MAC-Splat: Multi-Attribute Consistency for High-Fidelity Sparse-View Reconstruction

Jinqian Yang[1], Yichen Wu[2*], Wanhua Li[3], Haokun Lin[4], Renzhen Wang[5], Xiangchu Feng[1], and Xixi Jia[1*]

[1] Xidian University, China
[2] Harvard University, USA
[3] Nanyang Technological University, Singapore
[4] City University of Hong Kong, China
[5] Xi'an Jiaotong University, China

Abstract. Reconstructing high-fidelity 3D scenes from sparse views remains a central problem in generalizable neural rendering. Existing generalizable 3D Gaussian Splatting (3DGS) methods often exhibit geometric artifacts in sparse-view settings, since supervision based solely on 2D photometric losses cannot resolve depth and correspondence ambiguities. To address this issue, we propose MAC-Splat, a training framework built around direct 3D consistency supervision. MAC-Splat builds on the MASt3R geometric backbone and a frozen DINOv3 encoder to obtain semantically informed 2D correspondences, which serve as geometric anchors for 3D supervision. Using these anchors, we define the Multi-Attribute Consistency (MAC) loss. This objective jointly regularizes the 3D attributes of matched Gaussians, including their position, shape, and appearance, by enforcing agreement in a common world coordinate frame. The formulation is robust to outliers and respects the geometry of covariance matrices, which leads to stable training under sparse-view conditions. Experiments on ScanNet++ show that MAC-Splat outperforms strong baselines, with particularly large gains under different overlap regimes. In particular, it improves average PSNR over Splatt3R by more than 4.5 dB, reduces LPIPS, and maintains performance as the camera pose gap increases. These results indicate that a direct, multi-attribute 3D consistency objective, when combined with high-quality correspondences, is effective for addressing the ill-posed sparse-view reconstruction problem.

Keywords: 3D Gaussian Splatting · Sparse-View Reconstruction · Geometric Consistency · Semantic Guidance

1 Introduction

Capturing and rendering photorealistic 3D scenes is a central goal in computer vision [3, 37, 51, 57], with applications ranging from virtual reality [43] to

* Corresponding authors: Yichen Wu (wuyichen.am97@gmail.com) and Xixi Jia (xxjia@xidian.edu.cn).

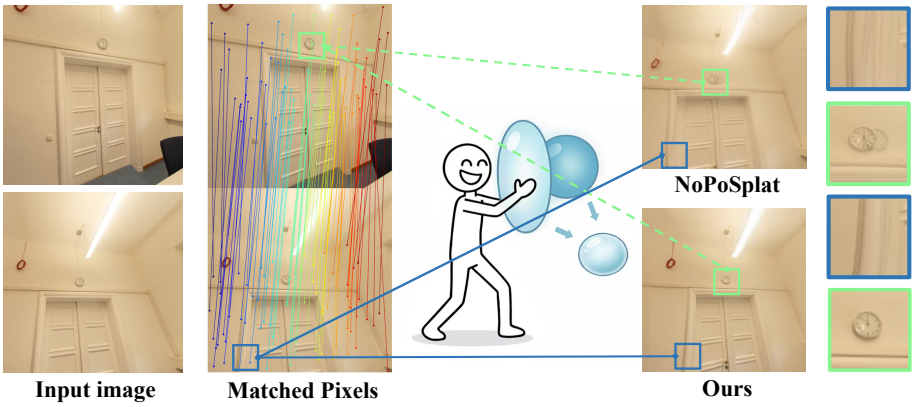

Fig. 1: Matched-pixel constraints for 3D Gaussians. Pixel correspondences are used as anchors to enforce consistent Gaussian attributes, which helps mitigate common sparse-view artifacts compared to prior generalizable 3DGS methods.

robotics [41]. Implicit neural representations such as NeRF [1, 32] have demonstrated striking visual quality; however, their training and inference costs remain high, which limits real-time deployment and scalability in practical systems.

To address this limitation, 3DGS and its variants [13, 20, 25, 36, 63] offer a powerful solution. They achieve real-time, high-quality rendering by using an explicit representation and an efficient GPU rasterization pipeline [49]. However, the strong performance of 3DGS usually relies on dense multi-view inputs with sufficient view overlap. In practice, when only a handful of images are available, reconstruction must proceed under sparse-view conditions, often with wide baselines and limited overlap [2, 9, 16, 21, 38, 44, 48, 52, 53, 56, 60, 67, 68]. This scarcity reduces geometric cues and thus impedes the formation of coherent 3D structures. Therefore, generalizing 3DGS to sparse-view regimes has become a key and rapidly evolving research focus [7, 29].

Despite the importance of this problem, existing mitigation strategies remain inadequate [7, 61]. First, methods that inject geometric priors from depth estimators [14, 22, 30, 45] or leverage foundation models for indirect perceptual guidance [42, 46] seldom provide pixel-level, cross-view anchors that directly tie corresponding 3D primitives, which are needed for strong geometric consistency. Second, even techniques that enforce intrinsic consistency [11, 34, 64] are fundamentally fragile, as their success depends on initial feature matches that are unreliable in sparse settings. Taken together, these shortcomings reflect a common difficulty: sparse-view reconstruction is highly under-constrained, and 2D photometric supervision alone is often insufficient to disambiguate geometry under occlusion and view-dependent appearance. Moreover, reliable cross-view correspondences are harder to obtain when overlap is limited, which further weakens any consistency regularization that depends on matches [50]. Consequently, as

shown in Fig. 1, reconstructions exhibit severe artifacts, including distorted 3D shapes and floating debris.

Our work addresses this limitation by complementing 2D rendering losses with correspondence-grounded 3D consistency constraints. Inspired by the observation that sparse-view ambiguities require a supervisory signal that understands what a scene is rather than only what it looks like, we leverage semantic understanding to guide the learning of geometrically coherent 3D structure. Accordingly, we design training to resolve geometric ambiguities by jointly enforcing cross-view consistency over 3D position, shape, and appearance.

Concretely, we present MAC-Splat, a framework grounded in Semantically-Guided Geometric Consistency. The approach integrates a robust correspondence stage with our core contribution, the MAC loss, which leverages semantically augmented correspondence features to enforce holistic 3D consistency. Specifically, we extract sparse 2D correspondences using a strong geometric backbone. We then enrich these matches with part-level features from a frozen DINOv3 encoder via a lightweight Residual Semantic Fusion module. This semantic augmentation improves correspondence reliability and confidence estimation, which in turn strengthens the downstream 3D consistency supervision. The loss then lifts these 2D anchors into 3D and enforces direct, confidence-weighted consistency on the corresponding Gaussians in a shared world frame, using ground-truth poses during training and MASt3R-estimated poses at test time. It jointly regularizes Gaussian centroids for positional consistency, covariance parameters for shape consistency, and opacity and color for appearance consistency. This supervision resolves ambiguities that evade conventional losses and yields reconstructions that are both geometrically coherent and semantically plausible. As shown in Fig. 1, matched pixels help maintain consistent Gaussian attributes and reduce common artifacts. In summary, our main contributions are threefold:

- We introduce a novel paradigm, Semantically-Guided Geometric Consistency, which strategically uses semantic information to resolve ambiguities in the 3D optimization process for sparse-view reconstruction.
- We propose the MAC loss, a comprehensive objective that is empowered by semantically-augmented correspondences to jointly regularize the 3D position, shape, and appearance of Gaussian primitives.
- We design a Residual Semantic Fusion module that enriches geometric features with semantic priors, providing the foundational representation that enables our semantically-aware consistency objective.

