GeoTrack-GS: Refining Depth Priors with Self-Supervised Geometric Constraints

Zhipeng Ye , Zihao Lu , Yuan Zhang , Wenjie Qin , Jiayi Hong , Yan Cao , Xuming Wu * , Zhibin Shao *

Posted Date: 5 May 2026

doi: 10.20944/preprints202605.0113.v1

Keywords: 3D Gaussian Splatting; novel view synthesis; self-supervised learning; multi-view geometry; 3D reconstruction

 doi:10.20944/preprints202605.0113.v1

Article

GeoTrack-GS: Refining Depth Priors with Self-Supervised Geometric Constraints

Zhipeng Ye [1], Zihao Lu [1], Yuan Zhang [1], Wenjie Qin [1], Jiayi Hong [1], Yan Cao [2], Xuming Wu [1,*] and Zhibin Shao [3,*]

[1] Lingnan Normal University, Zhanjiang 524048, China
[2] Luoyang Normal University, Luoyang 471934, China
[3] Shenzhen Polytechnic University, Shenzhen 518055, China
* Correspondence: wuxm@lingnan.edu.cn (X.W.); zhibin_shao@szpu.edu.cn (Z.S.)

Abstract

3D Gaussian Splatting (3DGS) degrades severely in sparse-view scenarios, often collapsing into artifacts due to under-constrained optimization. While incorporating monocular depth priors provides dense supervision, their inherent multi-view inconsistency frequently distorts geometry. To address this, we propose **GeoTrack-GS**, a geometry-first framework that refines noisy depth priors using reliable self-supervised constraints. Specifically, we leverage sparse feature tracks to enforce *macro-level* reprojection consistency and introduce a *micro-level* anisotropic regularizer via K-NN PCA to suppress rank-collapse. On this corrected geometry, we design **GT-DCA**, a geometry-guided deformable cross-attention module that captures view-dependent appearance without compromising structure. A Decoupled Constraint Stabilization strategy further balances these heterogeneous signals during training. Experiments on LLFF and DTU under 3–9 input views, and on Mip-NeRF 360 under 12 input views, demonstrate that GeoTrack-GS achieves **state-of-the-art geometric fidelity** while maintaining **competitive rendering quality** compared to existing baselines, effectively reducing floaters and "waxy" surfaces.

Keywords: 3D Gaussian Splatting; novel view synthesis; self-supervised learning; multi-view geometry; 3D reconstruction

1. Introduction

Obtaining precise 3D scene representations and high-quality novel view synthesis (NVS) from only a few posed images remains a fundamental challenge in computer vision and graphics [1,2], with broad relevance to applications such as AR/VR, robotics, and autonomous navigation. Neural Radiance Fields (NeRF) [3] demonstrate photorealistic NVS by learning a continuous volumetric radiance field, but their implicit MLP-based formulation typically incurs high training and rendering costs [4,5]. 3D Gaussian Splatting (3DGS) [6] advances this line by representing scenes with anisotropic 3D Gaussians and rasterizing them via a differentiable splat renderer, achieving NeRF-level visual quality with real-time rendering performance [7,8].

However, vanilla 3DGS critically relies on a high-quality, dense point cloud from Structure-from-Motion (SfM) [9,10] and a sufficiently large number of input views. In realistic sparse-view scenarios (e.g., 3–9 views), the optimization becomes severely under-constrained: gradients from different views conflict, adaptive densification becomes unstable, and Gaussians tend to overfit training images while degenerating into floating artifacts or needle-like structures [11–13]. As illustrated in Figure 1, even recent methods may produce visually plausible renderings while exhibiting distorted geometry and over-smoothed, "waxy" surfaces.

 doi:10.20944/preprints202605.0113.v1

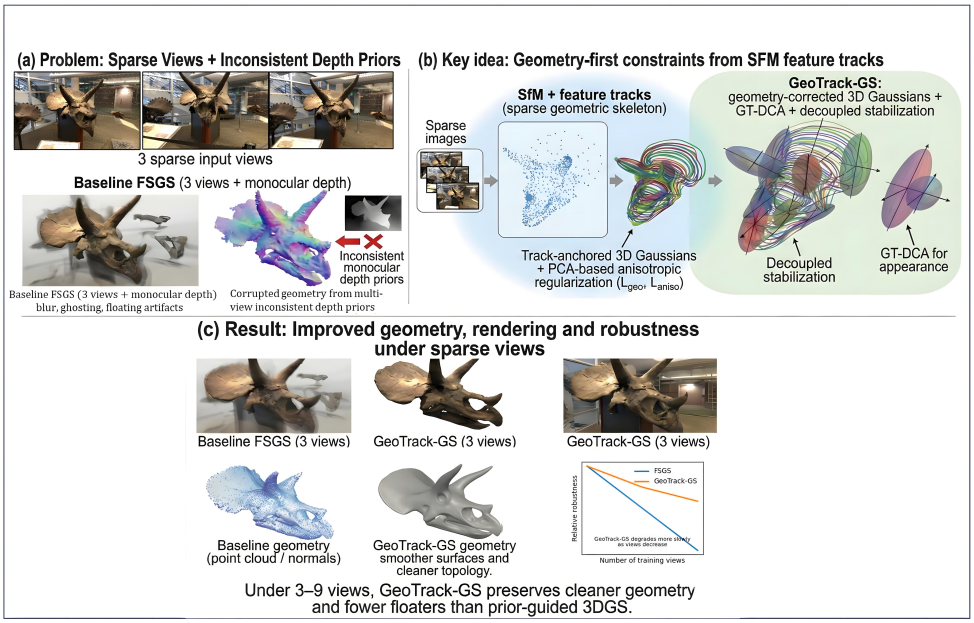

Figure 1. Overview of GeoTrack-GS. While prior-guided methods (a) suffer from multi-view inconsistent monocular depth artifacts, our geometry-first framework (b) leverages sparse feature tracks to correct geometry and guide appearance. (c) GeoTrack-GS improves geometric fidelity while maintaining competitive rendering quality under sparse views.

To mitigate the under-constrained nature of sparse-view reconstruction, existing methods mainly follow two paradigms. *Prior-guided* approaches incorporate external supervision, most commonly monocular depth [14–16], to provide dense geometric cues. Recent advancements like FSGS [17], DNGaussian [18], DN-Splatter [19], and DepthSplat [20] integrate depth or normal priors into the 3DGS pipeline. While such priors can stabilize training, they are predicted independently for each view and are therefore inherently multi-view inconsistent [21], which may distort geometry when enforced strongly.

In contrast, *self-regularized* approaches avoid external networks and instead rely on internal consistency or synthesized supervision, such as pseudo-view constraints or entropy regularization [22–24]. Although effective in reducing overfitting, these methods often introduce indirect constraints or complex training procedures that are not explicitly grounded in 3D scene geometry.

Our key observation is that under sparse-view, wide-baseline conditions, dense reconstruction is fundamentally unreliable [25,26], whereas sparse geometric correspondences remain robust. Dense multi-view stereo and standard 3DGS densification often fail due to occlusions and limited overlap. In contrast, modern feature matching methods can still recover *sparse but reliable* multi-view correspondences even under extreme viewpoint changes [27–29]. These sparse trajectories form a stable geometric skeleton that is physically grounded and significantly more reliable than per-view dense predictions.

Motivated by this observation, we propose **GeoTrack-GS**, a geometry-first framework for sparse-view 3D Gaussian Splatting. At the **macro** level, we leverage multi-view feature tracks to enforce reprojection consistency, anchoring Gaussian centers to reliable geometric observations and correcting inconsistent depth priors. At the **micro** level, we introduce a PCA-based anisotropic regularizer that aligns Gaussian covariances with local surface structure and suppresses rank-collapse artifacts, ensuring physically plausible surface representations.

On top of the stabilized geometry, we further develop **GT-DCA** (Geometry-guided Track-based Deformable Cross-Attention), an appearance module that enhances view-dependent color modeling without corrupting geometric structure. GT-DCA uses geometry-aware track queries to guide deformable attention over deep image feature maps [30,31], enabling adaptive aggregation of high-frequency appearance cues. Finally, to handle the heterogeneous noise characteristics of different

 doi:10.20944/preprints202605.0113.v1

supervision signals, we introduce a *Decoupled Constraint Stabilization* strategy that modulates geometric and pseudo-view constraints during training, improving robustness under sparse-view supervision.

In summary, our main contributions are:

- We propose **GeoTrack-GS**, a geometry-first framework for sparse-view 3DGS that treats monocular depth as a weak prior and explicitly corrects its multi-view inconsistency using robust feature-track-based supervision.
- We introduce a dual-level geometric constraint system consisting of a track-based reprojection loss and a PCA-based anisotropic regularizer, jointly addressing global inconsistency and local rank-collapse artifacts.
- We design **GT-DCA**, a geometry-guided deformable cross-attention module that improves view-dependent appearance modeling on top of corrected geometry.
- We develop a decoupled constraint stabilization strategy that mitigates conflicts among heterogeneous supervision signals, leading to more stable optimization in sparse-view settings.

