{avg}\left|D_r - \hat{D}_e\right|. \tag{12}$$

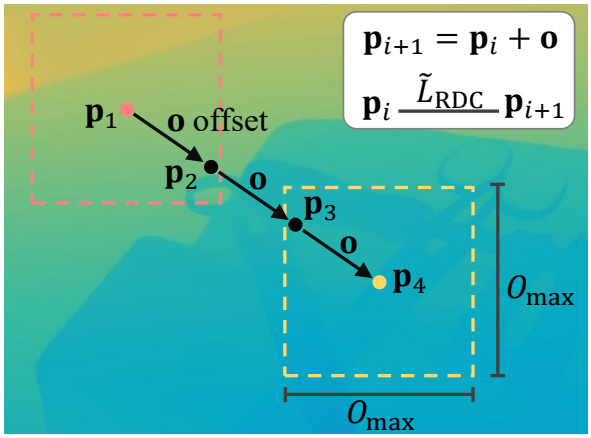


Figure 5: Relative Depth Consistency. The red and yellow dashed borders are local regions of $\mathbf{p}_1$ and $\mathbf{p}_4$. $O_{\max}$ is the maximal offset indicating the range of local region. The relative depth consistency $\tilde{L}_{\text{RDC}}$ is computed between each pixel pair ($\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, where $\mathbf{p}_{i+1} = \mathbf{p}_i + \text{o}$). Note that the depth order relationship can propagate from p_1 to p_4.

Relative Depth Consistency Loss

The relative depth consistency loss L_{RDC} aims to assist the depth loss L_{depth} in challenging cases where surface alignment is not accurate, such as complex scenes. Similar to the ranking scheme in previous methods (Wang et al. 2023; Xu et al. 2024), we penalize the rendered depth D_r if the depth order between any two pixels $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ is inconsistent to the depth prior D_p, as shown in Figure 5. Specifically, we first randomly sample an offset $\mathbf{o}$ at each iteration,

$$\mathbf{o} = (o_x, o_y), o_x \leq O_{\max} \text{ and } o_y \leq O_{\max}, \tag{13}$$

where $O_{\max}$ is the maximal offset indicating the range of local region. For every pixel $\mathbf{p}_i = (x, y)$ in the image $\mathcal{I}$, its relative depth consistency $\tilde{L}_{\text{RDC}}$ with the shifted pixel $\mathbf{p}_{i+1} = \mathbf{p}_i + \mathbf{o}$ in local region is given as,

$$\begin{aligned}\tilde{L}_{\text{RDC}}(\mathbf{p}_i) = \sigma(\text{sgn}(D_p(\mathbf{p}_i) - D_p(\mathbf{p}_{i+1}))* \\ (D_r(\mathbf{p}_{i+1}) - D_r(\mathbf{p}_i))),\end{aligned} \tag{14}$$

where σ is ReLU function. Next, the relative depth consistency loss L_{RDC} of the whole image can be written as,

$$L_{\text{RDC}} = \underset{\mathbf{p}_i \in \mathcal{I}}{\text{avg}} \tilde{L}_{\text{RDC}}(\mathbf{p}_i). \tag{15}$$

Note that previous methods (Wang et al. 2023; Huang et al. 2025) usually conduct the depth ranking in local patches to avoid unreliable depth ranking caused by long spatial distance in monocular depth prior. Differently, since we use the robust multi-view depth priors of VGGT, we propagate depth order (rank) in the entire image plane, leading to more continuous and consistent propagation of the consistency constraints across the whole image.

