PriSplat: Propagating Reliable Multi-view Information for Distractor-Free 3DGS

Yunseo Yang ⋆, Youngho Yoon ⋆, and Kuk-Jin Yoon

Visual Intelligence Lab., KAIST
{acorn, dudgh1732, kjyoon}@kaist.ac.kr

Abstract. Applying 3D Gaussian Splatting (3DGS) to uncontrolled, *in-the-wild* environments remains challenging due to transient distractors. Existing approaches typically mask these distractors and suppress their losses during training. However, this zero-masking strategy leaves such regions unsupervised, leading to per-view overfitting and causing severe floaters in novel views. To resolve this, it is essential to re-establish dense multi-view constraints by recovering the missing background information. In light of this, we propose **PriSplat**, a novel framework designed to propagate reliable multi-view information to restore these missing regions with high geometric integrity. Specifically, we repurpose a large-scale view synthesis prior into a 3D-aware inpainting engine, adapted through mask-aware fast-weight updates to prevent distractor leakage into the scene memory. To ensure the fidelity of this restoration, we introduce a geometry-aware support view selection algorithm based on information density and spatio-angular constraints. Ultimately, these synergistic advancements yield 3D-consistent pseudo-ground truth from masked regions, establishing the dense supervision necessary to eliminate artifacts. Extensive experiments show that our method outperforms state-of-the-art baselines in both synthesis quality and multi-view consistency. The code is available at `https://github.com/yun-seo/PriSplat.git`.

Keywords: 3DGS · Distractor-free · Multi-view information

1 Introduction

In modern 3D vision, Neural Radiance Fields (NeRF) [2,3,38] and 3D Gaussian Splatting (3DGS) [23] have become foundational frameworks for high-quality novel view synthesis (NVS). Notably, 3DGS has seen rapid adoption due to its efficient optimization and explicit scene representation, which facilitate a wide range of downstream applications [1,11,13,68].

Despite recent progress, the practical efficacy of these frameworks is limited by the strong assumption of a static scene [3,17,20,38,76,77]. In real-world scenes, however, transient distractors such as moving vehicles and pedestrians frequently appear, breaking the multi-view consistency essential for reliable reconstruction. Consequently, observations become mutually inconsistent across views, injecting noise into the optimization and ultimately degrading rendering quality.

⋆ Equal contribution.

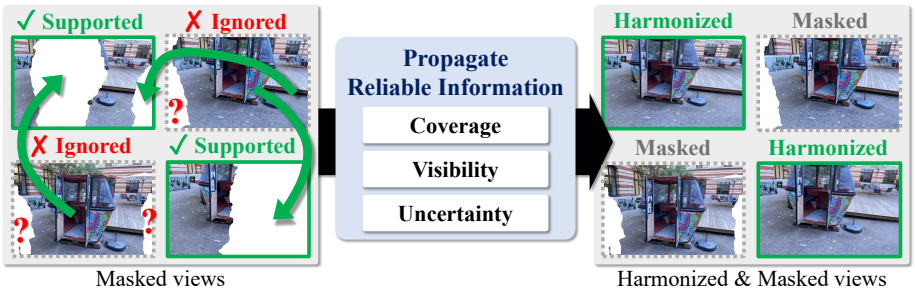

Fig. 1: Propagate reliable information. Instead of simply masking out distractors (left), our framework propagates reliable multi-view evidence by explicitly modeling coverage, visibility, and uncertainty (middle). This yields harmonized views (right) that provide consistent supervision and mitigate distractor-induced noise.

To achieve robust optimization in such scenarios, recent studies have proposed **distractor-free 3DGS** [15,16,26,32,43,48,59,61,70,79]. These methods commonly rely on reconstruction residuals to locate distractors and then ignore those areas during loss computation. However, this masking strategy removes critical supervision, weakening multi-view constraints and biasing optimization toward per-view fitting. Consequently, the resulting scenes often suffer from persistent artifacts, including floaters, ghosting, and localized geometry collapse.

To overcome this fundamental limitation, we exploit the inherent **multi-view redundancy** of unconstrained image collections. While distractors may fully obscure a background region in one view, the same region is often visible from other viewpoints, providing reliable multi-view evidence. This suggests that the masked regions can be recovered from other viewpoints where they remain unoccluded, restoring the dense multi-view constraints that masking removes.

A natural way to fill these masked regions is to borrow the power of generative inpainting models [4,8,27,28,34,47,51,53]. However, such models synthesize content purely in 2D, with no awareness of the underlying 3D geometry, so directly plugging their outputs into 3DGS optimization breaks multi-view consistency. To compensate, RogSplat [25] uses intermediate 3DGS renderings as reference and fills masked regions with generative priors [27, 28] during optimization. However, this approach may still face two issues. First, diffusion priors can generate 2D-plausible yet geometrically inconsistent content, reintroducing cross-view conflicts. Second, when inpainting is conditioned on rendered images, early rendering artifacts under heavy occlusions can corrupt the guidance, gradually reducing its reliability. As a result, 3DGS is supervised by unreliable inpainting results, often leading to floaters and hazy geometry.

To address these issues, we introduce **PriSplat**, a framework that achieves distractor-free 3DGS by propagating reliable multi-view information via 3D-aware inpainting. Our approach builds on the key insight that NVS is fundamentally a reference-based inpainting process, where unobserved regions are inferred from other views. We therefore repurpose a large-scale NVS prior [81]

as a multi-view harmonizer that aggregates unoccluded evidence across views to fill the masked regions. Unlike diffusion models that often hallucinate geometrically inconsistent details, our harmonizer enforces structural consistency through pose-conditioned spatial constraints. To further suppress transient distractor leakage at test time, we incorporate a **mask-aware fast-weight update**. Together, these mechanisms provide the stable pseudo-ground truth necessary for distractor-free 3DGS.

The quality of this harmonization, however, is sensitive to the evidence it is conditioned on. In our setting, this evidence comes from a small set of neighboring views, called support views, around a masked target (see Fig. 1, left). When these support views are chosen poorly or are themselves heavily occluded, the resulting inpainting can distort the reconstructed geometry. We therefore introduce a **geometry-aware support view selection** algorithm that supplies the harmonizer with reliable guidance. For each masked region, it ranks candidate views by a single selection score that combines Fisher-based information density and spatio-angular compatibility, retrieving the most informative neighbors.

Synergizing this selection strategy with the geometric constraints of our reformulated NVS prior, we transform empty masks into view-consistent pseudo-ground truth. Ultimately, this unified pipeline mitigates the spatial inconsistencies inherent to prior methods, enabling 3DGS optimization to faithfully converge toward a clean, distractor-free geometry.

Our contributions are summarized as follows:

- We propose **PriSplat**, shifting the paradigm of distractor-free 3DGS from simple masking to 3D-aware multi-view harmonization.
- We develop a **multi-view harmonizer** by repurposing a large-scale NVS prior, turning fragmented cross-view cues into 3D-consistent pseudo-ground truth for dense 3DGS supervision.
- We introduce a **geometry-aware support view selection** algorithm that combines Fisher information with spatio-angular metrics to identify reliable cross-view evidence and avoid injecting ambiguous signals.
- Our framework achieves consistent improvements over prior methods, removing floaters and artifacts in challenging *in-the-wild* scenes.

