GCGS: Obtaining Accurate Surface via Geometric-Consistent Gaussian Splatting from Sparse Input

Beiqi Chen, Chenxu Li, and Jinhe Su*
School of Computer Engineering
Jimei University
Xiamen 361021, China
{beiqichen, by_chenxu_li, sujh}@jmu.edu.cn

***Abstract*—3D surface reconstruction from sparse views is a challenging task in computer vision. Recently, there have been many improvements for Gaussian Splatting to extract high-quality geometry from multi-view inputs. However, this ability degrades when faced with sparse viewpoints. Lacking robust multi-view geometric constraints, it struggles to reconstruct complete surfaces in textureless regions and overfitting to the limited input views leads to geometric inaccuracies. This paper introduces GCGS, a method for acquiring accurate geometry from sparse views by introducing geometric consistency feature alignment and normal consistency constraints. Specifically, our method applies a multi-view consistency constraint to the projected depth features, with the weighted based on the geometric relationships between views. Building on this, we utilize monocular estimated normals to supervise the rendered normals, facilitating the reconstruction of detailed surfaces. These optimization strategies are built on a dense initial point cloud generated by the MASt3R model. Experimental results on the DTU and BlendedMVS datasets demonstrate that our method can reconstruct complete and detailed surfaces. It outperforms the leading NeRF-based method Neusurf by 0.04 in mean Chamfer Distance.**

***Index Terms*—Surface Reconstruction, Gaussian Splatting.**

I. Introduction

As a fundamental and persistent challenge in computer vision, 3D surface reconstruction is critical for a wide range of applications, including autonomous driving, robotics, and virtual reality. Traditional methods, such as Multi-View Stereo (MVS), have achieved considerable success, but they still struggle to establish reliable feature correspondences in texture-less regions. Recent progress in this domain has been largely propelled by breakthroughs in neural rendering. In particular, methods based on Neural Radiance Fields (NeRF) [1] and 3D Gaussian Splatting (3DGS) [2] have demonstrated exceptional efficacy. However, these pioneering methods [3]–[5] were designed primarily for novel view synthesis (NVS), which excels in generating photorealistic rendering results. Their native representations, namely NeRF's volumetric density field and the unorganized 3D Gaussians of 3DGS, are not inherently optimized for extracting precise surfaces, making it difficult to extract a high-quality surface. To address this limitation, a significant body of subsequent work has focused on adapting these frameworks specifically for high-quality geometry reconstruction. NeRF-based approaches [6], [7] evolved to optimize an implicit Signed Distance Function (SDF) to produce smooth surfaces. Meanwhile, methods based on Gaussian Splatting also led to variants for acquiring high-quality geometry, such as SuGaR [8], and 2DGS [9]—which employs planar primitives better suited for representing surfaces.

However, the performance of these advanced methods [6]–[8], [10]–[12] degrades significantly when faced with sparse input views. The core issue lies in the difficulty of establishing robust multi-view geometric consistency from limited observational data. Without sufficient constraints, NeRF-based methods tend to overfit to the few training images, often resulting in distorted geometry or surfaces with holes and incompleteness. Gaussian Splatting encounters the same problem. These methods suffer from a lack of multi-view consistency, which prevents the optimized Gaussians from forming a coherent surface, leading to fragmented and noisy results. Moreover, these Gaussian Splatting–based methods heavily rely on the initial point clouds provided by Structure-from-Motion [13]. In sparse-view scenarios, SfM struggles to establish reliable feature correspondences, leading to sparse and noisy initial geometry. Consequently, the Gaussian primitives overfit the training views, resulting in substantial geometric noise, floating artifacts, and incomplete surfaces.

In this paper, we introduce GCGS: Geometric-Consistent Gaussian Splatting, a framework designed to overcome these limitations and achieve accurate surface reconstruction from as few as three input views. Our approach tackles the inherent limitations of sparse-view reconstruction by utilizing a dense geometric prior from the MASt3R [14] model for initialization, and by introducing strong multi-view constraints to govern the optimization. These strategies ensure the generation of a complete and coherent surface. The contributions of our work are as follows.

- We propose GCGS, a framework for surface reconstruction from sparse input using 2D Gaussian splatting.
- We utilize geometrically coherent feature alignment to

*Corresponding author: sujh@jmu.edu.cn

enforce multi-view consistency for a complete geometric reconstruction. We then introduce a normal consistency constraints to further enhance this geometry, yielding finer surface details.
- Our method achieves a mean Chamfer Distance of 0.95 on the DTU dataset, surpassing the leading NeRF-based approach Neusurf (0.99).