Loss Function

Since we use the multi-view photometric consistency from PGSR (Chen et al. 2024a) for anchor selection, we implement VGGS based on PGSR. Several PGSR losses are kept for VGGS, including the image reconstrution loss L_{rgb}, the scale loss L_{s}, and the single-view normal loss L_{svgeom} that enforce consistency between the rendered normal and depth normal. The baseline loss is given as,

$$L_{\text{baseline}} = L_{\text{rgb}} + \lambda_s L_{\text{s}} + L_{\text{svgeom}} \tag{16}$$

Additionally, we use normal prior (Yin et al. 2023) to constrain the depth normal for better surface smoothness (L_{normal}). We define the overall loss function of VGGS as,

$$L_{\text{VGGS}} = \lambda_d L_{\text{depth}} + L_{\text{RDC}} + L_{\text{normal}} + L_{\text{baseline}}. \tag{17}$$

Experiments

Datasets and Implementation Details

We evaluate the surface reconstruction performance of our method on the DTU (object level) (Jensen et al. 2014), TNT (unbounded scene level) (Knapitsch et al. 2017) datasets, and self-captured data. The DTU dataset comprises of 15 selected scenes. Following previous methods, we take three views (22, 25, and 28) with small-overlap as inputs. For experiments on the TNT dataset, we follow MAtCha (Guédon et al. 2025) and take four scenes (Caterpillar, Ignatius, Truck, Barn) for comparison, where we sparsely sample 5, 10 and 20 input images to evaluate our method.

We implement VGGS based on the PGSR (Chen et al. 2024a) framework and use their loss functions (excluding the multi-view geometric and photometric constraints) as our baseline. Meanwhile, We follow MAtCha (Guédon et al. 2025) and use MASt3R-SfM (Duisterhof et al. 2024) for Gaussians initialization. The normal prior is obtained from Metric3D (Yin et al. 2023). For all experiments, We optimize VGGS for 3K iterations. In the first 1K iterations, only L_{baseline} loss is used to avoid erroneous anchor selection and misalignment in the early training stage; in the last 2K iterations, the full loss L_{VGGS} is adopted. Note that the anchor-calibrated depth estimation is time consuming, we merely estimate the depth map $\hat{D}_e$ for every 500 iterations, which substantially improve the training efficiency. All experiments are conducted on a single RTX 3090 GPU.

Experimental Comparisons

Evaluation on DTU. The CD results on DTU are shown in Table 1, where VGGS achieves state-of-the art reconstruction performance on average CD, substantially surpasses latest MAtCha (Guédon et al. 2025) and FatesGS (Huang et al. 2025). Particularly, VGGS improves surface reconstruction over the previous best MAtCha by 21.6% in terms of average CD. Moreover, since sparse Gaussian initialization may degrade the performance of surface reconstruction, we additionally follow previous methods (Huang et al. 2025) to use COLMAP (Schonberger and Frahm 2016) for sparse Gaussian initialization (denoted by *). The results indicate that although the COLMAP initialization has a negative impact on reconstruction performance, VGGS still achieves better average CD compared with previous methods, which can demonstrate the robustness of our method. Meanwhile, we present visual comparison of surface reconstruction in Figure 6, where VGGS predicts accurate surface with more detailed geometries.

Method	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Avg
2DGS	3.54	4.13	3.61	1.00	2.69	2.35	2.04	2.06	2.94	1.76	2.40	2.97	1.35	2.17	1.69	2.45
PGSR	4.01	5.19	3.65	0.93	2.96	2.84	1.62	2.16	3.24	1.42	2.35	1.91	0.57	1.55	1.26	2.38
PeTR	3.30	3.60	3.37	2.85	2.91	3.07	2.26	2.29	2.25	2.05	2.80	2.80	1.52	2.29	2.11	2.63
UFORecon	1.51	2.58	1.82	1.44	1.60	1.81	1.04	1.56	0.96	1.40	1.20	0.93	0.66	1.26	1.26	1.40
MonoSDF	3.47	3.61	2.10	1.05	2.37	1.38	1.41	1.85	1.74	1.10	1.46	2.28	1.25	1.44	1.45	1.86
NeuSurf	1.35	3.25	2.50	0.80	1.21	2.35	0.77	1.19	1.20	1.05	1.05	1.21	0.41	0.80	1.08	1.35
SparseCraft	2.42	2.79	2.78	0.74	1.44	2.51	1.26	1.42	1.65	1.10	1.34	5.24	0.65	0.88	1.16	1.83
FatesGS	1.32	2.85	2.71	0.80	1.44	2.08	1.11	1.19	1.33	**0.76**	1.49	0.85	**0.47**	1.05	1.06	1.37
MAtCha	0.88	2.79	1.40	0.83	**0.93**	1.38	0.87	1.25	1.51	0.95	1.15	1.02	0.56	1.01	0.88	1.16
VGGS*	0.82	**1.73**	1.10	0.79	1.44	**0.89**	0.79	1.21	**0.90**	0.87	**0.79**	**0.74**	0.49	0.81	0.86	0.95
VGGS	**0.60**	1.84	**1.05**	**0.73**	1.08	0.94	**0.71**	**1.06**	**0.90**	0.78	**0.79**	1.20	0.50	**0.74**	**0.78**	**0.91**