2 Related works

Distractors in NeRF. When NeRF [38] is applied to unconstrained *in-the-wild* collections [22], moving objects and occluders inevitably corrupt the reconstruction. To address this, several directions have emerged. A first group designs robust objectives or uncertainty models that down-weight distractors during fitting [9, 37, 49, 71]. A second group leverages pretrained priors [24, 41, 44] to segment or ignore distractor entities through heuristic or data-driven cues [7, 29, 46]. A third group explicitly separates static structure from dynamic content, recovering clean static geometry while isolating transient distractors [42, 67]. While effective at reducing distractor leakage, these methods often depend on scene-specific optimization or on the quality of auxiliary segmentation. Moreover, the

implicit nature of NeRF makes it difficult to localize and remove residual floaters explicitly, motivating a shift toward more explicit scene representations.

Distractors in 3DGS. Building on the success of 3DGS [23], recent works have extended the framework to unconstrained photo collections. One line of work focuses on appearance modeling by optimizing per-image embeddings [10, 14, 26, 54, 58, 64, 70, 79] or per-Gaussian parameters [12, 69], thereby allowing view-dependent color and tone shifts to be absorbed without distorting the scene geometry. Another line of work focuses on identifying distractor regions during training [16, 26, 36, 48, 56, 62]. To achieve this, these approaches compare rendered residuals with input images and measure inconsistencies in pretrained features [24,41,52] to down-weight unreliable pixels. SpotLessSplats [48] predicts per-pixel reliability with an MLP guided by self-supervised 2D semantics [52]. T-3DGS [36] fuses DINOv2 features and residual error, refines the result using an off-the-shelf segmenter [45], and propagates the mask temporally. RobustSplat [16] uses multiscale DINOv2 features to stabilize static structure and identify distractor leakage from early densification, and its extension [15] further mitigates this issue alongside *in-the-wild* illumination shifts. Concurrently, a new paradigm has emerged that not only represents distractors as per-view or deformable fields [31, 43, 59, 61] but also quantifies scene reliability through uncertainty modeling. OCSplats [32] and Hui et al. [19] formalize this by quantifying multi-view redundancy and joint uncertainties, respectively, to distinguish true noise from under-observed regions. Additionally, RogSplat [25] leverages a diffusion prior to actively refine unreliable regions by inpainting them from current renderings, and uses the revised views to continue optimization. Despite these advances, most works still focus on suppressing harmful regions.

Inpainting-assisted NVS. To restore occluded areas while preserving continuity across views, inpainting can exploit geometric cues such as depth, visibility, and reprojection [6, 33, 35, 40, 57, 72, 75]. Existing works largely follow two patterns. One line inpaints a small set of reference views and then lifts the result into 3D to propagate it to other viewpoints [39,63,73,78], sometimes inpainting only the views that preserve geometric consistency [18, 66]. Such methods are simple and efficient, but their quality hinges on the inpainted references and on sufficient view coverage. The other line trains or adapts the inpainting model under multi-view and geometric constraints, so that the prior itself internalizes multi-view consistency [5,30,50,60,74]. This incurs additional optimization cost but reduces reliance on any single reference, yielding a better balance between consistency and robustness. Our work follows the latter direction, but targets a distinct setting: rather than filling user-specified holes, we inpaint the regions corrupted by transient distractors to enable distractor-free NVS.

3 Methods

3.1 Overview

We target robust NVS in real-world scenes corrupted by transient distractors. Prior mask-based pipelines discard masked observations, leading to floaters and

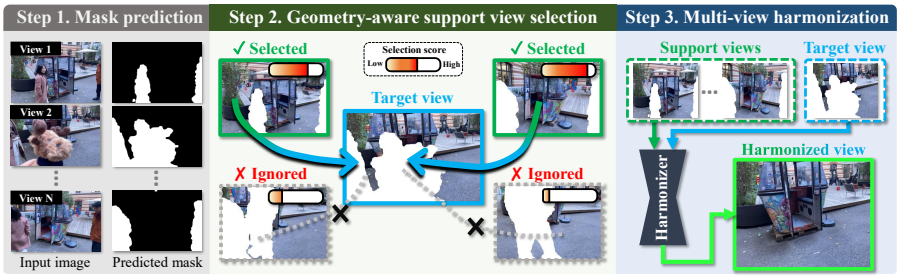

Fig. 2: Overview of PriSplat. Given distractor-corrupted inputs, PriSplat proceeds in three steps. (1) We estimate per-view distractor masks. (2) For each target view, we select geometry-aware support views that reliably observe the masked content, scored by reliability (high scores kept, low scores ignored). (3) Our harmonizer aggregates evidence from the selected support views to generate a 3D-consistent harmonized view, which serves as pseudo-ground truth. This harmonized supervision is then integrated into the 3DGS optimization loop, moving beyond simple mask-based loss suppression.

structural inconsistencies. As illustrated in Fig. 2, PriSplat addresses this through a harmonization stage embedded in the optimization loop. We begin with standard mask-based optimization. A mask predictor [16] estimates per-view distractor masks (**Step 1**), and we optimize 3DGS using only unmasked pixels to avoid fitting transient objects. Once the geometry becomes sufficiently stable, we temporarily invoke the harmonization stage. For each target view with masked regions, we first select support views that observe the missing background from nearby viewpoints (**Step 2**). The harmonizer then propagates reliable multi-view evidence from these supports into the occluded regions, producing a harmonized image as pseudo-ground truth (**Step 3**). Finally, we resume 3DGS optimization with this harmonized supervision, yielding denser and more consistent supervision than mask-only pipelines. The remainder of this section is organized as follows. Section 3.2 defines the problem setting, Sec. 3.3 details the multi-view harmonizer, Sec. 3.4 describes the geometry-aware support view selection, and Sec. 3.5 outlines the training objectives for 3DGS.

3.2 Problem setting

We follow the standard 3DGS formulation [23], where a static 3D scene is represented as a collection of K anisotropic Gaussian primitives $\mathcal{G} = \{g_k\}_{k=1}^{K}$. Each Gaussian g_k is characterized by its center position, an opacity, a 3D covariance matrix, and a set of spherical harmonics coefficients.

The model parameters θ are optimized using a set of N posed training images $\{(I_i, P_i)\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$ is the observed image and P_i denotes the corresponding calibrated camera parameters. To handle transient objects, we employ per-view binary distractor masks $M_i \in \{0,1\}^{H\times W}$. Here, $M_i(p) = 1$ indicates that pixel p in image I_i is identified as a distractor, while $M_i(p) = 0$ represents the static background. These masks are estimated online via a lightweight

MLP module following [16]. Under this setup, the standard optimization target for distractor-free 3DGS is the masked reconstruction loss $\mathcal{L}_{\text{mask}}$:

$$\mathcal{L}_{\text{mask}} = \sum_{i=1}^{N} \sum_{p \in \Omega} (1 - M_i(p)) \cdot \mathcal{D}(f_{p,i}(\theta), I_i(p)), \tag{1}$$

where $f_{p,i}(\theta)$ is the color rendered at pixel p for view i using the current parameters θ, Ω is the image pixel grid, and $\mathcal{D}(\cdot)$ is a distance metric, typically defined as a weighted combination of $\mathcal{L}_1$ and D-SSIM losses [65].