2 Related Work

NeRF [32] and 3DGS [20] have demonstrated remarkable performance in 3D reconstruction and novel view synthesis. However, both paradigms traditionally rely on densely captured, accurately posed images and per-scene optimization, which limits scalability and real-world deployment. In practical scenarios such as mobile AR/VR, robotics, or clinical imaging [19,31,55], only a few views may

be available, and camera poses can be unreliable. This challenge has driven the transition from per-scene optimization to generalizable feed-forward models capable of predicting 3D representations directly from sparse inputs with minimal test-time cost [6, 22, 61, 68].

Generalizable Feed-Forward 3DGS. Recent advances replace per-scene optimization with feed-forward networks that predict 3D Gaussians in a single pass. Early methods such as PixelSplat [2] and MVSplat [5] assume accurate SfM poses, while pose-free approaches (e.g., Splatt3R [40], NoPoSplat [58], AnySplat [18], SPFSplat [15]) extend the paradigm to uncalibrated inputs by leveraging foundation models such as DUSt3R [47] or MASt3R [23] to predict a canonical-space representation and infer or bypass poses. Despite improved practicality, sparse-view and wide-baseline reconstruction remains ill-posed: 2D rendering supervision is often insufficient to resolve occlusion and view-dependent ambiguity, and correspondence quality degrades with limited overlap, allowing matching errors to propagate into 3D Gaussians and produce floaters or duplicated surfaces. This motivates explicit 3D regularization beyond feed-forward prediction.

External Geometric Priors. Many such priors are applied either in posed pipelines or require pose estimates to align multi-view cues. A prominent line of work stabilizes sparse-view optimization by injecting explicit geometric cues. Depth and normal maps predicted by monocular estimators are frequently used to regularize Gaussian placement and suppress floaters [22, 30, 33, 45]. Some approaches further integrate multi-view stereo signals or surface extraction pipelines to improve structural coherence [5, 10, 27]. These priors effectively constrain macroscopic surface layout and improve reconstruction stability. However, they provide proxy geometry rather than explicitly binding corresponding 3D primitives across views, and their effectiveness depends on the reliability of external estimators and pose alignment. Monocular predictions may introduce scale ambiguity or bias in textureless or specular regions, while multi-view stereo cues degrade under sparse overlap or wide baselines.

Representation-Level and Foundation Priors. Another direction leverages large-scale foundation models to enhance reconstruction robustness. One line of work introduces diffusion-based priors to regularize rendered views toward realistic image statistics [8, 28, 49]. Another line uses vision transformer features or vision-language signals to provide richer semantic descriptors for reconstruction networks [4, 24, 26, 42, 46, 54, 62, 69]. These priors often improve global appearance coherence and feature distinctiveness, especially in visually ambiguous regions. However, they typically operate at the representation or image-projection level and do not explicitly couple corresponding 3D primitives across views. As a result, identity-level and attribute-level consistency is addressed only indirectly through image-space supervision.

Intrinsic Cross-View Regularization. Beyond injecting external cues, several works introduce intrinsic training objectives to promote structural robustness. Stochastic Gaussian dropout reduces overfitting [34, 35, 66], and binocular guidance or multi-view surface constraints encourage view-consistent predic-

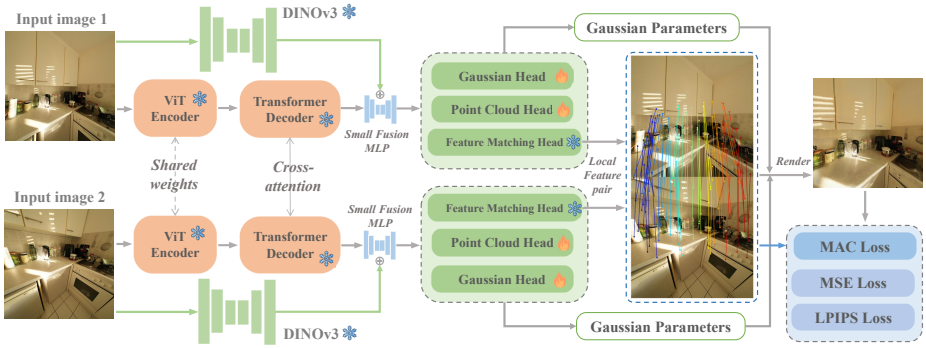

Fig. 2: Architectural overview of MAC-Splat. The framework augments a MASt3R backbone with features from a DINOv3 encoder through a residual fusion module to produce semantically enriched descriptors. A multi-stage filtering procedure yields sparse, high-confidence 2D correspondences. These matches anchor the Multi-Attribute Consistency (MAC) loss, which compares the world-space position, shape, and appearance of the associated Gaussians.

tions [11, 12]. Although effective, these objectives typically enforce agreement at the projection or surface level. Explicit identity-level coupling between matched 3D primitives is less commonly enforced, so primitive attributes are still not directly constrained to be consistent across views.

The preceding analysis reveals a common gap in sparse-view generalizable 3DGS: most methods improve reconstruction via proxy geometry, image-level priors, or projection-level consistency, but rarely enforce explicit identity-level coupling between corresponding 3D primitives across views. MAC-Splat addresses this gap by first improving correspondence reliability using a strong geometric backbone augmented with DINOv3 features, and then applying a Multi-Attribute Consistency loss that directly regularizes matched Gaussians in 3D, jointly aligning their position, shape, and appearance.

3 Methodology

Our methodology targets sparse-view 3D reconstruction, where limited viewpoints introduce geometric ambiguity. We decompose the task into coupled components: establishing high-fidelity 2D correspondences and leveraging these correspondences to enforce direct 3D geometric consistency. The proposed MAC-Splat pipeline (Fig. 2) integrates a feature fusion module, a correspondence filtering procedure, and a multi-attribute 3D consistency loss to realize this.