Table 1: Quantitative comparison on DTU in Chamfer Distance (CD ↓, * denotes Gaussian initialization with COLMAP).

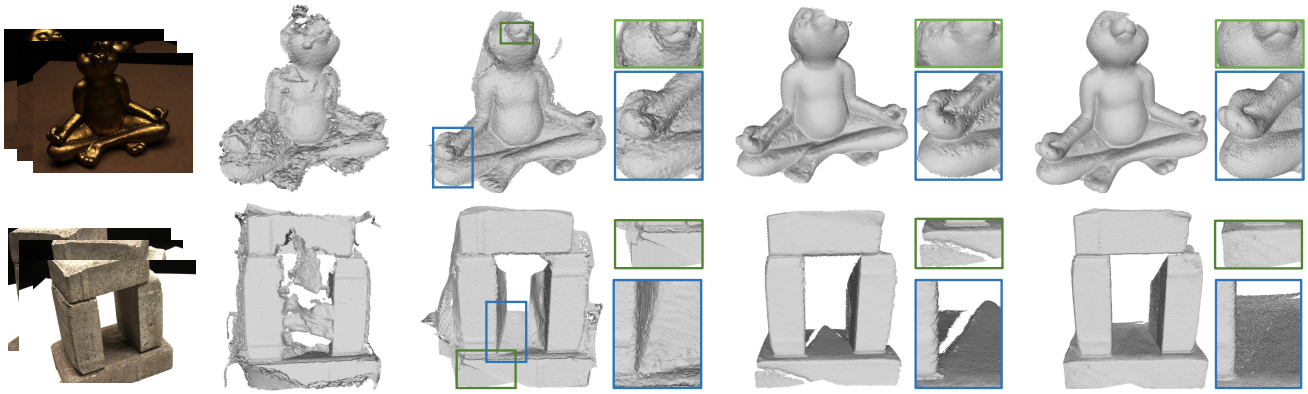


Figure 6: The visual comparison of sparse-view surface reconstruction on the DTU dataset.

	5 views	10 views	20 views
2DGS + MASt3R	0.052	0.121	-
GOF + MASt3R	0.054	0.144	-
MAtCha	0.072	0.156	0.218
VGGS (Ours)	**0.076**	**0.194**	**0.293**

Table 2: Quantitative comparison on TNT (F-Score ↑)

Exp	L_{depth}	L_{RDC}	L_{normal}	CD	F-Score
I				1.51	0.140
II	✓			0.96	0.174
III		✓		1.31	0.175
IV			✓	1.33	0.169
V	✓	✓		0.95	0.189
VI	✓	✓	✓	**0.91**	**0.194**

Table 3: Ablation study on the DTU and TNT datasets.

Evaluation on TNT. We show the quantitative comparison on the TNT dataset in Table 2. The F-Scores under 5-views and 10-views settings are taken from MAtCha paper, the result of 20 input views is obtained by running their source code. As shown in Table 2, using same Gaussian initialization (MASt3R-SfM (Duisterhof et al. 2024)), VGGS achieves the best results across different number of input views. We also present visual comparison between VGGS and MAtCha (Guédon et al. 2025) in Figure 7, where all the meshes are extracted with TSDF fusion. Compared with MAtCha, VGGS can reveal more faithful geometries. Moreover, as input views increase from 10 to 20, VGGS can substantially improve reconstruction quality and deliver more geometric details, showing substantially better surface quality using simple TSDF Fusion.The quantitative and visual results can demonstrate the generalization ability of VGGS on challenging unbounded scenes.