II. Related Work

A. Neural Implicit Reconstruction

Neural rendering learns implicit scene representations directly from images. Neural Radiance Fields [1] achieve photorealistic synthesis, but extracting surfaces via density thresholding often yields noisy results. Several variants have been proposed to address this issue. NeuS [6] unifies volume rendering with a Signed Distance Function, while VolSDF [7] models density as a transformed SDF for smoother surface extraction. To mitigate the reliance on dense viewpoints and extensive training time, recent works focus on sparse-view reconstruction. SparseNeuS [15] introduces 3D feature volumes to aggregate multi-view information. VolRecon [16] and ReTR [17] further leverage transformers to fuse global context. Alternatively, SparseNeRF [18] adopts local depth ranking losses from monocular predictors. While these methods improve robustness under limited observations, they often rely on complex external dependencies or struggle with extreme viewpoint changes.

B. Surface Reconstruction with Gaussian Splatting

3D Gaussian Splatting [2] provides an efficient explicit representation for high-fidelity scene synthesis, which has been increasingly adapted for surface reconstruction. Early attempts like SuGaR [8] introduce surface-alignment regularizations to facilitate mesh extraction via Poisson reconstruction. To ensure watertight geometry, a prominent trend integrates 3DGS with implicit neural representations: NeuSG [11] and GSDF [19] jointly optimize Gaussian primitives with an underlying SDF, leveraging the continuity of implicit fields to produce smooth surfaces. While effective, these hybrid approaches often reintroduce the computational overhead of volumetric sampling. To maintain rendering efficiency, explicit frameworks enhance geometry directly through stronger regularizations or external priors. PGSR [10] incorporates multi-view geometric constraints inspired by traditional MVS, while DN-Splatter [20] utilizes monocular depth and normal maps to guide Gaussian optimization. Alternatively, GOF [12] proposes a Gaussian opacity field to enable direct surface extraction from a learned level set. A pivotal development is the modification of the primitive itself; 2DGS [9] replaces volumetric ellipsoids with planar, disk-shaped primitives and employs a perspective-accurate splatting process to achieve multi-view consistency. However, despite these advancements, 2DGS remains susceptible to overfitting and geometric noise in sparse-view scenarios due to the lack of robust multi-view geometric constraints.

C. Learning-based Multi-View Stereo

Learning-based Multi-View Stereo, pioneered by MVSNet [21], revolutionized 3D reconstruction by introducing end-to-end differentiable architectures that construct 3D cost volumes via homography warping. While establishing a powerful paradigm, early methods relying on high-resolution cost volumes often suffer from significant memory and computational overhead. To improve efficiency and robustness, subsequent research has explored alternative architectures: TransMVSNet [22] adopts Transformers to aggregate long-range global context via inter-attention mechanisms, mitigating matching ambiguities in texture-less regions. More recently, the focus has transitioned toward visual foundation models such as DUSt3R [23] and MASt3R [14], which excel in sparse-view conditions where traditional geometry matching fails. These models leverage large-scale pre-training to learn robust 3D priors, typically employing transformer-based pairwise pointmap regression combined with sophisticated global alignment strategies. This pairwise-to-global methodology facilitates dense and geometrically accurate reconstruction even under extreme viewpoint variations. Given its superior performance in limited-view scenarios, MASt3R [14] serves as an ideal source for generating the dense initial point cloud required for our sparse-view reconstruction task.

III. Method

A. Preliminary: 2D Gaussian Splatting

Our framework is built upon 2D Gaussian Splatting [9]. Unlike traditional 3D Gaussian Splatting, which uses volumetric ellipsoids, 2DGS represents the scene with a set of planar, disk-shaped primitives. Each 2D Gaussian is defined by a central position $\mathbf{p}$, two principal tangential vectors $\mathbf{t}_u$ and $\mathbf{t}_v$, and a scaling vector (s_u, s_v). These parameters define a point $\mathbf{P}(u, v)$ on the local tangent plane as:

$$\mathbf{P}(u,v) = \mathbf{p} + s_u\mathbf{t}_u u + s_v\mathbf{t}_v v \tag{1}$$

The rendered depth $\hat{D}(\mathbf{r})$ for the ray is accumulated by alpha blending the intersection depths d_i, and the rendered normal $\hat{\mathbf{n}}(\mathbf{r})$ is also a weighted average of the individual primitive normals $\mathbf{n}_i$:

$$\hat{D}(\mathbf{r}) = \frac{\sum_{i=1} \omega_i d_i}{\sum_{i=1} \omega_i + \epsilon}, \quad \hat{\mathbf{n}}(\mathbf{r}) = \sum_{i=1} \mathbf{n}_i \omega_i \tag{2}$$

B. MASt3R-Guided Dense Initialization

High-quality Gaussian Splatting relies on accurate initial point clouds, yet traditional Structure-from-Motion often yields impoverished and noisy geometric scaffolds in sparse-view settings due to the failure of robust feature matching. As shown in Fig. 1, to circumvent this initialization bottleneck, we leverage MASt3R [14] to generate a dense and geometrically consistent foundation. Given N sparse views $\{I_i\}$ and poses $\{\mathbf{T}_i\}$, the process follows the workflow:

$$\text{MASt3R: } (\{I_i\}, \{\mathbf{T}_i\})_{i=1}^N \rightarrow \{D_i\}_{i=1}^N \rightarrow \mathcal{P} \tag{3}$$