2. Related Work

Our work is related to neural rendering, sparse-view novel view synthesis, multi-view geometry, and recent advances in geometric regularization and appearance modeling for 3D Gaussian Splatting.

2.1. From Neural Radiance Fields to 3D Gaussian Splatting

Neural Radiance Fields (NeRF) [3] represent scenes as continuous volumetric radiance fields parameterized by MLPs, achieving impressive novel view synthesis. However, they typically require dense input views and incur high computational costs. Subsequent works improve efficiency and quality through explicit structures, anti-aliasing, or encodings, such as Instant-NGP [4], Mip-NeRF [5], and Mip-NeRF 360 [2].

3D Gaussian Splatting (3DGS) [6] represents scenes using anisotropic 3D Gaussian primitives rendered via a differentiable splat renderer, enabling real-time performance. Following this paradigm, numerous works have explored improved initialization and regularization strategies [7,11,12]. However, as summarized in recent surveys [7,8], vanilla 3DGS relies heavily on accurate SfM initialization [9] and dense inputs. In sparse-view settings, the optimization becomes under-constrained, leading to floating artifacts and degenerate Gaussian shapes [11–13].

Our method builds upon 3DGS but explicitly injects geometry-aware constraints derived from multi-view feature tracks and local PCA analysis, improving stability when inputs are sparse.

2.2. Sparse-View Novel View Synthesis

Sparse-view synthesis is a core challenge in neural rendering. Existing methods can be categorized into prior-guided and self-regularized approaches.

Prior-guided approaches.

A common strategy incorporates external priors, particularly monocular depth. Representative methods include DS-NeRF [15], Dense Depth Priors [16], MonoSDF [32], and Geo-NeuS [33] in the NeRF family. In the 3DGS regime, FSGS [17], DNGaussian [18], DN-Splatter [19], and DepthSplat [20] utilize depth or normal priors to guide geometry. Recent works like HBSplat [34] and DWGS [35] further refine this with hybrid losses. While effective, monocular priors are predicted independently per view and can be multi-view inconsistent [14,21]. When enforced too strongly, they may distort geometry or produce over-smoothed surfaces.

GeoTrack-GS treats monocular depth as a weak prior, relying primarily on multi-view feature tracks for geometric supervision to directly address inconsistency in 3D space.

Self-regularized approaches.

Another line of work avoids external priors, relying on internal regularization. RegNeRF [25], FreeNeRF [26], DietNeRF [36], and InfoNeRF [37] introduce smoothness or semantic consistency for

 doi:10.20944/preprints202605.0113.v1

NeRF. Similarly, CoR-GS [22], FewViewGS [23], DropoutGS [24], SparseGS [13], and Self-ensembling GS [38] extend these ideas to 3DGS by enforcing consistency across models or pseudo-views. Although effective against overfitting, these schemes often lack explicit 3D geometric grounding.

Structure-consistent Gaussian Splatting (SCGaussian) [39] leverages feature correspondences for global consistency. GeoTrack-GS goes beyond this by combining track-based global supervision with a micro-level PCA-based anisotropic regularizer to explicitly suppress rank-collapse artifacts.

2.3. Multi-View Geometry, Regularization, and Appearance

Classical multi-view geometry has been reintroduced into neural rendering to improve stability [9,40]. Methods like BARF [41], TrackNeRF [42], and others [32,33] exploit geometric constraints for joint pose and scene optimization. Recent advancements in feature matching, such as SuperGlue [27], LoFTR [28], and LightGlue [29], facilitate robust correspondence estimation. Our geometric loss $\mathcal{L}_{geo}$ draws inspiration from these works but uses feature tracks solely to supervise Gaussian geometry with fixed poses.

Appearance modeling is also critical. Standard 3DGS struggles with high-frequency view-dependent effects. Recent works like Spec-Gaussian [43], 3iGS [44], and MS-GS [45] improve expressiveness through anisotropic reflectance or multi-appearance modeling. Our GT-DCA module complements these by leveraging geometry-guided attention to aggregate appearance cues, preserving geometric integrity while enhancing fidelity.

3. Methodology

We present **GeoTrack-GS**, a geometry-first framework that combines external depth priors with robust self-supervised constraints to enable high-quality sparse-view 3D Gaussian Splatting. Rather than treating monocular depth as ground truth, we regard it as a weak, dense prior and *correct* its multi-view inconsistency using feature-track-based supervision and anisotropic regularization. On top of this corrected geometric foundation, our **GT-DCA** module models high-frequency, view-dependent appearance. An overview is shown in Figure 2.

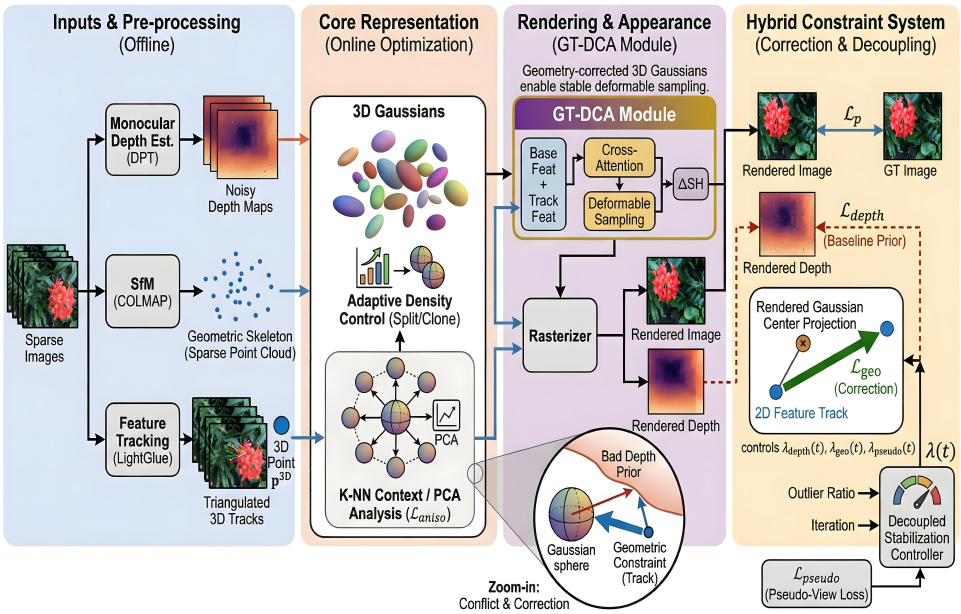

Figure 2. Overview of the GeoTrack-GS pipeline. Our framework treats monocular depth as a weak prior and corrects it using a sparse skeleton derived from multi-view feature tracks. A hybrid constraint system (macro-level reprojection and micro-level anisotropic regularization) ensures geometric fidelity, while the GT-DCA module models view-dependent appearance on top of the stabilized geometry.

 doi:10.20944/preprints202605.0113.v1

In this section, we detail the core components of our framework. We first describe the offline trajectory pre-processing and the track–Gaussian binding in Sec. 3.1. Next, we introduce the hybrid geometric constraint framework in Sec. 3.2 and present our decoupled constraint stabilization strategy in Sec. 3.3. Finally, we detail the GT-DCA appearance module in Sec. 3.4.

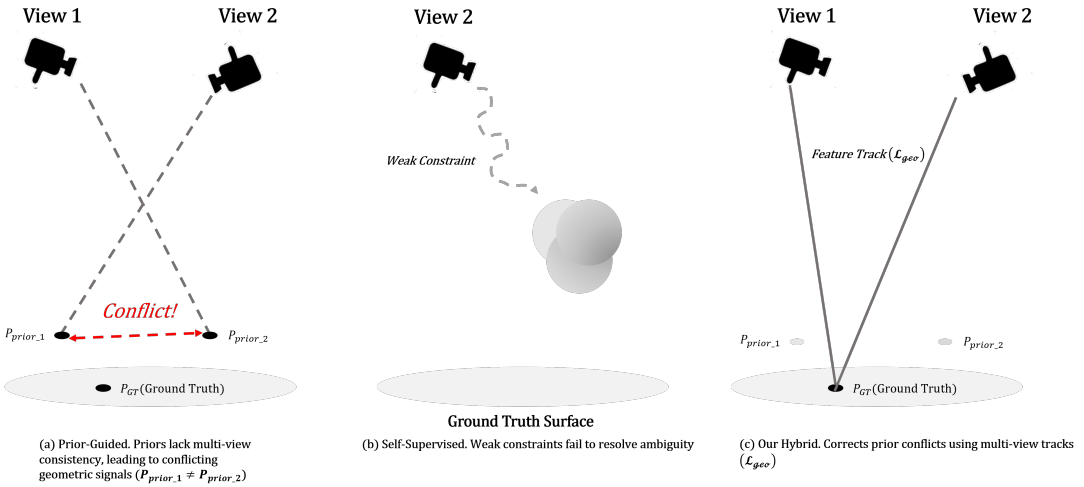

Figure 3. Components of the geometric constraint module. We employ three complementary constraints: (a) scale-invariant correlation for dense depth guidance ($\mathcal{L}_{depth}$), (b) track-based consistency to anchor global geometry ($\mathcal{L}_{geo}$), and (c) local PCA alignment to ensure physically plausible surface normals ($\mathcal{L}_{aniso}$).