Sparse-View Reconstruction on Self-Captured Data. In Figure 8, we present visualization of surface reconstruction on images captured by ourselves using mobile phones. The visual results show that our VGGS can predict faithful surface with detailed geometries, such as the shoelaces in the left example and the smooth face in the right example.

Ablation Study

To validate the effectiveness of each part in VGGS, we conduct ablation study on the DTU and TNT datasets, and report the CD and F-Score in Table 3, where Exp. I is the

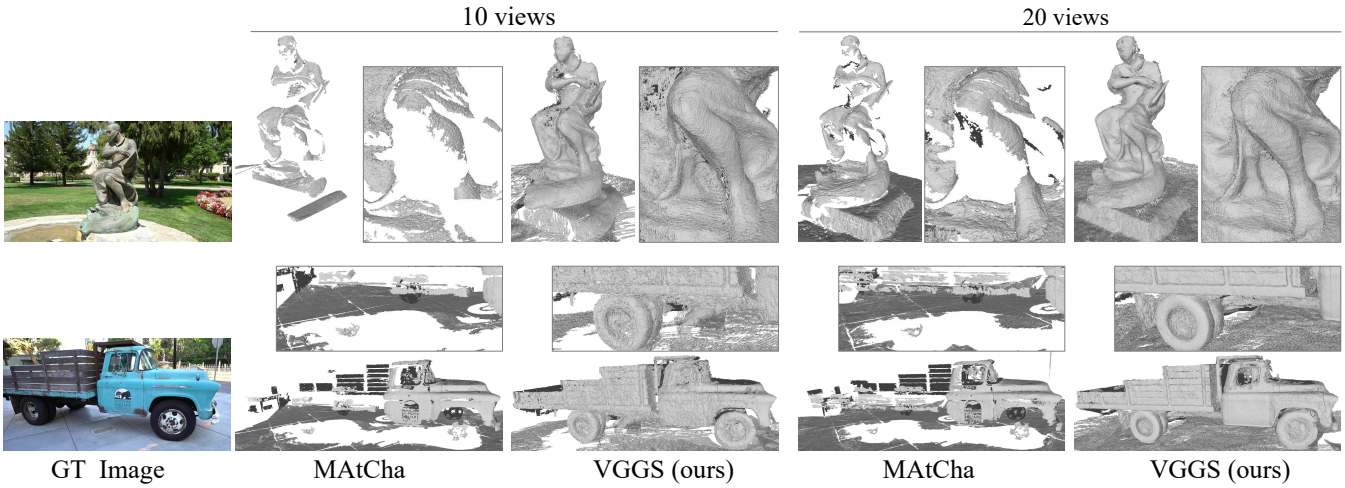


Figure 7: The visual comparison of sparse-view surface reconstruction on the TNT dataset (Knapitsch et al. 2017).

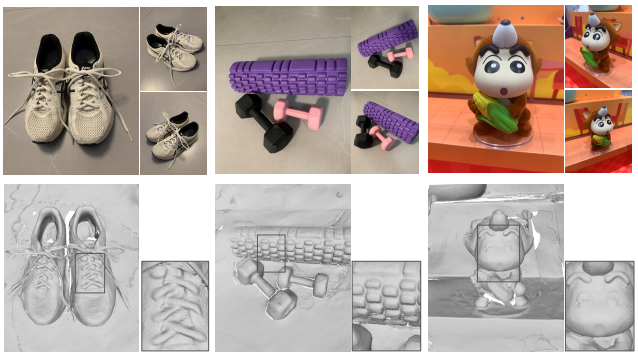


Figure 8: Surface reconstruction on self-captured data.