3.1 High-Fidelity Correspondence Establishment

To supply reliable geometric evidence for 3D consistency, we first construct a compact yet highly trustworthy set of 2D anchors that capture strong cross-

view relations. Because accurate 3D supervision hinges on these anchors, we design a multi-stage procedure that progressively filters and refines candidate matches, yielding a sparse set of high-confidence correspondences $\mathcal{M}$ for each image pair. Concretely, we use MASt3R [23] as the geometric matching backbone and enhance its descriptors with a lightweight residual semantic fusion module. This module injects dense semantic features from a frozen DINOv3 [39] encoder, providing corrective cues in ambiguous regions. Given a baseline descriptor f_{dec} from the MASt3R decoder and a spatially aligned DINOv3 feature f_{dino}, a small fusion MLP g_{fuse} predicts a residual $r = g_{\text{fuse}}([f_{\text{dec}}, f_{\text{dino}}])$, which is added to the baseline descriptor, $f' = f_{\text{dec}} + r$. DINOv3 features are resized to the decoder grid via bilinear interpolation and projected to the descriptor dimension with a linear layer before fusion. The resulting descriptor maps for the two views, $\mathcal{F}_1$ and $\mathcal{F}_2$, are more distinctive while remaining compatible with the original MASt3R features. Residual fusion is applied only at the final decoder layer, using a single MLP shared across spatial locations. We freeze the MASt3R matching head weights to preserve pretrained behavior, but reciprocal NN matches depend on the input descriptors, so feeding fused descriptors f' changes the correspondences.

Building on these enhanced descriptors, we generate dense candidate correspondences by identifying reciprocal nearest neighbors between $\mathcal{F}_1$ and $\mathcal{F}_2$. A pair of pixels $(\boldsymbol{p}_i^1, \boldsymbol{p}_j^2)$ from images $\mathbf{I}_1$ and $\mathbf{I}_2$ constitutes a reciprocal match if their respective descriptors $f'_{1,i} \in \mathcal{F}_1$ and $f'_{2,j} \in \mathcal{F}_2$ satisfy,

$$i = \arg\min_k D(f'_{2,j}, f'_{1,k}), \quad j = \arg\min_k D(f'_{1,i}, f'_{2,k}), \tag{1}$$

where $D(\cdot, \cdot)$ denotes cosine distance in feature space. This bidirectional check enforces a symmetry constraint and effectively prunes many ambiguous matches, especially in occluded or repetitive regions.

Finally, we refine the correspondences using the network's predicted descriptor confidence. Following MASt3R [23], we use a confidence head shared across all pixels, which is frozen and used only for correspondence gating. For each descriptor f', this head outputs a scalar confidence value $c(f')$ that estimates its distinctiveness. We define the joint confidence of a match as the product of the confidences at its two endpoints, $w_{ij} = c(f'_{1,i})\, c(f'_{2,j})$, and we retain only those matches that satisfy $w_{ij} > \tau_{\text{conf}}$, thereby discarding correspondences in regions where the network expresses low certainty. Because the retained anchors are softly weighted by their joint confidence in the subsequent MAC loss and residual outliers are suppressed by a Huber penalty, our framework is highly robust to the exact choice of τ_{conf}. Furthermore, if no match survives this gating in extremely ambiguous regions, the MAC loss gracefully falls back to zero, allowing the network to rely solely on photometric supervision. As a result, the final output is a sparse set of high-fidelity 2D matches, $\mathcal{M} = \{\boldsymbol{p}_k^1 \leftrightarrow \boldsymbol{p}_k^2\}_{k=1}^K$, which serve as geometric anchors for the subsequent MAC loss on 3D Gaussian attributes. We construct $\mathcal{M}$ for each sampled image pair; in practice, using multiple pairs per scene provides overlapping anchors for the same Gaussians.

3.2 Direct Multi-Attribute 3D Consistency Loss

Given the 2D correspondences $\mathcal{M}$ from Section 3.1, we define a Multi-Attribute Consistency loss on 3D Gaussians. The loss uses the matches in $\mathcal{M}$ as anchors and penalizes differences in position, shape, and appearance between the Gaussians associated with each matched pixel pair. During training, we use ground-truth camera-to-world transformations T_{wc}^v to compare all attributes in a common world coordinate system; at test time, T_{wc}^v is obtained from MASt3R pose estimation. An overview of this pipeline, from matched pixels to world-space Gaussians and the corresponding MAC loss terms, is provided in Fig. 3.

From matched pixels to Gaussian pairs. For each training pair $(\mathbf{I}_1, \mathbf{I}_2)$, our Gaussian prediction head predicts, for each view, a dense map of Gaussian parameters aligned with the descriptor grid. At every location (u, v) in view v, we obtain a 3D centroid $\boldsymbol{\mu}_{\text{cam}}^v(u,v) \in \mathbb{R}^3$, a covariance matrix $\boldsymbol{\Sigma}_{\text{cam}}^v(u,v) \in \text{Sym}^+(3)$, an opacity $\alpha^v(u,v)$, and spherical harmonic coefficients $\boldsymbol{c}^v(u,v)$. The reciprocal matches in $\mathcal{M}$ are defined on the same grid: a match $(\boldsymbol{p}_k^1, \boldsymbol{p}_k^2)$ specifies two pixel coordinates in the parameter maps of I_1 and I_2, together with a batch index. For these coordinates, we obtain the associated Gaussians G_k^1 and G_k^2 by bilinearly sampling the predicted parameter maps using differentiable grid sampling. This yields, for each correspondence k, a pair of Gaussians,

$$G_k^v = \big(\boldsymbol{\mu}_{\text{cam}}^{v,k},\, \boldsymbol{\Sigma}_{\text{cam}}^{v,k},\, \alpha^{v,k},\, \boldsymbol{c}^{v,k}\big), \quad v \in \{1,2\},$$

where the superscript k denotes sampling at $\boldsymbol{p}_k^v$. In the following, we transform these parameters into the world coordinate system using the camera-to-world transforms T_{wc}^1 and T_{wc}^2, and then apply the positional, shape, and appearance consistency terms.

(*Positional consistency.*) We encourage the 3D centroids of matched Gaussians to coincide in world space. Given the camera-to-world transform $T_{wc}^v = [\mathbf{R}^v \;\; \boldsymbol{t}^v]$, the world-space centroid of G_k^v is,

$$\boldsymbol{\mu}_{\text{world}}^{v,k} = \mathbf{R}^v \boldsymbol{\mu}_{\text{cam}}^{v,k} + \boldsymbol{t}^v. \tag{2}$$

For each correspondence, we define a robust positional loss,

$$\mathcal{L}_{\text{pos}}^{(k)} = \mathcal{H}\big(\|\boldsymbol{\mu}_{\text{world}}^{1,k} - \boldsymbol{\mu}_{\text{world}}^{2,k}\|_1\big), \tag{3}$$

which is evaluated only when both centroids lie in front of their cameras ($z_{\text{cam}} > 0$). Here, $\mathcal{H}$ denotes the Huber loss.