3.3 Multi-view harmonizer

Motivation. Beyond mask-based suppression, we aim to stabilize 3DGS optimization by actively filling masked regions with multi-view cues. The key challenge is aggregating fragmented evidence across views to reconstruct occlusions. While modern generative inpainting has advanced rapidly, it often introduces hallucinations that break multi-view consistency. Thus, we need an inpainting engine that can reliably gather and harmonize information across wide scenes.
Preliminaries. Here we briefly review the transformer-based NVS model [81] that we repurpose as our multi-view harmonizer. Given posed inputs $\{(I_i, P_i)\}_{i=1}^{N}$ and a target pose P_t, the model predicts the novel view as

$$\hat{I}_t = \texttt{NVS}\big(\{(I_i, P_i)\}_{i=1}^{N},\, P_t\big)\,. \tag{2}$$

Internally, the model patchifies each input image and projects patches into tokens, which are aggregated across views via self-attention. A crucial step is a *fast-weight update* performed at test time: instead of updating all parameters, the model updates scene-specific multi-view evidence into a small set of internal weights, forming a lightweight scene memory. This memory is written from the input views and later queried to synthesize the target view.

Formally, let K, V be key and value tokens extracted from the posed inputs and let Q_t denote query tokens for the target pose. The fast-weight W update mechanism can be summarized as

$$\Delta W = \texttt{Update}(W; K, V, \eta), \qquad W' \leftarrow W + \Delta W, \qquad \hat{I}_t \leftarrow \texttt{Apply}(W'; Q_t), \tag{3}$$

where η denotes token-wise learning rates. In `Update`, the fast weights are adapted using only the input views at their observed poses, and `Apply` uses the updated weights W' to render the target view given Q_t.
Repurposing view synthesis priors for inpainting. In our setting, unlike standard NVS benchmarks, the input images themselves may contain transient distractors. If the model naively aggregates all views, distractor-corrupted regions can be absorbed into the internal scene representation, thereby degrading target view synthesis. We therefore repurpose the NVS prior via (1) a mask-based reconstruction objective and (2) a mask-aware fast-weight update, making the adaptation robust to distractor-corrupted inputs.

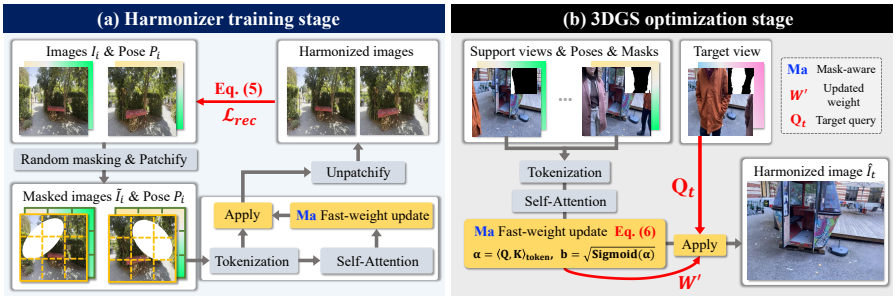

Fig. 3: Multi-view harmonizer. (a) During the harmonizer training stage, we train it to reconstruct clean content from masked inputs, encouraging cross-view evidence aggregation. (b) During the 3DGS optimization stage, a mask-aware fast-weight update suppresses distractor regions and propagates reliable multi-view information.

(1) Training objective under randomly masked inputs. In essence, the NVS model already performs a form of 3D-aware inpainting: it predicts missing target-view content by retrieving multi-view-consistent evidence from neighboring posed observations. We therefore fine-tune the NVS prior to operate as a 3D-aware inpainting model (Fig. 3(a)). Specifically, we synthetically mask each input view and train the model to reconstruct the corresponding clean view from the remaining multi-view context. Concretely, for each input view, we sample a binary mask $M_i \in \{0,1\}^{H\times W}$ and construct a masked observation

$$\tilde{I}_i = (1 - M_i) \odot I_i, \tag{4}$$

where $\odot$ denotes element-wise multiplication (masked pixels are set to zero). The masked image $\tilde{I}_i$ and its pose P_i are patchified and embedded, processed by self-attention, and then adapted via the mask-aware fast-weight update described in the next paragraph. We supervise the model to reconstruct the clean view via

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} \left\| \texttt{Harmonizer}\Big(\{(\tilde{I}_j, P_j)\}_{j=1}^{N},\, P_i\Big) - I_i \right\|_2^2 . \tag{5}$$

This objective encourages the model to aggregate unmasked, cross-view-consistent evidence to recover missing regions, rather than relying on image-space priors.

(2) Mask-aware fast-weight update. As discussed above, transient distractors can corrupt the standard fast-weight update by encoding distractor-corrupted evidence into the internal scene representation. To mitigate this, we propose a mask-aware fast-weight update. Formally, we compute

$$\alpha = \langle Q, K\rangle_{\text{token}}, \quad b = \sqrt{\sigma(\alpha)}, \quad \Delta W = \texttt{Update}(W;\ b \odot K,\ b \odot V,\ \eta)\,, \tag{6}$$

where α denotes the per-token alignment between queries and keys of input tokens, $\sigma(\cdot)$ is the sigmoid function, and $b \in (0,1)$ down-weights weakly aligned tokens while emphasizing well-aligned ones. This prevents distractor-corrupted

evidence from being incorporated into the fast-weight memory, improving inpainting robustness under masked observations.

In summary, our distractor-free 3DGS optimization proceeds as follows. We first predict per-view distractor masks using prior mask-based frameworks. Once the scene geometry stabilizes at an intermediate stage of optimization, we invoke our harmonizer to inpaint the target view at pose P_t using the posed inputs and predicted masks (Fig. 3(b)).

$$\hat{I}_t = \texttt{Harmonizer}\Big(\{(\tilde{I}_j, P_j)\}_{j=1}^{N},\, P_t\Big)\,. \tag{7}$$

The final harmonized view $\hat{I}_t$ is generated via the tokenization, self-attention, and fast-weight update steps described above. We then continue 3DGS optimization using the harmonized views as pseudo-ground truth, providing dense supervision.

3.4 Geometry-aware support view selection

Our harmonizer performs 3D-aware inpainting by aggregating cross-view evidence, reducing the unconstrained hallucinations typical of image-space priors. However, harmonization is reliable only when the masked region is well supported by other views. When such evidence is limited, the harmonizer may produce unstable or inconsistent outputs. Using these outputs as pseudo-ground truth can introduce inconsistent supervisory signals and hinder 3DGS optimization.

To mitigate this risk, we perform harmonization selectively when multi-view evidence is sufficiently reliable. Concretely, for each target view, we choose support views that (1) provide high information density in the masked region of the target and (2) satisfy spatio-angular constraints with respect to the target. This strategy ensures the harmonizer propagates only consistent, well-supported content, instead of amplifying view-specific artifacts.