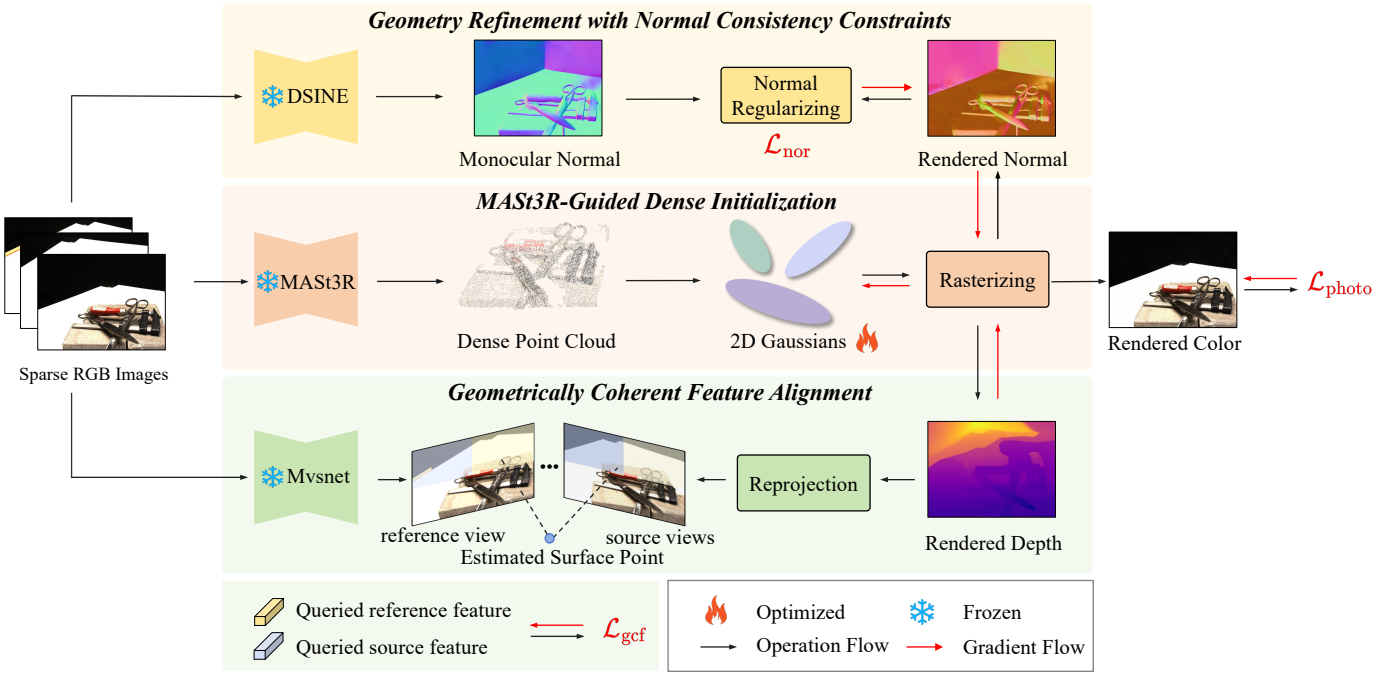

Fig. 1. Overview of the proposed GCGS pipeline. We leverage the MASt3R foundation model to generate a dense point cloud for geometric initialization. Then we propose a synergistic optimization strategy using geometrically coherent feature alignment for cross-view consistency and monocular normal priors to supervise primitive orientations.

Specifically, we utilize the pre-trained MASt3R to regress high-quality depth maps $\{D_i\}$ for each view. These depth pixels are then back-projected into 3D space and fused into a globally consistent point cloud $\mathcal{P} = \{\mathbf{p}_j\}_{j=1}^{M}$, defining the initial mean positions for our 2D Gaussian primitives. Building upon this scaffold, we sample initial colors $\mathbf{c}_j$ from the source images, set opacity α to a minimal value, and determine the initial scale $\mathbf{s}$ based on nearest-neighbor distances in $\mathcal{P}$.

C. Geometrically Coherent Feature Alignment

The optimization of 2D Gaussian primitives is fundamentally driven by photometric loss, making them prone to overfitting the colors of the training views. Under extremely sparse input conditions, the scene representation can easily memorize the appearance from these few perspectives, leading to significant misalignments in the underlying geometric structure. To address this issue, we introduce a viewpoint-aware feature consistency supervision to regularize the geometry and prevent this overfitting. While our initialization stage generates an initial structure for the scene, it requires further refinement. Inspired by MVSDF [24], we propose our geometrically coherent feature loss. The core insight is to dynamically prioritize feature correspondences that are geometrically more reliable. We first compute the geometric consistency between the reference view and the source view, which assigns higher weights to source views that are more geometrically consistent with the reference view when calculating multi-view feature discrepancies.