(*Shape Consistency.*) To promote consistent 3D shape across views, we compare the covariance matrices that define each Gaussian. Since each covariance $\boldsymbol{\Sigma} \in \text{Sym}^+(3)$ is predicted in the camera frame, we first transfer it to the world frame using the rotation $\mathbf{R}$ from T_{wc},

$$\boldsymbol{\Sigma}_{\text{world}} = \mathbf{R}\, \boldsymbol{\Sigma}_{\text{cam}}\, \mathbf{R}^\top. \tag{4}$$

A Euclidean distance between covariance matrices does not respect the geometry of $\text{Sym}^+(3)$ and is not invariant to rotations or uniform scaling. To address this,

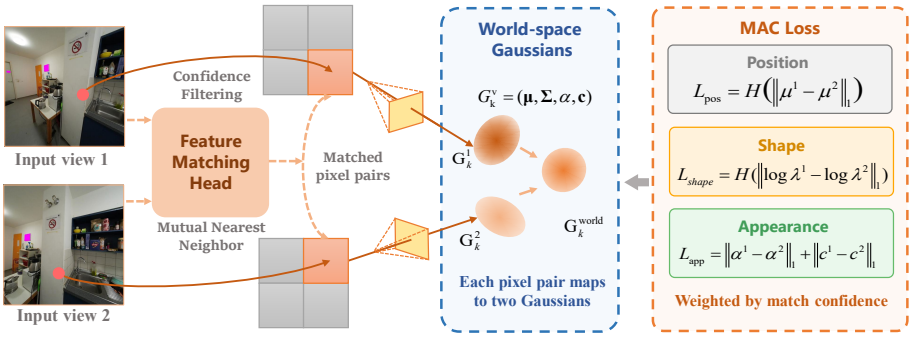

Fig. 3: From matched pixels to MAC loss. Each matched pixel pair selects two Gaussians in world space, and the MAC loss enforces consistency in their position, shape, and appearance.

we compare the logarithms of their eigenvalues. For a matched pair (G_k^1, G_k^2) with eigenvalues $\boldsymbol{\lambda}^{1,k}$ and $\boldsymbol{\lambda}^{2,k}$, the shape loss is,

$$\begin{aligned} \mathcal{L}_{\text{shape}}^{(k)} &= \mathcal{H}\left(\left\|\log \boldsymbol{\lambda}^{1,k} - \log \boldsymbol{\lambda}^{2,k}\right\|_1\right), \\ \boldsymbol{\lambda}^{v,k} &= \text{eig}\left(\boldsymbol{\Sigma}_{\text{world}}^{v,k}\right), \quad v \in \{1,2\}. \end{aligned} \tag{5}$$

This metric is rotation invariant because it depends only on the eigenvalues, and it measures relative differences in the principal axes through $\log(a/b) = \log(a) - \log(b)$. Consequently, this formulation is invariant to uniform global scaling and focuses exclusively on mismatches in anisotropy. By penalizing relative scale ratios rather than absolute scale differences, it actively prevents spurious stretching or flattening of the geometry across views.

(*Appearance Consistency.*) We regularize the photometric attributes by penalizing differences in opacity and spherical harmonic coefficients,

$$\mathcal{L}_{\text{app}}^{(k)} = \|\alpha^{1,k} - \alpha^{2,k}\|_1 + \|\boldsymbol{c}^{1,k} - \boldsymbol{c}^{2,k}\|_1. \tag{6}$$

Multi-Attribute Consistency Loss. The MAC loss combines the positional, shape, and appearance terms using the joint confidence w_k of each correspondence,

$$\mathcal{L}_{\text{MAC}} = \frac{\sum_{k=1}^{K} w_k \left(\lambda_p \mathcal{L}_{\text{pos}}^{(k)} + \lambda_s \mathcal{L}_{\text{shape}}^{(k)} + \lambda_a \mathcal{L}_{\text{app}}^{(k)}\right)}{\sum_{k=1}^{K} w_k + \epsilon}. \tag{7}$$

Here, w_k denotes the joint confidence of the k-th matched pixel pair, $\lambda_p, \lambda_s, \lambda_a$ are scalar weights for the three terms, and ϵ is a small constant that prevents division by zero.

Preventing Geometric Degeneracy. A critical theoretical property of our log-eigenvalue shape loss is its gradient behavior near geometric degeneracy. For a Huber-penalized log-difference, the magnitude of the gradient with respect to

a single eigenvalue λ scales inversely with the eigenvalue itself:

$$\left|\frac{\partial \mathcal{L}_{\text{shape}}}{\partial \lambda}\right| \propto \frac{1}{\lambda}. \tag{8}$$

As a Gaussian primitive becomes degenerate—meaning its smallest eigenvalue $\lambda_{\min}$ approaches zero and it collapses into a flat or needle-like shape—the gradient magnitude sharply increases. This creates a strong repulsive effect that actively pushes $\lambda_{\min}$ away from zero. This intrinsic anti-degeneracy mechanism heavily stabilizes the optimization landscape, explicitly discouraging the collapsed floating artifacts that frequently plague sparse-view reconstructions guided solely by photometric gradients.

3.3 Overall Training Objective

Having defined the MAC loss, we now describe the overall training objective of our model. We train the Gaussian prediction head and the residual fusion module with a combination of photometric loss and the proposed MAC loss, while keeping DINOv3 and the MASt3R matching heads frozen,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \lambda_{\text{MAC}} \mathcal{L}_{\text{MAC}}. \tag{9}$$

Here, $\mathcal{L}_{\text{photo}}$ is a weighted sum of an ℓ_2 reconstruction term and an LPIPS term [65], both applied to target views rendered from the predicted Gaussians. The scalar λ_{MAC} controls the strength of the 3D consistency regularization and is kept fixed across all experiments. Although the MASt3R correspondence loss can be added as an auxiliary term in principle, we set its weight to zero in all reported results.

4 Experiments

We evaluate MAC-Splat along four aspects: (i) quantitative and perceptual performance compared to state-of-the-art generalizable methods; (ii) the contribution of the MAC loss and residual semantic fusion; (iii) robustness under varying viewpoint separation; and (iv) zero-shot generalization from ScanNet++ to DTU.

4.1 Experimental Setup

Dataset and Difficulty Splits. We use ScanNet++ [59], a large-scale indoor dataset with accurate ground-truth geometry. Following Splatt3R [40], we group test pairs into four difficulty subsets using coverage thresholds (ϕ, ψ): **Close**, **Medium**, **Wide**, and **Very Wide**. These thresholds measure the fraction of depth-consistent visible pixels between context and target views. Larger (ϕ, ψ) values correspond to shorter baselines and higher view overlap, while lower values

Table 1: Quantitative results on ScanNet++. We report PSNR (↑), SSIM (↑), and LPIPS (↓) across four difficulty splits controlled by thresholds (ϕ, ψ) (higher means more overlap and smaller baseline). Best results are in **bold**.