3.1. Offline Trajectory Pre-Processing

Given N_{views} posed images $\{I_i\}_{i=1}^{N_{\text{views}}}$, GeoTrack-GS starts with an offline stage that extracts a geometric skeleton and high-quality feature tracks.

SfM-based geometric skeleton.

We run a standard Structure-from-Motion pipeline such as COLMAP [9], obtaining:

- camera intrinsics $\{\mathbf{K}_i\}$ and extrinsics $\{\mathbf{T}_i = [\mathbf{R}_i|\mathbf{t}_i]\}$,
- a sparse point cloud $\{\mathbf{p}_j^{3D}\}_{j=1}^{M}$ with associated 2D observations.

We use $\mathbf{p}_j^{3D}$ as initial Gaussian centers $\boldsymbol{\mu}_j$ and initialize scales, rotations, opacities, and SH coefficients as in [6].

Multi-view feature tracks.

SfM pipelines relying on handcrafted features (e.g., SIFT) may produce insufficient track lengths in wide-baseline sparse-view settings. To alleviate this, we employ state-of-the-art deep matchers (LightGlue [29] or LoFTR [28]) to extract robust pairwise correspondences. Unlike local descriptors, these transformer-based matchers leverage global context, enabling high-confidence matching even in texture-poor or repetitive regions. By chaining pairwise matches, we construct multi-view tracks

$$\boldsymbol{\tau}_k = \{(i_1, \mathbf{x}_{k,i_1}), \ldots, (i_{n_k}, \mathbf{x}_{k,i_{n_k}})\}, \tag{1}$$

where $\mathbf{x}_{k,i_j} \in \mathbb{R}^2$ is the 2D location of track k in view i_j. We triangulate each track using camera poses to obtain $\mathbf{p}_k^{3D}$ and compute its average reprojection error. Tracks with fewer than $N_{\min}$ observations, large reprojection error, or marked as outliers by RANSAC are discarded.

Implementation details.

To maximize connectivity under sparse views, we build a fully-connected matching graph and perform geometric verification using RANSAC. Unless otherwise specified, we set the RANSAC reprojection threshold to 4 pixels and discard tracks with fewer than $N_{\min} = 3$ observations.All tracks used by $\mathcal{L}_{geo}$ satisfy $|\mathcal{V}_k| \geq N_{\min}$.

 doi:10.20944/preprints202605.0113.v1

We further filter tracks by their bundle-adjusted reprojection error and remove tracks whose mean reprojection residual exceeds 2.0 pixels. All track anchors $\mathbf{p}_k^{3D}$ are triangulated using the fixed COLMAP camera poses and refined using the SfM **global bundle adjustment**.

Track–Gaussian association.

A key design choice is to treat tracks as *persistent anchors* for specific Gaussians. For each valid track τ_k with anchor $\mathbf{p}_k^{3D}$, we assign it to its nearest Gaussian:

$$\mathcal{G}_k^\star = \arg\min_{\mathcal{G}_i} \|\boldsymbol{\mu}_i - \mathbf{p}_k^{3D}\|_2. \tag{2}$$

This track–Gaussian association is established once and kept fixed during optimization. Even as Gaussians move, each track-constrained Gaussian is always pulled back towards its own anchor by our geometric loss. This fixed binding also provides a clean interface to GT-DCA (Sec. 3.4), guaranteeing that appearance aggregation along a track always refers to the same physical 3D point.

At the same time, the *strength* of this binding is adaptive: as detailed in Sec. 3.3, our quality-aware gating down-weights outlier tracks so that Gaussians are not permanently locked to erroneous anchors.

3.2. Hybrid Geometric Constraint Framework

We incorporate geometric supervision into 3DGS via a hybrid loss:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_p(t)\,\mathcal{L}_p + \lambda_{depth}(t)\,\mathcal{L}_{depth} + \lambda_{geo}(t)\,\mathcal{L}_{geo} \\ + \lambda_{aniso}(t)\,\mathcal{L}_{aniso} + \lambda_{pseudo}(t)\,\mathcal{L}_{pseudo},\end{aligned} \tag{3}$$

where t denotes the training iteration and $\lambda_\cdot(t)$ are adaptive weights. Figure 4 summarizes the roles of the different geometric losses.

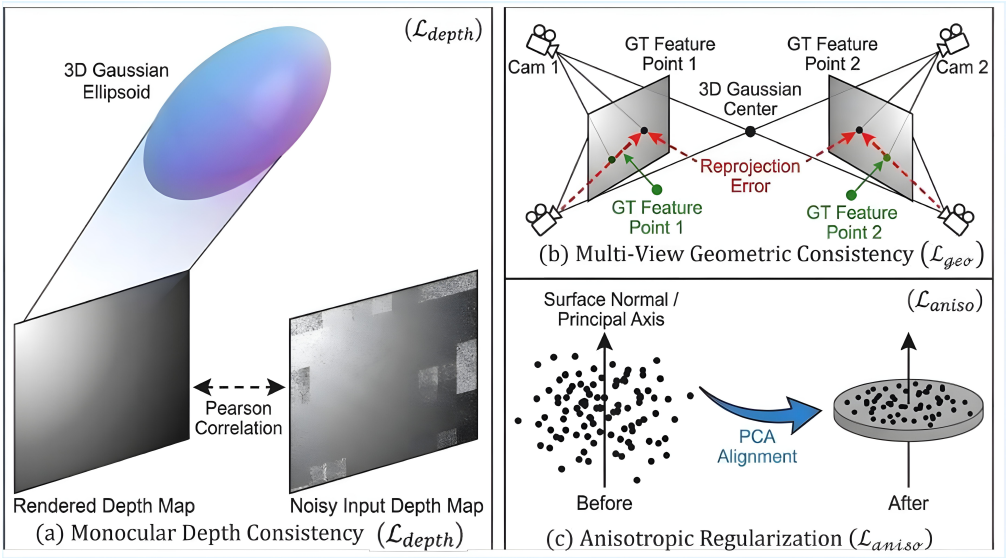

Figure 4. Visualization of our hybrid geometric constraint system. Our framework integrates (a) dense monocular depth priors for initialization, (b) *macro-level* feature track reprojection ($\mathcal{L}_{geo}$) to correct inconsistent depths, and (c) *micro-level* anisotropic regularization ($\mathcal{L}_{aniso}$) to prevent rank-collapse artifacts.

Photometric reconstruction loss.

Following [6], we use a combination of ℓ_1 and SSIM:

$$\mathcal{L}_p = (1 - \lambda_{ssim})\,\mathcal{L}_1 + \lambda_{ssim}\,\mathcal{L}_{ssim}, \tag{4}$$

with $\lambda_{ssim} = 0.2$, where $\mathcal{L}_1$ is the pixel-wise absolute difference and $\mathcal{L}_{ssim}$ is the structural similarity loss [46].

 doi:10.20944/preprints202605.0113.v1

Depth prior loss.

We pre-compute monocular depth maps D_{prior} using a pre-trained estimator such as DPT [14]. To mitigate scale ambiguity, we use a scale-invariant correlation loss between the rendered depth D_{render} and D_{prior}:

$$\mathcal{L}_{depth} = 1 - \frac{\text{Cov}(D_{\text{render}}, D_{\text{prior}})}{\sigma(D_{\text{render}})\,\sigma(D_{\text{prior}})}. \tag{5}$$

This provides dense geometric cues, but unlike prior-only methods, its gradients can be overruled by $\mathcal{L}_{geo}$ and $\mathcal{L}_{aniso}$ when they disagree.

Macro-level track-based geometric loss.