baseline. From Exp.II and Exp.V, we can find that the depth loss L_{depth} alone can bring substantial improvements over the baseline on both the DTU and TNT datasets, which can demonstrate the effectiveness of our anchor-calibrated depth estimation scheme. Meanwhile, as shown by Exp. III, the relative depth consistency loss L_{RDC} can also contribute to performance on both datasets. However, by comparing Exp. II and Exp. V, we can find that by combing with L_{depth}, the L_{RDC} cannot enhance performance on DTU, but bring substantial improvements on TNT. Because the object-level scans in the DTU dataset are easier for reconstruction, where the estimated depth $\hat{D}_e$ is quite close to the depth prior D_p in terms of relative depth consistency, and L_{RDC} cannot take effect. By contrast, on the more challenging unbounded scenes of TNT, L_{RDC} can assist L_{depth} to obtain more accurate depth. Additionally, as shown by IV and VI, the L_{normal} loss can also help to improve surface reconstruction. The combination of these three losses achieve the best performance across the DTU and TNT datasets.

Computing Resources Consumption

In Table 4, we show the comparison in computing resources consumption with state-of-the-art scene-specific methods on DTU. The GPU memory usage is recorded during training and the training time denotes averaged optimization time over 15 scenes. The results in Table 4 demonstrate that our VGGS can substantially improve the training efficiency to 2 minutes with less GPU memory consumption, achieving highly efficient sparse-view surface reconstruction.

Method	GPU Mem. Usage	Training Time
MonoSDF	14 GB	6 hours
NeuSurf	8 GB	14 hours
SparseCraft	10 GB	10 mins
FatesGS	4 GB	14 mins
MAtCha	3 GB	8 mins
VGGS (Ours)	**2 GB**	**2 mins**

Table 4: GPU memory and time consumption on the DTU.

Conclusion

In this paper, we propose VGGS, a GS-based method that achieves efficient and faithful sparse-view surface reconstruction by taking full advantages of multi-view geometric prior from VGGT. The core of VGGS is an anchor-calibrated depth estimation scheme that robustly aligns VGGT depth prior to the underlying surface, and obtains accurate depth map through relative depth estimation. Additionally, we propose a relative depth consistency loss to enforce overall depth consistency. Quantitative and qualitative experimental results on the DTU and TNT datasets demonstrate that VGGS substantially surpasses state-of-the-art methods in both efficiency and accuracy.

Limitation. While VGGS shows impressive performance on input views with small-overlap on the DTU and TNT, it cannot handle extreme cases with no overlap among different views, since VGGS relies on multi-view consistency to select anchors. Therefore, devising a more robust anchor selection scheme is a promising direction to improve VGGS.

Acknowledgements

This work was supported by Deep Earth Probe and Mineral Resources Exploration – National Science and Technology Major Project (2024ZD1003405), and the National Natural Science Foundation of China (62272263), and in part by Kuaishou.

References

Charatan, D.; Li, S. L.; Tagliasacchi, A.; and Sitzmann, V. 2024. 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*, 19457–19467.

Chen, D.; Li, H.; Ye, W.; Wang, Y.; Xie, W.; Zhai, S.; Wang, N.; Liu, H.; Bao, H.; and Zhang, G. 2024a. PGSR: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*.

Chen, Y.; Xu, H.; Zheng, C.; Zhuang, B.; Pollefeys, M.; Geiger, A.; Cham, T.-J.; and Cai, J. 2024b. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, 370–386. Springer.

Dai, P.; Xu, J.; Xie, W.; Liu, X.; Wang, H.; and Xu, W. 2024. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 conference papers*, 1–11.

Duisterhof, B.; Zust, L.; Weinzaepfel, P.; Leroy, V.; Cabon, Y.; and Revaud, J. 2024. MAst3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Fu, Q.; Xu, Q.; Ong, Y. S.; and Tao, W. 2022. Geo-Neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35: 3403–3416.

Gao, Z.; Bian, J.-W.; Lin, G.; Chen, H.; and Shen, C. 2025. SurfaceSplat: Connecting Surface Reconstruction and Gaussian Splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 28525–28534.

Guédon, A.; Ichikawa, T.; Yamashita, K.; and Nishino, K. 2025. MAtCha Gaussians: Atlas of Charts for High-Quality Geometry and Photorealism From Sparse Views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6001–6011.

Guédon, A.; and Lepetit, V. 2024. SuGaR: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5354–5363.