Method	Close (ϕ=0.9, ψ=0.9)			Medium (ϕ=0.7, ψ=0.7)			Wide (ϕ=0.5, ψ=0.5)			Very Wide (ϕ=0.3, ψ=0.3)		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Splatt3R [40]	18.89	0.757	0.233	18.79	0.760	0.228	16.71	0.686	0.289	16.16	0.662	0.289
MASt3R [23]	18.44	0.722	0.292	15.23	0.648	0.359	15.72	0.719	0.295	15.82	0.608	0.278
DUSt3R [47]	16.46	0.693	0.279	15.96	0.652	0.312	15.64	0.704	0.323	15.23	0.612	0.262
NoPoSplat [58]	19.97	0.722	0.251	19.66	0.801	0.263	18.62	0.745	0.276	18.74	0.730	0.239
MVSplat [5]	22.64	0.802	0.192	18.32	0.757	0.287	16.11	0.710	0.295	13.91	0.662	0.371
PixelSplat [2]	**23.98**	**0.817**	0.169	20.31	0.783	0.227	18.46	0.770	0.235	15.36	0.673	0.329
MAC-Splat (Ours)	22.76	0.810	**0.146**	**22.60**	**0.816**	**0.135**	**22.09**	**0.818**	**0.142**	**21.34**	**0.818**	**0.158**

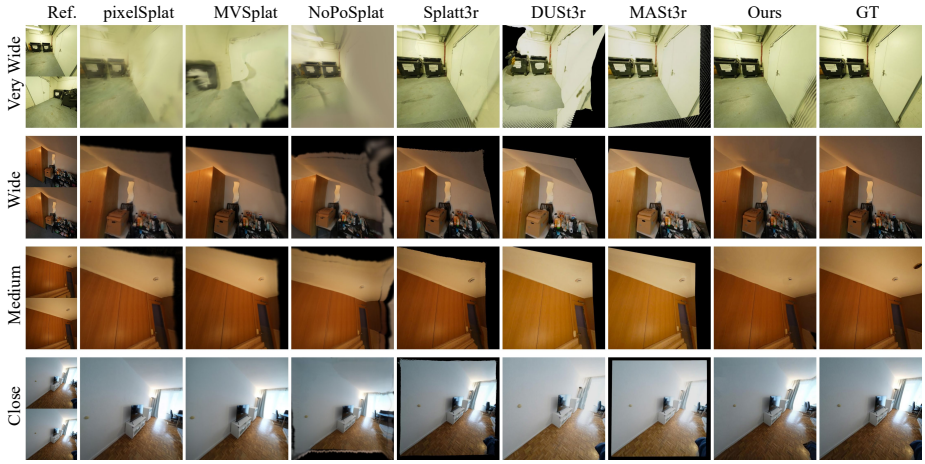

Fig. 4: Qualitative comparison on ScanNet++. Compared to recent methods, MAC-Splat significantly reduces severe floaters and ghosting artifacts. By enforcing 3D multi-attribute consistency on high-fidelity anchors, it produces sharper textures and cleaner occlusion boundaries under challenging wide-baseline conditions.

indicate significantly wider baselines and minimal overlap, providing a rigorous testbed for geometric robustness.

Inference protocol. For MAC-Splat, during inference the camera poses for the input context views are estimated by the MASt3R backbone from the input views, without ground-truth poses or external SfM for the input; all baselines are evaluated under their authors' official inference protocols, including their pose requirements. Ground-truth poses are used only to define target novel-view cameras for metric computation, which is standard for evaluating novel view synthesis.

Cross-dataset protocol. For DTU [17] evaluation, MAC-Splat operates in a strictly zero-shot setting, using the model trained exclusively on ScanNet++

Table 2: Cross-dataset results on DTU. All methods are evaluated zero-shot on DTU. We report PSNR (↑), SSIM (↑), and LPIPS (↓), averaged over all target views. Best results are in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
PixelSplat [2]	13.33	0.517	0.329
MVSplat [5]	14.03	0.522	0.334
Splatt3R [40]	12.13	0.454	0.437
MAC-Splat (Ours)	**17.32**	**0.605**	**0.311**

without any dataset-specific fine-tuning. All baseline methods are similarly evaluated using their authors' official pretrained models. To control evaluation variables, all methods use the official DTU camera intrinsics and a common input resolution, and we report metrics averaged over all target views.

Evaluation metrics. We report peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and LPIPS [65], averaged over all target views and test scenes. All metrics are computed in sRGB space on the full image unless otherwise specified. Furthermore, following Splatt3R [40], we also report **masked metrics** evaluated exclusively on observable regions.

Comparison methods. We compare MAC-Splat with correspondence-based baselines DUSt3R [47] and MASt3R [23], which are rendered as colored point clouds. We also evaluate generalizable 3D Gaussian splatting methods, including Splatt3R [40], MVSplat [5], PixelSplat [2], and NoPoSplat [58], using their official implementations under a common evaluation protocol.

Implementation details. MAC-Splat is trained on ScanNet++ using four NVIDIA RTX 4090 GPUs and AdamW with a learning rate of 10^{-5}. The overall MAC loss weight is set to $\lambda_{\text{MAC}} = 0.25$, with internal weights $\lambda_p = 1.0$, $\lambda_s = 0.02$, and $\lambda_a = 0.1$ for the positional, shape, and appearance terms. PixelSplat, MVSplat, and NoPoSplat are evaluated using the publicly released checkpoints provided by the authors. Splatt3R, DUSt3R and MASt3R are used with their official pretrained weights. For all baseline methods, we adopt the authors' recommended inference settings without modification.

4.2 Quantitative Comparison

Table 1 reports results on ScanNet++. Fig. 4 provides qualitative comparisons on representative ScanNet++ scenes. Under our colored point-cloud rendering protocol, DUSt3R and MASt3R obtain lower image metrics than 3D Gaussian splatting methods. Among 3DGS methods, PixelSplat achieves the highest PSNR on **Close** (23.98 dB) but drops to 15.36 dB on **Very Wide**. In contrast, MAC-Splat decreases from 22.76 dB to 21.34 dB and achieves the best PSNR, SSIM, and LPIPS on **Medium**, **Wide**, and **Very Wide**, while attaining the best LPIPS on **Close**. NoPoSplat remains competitive across subsets but still lags behind MAC-Splat on **Wide** and **Very Wide**, indicating that wide-baseline generalization remains challenging under limited overlap. Compared to Splatt3R,