For each track k and its associated Gaussian $\mathcal{G}_k^\star$ with center $\boldsymbol{\mu}_k^\star$, we project $\boldsymbol{\mu}_k^\star$ into all views where the track is observed:

$$\hat{\mathbf{x}}_{k,i} = \boldsymbol{\pi}(\mathbf{K}_i, \mathbf{T}_i, \boldsymbol{\mu}_k^\star), \tag{6}$$

and minimize the robust reprojection error:

$$\mathcal{L}_{geo} = \frac{1}{M} \sum_{k=1}^{M} \sum_{i \in \mathcal{V}_k} \rho\big(\|\hat{\mathbf{x}}_{k,i} - \mathbf{x}_{k,i}\|_2\big), \tag{7}$$

where $\mathcal{V}_k$ is the set of views where track k is visible, M is the number of tracks, and ρ is a Huber loss with threshold δ:

$$\rho(r) = \begin{cases} \frac{1}{2} r^2, & |r| \le \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta. \end{cases} \tag{8}$$

We apply this loss at multiple image scales and average the results, improving robustness to noise and outliers.

Micro-level anisotropic regularization.

While $\mathcal{L}_{geo}$ constrains *where* Gaussians should be, it does not constrain *how* they should look in 3D. In sparse-view settings, unconstrained Gaussians tend to become highly elongated, causing rank-collapse artifacts [11]. We counteract this with a local PCA-based anisotropic regularizer.

For each Gaussian $\mathcal{G}_k$ at center $\boldsymbol{\mu}_k$, we find its K nearest neighbors and compute the covariance of their positions:

$$\mathbf{C}_k = \frac{1}{K} \sum_{j \in \mathcal{N}_k} (\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}_k)(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}_k)^\top, \tag{9}$$

where $\bar{\boldsymbol{\mu}}_k$ is the neighbor mean. We perform PCA on $\mathbf{C}_k$ to obtain eigenpairs $(\lambda_{k,i}, \mathbf{v}_{k,i})$ with $\lambda_{k,1} \ge \lambda_{k,2} \ge \lambda_{k,3}$. The smallest eigenvector $\mathbf{v}_{k,3}$ approximates the local surface normal.

We also diagonalize the Gaussian covariance $\boldsymbol{\Sigma}_k$ as

$$\boldsymbol{\Sigma}_k = \mathbf{U}_k \,\text{diag}(\sigma_{k,1}^2, \sigma_{k,2}^2, \sigma_{k,3}^2)\, \mathbf{U}_k^\top, \tag{10}$$

with eigenvectors $\mathbf{u}_{k,i}$ and eigenvalues $\sigma_{k,i}^2$. We define

$$\mathcal{L}_{align} = \frac{1}{N} \sum_k \Big(1 - |\mathbf{u}_{k,3}^\top \mathbf{v}_{k,3}|\Big), \tag{11}$$

$$\mathcal{L}_{scale} = \frac{1}{N} \sum_k \sum_{i=1}^{3} \left| \log \frac{\sigma_{k,i}}{\sqrt{\lambda_{k,i} + \epsilon}} \right|, \tag{12}$$

$$\mathcal{L}_{ratio} = \frac{1}{N} \sum_k \max\left(0, \frac{\sigma_{k,1}}{\sigma_{k,3}} - \theta_{aniso}\right)^2, \tag{13}$$

 doi:10.20944/preprints202605.0113.v1

where ϵ avoids division by zero and θ_{aniso} is a maximum anisotropy ratio. $\mathcal{L}_{align}$ encourages the shortest axis of each Gaussian to align with the estimated surface normal. Theoretically, $\mathcal{L}_{scale}$ is grounded in the assumption that the local distribution of geometric anchors serves as a reliable statistical proxy for the physical surface extent; matching $\sigma_{k,i}$ to $\sqrt{\lambda_{k,i}}$ ensures unit consistency between the Gaussian's covariance and the local point density. Finally, $\mathcal{L}_{ratio}$ penalizes extremely elongated Gaussians to suppress rank-collapse.The total anisotropic regularizer is

$$\mathcal{L}_{aniso} = \lambda_{align}\mathcal{L}_{align} + \lambda_{scale}\mathcal{L}_{scale} + \lambda_{ratio}\mathcal{L}_{ratio}. \tag{14}$$

We activate $\mathcal{L}_{aniso}$ after a short warm-up to allow the initial geometry to form.

Pseudo-view Regularizer.

To regularize geometry in unobserved regions, we employ a **perturbation-based consistency** strategy. Specifically, we generate pseudo-views by perturbing training camera poses with random rotation ($\delta\mathbf{R} \in [-10°, 10°]$) and translation ($\delta\mathbf{t}$ within 10% of scene bounds). We warp the training view to the pseudo-view using the current depth prediction and enforce consistency loss:

$$\mathcal{L}_{pseudo} = \|\mathbf{I}_{pseudo} - \mathbf{I}_{warp}\|_1 + \lambda_{smooth}\mathcal{L}_{TV}(D_{pseudo}), \tag{15}$$

where $\mathbf{I}_{warp}$ is the warped image and $\mathcal{L}_{TV}$ is a total variation smoothness term on the depth map. Crucially, since pseudo-labels are noisy during early training, this loss is modulated by our Decoupled Constraint Stabilization (DCS) schedule (see Sec. 3.3), which gradually increases the weight λ_{pseudo} only after the geometric skeleton has stabilized.

3.3. Decoupled Constraint Stabilization

The auxiliary losses in Eq. (3) have heterogeneous noise profiles. Feature tracks contain spatially sparse outliers, while pseudo-view supervision is temporally unstable in early training when geometry is immature. Naively summing all losses with fixed weights often yields optimization conflicts.

We propose a **Decoupled Constraint Stabilization** (DCS) strategy that assigns tailored modulation logic to each constraint branch (Figure 5).

Quality-aware geometric gating.

For the geometric branch, we compute the current reprojection outlier ratio $r_{\text{out}}(t)$, i.e., the fraction of tracks whose reprojection error exceeds a threshold. We set

$$\lambda_{geo}(t) = \lambda_{geo}^{\max}\,(1 - r_{\text{out}}(t))^{\gamma}, \tag{16}$$

with exponent $\gamma > 0$. When track quality is high, $\lambda_{geo}(t) \approx \lambda_{geo}^{\max}$ and $\mathcal{L}_{geo}$ strongly corrects depth prior errors. When many tracks are unreliable, $\lambda_{geo}(t)$ is automatically down-weighted, preventing noisy tracks from dominating gradients. In addition, we down-weight individual tracks whose error remains high, further improving robustness.

This gating mechanism effectively acts as a *dynamic unbinding* operation. When a track is identified as an outlier and its weight drops towards zero, the associated Gaussian is no longer effectively pulled towards the erroneous 3D anchor and is free to update its position under $\mathcal{L}_p$ and other reliable losses, allowing it to escape bad local minima.

 doi:10.20944/preprints202605.0113.v1

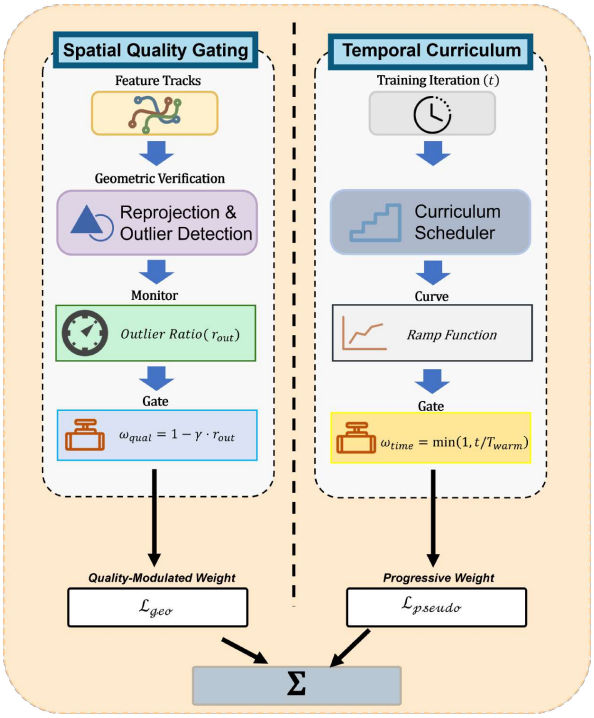

Figure 5. Decoupled Constraint Stabilization (DCS) strategy. We employ (left) *spatial quality gating* to dynamically down-weight unreliable track constraints based on outlier ratios, and (right) a *temporal curriculum* to gradually introduce pseudo-view supervision only after the geometry has stabilized.

Curriculum for pseudo-view supervision.