Information density. For each target view t, we seek a support set $\mathcal{S}_t^{\text{sup}}$ that provides reliable observations for the regions masked in t, while remaining geometrically compatible with it. Let $\mathcal{S}_t = \{1, \ldots, N\} \setminus \{t\}$ be the candidate view set excluding t and $\mathcal{G}_t^{\text{rel}}$ denote the set of Gaussians whose projections fall inside the masked region of t. For each candidate view $s \in \mathcal{S}_t$, a desirable support should provide *clean* and *informative* observations for many Gaussians in $\mathcal{G}_t^{\text{rel}}$. We thus define an information density score:

$$r_{t,s} = \frac{1}{|\mathcal{G}_t^{\text{rel}}|} \sum_{k \in \mathcal{G}_t^{\text{rel}}} \mathbb{I}[g_k \text{ is well-observed in view } s]\,, \tag{8}$$

where $\mathbb{I}[\cdot]$ is an indicator function. Intuitively, $r_{t,s}$ is high when s provides reliable evidence for most of the masked content of the target. To instantiate $\mathbb{I}[\cdot]$, we exploit a Fisher-based score [21] computed from the current 3DGS model. Formally, given a Gaussian g_k and the relevant pixel set $\mathcal{U}$ in view s, we define $\mathcal{F}_{k,s}(\mathcal{U}) = \sum_{p \in \mathcal{U}} \left\| \frac{\partial f_{p,s}(\theta)}{\partial \theta_k} \right\|_2^2$, where $f_{p,s}(\theta)$ is the rendered color at pixel p and θ_k are the parameters of g_k. This score measures how sensitively pixels in view

s respond to perturbations of g_k, *i.e.*, how strongly that view supervises the underlying 3D primitive. We use $\mathcal{F}_{k,s}$ to decide whether g_k is well-observed in view s. Further implementation details are provided in the *Suppl.*

Spatio-angular constraints. To select geometrically compatible support views, we impose spatio-angular constraints that penalize candidate views far from the target view or with significantly different viewing directions:

$$w_{\text{dist}}(t,s) = \exp\left(-\frac{\|c_t - c_s\|_2}{\sigma_{\text{dist}}}\right), \qquad w_{\text{ang}}(t,s) = \max(0, v_t \cdot v_s)^{k_{\text{ang}}}, \tag{9}$$

where c_t, c_s are camera centers and v_t, v_s are unit forward vectors. Here σ_{dist} and k_{ang} are fixed hyperparameters.

Selection. We combine the three cues into a single reliability score:

$$\rho_{t,s} = r_{t,s} \cdot w_{\text{dist}}(t,s) \cdot w_{\text{ang}}(t,s). \tag{10}$$

We select support views using an absolute threshold τ and a relative threshold τ_{rel} to discard candidates that score far below the best one:

$$\rho_t^{\max} = \max_{s \in \mathcal{S}_t} \rho_{t,s}, \qquad \mathcal{S}_t^{\text{sup}} = \left\{ s \in \mathcal{S}_t \;\middle|\; \rho_{t,s} > \tau \;\wedge\; \rho_{t,s} \geq \tau_{\text{rel}}\, \rho_t^{\max} \right\}. \tag{11}$$

We execute harmonization for view t only if a sufficient number of reliable supports are available, i.e., harmonize$(t) = \mathbb{I}[\,|\mathcal{S}_t^{\text{sup}}| \geq K_{\min}]$. If fewer than $K_{\min}$ support views are available, we skip harmonization for t to avoid producing unreliable pseudo-targets under weak multi-view evidence. Through this, we select the target and support set to pass through the harmonizer.

3.5 3DGS Training objective

To optimize the 3DGS parameters θ, we adopt a simple three-stage schedule: (1) warm up the distractor masks with prior methods, (2) temporarily use harmonized results as pseudo-ground truth when reliable supports exist, and (3) return to masked-only training for final refinement.

Stage 1 (mask warm-up). We first train 3DGS with the standard masked reconstruction loss in Eq. (1) until the geometry becomes sufficiently stable.

Stage 2 (harmonized supervision). When harmonization is enabled for view i (i.e., harmonize$(i) = 1$; Sec. 3.4), we generate a pseudo-ground truth $\hat{I}_i$ and apply dense supervision, otherwise we keep masked supervision:

$$\mathcal{L}_i^{\text{stage2}} = \begin{cases} \sum_{p \in \Omega} \mathcal{D}\Big(f_{p,i}(\theta), \hat{I}_i(p)\Big), & \text{if harmonize}(i) = 1, \\ \mathcal{L}_{\text{mask}}^i, & \text{otherwise,} \end{cases} \tag{12}$$

where $\mathcal{L}_{\text{mask}}^i$ denotes the masked reconstruction loss for view i in Eq. (1). We optimize 3DGS using $\sum_i \mathcal{L}_i^{\text{stage2}}$ during this stage.

Stage 3 (masked refinement). Finally, we switch back to the masked reconstruction loss in Eq. (1) to refine the final appearance.

Table 1: Quantitative results on the *On-the-go* dataset across occlusion levels. Colors indicate rank: 1st (red), 2nd (peach), 3rd (yellow).

Method	Low Occlusion						Medium Occlusion						High Occlusion					
	Mountain			Fountain			Corner			Patio			Spot			Patio-high		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SLS-mlp [48]	21.16	0.6571	0.2372	21.07	0.6702	0.2096	23.84	0.8130	0.1349	20.74	0.7365	0.1367	22.63	0.7241	0.2357	21.27	0.6920	0.2441
WildGaussians [26]	20.97	0.6746	0.2338	20.91	0.6694	0.2142	23.49	0.8105	0.1540	21.27	0.8061	0.1347	23.93	0.7819	0.1642	22.01	0.7343	0.2049
OCSplats [32]	18.48	0.4237	0.2685	18.01	0.4018	0.2901	20.80	0.6764	0.1895	19.57	0.6786	0.1431	21.77	0.4780	0.2360	19.98	0.5282	0.2202
ForestSplats [43]	20.43	0.6605	0.2185	20.69	0.6660	0.2119	22.34	0.7830	0.1885	18.03	0.7276	0.2011	22.58	0.7326	0.2568	19.89	0.6739	0.2948
HybridGS [31]	21.26	0.6217	0.3480	20.68	0.6364	0.2515	24.84	0.8219	0.1193	21.90	0.7876	0.1490	23.81	0.7186	0.1906	21.59	0.6962	0.2155
DeSplat [61]	19.92	0.6696	0.1780	20.66	0.6830	0.1673	24.61	0.8178	0.1312	21.63	0.8122	0.1093	24.53	0.7830	0.1521	21.34	0.7377	0.1880
DeGauss [59]	21.71	0.7204	0.1790	20.32	0.6809	0.1942	23.17	0.8026	0.1440	21.91	0.8206	0.1058	24.02	0.7900	0.1391	21.67	0.7403	0.1877
RobustSplat [16]	21.42	0.6934	0.1761	20.97	0.6967	0.1603	25.76	0.8513	0.1077	21.94	0.8182	0.1054	24.95	0.8035	0.1179	22.08	0.7538	0.1622
Ours	21.65	0.7080	0.1688	21.42	0.7066	0.1590	25.98	0.8569	0.0992	21.97	0.8215	0.1052	25.15	0.8070	0.1149	22.48	0.7559	0.1593