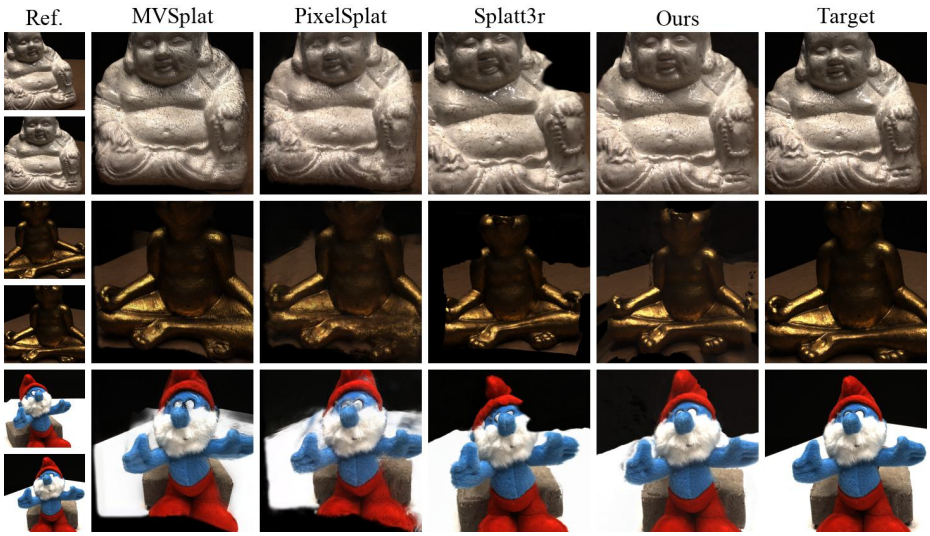

Fig. 5: Cross-dataset generalization: ScanNet++→DTU. Compared to other generalizable 3DGS methods, MAC-Splat better preserves geometry and edges and produces fewer artifacts under the controlled lighting conditions of DTU.

MAC-Splat improves the average PSNR across the four subsets by more than 4.5 dB and reduces the average LPIPS by about 44%.

Robustness to wide baselines. A key trend in Table 1 is the behavior across subsets with decreasing overlap (lower (ϕ, ψ)). Across the four subsets from **Close** to **Very Wide** defined by (ϕ, ψ), PixelSplat and MVSplat lose more than 8 dB in PSNR, while MAC-Splat drops by only 1.42 dB and maintains the best scores in all three metrics on the harder splits. LPIPS follows a similar pattern: for PixelSplat and MVSplat it roughly doubles on **Very Wide**, whereas for MAC-Splat it changes only from 0.146 to 0.158, indicating that perceptual quality is largely preserved under long baselines. We attribute this robustness to the MAC-Splat design, which uses high-confidence correspondences to couple Gaussians across views and enforces consistency in world space on position, shape, and appearance, rather than relying solely on multi-view photometric supervision.

Evaluation on Observable Regions. To further verify that the gains reflect improved reconstruction of geometry that is actually constrained by the inputs, we additionally evaluate using the visibility-mask definition proposed by Splatt3R [40]. This mask restricts evaluation to pixels that are visible from at least one context view and satisfy depth-consistency, thereby reducing the influence of unobserved regions where image-space completion can dominate the full-image metrics. As shown in Table 3, MAC-Splat outperforms Splatt3R and MASt3R on masked PSNR/LPIPS across all difficulty splits. Notably, on the **Very Wide** subset, Splatt3R drops to 13.04 dB masked PSNR, whereas MAC-

Table 3: Quantitative comparison on masked metrics (observable regions). Metrics are calculated strictly within the Splatt3R visibility mask to evaluate the reconstruction fidelity of observed geometry.

Method	Close (ϕ=0.9, ψ=0.9)		Medium (ϕ=0.7, ψ=0.7)		Wide (ϕ=0.5, ψ=0.5)		Very Wide (ϕ=0.3, ψ=0.3)	
	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
Splatt3R [40]	14.73	0.233	14.44	0.242	13.79	0.251	13.04	0.258
MASt3R [23]	13.57	0.283	12.96	0.280	12.50	0.293	11.27	0.322
MAC-Splat (Ours)	**21.56**	**0.159**	**21.06**	**0.163**	**20.26**	**0.178**	**19.06**	**0.189**

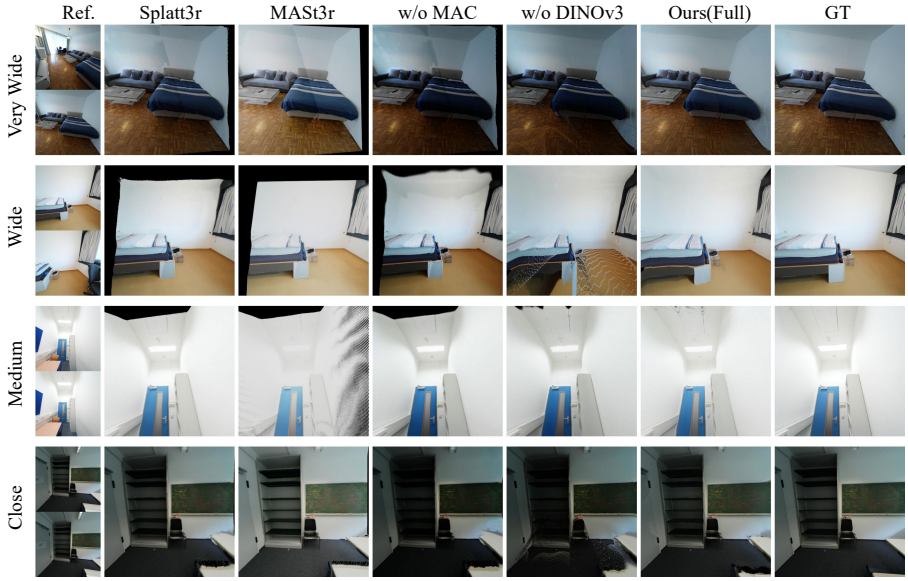

Fig. 6: Qualitative ablation on ScanNet++. Pure 2D photometric supervision (*w/o MAC Loss*) yields severe geometric distortions and floaters. While the MAC loss alone (*w/o DINOv3*) corrects global 3D geometry, our full model uses semantically enriched correspondences to further refine local details and cleanly preserve thin structures.

Splat reaches 19.06 dB, yielding a +6.02 dB gap in regions that are jointly observable. This improvement is consistent with our design goal: the MAC loss explicitly couples matched Gaussians across views and enforces multi-attribute consistency in 3D, which reduces geometric drift and suppresses double surfaces even when overlap is minimal.

Table 4: Ablations on ScanNet++. We report PSNR (↑), SSIM (↑), and LPIPS (↓) across the four (ϕ, ψ) splits. Splatt3R and MASt3R are included for reference. Best results per column are in **bold**.