For the pseudo-view branch, the primary issue is temporal maturity: early in training, pseudo-views are of low quality and enforcing consistency can freeze the model in bad local minima. We therefore use a curriculum schedule for $\lambda_{pseudo}(t)$:

$$\lambda_{pseudo}(t) = \lambda_{pseudo}^{\max} \cdot \min\left(1, \frac{t - t_0}{t_{\text{ramp}}}\right)_+, \tag{17}$$

where t_0 is the start iteration, t_{ramp} controls ramp length, and $(\cdot)_+$ clamps negatives to zero. This keeps $\mathcal{L}_{pseudo}$ negligible during the volatile early phase and gradually introduces pseudo-view consistency only after the geometry has been stabilized by $\mathcal{L}_{depth}$ and $\mathcal{L}_{geo}$.

Empirically (Sec. 4), this decoupled modulation yields faster convergence and better final metrics than fixed or hand-tuned schedules, and is crucial for stable sparse-view training.

3.4. Geometry-Guided Track-Based Deformable Cross-Attention (GT-DCA)

Standard 3DGS relies on Spherical Harmonics, which struggle to model high-frequency view-dependent effects. As illustrated in Figure 6, we propose GT-DCA, which explicitly utilizes deep visual context to enhance appearance modeling.

 doi:10.20944/preprints202605.0113.v1

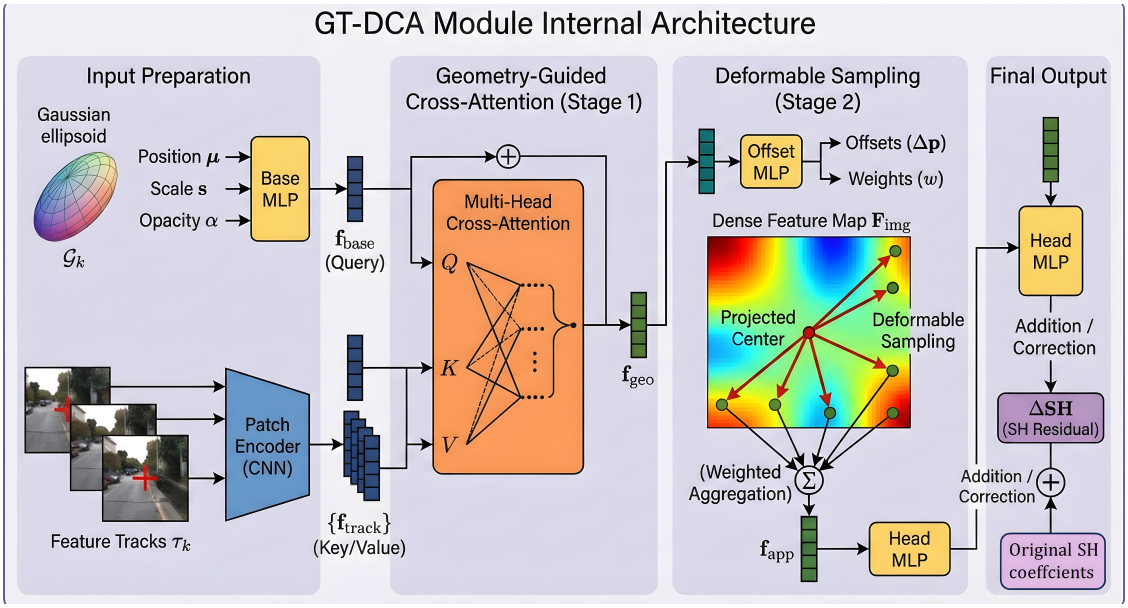

Figure 6. Architecture of the GT-DCA module. By utilizing feature tracks as geometry-aware queries and applying deformable cross-attention on deep image features, GT-DCA adaptively aggregates view-dependent appearance cues (e.g., highlights) without compromising the underlying geometric structure.

Learnable Feature Extraction.

To capture fine-grained texture and specular cues, we employ a shared **lightweight CNN encoder** (Patch Encoder) to extract dense feature maps $\mathbf{F}_{img}$ from input images. This encoder consists of 3 shallow residual blocks and maps RGB images to a feature space of dimension $D = 64$. (Detailed architecture provided in Supp. Material).

Geometry-Guided Sampling.

For each feature track τ_k associated with a Gaussian, we project its 3D anchor to the image plane to obtain observations $\{(i, \mathbf{u}_{k,i})\}$. We then perform **bilinear interpolation** on the feature map $\mathbf{F}_{img}^{(i)}$ at coordinates $\mathbf{u}_{k,i}$. These sampled features serve as *geometry-anchored keys and values* for the attention mechanism ($\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times D}$), ensuring the network learns from actual image content rather than just spatial positions.

Deformable Aggregation.

To robustly aggregate multi-view information, we employ a multi-head cross-attention layer ($H = 4$ heads). The query $\mathbf{Q} \in \mathbb{R}^{1\times D}$ is derived from the Gaussian's geometric attributes via a linear projection. The attention mechanism dynamically weights the contributions from different views based on their feature similarity, producing a context-aware appearance embedding. To handle minor misalignment, we further apply deformable sampling with $M = 8$ offset points around the projected center. The final output predicts an SH residual $\Delta\mathbf{c}$, refining the base color.

SH update.

Finally, the output embedding $\mathbf{f}_{\text{app}}^{(k)}$ predicts an SH residual:

$$\Delta\mathbf{c}_k = \text{MLP}_{\text{SH}}(\mathbf{f}_{\text{app}}^{(k)}), \quad \mathbf{c}_k^{\text{eff}} = \mathbf{c}_k^{\text{base}} + \Delta\mathbf{c}_k, \tag{18}$$

which refines the base appearance in the differentiable renderer.

4. Experiments

We now evaluate **GeoTrack-GS** on three standard benchmarks (DTU, LLFF, and Mip-NeRF 360) under challenging sparse-view settings. We first describe the experimental setup and baselines, then

 doi:10.20944/preprints202605.0113.v1

present state-of-the-art (SOTA) comparisons, followed by robustness analysis, ablations, and additional diagnostics. For Mip-NeRF 360, we follow a 12-view sparse-input protocol (Table 3).

4.1. Experimental Setup

Datasets and metrics.

We evaluate GeoTrack-GS on three standard benchmarks: **DTU** [47], **LLFF** [1], and **Mip-NeRF 360** [2]. Following sparse-view protocols, we train with 3, 6, or 9 input views on LLFF and DTU and evaluate on held-out views. For Mip-NeRF 360, we follow a 12-view protocol to ensure sufficient scene coverage under unbounded, wide-baseline capture, and we train/evaluate all methods under the same 12-view setting (Table 3). LLFF images are downsampled by $8\times$, and DTU and Mip-NeRF 360 by $4\times$. We report PSNR, SSIM, and LPIPS, and additionally evaluate geometric quality on DTU using Chamfer Distance (CD-L1) and F-score.

Evaluation protocol.

PSNR/SSIM are computed in sRGB space; LPIPS uses the VGG backbone (official implementation). For DTU, we apply the provided foreground mask for photometric metrics and use the standard DTU evaluation code for geometry. All renderings are resized to match the ground-truth resolution using bilinear interpolation. Full details (library versions, DTU thresholds, and point-cloud extraction) are provided in the Supplementary Material.

Baselines and protocol.

We compare against representative prior-guided and self-regularized 3DGS methods, including **3DGS**, **FSGS**, **DNGaussian**, **CoR-GS**, **FewViewGS**, and **SCGaussian**. To ensure a fair comparison, all depth-guided methods use the *same* monocular depth maps (DPT-Hybrid), and all methods share the *same* COLMAP initialization.

Implementation details.

All experiments are conducted on a single NVIDIA RTX 4090 GPU with 30k training iterations. We set the number of neighbors for PCA to $K = 16$, the anisotropy threshold to $\theta_{\text{aniso}} = 5.0$, and the geometric gating exponent to $\gamma = 2$. Additional hyperparameters, learning rates, and dataset-specific statistics are provided in Tables S2 and S3 of the Supplementary Material.

4.2. Quantitative Comparison with Existing Methods

We compare GeoTrack-GS against prior methods on DTU, LLFF, and Mip-NeRF 360 under the same *per-dataset* sparse-input protocol; Tables 1–3 summarize the results (DTU/LLFF: 3 views; Mip-NeRF 360: 12 views).

Metric discrepancies in prior works often come from different evaluation settings (LPIPS backbone, color space, masking/cropping, resizing). We therefore re-evaluate *all* methods with a single protocol and the same scripts and report only these unified results.