4 Experiments

4.1 Experimental settings

Datasets. We evaluate on the *On-the-go* dataset [46]. Following prior work, we evaluate on 6 standard benchmark scenes. We additionally evaluate on the RobustNeRF dataset [49], and report detailed settings and results in the *Suppl.*
Metrics. We report PSNR, SSIM [65], and LPIPS [80]. PSNR and SSIM quantify photometric fidelity (higher is better), while LPIPS measures perceptual similarity using an AlexNet backbone (lower is better).
Baselines. We compare against state-of-the-art baselines with publicly available official implementations. We evaluate SLS-mlp [48], WildGaussians [26], HybridGS [31], DeSplat [61], DeGauss [59], RobustSplat [16], OCSplats [32], and ForestSplats [43] under a unified evaluation protocol.
Implementation details. We build on the official 3DGS implementation [23]. All scenes are trained for 30k iterations. We optimize a per-view exposure parameter that scales the rendered image, with a learning rate linearly decayed from 5×10^{-3} to 1×10^{-3}. Multi-view harmonization is applied between 10k and 20k iterations. Throughout 3DGS optimization, the distractor mask predictor is jointly trained following RobustSplat [16]. We set σ_{dist}=0.2 proportional to the scene extent and k_{ang}=2. For support view selection, we set τ=0.3 and τ_{rel}=0.4, and we apply harmonization to a target view only when at least $K_{\min}$=8 supports are available. Further details are provided in the *Suppl.*

4.2 Comparisons with the state-of-the-arts

Quantitative evaluation. We compare our method with recent distractor-free 3DGS baselines across different occlusion levels using PSNR, SSIM, and LPIPS (Tab. 1). Our method achieves the best overall mean performance across metrics, with the highest PSNR, the highest SSIM, and the lowest LPIPS, indicating high-fidelity reconstruction and strong perceptual consistency. In low-occlusion scenes, our method is competitive with the best baseline, while in medium- and high-occlusion scenes it yields clearer improvements. Overall, these results show that replacing mask-based suppression with our multi-view harmonization yields consistent gains in distractor-free 3DGS across all occlusion levels.

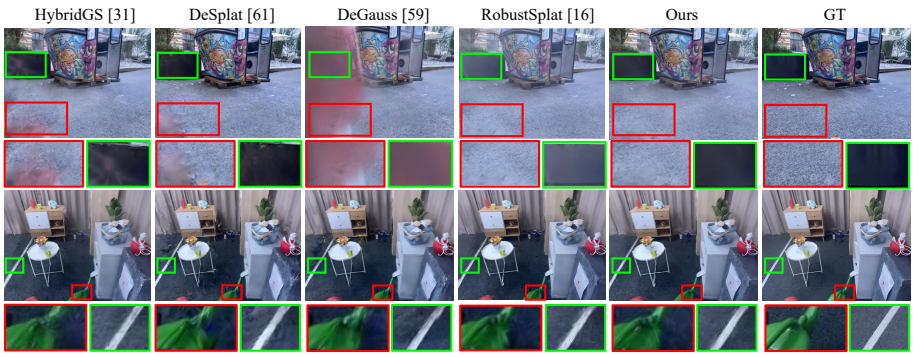

Fig. 4: Qualitative results on *Patio-high* and *Corner* from *On-the-go* dataset.

Qualitative evaluation. We qualitatively evaluate our method on challenging scenes (*Patio-high*, *Corner*). As shown in Fig. 4, prior distractor-free 3DGS baselines exhibit view-dependent artifacts and under-optimized regions around masked areas, whereas ours produces cleaner and more coherent results. On *Patio-high*, baselines produce floaters and ghosting artifacts near the distractor, and the entire rendering appears hazy due to sparsely observed regions. In contrast, ours removes these artifacts without oversmoothing, yielding sharper and more stable novel views. On *Corner*, baselines distort straight structures and leave spiky Gaussian artifacts on the watering can, whereas our method preserves sharp lines and produces smooth, artifact-free reconstructions. These observations indicate that our harmonization effectively suppresses distractor-induced artifacts and injects helpful cross-view information.

4.3 Ablation studies

Effect of geometry-aware support view selection. To demonstrate the selection strategy, we ablate two factors. First, we compare our selection with a simple KNN strategy that picks nearby k-nearest views based on camera position and viewing direction ("Selection" in Tab. 2). Second, we analyze our threshold τ by either harmonizing only well-supported views or harmonizing all views ("Threshold" in Tab. 2). Table 2 highlights four configurations: (a) mask-only baseline without harmonization, (b) harmonizing all views with KNN supports and no thresholding, (c) harmonizing all views with geometry-aware selection but no thresholding, and (g) our full model with geometry-aware selection and thresholding. Settings (b) and (c) achieve PSNR/SSIM similar to or worse than (a), indicating that harmonizing views with insufficient or unreliable support injects inconsistent inpainting artifacts and harms multi-view consistency. As shown in Fig. 5, relying only on the nearest camera often fails to adequately observe the distractor region. In contrast, our model (g) yields clear gains, confirming that selectively harmonizing is crucial for improving performance.

Table 2: Core ablations on PriSplat. We evaluate the effect of enabling harmonization on the target views, the support-view selection strategy (KNN vs Geometry-aware), joint mask updates, and mask-aware fast-weight updates on the mean performance on the *On-the-go* dataset. Columns "Harmonization", "Threshold", "Selection", "Mask", and "Mask-aware" indicate which components are enabled in each variant (rows (a)–(g)). Colors indicate rank: 1st (red), 2nd (peach), 3rd (yellow).

	Harmonization	Threshold	Selection	Mask	Mask-aware	PSNR↑	SSIM↑	LPIPS↓
(a)	✗	✗	✗	✓	✗	22.86	0.7710	0.1387
(b)	✓	✗	KNN	✓	✓	22.72	0.7673	0.1415
(c)	✓	✗	Geometry-aware	✓	✓	22.85	0.7711	0.1375
(d)	✓	✓	KNN	✓	✓	22.89	0.7706	0.1386
(e)	✓	✓	Geometry-aware	✗	✓	22.89	0.7712	0.1404
(f)	✓	✓	Geometry-aware	✓	✗	22.93	0.7736	0.1373
(g)	✓	✓	Geometry-aware	✓	✓	23.11	0.7760	0.1344

Fig. 5: KNN vs. Geometry-aware selection. Given a target view (left), the top row shows KNN-selected support views, while the bottom row shows support views selected by our geometry-aware strategy. Our selection better covers the distractor region and provides more informative supervision for harmonization (middle). As a result, the harmonized view is clearer with our selection (right).

Effect of continuous mask training. We perform harmonization using distractor masks obtained at an intermediate stage of 3DGS training, and continue updating the mask predictor in later iterations. This allows the masks to correct residual inconsistencies and under-masked distractor regions. To evaluate the benefit of this strategy, we compare settings (e) and (g) in Tab. 2. The only difference between them is whether the mask predictor is kept trainable, yet we observe a noticeable performance gap. This indicates that accounting for initially under-masked or inconsistent regions by continuously training the mask predictor is beneficial for overall reconstruction quality.

Effect of mask-aware fast-weight updates. We validate the benefit of our mask-aware fast-weight update by comparing it to the original update [81] that does not use the alignment guidance score b. This comparison corresponds to settings (f) and (g) in Tab. 2. At test time, the NVS prior adapts its fast weights from the selected support views. If some supports are weakly related to the target view or still contain residual distractors, the updates can inject harmful biases into the harmonization. By downweighting unreliable regions during fast-weight adaptation, our mask-aware update prevents such adverse updates.

