RECTIFY POSE AND CAPTURE AMBIGUITY: 3D SPATIAL-AWARE GS IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent 3DGS and NeRF variants have significantly advanced for real-world scenarios. However, they still struggle to capture ambiguous distractors that are visually or semantically similar to static regions and remain susceptible to inaccurate camera poses. We found that these two challenges are inherently coupled, making them difficult to address independently and leading to artifact-prone 3D reconstruction. Motivated by these findings, we introduce PAGauss, a novel framework that jointly addresses these challenges. Specifically, unlike previous works that focus on 2D representations, PAGauss leverages 3D spatial relationships across viewpoints. PAGauss consists of two novel components: a spatial-aware adaptive mask and a spatial-aware camera pose alignment. The spatial-aware adaptive mask leverages 3D multi-view consistency to robustly capture ambiguous distractors regardless of their colors or semantic similarity. The spatial-aware camera pose alignment rectifies imprecise camera poses using a source refinement strategy with distractor-free images. To stabilize camera pose alignment, we introduce a multi-stage training strategy, improving 3D Gaussian alignment. We also create and release the WildArena dataset that includes challenging scenarios with corresponding mask annotations. To our knowledge, WildArena is the first dataset to provide such mask annotations, enabling robustness analysis over baselines. Through extensive experiments, we demonstrate that PAGauss achieves superior results and robustness across in-the-wild datasets. Our source code and dataset will be released upon acceptance.

1 INTRODUCTION

Perceiving and reconstructing 3D scenes in real-world scenarios is crucial for computer vision. Recent Neural Radiance Field (NeRF) (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) variants can tackle distractors and appearance variations, showing promising results with camera poses obtained from Structure-from-Motion (SfM).

However, previous methods (Wang et al., 2026c; Li et al., 2026a; Park et al., 2025; Sabour et al., 2024; Zhang et al., 2024) still neglect two formidable challenges in unconstrained scenarios. Firstly, prior methods that leverage photometric errors or semantic similarity fail to distinguish ambiguous distractors from static elements with similar colors or semantics, leading to artifact-prone 3D reconstructions. Secondly, they overlook the inherent deficiencies in camera poses estimated from unconstrained images containing distractors. To gain further insight, we analyze the relationship between distractors and camera poses, as shown in Fig. 1. The results indicate that imprecise masks lead to artifacts and hinder precise pose estimation, while inaccurate camera poses misalign 3D Gaussians and undermine robustness. These findings underscore the necessity of addressing both challenges on equal footing within a unified framework. Although prior methods (Zhang et al., 2026b; Furutani et al., 2026; Li et al., 2026b; Zhong et al., 2026; Zheng et al., 2025) have been proposed, they still struggle to identify ambiguous distractors and precisely estimate camera poses. Overall, extant methods suffer from two critical bottlenecks, yet no prior methods have attempted to address both.

Inspired by these, we introduce PAGauss that leverages 3D spatial relationships to identify ambiguous distractors and refine camera poses. PAGauss seamlessly integrates the spatial-aware adaptive mask and the spatial-aware camera pose refinement within the unified framework. Concretely, the spatial-aware adaptive mask considers 3D multi-view consistency to uncover static regions. By aggregating 3D uncertainty scores across viewpoints, PAGauss captures ambiguous distractors that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

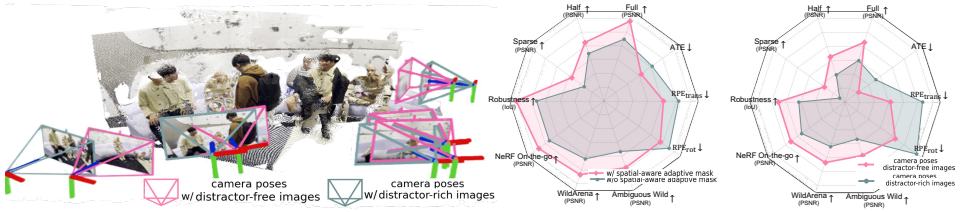


Figure 1: We visualize the discrepancy between camera poses estimated from unconstrained and distractor-free images. We observe that imprecise masks cause artifacts and impede accurate pose estimation, while inaccurate camera poses misalign 3D Gaussians and deteriorate robustness, implying the importance of jointly addressing distractors and camera poses within a unified framework.

are difficult to distinguish from static regions using colors or semantic similarity. Notably, against prior works, the spatial-aware adaptive mask does not require auxiliary elements (*e.g.*, Dual 3DGS, MLPs) and additional pre-training. To refine imprecise camera poses, the spatial-aware camera pose refinement utilizes distractor-free images derived from the spatial-aware adaptive mask through a adaptive source refinement strategy, improving 3D Gaussian alignment. Furthermore, we create and release the WildArena dataset to rigorously evaluate PAGauss, comprising 82 challenging scenarios with diverse viewpoints (*e.g.*, bird's-eye, frontal, and 360°) along with corresponding mask annotations. We further extend existing in-the-wild datasets with previously unavailable mask annotations for distractors. Unlike prior datasets, which lack mask annotations for assessing robustness across baselines, the WildArena dataset is, to the best of our knowledge, the first to provide mask annotations that enable robustness analysis against transient elements.