On DTU (Table 1), GeoTrack-GS attains PSNR comparable to strong depth-guided baselines such as FSGS and DNGaussian, while often yielding lower Chamfer Distance and higher F-Score than both prior-guided and self-regularized methods. This indicates that our geometry-first design improves 3D reconstruction quality while keeping photometric accuracy competitive under the same protocol. The geometric benefits are further illustrated by qualitative comparisons on *Scan 8* (see Figure 9), where our method effectively eliminates the surface noise and wavy distortions common in prior works.

 doi:10.20944/preprints202605.0113.v1

Table 1. Sparse-view comparison on **DTU** with 3 input views at 1/4 image resolution. We report rendering quality (PSNR ↑, SSIM ↑, LPIPS (VGG) ↓) and geometric quality (F-score ↑, CD-L1 (mm) ↓). All methods are evaluated using the same scripts and settings under our unified protocol. We do not mix quoted numbers from prior papers in the main tables.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	F-score ↑	CD-L1 (mm) ↓
3DGS (Original)	17.65	0.779	0.146	20.8	0.96
FSGS (Baseline)	18.12	0.811	0.131	32.4	0.75
DNGaussian	18.91	0.790	0.102	38.5	0.62
CoR-GS	19.21	0.853	0.119	40.5	0.58
SparseGS	18.89	0.834	0.178	31.2	0.82
SE-GS	19.24	0.857	0.132	36.8	0.68
DropoutGS	20.22	0.830	0.150	32.5	0.95
SCGaussian	20.56	0.864	0.122	42.0	0.52
FewViewGS	19.74	0.861	0.127	39.0	0.65
Ours (GeoTrack-GS)	**19.96**	0.847	0.126	**43.0**	**0.49**

On LLFF (Table 2), vanilla 3DGS degrades noticeably under sparse views, with visible artifacts and reduced PSNR/SSIM. Depth-regularized methods (FSGS, DNGaussian) and self-regularized ones (CoR-GS, SparseGS, SE-GS, DropoutGS, FewViewGS, SCGaussian) substantially improve visual quality over the vanilla baseline. GeoTrack-GS achieves competitive PSNR/SSIM while offering stronger geometric stability (fewer floaters and less "waxy" surfaces); LPIPS is comparable to baselines but not consistently better, indicating competitive perceptual quality. Qualitative comparisons (see Figure 8 and Figure 7) further show fewer floaters and less "waxy" geometry, especially around depth discontinuities and textureless regions.

Table 2. Sparse-view comparison on **LLFF** with 3 input views at 1/8 image resolution. We report PSNR ↑, SSIM ↑, and LPIPS (VGG) ↓. All numbers are re-evaluated under the same protocol using our unified evaluation scripts.

Method	PSNR ↑	SSIM ↑	LPIPS (VGG) ↓
3DGS (Original)	18.22	0.591	0.229
FSGS (Baseline)	20.31	0.652	0.288
DNGaussian	19.12	0.649	0.294
SE-GS	20.79	0.724	0.183
SparseGS	19.86	0.668	0.322
DropoutGS	19.35	0.622	0.282
FewViewGS	20.54	0.693	0.214
CoR-GS	20.45	0.712	0.196
SCGaussian	20.77	0.705	0.218
Ours (GeoTrack-GS)	20.52	0.691	0.231

On Mip-NeRF 360 (Table 3), GeoTrack-GS achieves competitive PSNR/SSIM while offering strong geometric stability; LPIPS is comparable to baselines but not consistently better, particularly in scenes with complex backgrounds and large camera baselines. These results suggest that our hybrid geometric constraints and track-guided appearance modeling generalize beyond forward-facing scenes to more challenging unbounded environments.

 doi:10.20944/preprints202605.0113.v1

Table 3. Sparse-view comparison on **Mip-NeRF 360** at 1/4 image resolution under **12 input views**. We report PSNR ↑, SSIM ↑, and LPIPS (VGG) ↓. All methods (including 3DGS) are trained and evaluated with the same view count and the same protocol.

Method	PSNR ↑	SSIM ↑	LPIPS (VGG) ↓
3DGS (12 views)	17.49	0.413	0.499
FSGS	18.46	0.515	0.479
SparseGS	19.37	0.577	0.398
SE-GS	19.91	0.596	0.400
CoR-GS	19.52	0.558	0.418
Ours (GeoTrack-GS)	19.45	0.575	0.435

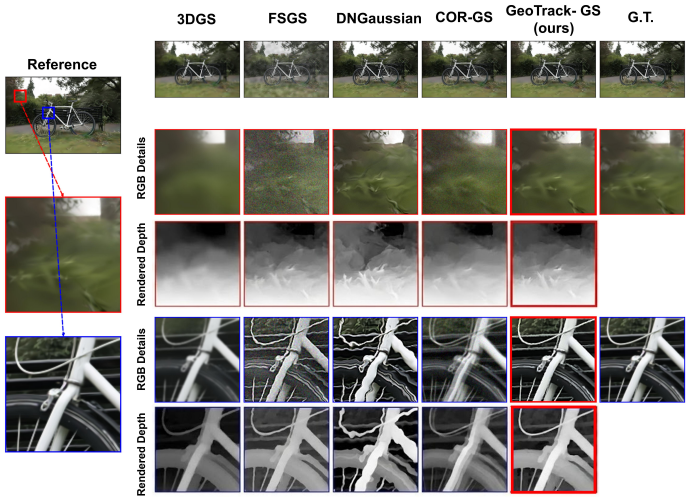

Figure 7. Qualitative comparison on Mip-NeRF 360 *Bicycle*. Baselines like 3DGS and FSGS fail to reconstruct thin structures (spokes) or over-smooth textures (grass). GeoTrack-GS recovers sharp topology and fine details consistent with the Ground Truth.

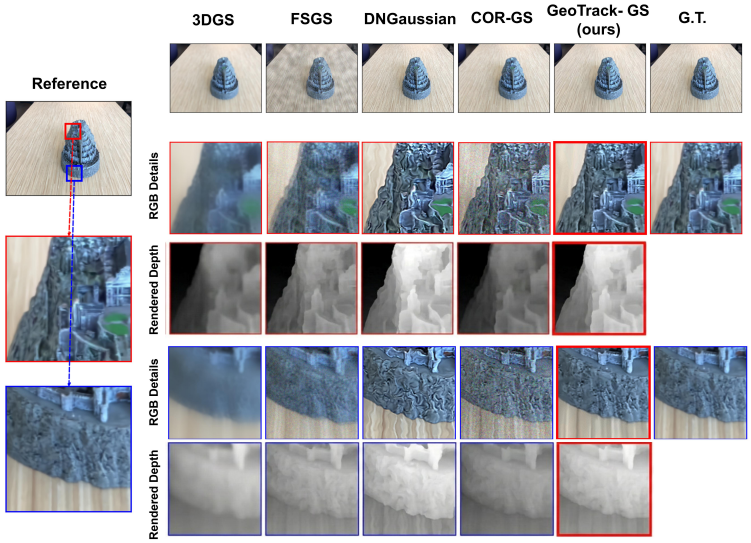

Figure 8. Visual comparison on LLFF *Fortress*. Our method corrects the severe geometric distortions (wavy artifacts) observed in CoR-GS and the over-smoothing in DNGaussian. By leveraging feature tracks, GeoTrack-GS produces consistent depth maps and preserves high-frequency texture details lost in prior-guided baselines.

 doi:10.20944/preprints202605.0113.v1

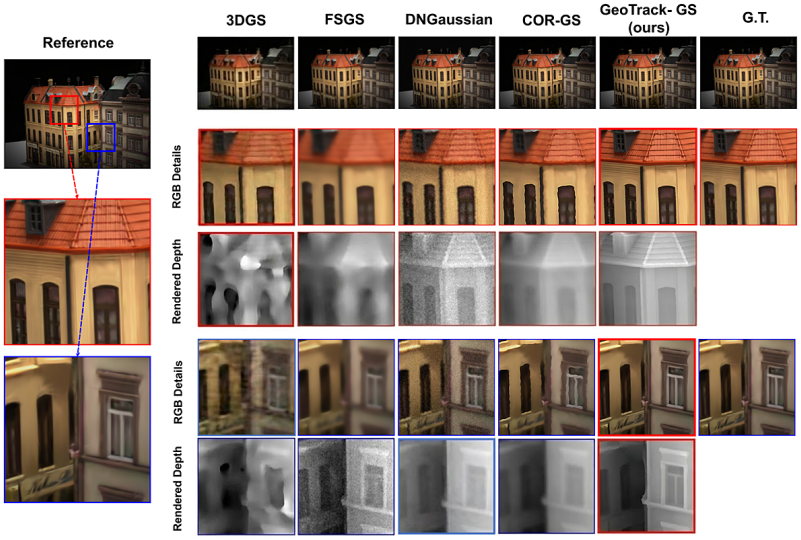

Figure 9. Visual comparison on DTU *Scan 8*. We compare RGB renderings and depth maps. GeoTrack-GS demonstrates superior geometric fidelity, effectively eliminating high-frequency surface noise (FSGS) and wavy distortions (CoR-GS) while accurately preserving view-dependent specular highlights compared to depth-guided baselines.