As illustrated in Fig. 2, Our process begins with a pixel p_{ref} in the reference view image I_{ref}. With its rendered depth

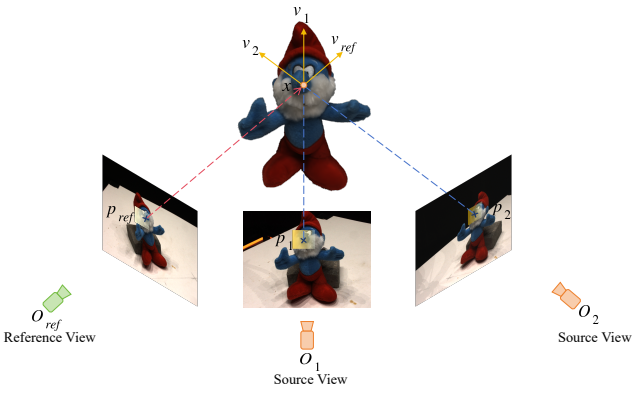

Fig. 2. Illustration of the geometrically coherent feature alignment. Features corresponding to a 3D point $\mathbf{x}$ are compared between a reference view (p_{ref}) and a source view (p_i). This comparison is weighted by the geometric coherence, which is derived from their respective viewing vectors ($\mathbf{v}_{\text{ref}}$, $\mathbf{v}_i$).

$\hat{D}_{p_{\text{ref}}}$, we can calculate the corresponding spatial point $\mathbf{x}$ in the world coordinate system by:

$$\mathbf{x} = O_{\text{ref}} + \hat{D}_{p_{\text{ref}}} \cdot \mathbf{d}_{p_{\text{ref}}} \tag{4}$$

where O_{ref} is the optical center of the reference camera, and $\mathbf{d}_{p_{\text{ref}}}$ is the normalized direction vector of the ray passing through pixel p_{ref}.

With the 3D point $\mathbf{x}$ established, we then quantify the geometric consistency between the reference view and each source view I_i. To do this, we first define the normalized

viewing vectors from each camera center to this point:

$$\mathbf{v}_{\text{ref}} = \frac{\mathbf{x} - O_{\text{ref}}}{\|\mathbf{x} - O_{\text{ref}}\|}, \quad \mathbf{v}_i = \frac{\mathbf{x} - O_i}{\|\mathbf{x} - O_i\|} \tag{5}$$

where O_{ref} and O_i are the optical centers of the reference and the i-th source camera, respectively. From these vectors, we define a geometric weight $w(\mathbf{x}, I_i)$ as the clamped dot product:

$$w(\mathbf{x}, I_i) = \max(0, \mathbf{v}_{\text{ref}} \cdot \mathbf{v}_i) \tag{6}$$

This weight approaches 1 when the viewing directions are aligned and 0 when they are orthogonal or opposing, thereby prioritizing views with high angular consistency.

Next, to compute the feature consistency, we project the 3D point $\mathbf{x}$ onto the image plane of a source view I_i to find the corresponding pixel coordinate p_i:

$$p_i = K_i P_i^{-1} \mathbf{x} \tag{7}$$

where K_i and P_i represent the intrinsic matrix and camera pose of the source view I_i, respectively.

The feature loss for a single source view is then defined by comparing the features sampled at the reference pixel p_{ref} and the projected source pixel p_i:

$$\mathcal{L}_{\text{feat}}(\mathbf{x}, I_i) = 1 - \frac{f_{\text{ref}}(p_{\text{ref}}) \cdot f_i(p_i)}{\|f_{\text{ref}}(p_{\text{ref}})\| \cdot \|f_i(p_i)\|} \tag{8}$$

where $f_{\text{ref}}(\cdot)$ and $f_i(\cdot)$ are the feature vectors sampled from their respective feature maps.

Finally, we combine the geometric weight and the feature loss into a single objective function. In practice, our final objective $\mathcal{L}_{\text{gcf}}$ is computed as the mean loss over a batch of 3D points $\mathcal{P}$. The total loss is formulated as the average of the geometrically-weighted feature consistency losses summed over all source views:

$$\mathcal{L}_{\text{gcf}} = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{x} \in \mathcal{P}} \sum_{i=1}^{N} w(\mathbf{x}, I_i) \cdot \mathcal{L}_{\text{feat}}(\mathbf{x}, I_i) \tag{9}$$

By minimizing this objective function, our model is encouraged to learn a feature representation that is not only visually consistent but also geometrically plausible.

D. Geometry Refinement with Normal Consistency Constraints

While the geometrically coherent feature alignment establishes a globally consistent structure, it is insufficient for recovering fine-grained surface details. We address this by introducing an explicit normal supervision signal derived from a pre-trained monocular normal estimation network, which provides high-quality monocular normals. We then propose a dual-component loss strategy to align our rendered normals with these monocular normals. This approach ensures both point-wise accuracy, enforced by a normal consistency loss, and local structural fidelity, enforced by a normal gradient consistency loss.