Method	Close (ϕ=0.9, ψ=0.9)			Medium (ϕ=0.7, ψ=0.7)			Wide (ϕ=0.5, ψ=0.5)			Very Wide (ϕ=0.3, ψ=0.3)		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Splatt3R [40]	18.89	0.757	0.233	18.79	0.760	0.228	16.71	0.686	0.289	16.16	0.662	0.289
MASt3R [23]	18.44	0.722	0.292	15.23	0.648	0.359	15.72	0.719	0.295	15.82	0.608	0.278
w/o MAC Loss	19.94	0.791	0.193	19.45	0.756	0.203	20.10	0.780	0.186	17.12	0.695	0.278
w/o DINOv3	20.57	0.749	0.237	20.83	0.766	0.204	20.71	0.776	0.199	20.23	0.785	0.203
Ours (Full)	**22.76**	**0.810**	**0.146**	**22.60**	**0.816**	**0.135**	**22.09**	**0.818**	**0.142**	**21.34**	**0.818**	**0.158**

4.3 Cross-Dataset Generalization

Settings. DTU is object-centric with controlled lighting and fixed intrinsics, whereas ScanNet++ is scene-centric with wider fields of view and larger viewpoint variations. This results in a noticeable domain shift in both scene geometry and appearance statistics.

Results. Table 2 reports metrics on DTU [17]. MAC-Splat outperforms PixelSplat, MVSplat, and Splatt3R on all three metrics, despite being trained only on ScanNet++. These results suggest that MAC-Splat transfers favorably under the different imaging conditions of DTU. Fig. 5 shows qualitative examples.

4.4 Ablation Studies

We evaluate two ablated variants and the full model on ScanNet++ to isolate the contributions of the MAC loss and the semantic fusion module. All other losses, training schedules, and hyperparameters are kept fixed.

Variants. (1) *w/o MAC Loss*: the MAC loss is disabled and training uses only the standard photometric reconstruction losses. (2) *w/o DINOv3*: the MAC loss is enabled, but the Residual Semantic Fusion branch is removed, so descriptors are taken from the backbone only. (3) *Ours (Full)*: the complete MAC-Splat model with both MAC and DINOv3-based semantic fusion.

Table 4 reports the quantitative ablations. Removing the MAC loss (*w/o MAC Loss*) severely degrades performance, yielding only 17.12 dB on the **Very Wide** split. Applying the MAC loss without semantic fusion (*w/o DINOv3*) restores the PSNR to 20.23 dB. By outperforming PixelSplat by 4.87 dB, this MAC-only variant indicates that explicit 3D geometric regularization, rather than DINOv3 alone, primarily drives our wide-baseline robustness.

Finally, *Ours (Full)* achieves the best results across all splits. While MAC secures the global 3D geometry, DINOv3 enhances descriptor distinctiveness, further reducing LPIPS by ∼22% and preserving cleaner thin structures (Fig. 6).

5 Conclusion

We presented MAC-Splat, a 3D Gaussian splatting framework for sparse-view reconstruction. The method combines an MASt3R backbone with a frozen DINOv3 encoder through a Residual Semantic Fusion module to obtain semantically informed descriptors and a sparse set of high-confidence correspondences. Based on these correspondences, we introduce the MAC loss, which enforces world-space consistency on the position, shape, and appearance of matched Gaussians. Together, these components form a unified pipeline for sparse-view 3D Gaussian splatting. This supervision improves stability under low-overlap and wide-baseline conditions. In such regimes, photometric losses often fail due to parallax and occlusions. By anchoring Gaussians with high-confidence correspondences and regularizing key 3D attributes, MAC-Splat preserves coherent geometry and appearance under large viewpoint changes. Experiments on ScanNet++ show clear gains over correspondence-based and generalizable 3DGS baselines, with particularly strong gains on lower-overlap splits, and zero-shot DTU results further demonstrate strong cross-dataset generalization. We believe that multi-attribute 3D consistency is a promising direction for robust sparse-view reconstruction and can extend to scenarios with unknown poses or dynamic scenes.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62372359 and 62306233, the Xidian University Specially Funded Project for Interdisciplinary Exploration under Grant No. TZJHF202513, and the NTU Nanyang Assistant Professorship Startup Grant under Grant No. 025661-00012. The authors thank the anonymous reviewers and area chairs for their constructive comments.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV. pp. 5855–5864 (2021)
2. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR. pp. 19457–19467 (2024)
3. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2025)
4. Chen, K., Dai, B., Qin, M., Zhang, D., Li, P., Zou, Y., Wang, H.: Slgaussian: Fast language gaussian splatting in sparse views. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3047–3056 (2025)
5. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: ECCV. pp. 370–386. Springer (2024)

6. Chung, J., Oh, J., Lee, K.M.: Depth-regularized optimization for 3d gaussian splatting in few-shot images. In: CVPR. pp. 811–820 (2024)
7. Dalal, A., Hagen, D., Robbersmyr, K.G., Knausgård, K.M.: Gaussian splatting: 3d reconstruction and novel view synthesis: A review. IEEE Access **12**, 96797–96820 (2024)
8. Du, B., Meng, L., Hu, W.: Auggs: Self-augmented gaussians with structural masks for sparse-view 3d reconstruction. arXiv preprint arXiv:2408.04831 (2024)
9. Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., et al.: Instantsplat: Sparse-view gaussian splatting in seconds. arXiv preprint arXiv:2403.20309 (2024)
10. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In: CVPR. pp. 5354–5363 (2024)
11. Han, L., Zhou, J., Liu, Y.S., Han, Z.: Binocular-guided 3d gaussian splatting with view consistency for sparse view synthesis. NeurIPS **37**, 68595–68621 (2024)
12. Hou, C., Yeo, Q.X., Guo, M., Su, Y., Li, Y., Lee, G.H.: Mvgsr: Multi-view consistency gaussian splatting for robust surface reconstruction. arXiv preprint arXiv:2503.08093 (2025)
13. Huang, C., Zhang, Q., Zhang, Q., Li, N., Gong, Y., Wang, X., Feng, W.: Trigs: Triconsistency 3d gaussian splatting from sparse and unposed views. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 699–708 (2025)
14. Huang, H., Wu, Y., Deng, C., Gao, G., Gu, M., Liu, Y.S.: Fatesgs: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency. In: AAAI. pp. 3644–3652 (2025)
15. Huang, R., Mikolajczyk, K.: No pose at all: Self-supervised pose-free 3d gaussian splatting from sparse views. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27947–27957 (2025)
16. Jena, S., Vutukur, S.R., Boukhayma, A.: Sparsplat: Fast multi-view reconstruction with generalizable 2d gaussian splatting. arXiv preprint arXiv:2505.02175 (2025)
17. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: CVPR. pp. 406–413 (2014)
18. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
19. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In: CVPR. pp. 21357–21366 (2024)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics **42**(4), 139:1–139:14 (2023)
21. Kong, H., Yang, X., Wang, X.: Generative sparse-view gaussian splatting. In: CVPR. pp. 26745–26755 (2025)
22. Kumar, R., Vats, V.: Few-shot novel view synthesis using depth aware 3d gaussian splatting. In: ECCV. pp. 1–13. Springer (2024)
23. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV. pp. 71–91. Springer (2024)
24. Li, Y., Lyu, C., Di, Y., Zhai, G., Lee, G.H., Tombari, F.: Geogaussian: Geometry-aware gaussian splatting for scene rendering. In: European conference on computer vision. pp. 441–457. Springer (2024)