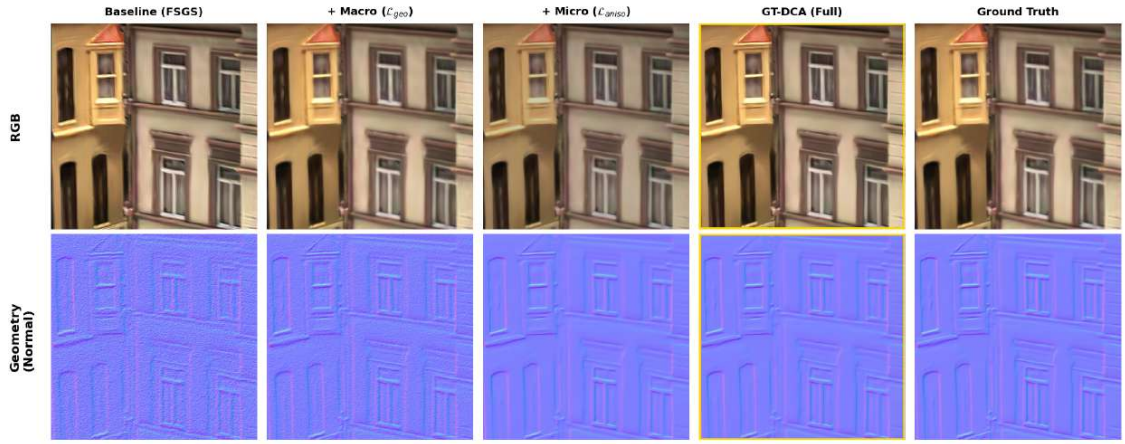

Figure 10. Visual ablation on DTU *Scan 24*. Left: Excluding geometric constraints results in noisy surfaces. Adding $\mathcal{L}_{aniso}$ smooths geometry. **Right:** The full model with GT-DCA further enhances high-frequency texture details (e.g., window frames) compared to the geometry-only variant.

4.3. Robustness to View Sparsity

To further evaluate robustness under extreme sparsity, we conduct a 3/6/9-view analysis on both LLFF and DTU datasets. For each scene, we randomly sample 3, 6, or 9 input views as training images and use the remaining views for evaluation.

Detailed numerical results for all view settings are provided in the **Supplementary Material**, which demonstrate that GeoTrack-GS degrades gracefully compared to baselines as the number of views decreases.

Specifically, on LLFF, prior-guided methods tend to trade off sharp geometry for smoother surfaces when views are scarce, while purely self-regularized methods lose high-frequency details. In contrast, GeoTrack-GS maintains competitive perceptual quality, and its main advantage lies in strong geometric stability under extreme sparsity. Furthermore, on DTU, our method attains consistently lower Chamfer Distance across different view settings, confirming that the decoupled constraints effectively preserve 3D structure even under extreme 3-view supervision.

 doi:10.20944/preprints202605.0113.v1

4.4. Ablation Study

We now analyze the contribution of each component of GeoTrack-GS. Unless otherwise noted, we perform the ablation on DTU in a sparse-view setting and report averaged metrics over the evaluated scans.

We start from a depth-guided 3DGS baseline (FSGS) and progressively add our proposed modules: macro-level geometry ($\mathcal{L}_{geo}$), micro-level anisotropic regularization ($\mathcal{L}_{aniso}$), the GT-DCA appearance module, and the Decoupled Constraint Stabilization (DCS).

From Table 4, we observe a clear geometry–appearance trade-off that is progressively resolved as modules are added. (i) Introducing the macro-level geometric loss $\mathcal{L}_{geo}$ yields a noticeable reduction in CD-L1 and an increase in F-score, indicating that anchoring Gaussians to feature-track geometry corrects large-scale errors introduced by noisy monocular priors. Notably, this step may temporarily degrade SSIM/LPIPS, as stronger geometric correction can reduce photometric fit before local shape is stabilized. (ii) Adding the micro-level anisotropic regularizer $\mathcal{L}_{aniso}$ further improves geometric fidelity and partially recovers photometric/perceptual metrics by regularizing local surface structure and suppressing rank-collapse artifacts. (iii) On top of the stabilized geometry, GT-DCA significantly improves LPIPS while keeping PSNR/SSIM competitive, suggesting that geometry-guided appearance aggregation better captures view-dependent effects without distorting the underlying 3D structure. (iv) Finally, DCS further improves the overall trade-off and stabilizes training, yielding the best combined photometric and geometric performance.

Table 4. Component-wise ablation on **DTU** dataset under 3-view supervision. Starting from the FSGS baseline, we progressively incorporate the Macro Geometry Loss ($\mathcal{L}_{geo}$), the Micro Anisotropic Regularizer ($\mathcal{L}_{aniso}$), the GT-DCA appearance module, and the Decoupled Constraint Stabilization (DCS). **Note:** By splitting metrics into separate columns, we ensure precise alignment.

	Components					**Metrics**				
Model	$\mathcal{L}_{depth}$	$\mathcal{L}_{geo}$	$\mathcal{L}_{aniso}$	GT-DCA	DCS	PSNR↑	SSIM↑	LPIPS↓	CD-L1↓	F-Score↑
Baseline (FSGS)	✓					18.12	0.811	0.131	0.75	32.4
+ Macro geometry	✓	✓				18.65	0.745	0.210	0.60	37.5
+ Micro geometry	✓	✓	✓			18.95	0.810	0.185	0.55	40.8
+ Geometry + GT-DCA	✓	✓	✓	✓		19.25	0.835	0.135	0.54	41.0
Full (GeoTrack-GS)	✓	✓	✓	✓	✓	**19.96**	**0.852**	**0.128**	**0.49**	**43.0**

Additional diagnostic analyses.

We provide detailed visual diagnostics on monocular depth prior inconsistency and geometry–appearance decoupling in the Supplementary Material.

4.5. Robustness to Track Sparsity and Training Stability

We observe that GeoTrack-GS remains robust when a large portion of feature tracks are removed, and DCS stabilizes early training. Detailed results and curves are provided in the Supplementary Material.(Section S1, Fig. S3).

4.6. Efficiency Analysis

GeoTrack-GS takes approximately 25 minutes per LLFF scene (3 views) on an RTX 4090, compared to 18 minutes for FSGS and 45 minutes for CoR-GS.(see Table S7 in Supplementary Material). The additional overhead mainly comes from the geometric constraints and GT-DCA module, while remaining practical for sparse-view reconstruction.

5. Conclusion and Discussion

In this paper, we presented **GeoTrack-GS**, a geometry-first framework for sparse-view 3D Gaussian Splatting. By treating monocular depth as a noisy prior and correcting it with robust multi-view feature tracks, we effectively address the geometric inconsistency issues prevalent in prior-guided

 doi:10.20944/preprints202605.0113.v1

methods. Our dual-level geometric constraints—combining a macro-level track-based reprojection loss with a micro-level anisotropic regularizer—provide a stable geometric skeleton. Furthermore, the proposed GT-DCA module enables high-fidelity appearance modeling without compromising the corrected geometry. Extensive experiments verify that GeoTrack-GS achieves state-of-the-art geometric fidelity while keeping rendering quality competitive under sparse inputs, particularly in challenging sparse-view scenarios.

Limitations and Future Work. Despite these improvements, our method relies on the availability of reliable feature tracks from SfM. In scenes with extreme texture-less regions or very wide baselines where feature matching fails, the corrective capability of $\mathcal{L}_{geo}$ diminishes. Additionally, the GT-DCA module introduces a moderate computational overhead. Future work will explore end-to-end joint optimization of poses and Gaussians to reduce dependency on offline SfM, and investigate lightweight distillation techniques to improve real-time efficiency.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Zhipeng Ye: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft. **Zihao Lu:** Investigation, Visualization, Validation, Formal analysis. **Yuan Zhang:** Formal analysis, Writing—review & editing. **Wenjie Qin:** Investigation, Resources, Data curation. **Jiayi Hong:** Software, Validation, Writing—review & editing. **Xuming Wu:** Conceptualization, Supervision, Resources, Funding acquisition, Writing—review & editing. **Zhibin Shao:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing—review & editing.