Table 3: Hyperparameter analysis. PSNR/SSIM ablations on harmonization schedule and number of support views. Colors indicate rank: 1st, 2nd, 3rd.

(a) Harmonization schedule

5-20k		5-30k		10-20k (**ours**)		10-30k	
PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
23.00	0.7727	22.86	0.7718	23.11	0.7760	22.92	0.7729

(b) Number of support views ($K_{\min}$)

2		4		8 (**ours**)		16	
PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
22.91	0.7719	22.92	0.7736	23.11	0.7760	22.95	0.7743

Table 4: Computational cost comparison.

Method	Peak(GB)	Avg.(GB)	Time(h)
SLS-mlp [48]	5.4	4.8	0.38
WildGaussians [26]	31.0	25.8	0.81
DeSplat [61]	2.9	1.9	0.20
DeGauss [59]	20.3	17.1	7.41
RobustSplat [16]	4.4	3.4	0.36
Ours	7.3	3.6	0.38

Table 5: Prior-based approach. Colors indicate rank: 1st, 2nd, 3rd.

	Prior	Size(B)	PSNR↑	SSIM↑	LPIPS↓
Inpaint360GS [60]	[51]	0.05	20.93	0.7052	0.2392
RogSplat [25]	[27]	8.00	21.62	0.7219	0.2298
3DGIC [18]	[51]	0.05	22.26	0.7530	0.1611
LeftRefill [4]	[47]	5.22	22.78	0.7698	0.1458
Ours	[81]	0.31	23.11	0.7760	0.1344

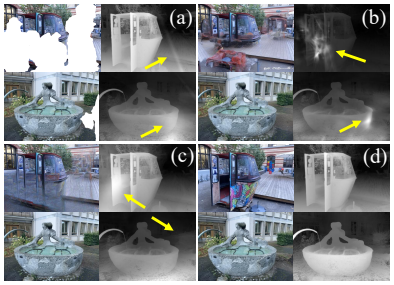

Fig. 6: Comparison of prior-based approach. (a) mask-only [16], (b) LDM [47], (c) LaMa [51], and (d) ours.

Overall, multi-view harmonization already improves prior 3DGS baselines, and our mask-aware update further enhances performance.

4.4 Discussion

Hyperparameter analysis. We analyze two key hyperparameters of our harmonization pipeline. First, we vary the iteration range over which harmonized views supervise 3DGS training, as shown in Tab. 3(a). We observe that restricting harmonized supervision to the middle of training, rather than using it until the end, yields the best performance. Consistent with recent oracle-guided approaches [55], additional cues are most effective when injected at an intermediate stage, while later iterations are better reserved for refining the model with cleaner, self-consistent signals. Second, we evaluate the effect of the number of support views $K_{\min}$ in Tab. 3(b). Increasing the number of supports raises the chance of strong supervision but also the risk of incorporating distracting or inconsistent evidence. In practice, $K_{\min}$=8 strikes a favorable balance and is used as our default. Additional hyperparameter analyses are provided in the *Suppl.*

Computational cost. We measure per-scene training time on *Patio-high* from the *On-the-go* dataset, the most heavily occluded and challenging scene in our benchmark. All methods use the same GPU and training schedule. As summarized in Tab. 4, our method introduces only a modest training-time overhead compared to existing distractor-free 3DGS baselines, despite the additional multi-view harmonization stage, and remains within a practical training budget while improving robustness. In addition, our multi-view harmonizer runs at

Table 6: Component-wise ablation. Each score is evaluated alone and in combination: information density (Inform.), spatial (Spa.), and angular (Ang.) constraints. Colors indicate rank: 1st, 2nd, 3rd.

Exp.	Inform.	Spa.	Ang.	PSNR↑	SSIM↑	LPIPS↓
A	✓			22.87	0.7728	0.1390
B		✓		22.90	0.7732	0.1393
C			✓	22.79	0.7709	0.1407
D	✓	✓		22.88	0.7738	0.1383
E		✓	✓	22.97	0.7737	0.1385
F	✓		✓	22.94	0.7739	0.1373
G	✓	✓	✓	23.11	0.7760	0.1344

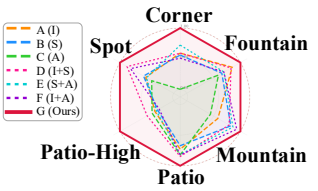

Fig. 7: Component-wise ablation across scenes. Radar plot comparing each configuration (A-G) over diverse *On-the-go* scenes.

about 4.36 s per target view at 512×512 resolution, which makes harmonization overhead negligible at inference time compared to 3DGS rendering. Overall training time is comparable to existing methods, with details in *Suppl.*

Prior-based models. We conducted an additional study to demonstrate whether distractor-free 3DGS performance simply stems from using a large model prior (Tab. 5). For a fair comparison, we finetuned all baselines on the same data to accept masked inputs, following our training setup (see *Suppl.* for details). We found that Inpaint360GS [60] and 3DGIC [18] (LaMa-based priors [51]) often remove objects or copy visible textures into missing regions, whereas LeftRefill [4] (LDM [47]) frequently introduces hallucinated content. RogSplat [25], which uses FLUX [27], also exhibits the multi-view inconsistencies typical of diffusion-based inpainting, which can in turn degrade training. In Fig. 6, we visualize artifacts induced by training on inconsistent inpainted views (left) using depth maps: (a) mask-only [16], (b) LDM [47], (c) LaMa [51], and (d) ours. Overall, these results suggest that improvements are not guaranteed by the prior alone. Instead, distractor-free 3DGS requires careful, geometry-consistent use of priors.

Component-wise ablation. In Tab. 6, we detail the individual and combined versions of each reliability score (Exp. A-G), and Fig. 7 visualizes the performance of each configuration relative to our full model (Exp. G) across diverse scenes from the *On-the-go* dataset. In high-occlusion scenes (Spot), the spatial-distance term (Exp. B) fails because neighboring cameras often share the same distractors, and the information-density term (Exp. D, G) is needed to prioritize unoccluded views. In outdoor scenes (Mountain), relying solely on the density or angular term (Exp. A, C) can retrieve excessively distant views that suffer from low resolution and severe projection distortion, so a spatial term is essential to maintain local fidelity. In narrow indoor scenes (Corner), highly constrained trajectories yield densely sampled views with nearly identical orientations, making the angular term (Exp. C) alone redundant. These varied conditions show that no single score suffices. By combining all three (Exp. G), our method resolves these diverse geometric challenges, yielding robust, high-quality reconstruction across diverse environments.