We enforce a normal consistency loss. This loss minimizes the L1 distance between the rendered normal map and the monocular normal map over a set of valid pixels, encouraging the rendered normals to match the monocular normal values. The loss is formulated as:

$$\mathcal{L}_{\text{nc}} = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} \|\hat{\mathbf{n}}_k - \mathbf{n}_{\text{m},k}\|_1 \tag{10}$$

where $\mathcal{M}$ is the set of pixels where a valid monocular normal exists, $\hat{\mathbf{n}}_k$ is the rendered normal at pixel k, and $\mathbf{n}_{\text{m},k}$ is the corresponding monocular normal.

To ensure local smoothness and preserve sharp features, we introduce a normal gradient consistency loss. This loss enforces that the spatial gradients of the rendered normal map match those of the monocular normals. We apply a Sobel operator to compute the gradients and define the loss as the Mean Squared Error between them:

$$\mathcal{L}_{\text{ngc}} = \mathbb{E}\left[\|\nabla\hat{\mathbf{n}} - \nabla\mathbf{n}_{\text{m}}\|^2\right] \tag{11}$$

This preserves local curvature and enhances fine details. Our final normal supervision loss is a weighted sum of these two components:

$$\mathcal{L}_{\text{nor}} = \lambda_{\text{nc}}\mathcal{L}_{\text{nc}} + \lambda_{\text{ngc}}\mathcal{L}_{\text{ngc}} \tag{12}$$

E. Loss Function

Our final training objective is a composite loss function that combines a photometric reconstruction term with our proposed geometric regularization terms, as well as the original regularizers from 2DGS. The overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_{\text{gcf}}\mathcal{L}_{\text{gcf}} + \lambda_{\text{nor}}\mathcal{L}_{\text{nor}} + \lambda_d\mathcal{L}_d \tag{13}$$

where $\mathcal{L}_{\text{gcf}}$ is the Geometrically Coherent Feature Loss from Section 3.3, and $\mathcal{L}_{\text{nor}}$ denotes the normal supervision loss from Section 3.4. The hyperparameters λ_{gcf}, λ_{nor}, and λ_d, control the weight of the geometric constraints.

The term $\mathcal{L}_{\text{photo}}$ is the photometric loss, which is a combination of an $\mathcal{L}_1$ loss and a D-SSIM term to ensure visual fidelity. Furthermore, we incorporate the depth distortion loss $\mathcal{L}_d$ from the original 2DGS framework [9] to encourage the concentration of Gaussians along rays.

IV. Experiments

A. Benchmark Comparisons

In this section, we validate the effectiveness of our proposed method for sparse-view surface reconstruction. We conduct extensive qualitative and quantitative experiments on the public DTU and BlendedMVS datasets. Subsequently, we perform ablation study to verify the contribution of each individual component of our framework.

B. Datasets and Metrics

DTU Dataset. We conduct our evaluation on the widely-used DTU dataset [25], which serves as a standard benchmark for multi-view 3D reconstruction tasks due to its high-quality ground-truth scans and controlled lighting conditions. To ensure a fair and direct comparison, we use the same experimental setup as previous sparse-view works, such as SparseNeuS

Table I
QUANTITATIVE COMPARISON OF CHAMFER DISTANCE (CD↓) ON THE DTU DATASET. THE BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN BOLD.

Scan ID	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
COLMAP	0.90	2.89	1.63	1.08	2.18	1.94	1.61	1.30	2.34	1.28	1.10	1.42	0.76	1.17	1.14	1.52
VolRecon	1.20	2.59	1.56	1.08	1.43	1.92	1.11	1.48	1.42	1.05	1.19	1.38	0.74	1.23	1.27	1.38
ReTR	1.05	2.31	1.44	0.98	1.18	1.52	0.88	1.35	1.30	0.87	1.07	0.77	0.59	1.05	1.12	1.17
C2F2NeuS	1.12	2.42	1.40	**0.75**	1.41	1.77	0.85	1.16	1.26	0.76	0.91	0.60	0.46	0.88	0.92	1.11
NeuS	4.57	4.49	3.97	4.32	4.63	1.95	4.68	3.83	4.15	2.50	1.52	6.47	1.26	5.57	6.11	4.00
VolSDF	4.03	4.21	6.12	0.91	8.24	1.73	2.74	1.82	5.14	3.09	2.08	4.81	0.60	3.51	2.18	3.41
MonoSDF	2.85	3.91	2.26	1.22	3.37	1.95	1.95	5.53	5.77	1.10	5.99	2.28	0.65	2.65	2.44	2.93
SparseNeus	1.29	2.27	1.57	0.88	1.61	1.86	1.06	1.27	1.42	1.07	0.99	0.87	0.54	1.15	1.18	1.27
NeuSurf	0.78	2.35	1.55	**0.75**	1.04	1.68	**0.60**	1.14	**0.98**	**0.70**	**0.74**	0.49	**0.39**	**0.75**	**0.86**	0.99
PGSR	3.33	3.89	2.77	1.06	4.38	2.10	1.00	1.48	2.21	0.96	1.83	2.33	0.68	1.19	1.21	2.03
2DGS	2.06	3.36	3.47	1.11	3.43	1.82	1.54	2.11	1.99	1.37	2.31	3.95	1.22	3.55	2.16	2.36
Ours	**0.70**	**1.89**	**1.10**	0.81	**1.02**	**1.37**	0.64	**1.12**	1.28	0.76	0.89	**0.48**	0.43	0.83	0.97	**0.95**