Han, L.; Zhang, X.; Song, H.; Shi, K.; Liu, Y.-S.; and Han, Z. 2025. SparseRecon: Neural implicit surface reconstruction from sparse views with feature and depth consistencies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 28514–28524.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024a. 2D gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, 1–11.

Huang, H.; Wu, Y.; Deng, C.; Gao, G.; Gu, M.; and Liu, Y.-S. 2025. FatesGS: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3644–3652.

Huang, H.; Wu, Y.; Zhou, J.; Gao, G.; Gu, M.; and Liu, Y.-S. 2024b. NeuSurf: On-surface priors for neural surface reconstruction from sparse input views. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 2312–2320.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 406–413.

Jiang, K.; Sivaram, V.; Peng, C.; and Ramamoorthi, R. 2025. Geometry Field Splatting with Gaussian Surfels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5752–5762.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Li, Q.; Feng, H.; Gong, X.; and Liu, Y.-S. 2025. VA-GS: Enhancing the geometric representation of gaussian splatting via view alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Li, Z.; Müller, T.; Evans, A.; Taylor, R. H.; Unberath, M.; Liu, M.-Y.; and Lin, C.-H. 2023. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8456–8465.

Liu, Y.-C.; Höllein, L.; Nießner, M.; and Dai, A. 2025. QuickSplat: Fast 3D Surface Reconstruction via Learned Gaussian Initialization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 27851–27861.

Long, X.; Lin, C.; Wang, P.; Komura, T.; and Wang, W. 2022. SparseNeus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, 210–227. Springer.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. 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.

Ren, Y.; Wang, F.; Zhang, T.; Pollefeys, M.; and Süsstrunk, S. 2023. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16685–16695.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Wang, G.; Chen, Z.; Loy, C. C.; and Liu, Z. 2023. SparseNeRF: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9065–9076.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, J.; Wang, P.; Long, X.; Theobalt, C.; Komura, T.; Liu, L.; and Wang, W. 2022. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European conference on computer vision*, 139–155. Springer.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS*.

Wu, J.; Li, R.; Zhu, Y.; Guo, R.; Sun, J.; and Zhang, Y. 2025. Sparse2DGS: Geometry-prioritized gaussian splatting for surface reconstruction from sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11307–11316.

Xiang, H.; Li, X.; Cheng, K.; Lai, X.; Zhang, W.; Liao, Z.; Zeng, L.; and Liu, X. 2025. Gaussianroom: Improving 3D gaussian splatting with SDF guidance and monocular cues for indoor scene reconstruction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2686–2693. IEEE.

Xu, H.; Peng, S.; Wang, F.; Blum, H.; Barath, D.; Geiger, A.; and Pollefeys, M. 2025. DepthSplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16453–16463.

Xu, W.; Gao, H.; Shen, S.; Peng, R.; Jiao, J.; and Wang, R. 2024. MVPGS: Excavating multi-view priors for gaussian splatting from sparse input views. In *European Conference on Computer Vision*, 203–220. Springer.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10371–10381.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything V2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9043–9053.

Younes, M.; Ouasfi, A.; and Boukhayma, A. 2024. SparseCraft: Few-shot neural reconstruction through stereopsis guided geometric linearization. In *European Conference on Computer Vision*, 37–56. Springer.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35: 25018–25032.

Yu, Z.; Sattler, T.; and Geiger, A. 2024. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (ToG)*, 43(6): 1–13.

Zhang, J.; Zhang, Y.; Tosi, F.; Gu, M.; Li, J.; Yu, X.; Zheng, J.; Bai, X.; and Poggi, M. 2025a. Eve3D: Elevating Vision Models for Enhanced 3D Surface Reconstruction via Gaussian Splatting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhang, W.; Yang, Y.; Huang, H.; Han, L.; Shi, K.; Liu, Y.-S.; and Han, Z. 2025b. MonoInstance: Enhancing Monocular Priors via Multi-view Instance Alignment for Neural Rendering and Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, W.; Zhou, J.; Geng, H.; Zhang, W.; and Liu, Y.-S. 2025c. GAP: Gaussianize Any Point Clouds with Text Guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*.