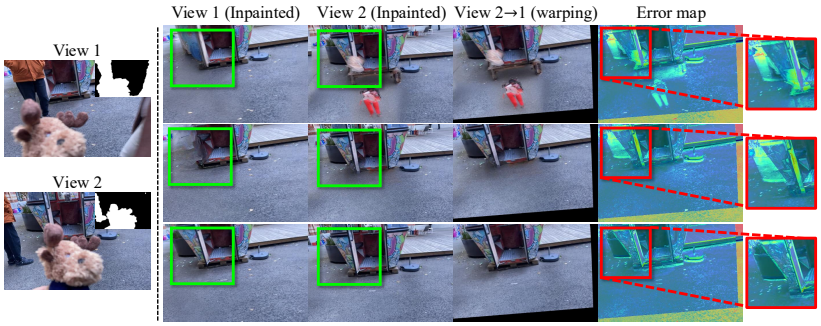

Fig. 8: Cross-view consistency analysis. We compare the geometric consistency of different inpainting strategies by warping restored content across viewpoints. Compared to diffusion-based and 2D-centric priors (top [4], middle [51] row) which exhibit inconsistent textures and high warping errors, our method (bottom row) produces 3D-aware completions that align accurately across views. This demonstrates the effectiveness of our multi-view harmonizer in preserving scene integrity.

Geometric consistency. To verify that our harmonizer performs 3D-aware inpainting, we conduct qualitative experiments as illustrated in Fig. 8. Specifically, we warp the inpainting result from View 2 to the viewpoint of View 1 via depth-based reprojection. By generating an error map comparing View 1 and the warped View 2→1 result, we evaluate the cross-view consistency of the inpainting. Our approach demonstrates superior consistency compared to existing 2D prior-based methods, confirming it preserves underlying scene geometry.

5 Conclusion

We introduce PriSplat, replacing simple mask-based suppression in distractor-free 3DGS with 3D-aware multi-view harmonization. We repurpose a large-scale NVS prior into a robust inpainting engine. To handle corrupted reference views, we introduce a mask-aware fast-weight update and geometry-aware view selection. This aggregates reliable cross-view evidence to generate dense pseudo-ground truth. Ultimately, PriSplat achieves state-of-the-art performance in the distractor-free 3DGS task across challenging benchmarks.

Limitations. Our current formulation does not explicitly model illumination changes or complex appearance factors. In future work, we plan to extend the harmonization framework to handle lighting and broader appearance variations.

Acknowledgements

This work was supported by the InnoCORE program of the Ministry of Science and ICT(N10250156). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963).

References

1. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology (2025)
2. 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. pp. 5855–5864 (2021)
3. 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. pp. 5470–5479 (2022)
4. Cao, C., Cai, Y., Dong, Q., Wang, Y., Fu, Y.: Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7705–7715 (2024)
5. Cao, C., Yu, C., Wang, F., Xue, X., Fu, Y.: Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. Advances in Neural Information Processing Systems **37**, 99299–99332 (2024)
6. Chen, H., Loy, C.C., Pan, X.: Mvip-nerf: Multi-view 3d inpainting on nerf scenes via diffusion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5344–5353 (2024)
7. Chen, J., Qin, Y., Liu, L., Lu, J., Li, G.: Nerf-hugs: Improved neural radiance fields in non-static scenes using heuristics-guided segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19436–19446 (2024)
8. Chen, X., Feng, Y., Chen, M., Wang, Y., Zhang, S., Liu, Y., Shen, Y., Zhao, H.: Zero-shot image editing with reference imitation. Advances in Neural Information Processing Systems **37**, 84010–84032 (2024)
9. Chen, X., Zhang, Q., Li, X., Chen, Y., Feng, Y., Wang, X., Wang, J.: Hallucinated neural radiance fields in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12943–12952 (2022)
10. Dahmani, H., Bennehar, M., Piasco, N., Roldao, L., Tsishkou, D.: Swag: Splatting in the wild images with appearance-conditioned gaussians. In: European Conference on Computer Vision. pp. 325–340. Springer (2024)
11. 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)
12. Darmon, F., Porzi, L., Rota-Bulò, S., Kontschieder, P.: Robust gaussian splatting. arXiv preprint arXiv:2404.04211 (2024)
13. Fei, B., Xu, J., Zhang, R., Zhou, Q., Yang, W., He, Y.: 3d gaussian splatting as new era: A survey. IEEE Transactions on Visualization and Computer Graphics (2024)
14. Fischer, T., Kulhanek, J., Rota Bulò, S., Porzi, L., Pollefeys, M., Kontschieder, P.: Dynamic 3d gaussian fields for urban areas. Advances in Neural Information Processing Systems **37**, 80466–80494 (2024)
15. Fu, C., Chen, G., Zhang, Y., Yao, K., Xiong, Y., Huang, C., Cui, S., Matsushita, Y., Cao, X.: Robustsplat++: Decoupling densification, dynamics, and illumination for in-the-wild 3dgs. arXiv preprint arXiv:2512.04815 (2025)

16. Fu, C., Zhang, Y., Yao, K., Chen, G., Xiong, Y., Huang, C., Cui, S., Cao, X.: Robustsplat: Decoupling densification and dynamics for transient-free 3dgs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27126–27136 (2025)
17. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (ToG) **37**(6), 1–15 (2018)
18. Huang, S.Y., Chou, Z.T., Wang, Y.C.F.: 3d gaussian inpainting with depth-guided cross-view consistency. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26704–26713 (2025)
19. Hui, Z., Gostar, A.K., Chuah, W., Bab-Hadiashar, A., Tennakoon, R.: Joint modeling of corruption-driven and information-limited uncertainty for robust 3d gaussian splatting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 688–697 (2026)
20. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014)
21. Jiang, W., Lei, B., Daniilidis, K.: Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. arXiv preprint arXiv:2311.17874 (2023)
22. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image matching across wide baselines: From paper to practice. International Journal of Computer Vision **129**(2), 517–547 (2021)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
25. Kong, H., Yang, X., Wang, X.: Rogsplat: Robust gaussian splatting via generative priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25735–25745 (2025)
26. Kulhanek, J., Peng, S., Kukelova, Z., Pollefeys, M., Sattler, T.: Wildgaussians: 3d gaussian splatting in the wild. Advances in Neural Information Processing Systems **37**, 21271–21288 (2024)
27. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
28. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), `https://arxiv.org/abs/2506.15742`
29. Lee, J., Kim, I., Heo, H., Kim, H.J.: Semantic-aware occlusion filtering neural radiance fields in the wild. arXiv preprint arXiv:2303.03966 (2023)
30. Lin, C.H., Kim, C., Huang, J.B., Li, Q., Ma, C.Y., Kopf, J., Yang, M.H., Tseng, H.Y.: Taming latent diffusion model for neural radiance field inpainting. In: European Conference on Computer Vision. pp. 149–165 (2024)
31. Lin, J., Gu, J., Fan, L., Wu, B., Lou, Y., Chen, R., Liu, L., Ye, J.: Hybridgs: Decoupling transients and statics with 2d and 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 788–797 (2025)
32. Ling, H., Xu, X., Sun, Y., Sun, Q.: Ocsplats: Observation completeness quantification and label noise separation in 3dgs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25680–25689 (2025)