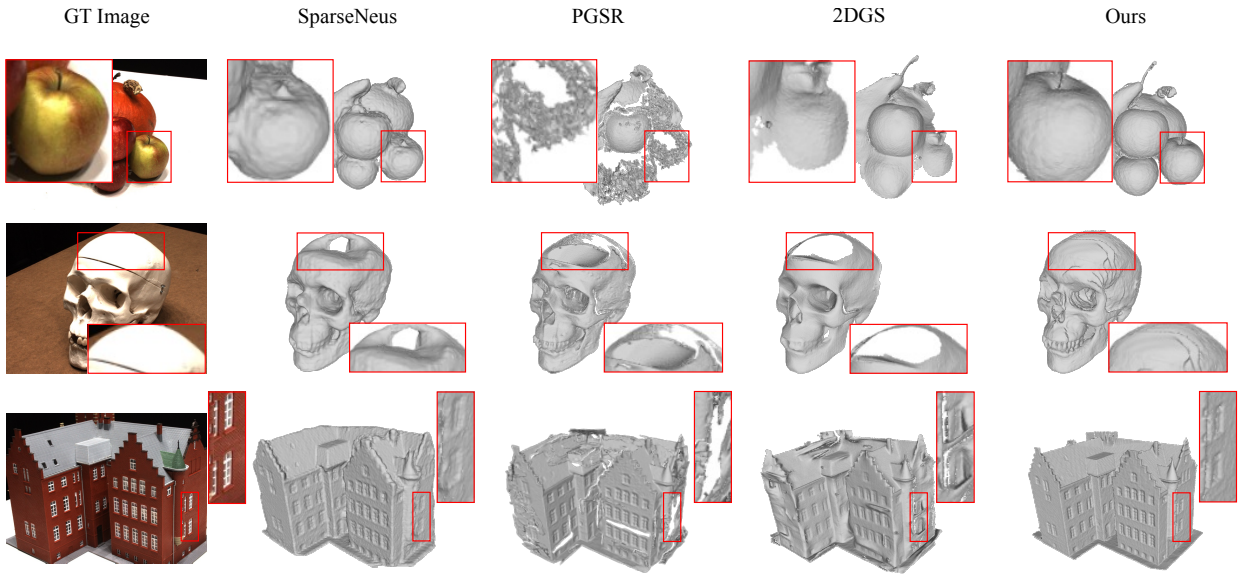

Fig. 3. Reconstruction completeness on the DTU dataset. Our method produces a complete and coherent surface in textureless regions

[15]. Specifically, we select the 15 scenes commonly used for this evaluation protocol. For each of these scenes, we test our method under a 3-view reconstruction scenario, utilizing views 23, 24, and 33. All experiments are performed using the original, high-resolution images of 1600×1200 to validate our method's robustness.

Evaluation Metrics. For quantitative evaluation of the reconstruction quality, we follow standard practice and report the Chamfer Distance (CD) in mm between the reconstructed mesh and the ground-truth point cloud.

C. Implementation Details

Our method is built upon the 2DGS framework. We use the visual foundation model MASt3R to generate a dense initial point cloud, while all other parameters of the Gaussian primitives are initialized consistently with the original 2DGS. We set the number of training iterations to 10,000 for all experiments. During training, we downsample the images to a resolution of 800×600. For final mesh extraction, we render depth maps from the optimized primitives and fuse them using Truncated Signed Distance Fusion (TSDF). The hyperparameters for our loss functions are set as follows: $\lambda_{\text{gcf}} = 2.0$, $\lambda_{\text{nor}} = 0.25$, and $\lambda_d = 10000$. To extract deep features for our geometrically coherent feature alignment loss, we employ a pre-trained Vismvsnet [26] feature extractor. The normal priors are obtained from a pre-trained DSINE [27] monocular normal estimation network. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB of memory.