Funding: This work was supported by the the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515012277), the Research Foundation for Advanced Talents of Lingnan Normal University (No. ZL22001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We will release the code upon acceptance, and an anonymized version can be provided for review if required. All datasets used are publicly available (DTU/LLFF/Mip-NeRF 360).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Mildenhall, B.; Srinivasan, P.P.; Ortiz-Cayon, R.; Kalantari, N.K.; Ramamoorthi, R.; Ng, R.; Kar, A. Local Light Field Fusion: Practical View Synthesis for Wide-Baseline Images. In Proceedings of the ACM Transactions on Graphics (SIGGRAPH), 2019, Vol. 38, pp. 1–14. https://doi.org/10.1145/3306346.3323028.
2. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5470–5479. https://doi.org/10.1109/CVPR52688.2022.00541.
3. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 405–421.
4. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. In Proceedings of the ACM Transactions on Graphics (SIGGRAPH), 2022, Vol. 41, pp. 1–15. https://doi.org/10.1145/3528223.3530127.
5. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5855–5864. https://doi.org/10.1109/ICCV48922.2021.00582.
6. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. In Proceedings of the ACM Transactions on Graphics (SIGGRAPH), 2023, Vol. 42, pp. 1–14. https://doi.org/10.1145/3592403.3592423.

 doi:10.20944/preprints202605.0113.v1

7. Wu, T.; Yuan, Y.J.; Zhang, L.X.; Yang, J.; Cao, Y.P.; Yan, L.Q.; Gao, L. Recent Advances in 3D Gaussian Splatting. *arXiv preprint arXiv:2403.11134* **2024**. https://doi.org/10.48550/arXiv.2403.11134.
8. Gu, J.; Liu, B.; Wang, L.; Huang, Y.; Zhu, J.; Cheng, S.; Zhao, J.; Xu, Y.; Liu, Y.; Liu, R.; et al. A Survey on 3D Gaussian Splatting. *arXiv preprint arXiv:2401.03890* **2024**. https://doi.org/10.48550/arXiv.2401.03890.
9. Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104–4113. https://doi.org/10.1109/CVPR.2016.445.
10. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle Adjustment—A Modern Synthesis. In Proceedings of the International Workshop on Vision Algorithms (ICCV WA '99). Springer, 1999, pp. 298–372.
11. Hyung, J.; Hong, S.; Hwang, S.; Lee, J.; Kim, J.H.; Choo, J. Effective Rank Analysis and Regularization for Enhanced 3D Gaussian Splatting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37.
12. Jo, K.M.; Lee, D.H.; Han, D.H.; Lee, S.H.; Lee, H.G. GS-Reg: 3D Gaussian Splatting with Regularization. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2008–2017.
13. Xiong, H.; Muttukuru, S.; Xiao, H.; Upadhyay, R.; Chari, P.; Zhao, Y.; Kadambi, A. Sparsegs: Sparse view synthesis using 3d gaussian splatting. In Proceedings of the 2025 International Conference on 3D Vision (3DV). IEEE, 2025, pp. 1032–1041.
14. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022, Vol. 44, pp. 1537–1553. https://doi.org/10.1109/TPAMI.2020.3013691.
15. Deng, K.; Liu, A.; Zhu, J.Y.; Ramanan, D. Depth-Supervised NeRF: Fewer Views and Faster Training for Free. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12882–12891. https://doi.org/10.1109/CVPR52688.2022.01254.
16. Roessle, B.; Barron, J.T.; Mildenhall, B.; Sajjadi, M.S.M.; Gijsenij, A.; Radwan, N. Dense Depth Priors for Neural Radiance Fields from Sparse Input Views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12892–12901. https://doi.org/10.1109/CVPR52688.2022.01255.
17. Zhu, Z.; Fan, Z.; Jiang, Y.; Wang, Z. FSGS: Real-Time Few-Shot View Synthesis using Gaussian Splatting. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2024.
18. Li, J.; Zhang, J.; Yu, X.; Huang, L.; Gu, L.; Zheng, J.; Bai, X. DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5703–5712. https://doi.org/10.1109/CVPR52726.2024.00551.
19. Turkulainen, M.; Ren, X.; Melekhov, I.; Seiskari, O.; Rahtu, E.; Kannala, J. DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025, pp. 2421–2431. https://doi.org/10.1109/WACV61041.2025.00241.
20. Xu, H.; Peng, S.; Wang, F.; Blum, H.; Barath, D.; Geiger, A.; Pollefeys, M. DepthSplat: Connecting Gaussian Splatting and Depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
21. Shi, D.; Wang, W.; Chen, D.Y.; Zhang, Z.; Bian, J.W.; Zhuang, B.; Shen, C. Revisiting Depth Representations for Feed-Forward 3D Gaussian Splatting. *arXiv preprint arXiv:2506.05327* **2025**.
22. 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 Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2024.
23. Yin, R.; Yugay, V.; Li, Y.; Gevers, T.; Karaoglu, S. FewViewGS: Gaussian Splatting with Few View Matching and Multi-stage Training. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37, pp. 127204–127225.
24. Xu, Y.; Wang, L.; Chen, M.; Ao, S.; Li, L.; Guo, Y. DropoutGS: Dropping Out Gaussians for Better Sparse-view Rendering. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 701–710.
25. Niemeyer, M.; Barron, J.T.; Mildenhall, B.; Sajjadi, M.S.M.; Geiger, A.; Radwan, N. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In Proceedings of the IEEE/CVF Conference

 doi:10.20944/preprints202605.0113.v1

on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5491–5500. https://doi.org/10.1109/CVPR52688.2022.00543.
26. Yang, J.; Wang, B.; Guo, Y.; Chen, K.; Zhou, Z.; Wang, X. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20556–20565.
27. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching with Graph Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3204–3214. https://doi.org/10.1109/CVPR42600.2020.00327.
28. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2021, pp. 8922–8931.
29. Lindenberger, P.; Sarlin, P.E.; Pollefeys, M. LightGlue: Local Feature Matching at Light Speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17581–17592. https://doi.org/10.1109/ICCV51070.2023.01602.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30.
31. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.
32. Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; Geiger, A. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022, Vol. 35, pp. 15951–15963.
33. Fu, Q.; Xu, Q.; Ong, Y.S.; Tao, W. Geo-NeuS: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022, Vol. 35, pp. 8637–8649.
34. Ma, Y.; Wei, G.; Xiao, H.; Cheng, Y. HBSplat: Robust Sparse-View Gaussian Reconstruction with Hybrid-Loss Guided Depth and Bidirectional Warping. *arXiv preprint arXiv:2509.24893* **2025**.
35. Ma, Y.; Wei, G.; Cheng, Y. DWGS: Enhancing Sparse-View Gaussian Splatting with Hybrid-Loss Depth Estimation and Bidirectional Warping. *arXiv e-prints* **2025**, pp. arXiv–2509.
36. Jain, A.; Tancik, M.; Abbeel, P. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5885–5894. https://doi.org/10.1109/ICCV48922.2021.00583.
37. Kim, M.; Kweon, I.S.; Heo, M. InfoNeRF: Ray-Based Information Maximization for Few-Shot View Synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16723–16732.
38. Zhao, C.; Wang, X.; Zhang, T.; Javed, S.; Salzmann, M. Self-ensembling gaussian splatting for few-shot novel view synthesis. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 4940–4950.
39. 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. *Advances in Neural Information Processing Systems* **2024**, *37*, 97328–97352.
40. Hartley, R.; Zisserman, A. *Multiple view geometry in computer vision*; Cambridge university press, 2003.
41. Lin, C.H.; Ma, W.C.; Torralba, A.; Lucey, S. BARF: Bundle-Adjusting Neural Radiance Fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5741–5751. https://doi.org/10.1109/ICCV48922.2021.00571.
42. Mai, J.; Zhu, W.; Rojas, S.; Zarzar, J.; Hamdi, A.; Qian, G.; Li, B.; Giancola, S.; Ghanem, B. TrackNeRF: Bundle Adjusting NeRF from Sparse and Noisy Views via Feature Tracks. In Proceedings of the Computer Vision – ECCV 2024, 2024, pp. 470–489. https://doi.org/10.1007/978-3-031-73254-6_27.
43. Yang, Z.; Gao, X.; Sun, Y.T.; Huang, Y.H.; Lyu, X.; Zhou, W.; Jiao, S.; Qi, X.; Jin, X. Spec-Gaussian: Anisotropic View-Dependent Appearance for 3D Gaussian Splatting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37.
44. Tang, Z.J.; Cham, T.J. 3iGS: Factorised Tensorial Illumination for 3D Gaussian Splatting. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2024.

 doi:10.20944/preprints202605.0113.v1

45. Li, D.; Jiang, K.; Tang, Y.; Ramamoorthi, R.; Chellappa, R.; Peng, C. MS-GS: Multi-Appearance Sparse-View 3D Gaussian Splatting in the Wild. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
46. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity (SSIM). *IEEE Transactions on Image Processing* **2004**, *13*, 600–612. https://doi.org/10.1109/TIP.2003.819861.
47. Jensen, R.; Dahl, A.; Aanaes, H.; Dahl, V.A.; Stegmann, M.B.; Bærentzen, J.A.; Philipsen, P. Large Scale Multi-View Stereopsis Evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2014, pp. 446–453. https://doi.org/10.1109/CVPRW.2014.73.