33. Liu, H.K., Shen, I., Chen, B.Y., et al.: Nerf-in: Free-form nerf inpainting with rgb-d priors. arXiv preprint arXiv:2206.04901 (2022)
34. Liu, K.H., Yang, C.K., Chen, M.H., Liu, Y.L., Lin, Y.Y.: Corrfill: Enhancing faithfulness in reference-based inpainting with correspondence guidance in diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1618–1627. IEEE (2025)
35. Liu, Z., Ouyang, H., Wang, Q., Cheng, K.L., Xiao, J., Zhu, K., Xue, N., Liu, Y., Shen, Y., Cao, Y.: Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. arXiv preprint arXiv:2404.11613 (2024)
36. Markin, A., Pryadilshchikov, V., Komarichev, A., Rakhimov, R., Wonka, P., Burnaev, E.: T-3dgs: Removing transient objects for 3d scene reconstruction. arXiv preprint arXiv:2412.00155 (2024)
37. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7210–7219 (2021)
38. 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)
39. Mirzaei, A., Aumentado-Armstrong, T., Brubaker, M.A., Kelly, J., Levinshtein, A., Derpanis, K.G., Gilitschenski, I.: Reference-guided controllable inpainting of neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 17815–17825 (2023)
40. Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K.G., Kelly, J., Brubaker, M.A., Gilitschenski, I., Levinshtein, A.: Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20669–20679 (2023)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
42. Otonari, T., Ikehata, S., Aizawa, K.: Entity-nerf: Detecting and removing moving entities in urban scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20892–20901 (2024)
43. Park, W., Nam, M., Kim, S., Jo, S., Lee, S.: Forestsplats: Deformable transient field for gaussian splatting in the wild. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6978–6987 (2026)
44. Qi, L., Kuen, J., Shen, T., Gu, J., Li, W., Guo, W., Jia, J., Lin, Z., Yang, M.H.: High quality entity segmentation. In: ICCV (2023)
45. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
46. Ren, W., Zhu, Z., Sun, B., Chen, J., Pollefeys, M., Peng, S.: Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8931–8940 (2024)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
48. Sabour, S., Goli, L., Kopanas, G., Matthews, M., Lagun, D., Guibas, L., Jacobson, A., Fleet, D., Tagliasacchi, A.: Spotlesssplats: Ignoring distractors in 3d gaussian splatting. ACM Transactions on Graphics **44**(2), 1–11 (2025)

49. Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D.J., Tagliasacchi, A.: Robustnerf: Ignoring distractors with robust losses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20626–20636 (2023)
50. Shi, Z., Huo, D., Zhou, Y., Min, Y., Lu, J., Zuo, X.: Imfine: 3d inpainting via geometry-guided multi-view refinement. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26694–26703 (2025)
51. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2149–2159 (2022)
52. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023)
53. Tang, L., Ruiz, N., Chu, Q., Li, Y., Holynski, A., Jacobs, D.E., Hariharan, B., Pritch, Y., Wadhwa, N., Aberman, K., et al.: Realfill: Reference-driven generation for authentic image completion. ACM Transactions on Graphics (TOG) **43**(4), 1–12 (2024)
54. Tang, Y., Xu, D., Hou, Y., Wang, Z., Jiang, M.: Nexussplats: Efficient 3d gaussian splatting in the wild. arXiv preprint arXiv:2411.14514 (2024)
55. Topaloglu, A., Li, K., Niemeyer, M., Navab, N., Tekalp, A.M., Tombari, F.: Oraclegs: Grounding generative priors for sparse-view gaussian splatting. arXiv (2025)
56. Ungermann, P., Ettenhofer, A., Nießner, M., Roessle, B.: Robust 3d gaussian splatting for novel view synthesis in presence of distractors. In: DAGM German Conference on Pattern Recognition. pp. 153–167. Springer (2024)
57. Wang, D., Zhang, T., Abboud, A., Süsstrunk, S.: Innerf360: Text-guided 3d-consistent object inpainting on 360-degree neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12677–12686 (2024)
58. Wang, H., Anantrasirichai, N., Zhang, F., Bull, D.: Uw-gs: Distractor-aware 3d gaussian splatting for enhanced underwater scene reconstruction. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3280–3289. IEEE (2025)
59. Wang, R., Lohmeyer, Q., Meboldt, M., Tang, S.: Degauss: Dynamic-static decomposition with gaussian splatting for distractor-free 3d reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6294–6303 (2025)
60. Wang, S., Zhang, S., Millerdurai, C., Westermann, R., Stricker, D., Pagani, A.: Inpaint360gs: Efficient object-aware 3d inpainting via gaussian splatting for 360deg scenes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 117–127 (2026)
61. Wang, Y., Klasson, M., Turkulainen, M., Wang, S., Kannala, J., Solin, A.: Desplat: Decomposed gaussian splatting for distractor-free rendering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 722–732 (2025)
62. Wang, Y., Li, K., Chen, M., Wang, L., Zhou, S., Xue, K., Guo, Y.: Distractor-free novel view synthesis via exploiting memorization effect in optimization. In: European Conference on Computer Vision. pp. 477–493. Springer (2024)
63. Wang, Y., Wu, Q., Zhang, G., Xu, D.: Learning 3d geometry and feature consistent gaussian splatting for object removal. In: European Conference on Computer Vision. pp. 1–17 (2024)
64. Wang, Y., Wang, J., Qi, Y.: We-gs: An in-the-wild efficient 3d gaussian representation for unconstrained photo collections. arXiv preprint arXiv:2406.02407 (2024)

65. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
66. Weder, S., Garcia-Hernando, G., Monszpart, A., Pollefeys, M., Brostow, G.J., Firman, M., Vicente, S.: Removing objects from neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16528–16538 (2023)
67. Wu, T., Zhong, F., Tagliasacchi, A., Cole, F., Oztireli, C.: Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. Advances in neural information processing systems **35**, 32653–32666 (2022)
68. 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)
69. Xu, C., Kerr, J., Kanazawa, A.: Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections. arXiv preprint arXiv:2407.12306 (2024)
70. Xu, J., Mei, Y., Patel, V.: Wild-gs: Real-time novel view synthesis from unconstrained photo collections. Advances in Neural Information Processing Systems **37**, 103334–103355 (2024)
71. Yang, Y., Zhang, S., Huang, Z., Zhang, Y., Tan, M.: Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15901–15911 (2023)
72. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European Conference on Computer Vision. pp. 162–179 (2024)
73. Yin, Y., Fu, Z., Yang, F., Lin, G.: Or-nerf: Object removing from 3d scenes guided by multiview segmentation with neural radiance fields. arXiv preprint arXiv:2305.10503 (2023)
74. Yoon, Y., Cho, W., Ha, H., Kim, S., Yoon, K.J.: Expose: Reinforcing video generation models for extreme pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32636–32646 (2026)
75. Yoon, Y., Jang, H.K., Yoon, K.J.: Gmt: Enhancing generalizable neural rendering via geometry-driven multi-reference texture transfer. In: European Conference on Computer Vision. pp. 274–292. Springer (2024)
76. Yoon, Y., Yoon, K.J.: Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12428–12438 (2023)
77. Yoon, Y., Yoon, K.J.: Resplat: Degradation-agnostic feed-forward gaussian splatting via self-guided residual diffusion. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=461VpgnLsi
78. You, J., Lin, C.H., Lyu, W., Zhang, Z., Yang, M.H.: Instainpaint: Instant 3d-scene inpainting with masked large reconstruction model. arXiv preprint arXiv:2506.10980 (2025)
79. Zhang, D., Wang, C., Wang, W., Li, P., Qin, M., Wang, H.: Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In: European Conference on Computer Vision. pp. 341–359. Springer (2024)

80. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
81. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025)