Comparison on DTU. We compare the performance of our method against both Gaussian-based and NeRF-based approaches for 3-view surface reconstruction on the DTU dataset. We benchmark against Gaussian-based methods such as PGSR [10] and our baseline 2DGS [9], as well as leading

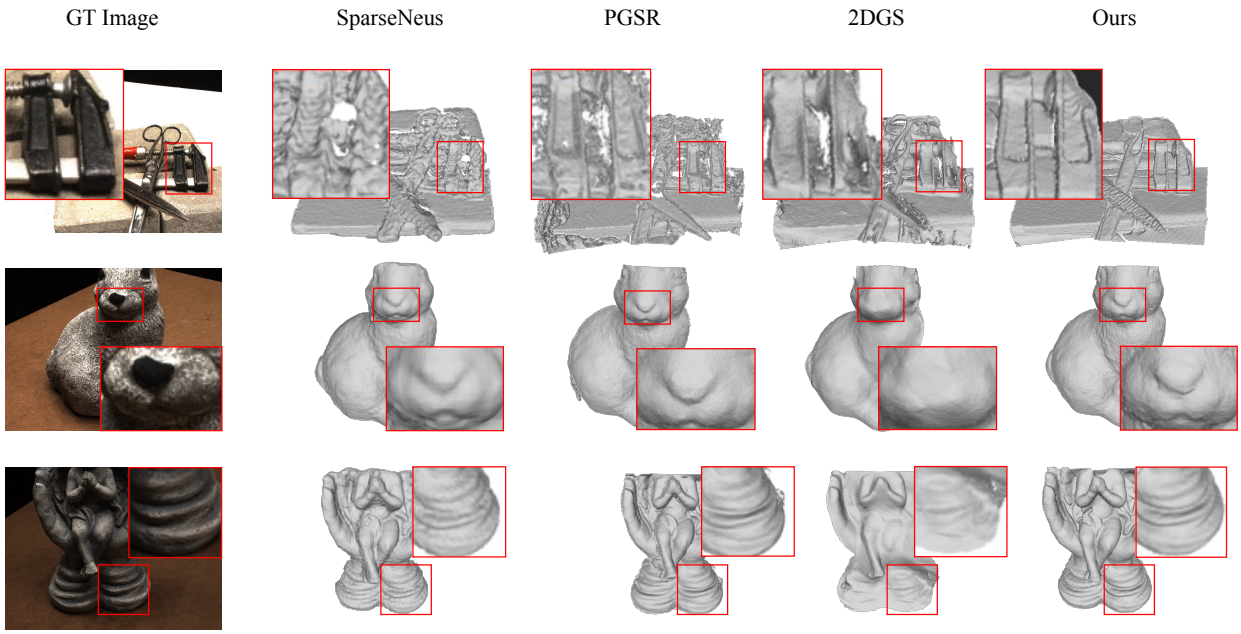

Fig. 4. Reconstruction of fine details on the DTU dataset. Our method produces surfaces rich in intricate geometric details and sharp edges.

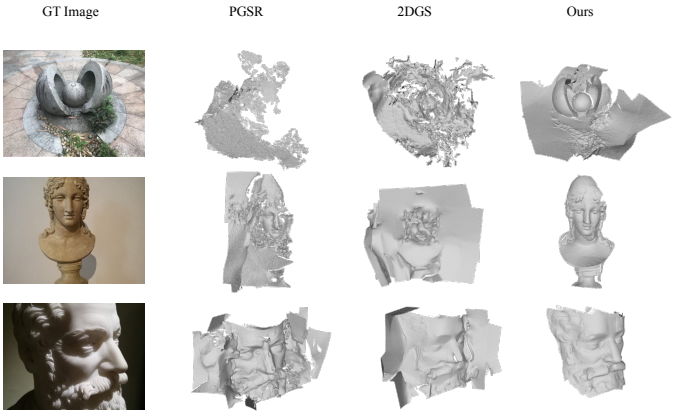

Fig. 5. Qualitative comparison on the BlendedMVS dataset.

NeRF-based methods including NeuSurf [28], SparseNeus [15], VolRecon [16], and C2F2NeuS [29]. As shown in Table I, our method demonstrates a substantial improvement over other Gaussian-based methods and outperforms our baseline, 2DGS. Our method demonstrates superior overall performance, achieving the best mean Chamfer Distance. Fig. 3 and Fig. 4 provide a qualitative comparison against prior methods. As illustrated in Fig. 3, our approach reconstructs a more complete geometry, whereas competing methods often yield fragmented results. Furthermore, Fig. 4 highlights our ability to recover finer details.

Generalization on BlendedMVS. We also present qualitative results on the BlendedMVS dataset in Fig. 5 to demonstrate the generalization ability of our method. The results show that our method produces reconstructions that are both more complete and richer in detail.

Table II
ABLATION STUDY ON THE KEY COMPONENTS OF OUR METHOD ON THE DTU DATASET.

MASt3R Init	$\mathcal{L}_{\text{gcf}}$	$\mathcal{L}_{\text{nor}}$	Mean CD (mm) ↓
✗	✗	✗	2.36
✓	✗	✓	2.11
✓	✓	✗	0.99
✓	✓	✓	**0.95**

D. Ablation Study

We evaluate the effectiveness of our proposed components on the DTU dataset. By isolating the contributions of our geometric constraints and dense initialization, we demonstrate how GCGS effectively mitigates overfitting in sparse-view scenarios. The baseline 2DGS with SfM initialization yields a mean Chamfer Distance (CD) of 2.36 mm. Substituting this with our MASt3R-guided dense initialization and proposed losses significantly improves the reconstruction quality.