25. Li, Y., Ma, Q., Yang, R., Li, H., Ma, M., Ren, B., Popovic, N., Sebe, N., Konukoglu, E., Gevers, T., Gool, L.V., Oswald, M.R., Paudel, D.P.: Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. arXiv preprint arXiv:2503.18052 (2025)
26. Liu, C., Ouyang, C., Chen, Y., Quilodrán-Casas, C., Ma, L., Fu, J., Guo, Y., Shah, A., Bai, W., Arcucci, R.: T3d: Advancing 3d medical vision-language pre-training by learning multi-view visual consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 6763–6773 (October 2025)
27. Liu, T., Wang, G., Hu, S., Shen, L., Ye, X., Zang, Y., Cao, Z., Li, W., Liu, Z.: Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In: ECCV. pp. 37–53. Springer (2024)
28. Liu, X., Zhou, C., Huang, S.: 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. NeurIPS **37**, 133305–133327 (2024)
29. Luo, J., Huang, T., Wang, W., Feng, W.: A review of recent advances in 3d gaussian splatting for optimization and reconstruction. Image and Vision Computing **151**, 105304 (2024)
30. Lyu, X., Sun, Y.T., Huang, Y.H., Wu, X., Yang, Z., Chen, Y., Pang, J., Qi, X.: 3dgsr: Implicit surface reconstruction with 3d gaussian splatting. ACM Transactions on Graphics (TOG) **43**(6), 1–12 (2024)
31. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: CVPR. pp. 18039–18048 (2024)
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
33. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: A review. Neurocomputing **438**, 14–33 (2021)
34. Park, H., Ryu, G., Kim, W.: Dropgaussian: Structural regularization for sparse-view gaussian splatting. In: CVPR. pp. 21600–21609 (2025)
35. Peng, R., Xu, W., Tang, L., Liao, L., Jiao, J., Wang, R.: Structure consistent gaussian splatting with matching prior for few-shot novel view synthesis. NeurIPS **37**, 97328–97352 (2024)
36. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: CVPR. pp. 20051–20060 (2024)
37. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016)
38. Shen, J., Feng, T., Dong, H., Ji, J., Wang, X., Shao, T.: Surags: Toward efficient few-shot novel view synthesis via surface-aware gaussian splatting. Computers & Graphics p. 104349 (2025)
39. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
40. Smart, B., Zheng, C., Laina, I., Prisacariu, V.A.: Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. arXiv preprint arXiv:2408.13912 (2024)
41. Sucar, E., Liu, S., Ortiz, J., Davison, A.J.: imap: Implicit mapping and positioning in real-time. In: ICCV. pp. 6229–6238 (2021)
42. Sun, X., Chen, R., Gong, M., Xu, D., Liu, T.: Intern-gs: Vision model guided sparse-view 3d gaussian splatting. arXiv preprint arXiv:2505.20729 (2025)

43. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: CVPR. pp. 8248–8258 (2022)
44. Tang, S., Ye, W., Ye, P., Lin, W., Zhou, Y., Chen, T., Ouyang, W.: Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. arXiv preprint arXiv:2410.06245 (2024)
45. Turkulainen, M., Ren, X., Melekhov, I., Seiskari, O., Rahtu, E., Kannala, J.: Dn-splatter: Depth and normal priors for gaussian splatting and meshing. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2421–2431. IEEE (2025)
46. Ververas, E., Potamias, R.A., Song, J., Deng, J., Zafeiriou, S.: Sags: Structure-aware 3d gaussian splatting. In: ECCV. pp. 221–238. Springer (2024)
47. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024)
48. Wang, Y., Huang, T., Chen, H., Lee, G.H.: Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. NeurIPS **37**, 107326–107349 (2024)
49. Wang, Z., Li, D., Wu, Y., He, T., Bian, J., Jiang, R.: Diffusion models in 3d vision: A survey. arXiv preprint arXiv:2410.04738 (2024)
50. Wu, J., Li, H., Jiang, X., Yao, Y., Zhang, L.: Coca-splat: Collaborative optimization for camera parameters and 3d gaussians. arXiv preprint arXiv:2504.00639 (2025)
51. Wu, T., Yuan, Y.J., Zhang, L.X., Yang, J., Cao, Y.P., Yan, L.Q., Gao, L.: Recent advances in 3d gaussian splatting. Computational Visual Media **10**(4), 613–642 (2024)
52. Xiong, H., Muttukuru, S., Xiao, H., Upadhyay, R., Chari, P., Zhao, Y., Kadambi, A.: SparseGS: Sparse View Synthesis using 3D Gaussian Splatting. arXiv preprint arXiv:2312.00206 (2023)
53. Xu, J., Gao, S., Shan, Y.: Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. In: ICCV. pp. 25442–25452 (2025)
54. Xu, Y., Shi, Z., Yifan, W., Chen, H., Yang, C., Peng, S., Shen, Y., Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In: ECCV. pp. 1–20. Springer (2024)
55. Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., Li, X.: Gs-slam: Dense visual slam with 3d gaussian splatting. In: CVPR. pp. 19595–19604 (2024)
56. Yang, C., Li, S., Fang, J., Liang, R., Xie, L., Zhang, X., Shen, W., Tian, Q.: Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. arXiv preprint arXiv:2402.10259 (2024)
57. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV. pp. 767–783 (2018)
58. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207 (2024)
59. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV. pp. 12–22 (2023)
60. Yin, R., Yugay, V., Li, Y., Karaoglu, S., Gevers, T.: Fewviewgs: Gaussian splatting with few view matching and multi-stage training. NeurIPS **37**, 127204–127225 (2024)
61. Younis, T., Cheng, Z.: Sparse-view 3d reconstruction: Recent advances and open challenges. arXiv preprint arXiv:2507.16406 (2025)
62. Yu, H., Long, X., Tan, P.: Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors. arXiv preprint arXiv:2409.03456 (2024)

63. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: CVPR. pp. 19447–19456 (2024)
64. Zhang, J., Li, J., Yu, X., Huang, L., Gu, L., Zheng, J., Bai, X.: Cor-gs: sparse-view 3d gaussian splatting via co-regularization. In: European conference on computer vision. pp. 335–352. Springer (2024)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
66. Zhao, C., Wang, X., Zhang, T., Javed, S., Salzmann, M.: Self-ensembling gaussian splatting for few-shot novel view synthesis. In: ICCV. pp. 4940–4950 (2025)
67. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In: CVPR. pp. 19680–19690 (2024)
68. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. In: ECCV. pp. 145–163. Springer (2024)
69. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In: CVPR. pp. 10324–10335 (2024)