As shown in Table II, our complete method achieves the best performance with the lowest Mean CD of 0.95 mm. Omitting the normal supervision prior ($\mathcal{L}_{\text{nor}}$) slightly increases the CD to 0.99 mm, indicating its effectiveness in smoothing the geometry and recovering fine details. More significantly, removing the geometrically coherent feature loss ($\mathcal{L}_{\text{gcf}}$) drastically degrades the CD to 2.11 mm, highlighting its crucial foundational role in aligning 2D Gaussians with true surfaces.

In summary, the ablation study demonstrates that the MASt3R initialization, $\mathcal{L}_{\text{gcf}}$, and $\mathcal{L}_{\text{nor}}$ provide distinct yet complementary improvements, collectively enabling more robust and accurate reconstruction.

V. Conclusion

In this paper, we introduced GCGS, a framework designed to achieve accurate 3D surface reconstruction from sparse views. The core of our method is a synergistic optimization strategy that introduces a geometrically coherent feature alignment loss, which ensures multi-view consistency, and a monocular normal prior, which facilitates the recovery of fine-grained details. Extensive experiments conducted on the DTU and BlendedMVS datasets demonstrate the superiority of our approach. GCGS is shown to significantly improve both the completeness and detail of reconstructed surfaces, achieving leading performance. Our work provides a fast, detailed, and robust solution to the persistent challenge of 3D surface reconstruction in challenging sparse-view scenarios.

Acknowledgment

This research was funded in part by the Natural Science Foundation of Xiamen, China, under Grant 3502Z202373036; in part by the Open Competition for Innovative Projects of Xiamen, 3502Z20251012, in part by the National Natural Science Foundation of China under Grant 42371457; in part by the Natural Science Foundation of Fujian Province, China, under Grant 2025J01345.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5470–5479.

[4] Z. Yu, A. Chen, B. Huang, T. Sattler, and A. Geiger, "Mip-splatting: Alias-free 3d gaussian splatting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19 447–19 456.

[5] T. Lu, M. Yu, L. Xu, Y. Xiangli, L. Wang, D. Lin, and B. Dai, "Scaffold-gs: Structured 3d gaussians for view-adaptive rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 654–20 664.

[6] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[7] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," *Advances in neural information processing systems*, vol. 34, pp. 4805–4815, 2021.

[8] A. Guédon and V. Lepetit, "Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering," *CVPR*, 2024.

[9] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2d gaussian splatting for geometrically accurate radiance fields," in *ACM SIGGRAPH 2024 conference papers*, 2024, pp. 1–11.

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

[11] H. Chen, C. Li, and G. H. Lee, "Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance," *arXiv preprint arXiv:2312.00846*, 2023.

[12] Z. Yu, T. Sattler, and A. Geiger, "Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes," *ACM Transactions on Graphics*, 2024.

[13] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[14] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.

[15] X. Long, C. Lin, P. Wang, T. Komura, and W. Wang, "Sparseneus: Fast generalizable neural surface reconstruction from sparse views," in *European Conference on Computer Vision*. Springer, 2022, pp. 210–227.

[16] Y. Ren, F. Wang, T. Zhang, M. Pollefeys, and S. Süsstrunk, "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*, 2023, pp. 16 685–16 695.

[17] Y. Liang, H. He, and Y. Chen, "Retr: Modeling rendering via transformer for generalizable neural surface reconstruction," *Advances in Neural Information Processing Systems*, vol. 36, pp. 62 332–62 351, 2023.

[18] G. Wang, Z. Chen, C. C. Loy, and Z. Liu, "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[19] M. Yu, T. Lu, L. Xu, L. Jiang, Y. Xiangli, and B. Dai, "Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction," *Advances in Neural Information Processing Systems*, vol. 37, pp. 129 507–129 530, 2024.

[20] M. Turkulainen, X. Ren, I. Melekhov, O. Seiskari, E. Rahtu, and J. Kannala, "Dn-splatter: Depth and normal priors for gaussian splatting and meshing," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 2421–2431.

[21] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[22] Y. Ding, W. Yuan, Q. Zhu, H. Zhang, X. Liu, Y. Wang, and X. Liu, "Transmvsnet: Global context-aware multi-view stereo network with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8585–8594.

[23] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[24] J. Zhang, Y. Yao, and L. Quan, "Learning signed distance field for multi-view surface reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6525–6534.

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

[26] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network," *arXiv preprint arXiv:2008.07928*, 2020.

[27] G. Bae and A. J. Davison, "Rethinking inductive biases for surface normal estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9535–9545.

[28] H. Huang, Y. Wu, J. Zhou, G. Gao, M. Gu, and Y.-S. Liu, "Neusurf: On-surface priors for neural surface reconstruction from sparse input views," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 2312–2320.

[29] L. Xu, T. Guan, Y. Wang, W. Liu, Z. Zeng, J. Wang, and W. Yang, "C2f2neus: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 291–18 301.