Through extensive experiments on challenging in-the-wild datasets (Park et al., 2026; Tang et al., 2025; Ren et al., 2024; Snavely et al., 2006), we demonstrate that PAGauss robustly identifies ambiguous distractors (Accuracy - 8.7%↑, Precision - 8.4%↑) and refines camera poses (RPE_{trans} - 15.8%↓, RPE_{rot} - 13.3%↓), outperforming prior methods (PSNR - 11.8%↑, SSIM - 4.1%↑, LPIPS - 9.5%↓). In summary, our key contributions w.r.t. the prior works are as follows: (i) We propose the novel PAGauss framework that considers 3D spatial relationships across viewpoints, achieving high-quality results in unconstrained scenarios. (ii) We design two novel components, including the spatial-aware adaptive mask that identifies ambiguous distractors regardless of colors or semantic similarity and the spatial-aware camera pose alignment that refines camera poses through source refinement with distractor-free images, improving 3D Gaussian alignment. (iii) We establish and release the WildArena dataset, comprising challenging scenarios and mask annotations for distractors, to facilitate robustness analysis across baselines.

2 Related Work

PAGauss is designed to identify ambiguous distractors and rectify imprecise camera poses. We briefly highlight the key distinctions between our method and prior methods, as presented in Tab. 1.

3D geometry models. Recent SfM methods (Schönberger et al., 2016; Pan et al., 2024; Lee & Yoo, 2025) have long been regarded as fundamental approaches for estimating camera poses and 3D geometry. Despite their promise, prior methods struggle with severe camera parallax in real-world scenarios, while increasing time-complexity. To overcome these, transformer-based methods have emerged. These methods primarily fall into three paradigms, depending on the modality: (i) image pair-based methods (Wang et al., 2024a; Weinzaepfel et al., 2022; Leroy et al., 2024), which estimate point clouds and confidence; (ii) multi-view-based methods (Zuo et al., 2026; Wang et al., 2025a; Li et al., 2026c; Zhang et al., 2025), which leverage global and frame attention to tackle permutation issues; (iii) image sequence-based methods (Li et al., 2026b; Maggio et al., 2025; Jin et al., 2026; Xie et al., 2026), which utilize current and prior frames, estimating next-frame geometry.

Despite their promising results, they struggle to estimate poses and 3D geometry in unconstrained scenarios containing distractors, appearance variations, and irrelevant views within the captured image set. Motivated by this, RobustVGGT (Han et al., 2025) employs an outlier rejection strategy, while Wild3R (Furutani et al., 2026) and GenWildSplat (Gupta et al., 2026) address appearance variations and minor distractors for 3D reconstruction. However, they rely on the strong assumption that camera poses obtained from unconstrained images are accurate. We challenge this prevailing assumption by rectifying imprecise camera poses, improving 3D Gaussian alignment.

Methods	Rectify pose	Capture ambiguity	w/o auxiliary	Frozen VFM	In the Wild
AsymGS (Li et al., 2026a)	✗	✗	✗	✓	✓
Wild3R (Furutani et al., 2026)	✗	✗	✓	✗	▲
WildSplat (Zhang et al., 2026b)	✗	✗	✓	✓	▲
MVFusion-GS (Hu et al., 2026)	✗	✗	✗	✓	✓
RefineSplat (Park et al., 2026)	✗	✓	✗	✓	✓
PAGauss (Ours)	✓	✓	✓	✓	✓

Table 1: PAGauss captures ambiguous distractors and rectifies poses without fine-tuned Vision Foundation Model (VFM) and auxiliary elements (*e.g.*, dual 3DGS, MLPs). ▲ denotes methods that use datasets with a low proportion of distractors that have little effect on the rendering results.

Robustness of novel view synthesis. While prior methods (Long et al., 2026; Topaloğlu et al., 2026; Kim et al., 2026; Wang et al., 2026b) show favorable results on constrained scenes, they struggle with unconstrained scenes. To address this, recent methods have been proposed. These methods can be primarily divided into three paradigms: (i) residual-based approaches (Martin-Brualla et al., 2021; Park et al., 2025; Zhang et al., 2024; Yang et al., 2023), which leverage photometric errors with MLP layers to identify distractors; (ii) semantic-based methods (Ren et al., 2024; Kulhanek et al., 2024), which use extracted semantic features from a 2D VFM (*e.g.*, DINOv2 (Oquab et al., 2023), FeatUp (Fu et al., 2024a), DIFT (Tang et al., 2023)); and (iii) bootstrapping-based methods (Wang et al., 2026c; Li et al., 2026a), which utilize both photometric errors and semantic features.

Nevertheless, these methods remain limited in identifying ambiguous distractors that share similar colors or semantic similarity with static regions in Fig. 2. To tackle this, RefineSplat (Park et al., 2026) leverages entropy values to capture ambiguous distractors via auxiliary MLPs, albeit at the cost of additional pre-training time. However, in contrast to prior works that focus on 2D representations, PAGauss harnesses 3D multi-view consistency to robustly detect ambiguous distractors without requiring auxiliary elements (*e.g.*, Dual 3DGS, MLPs) or additional pre-training cost.

Camera pose optimization. Most 3DGS variants (Zhang et al., 2026a; Polansky et al., 2026; Wang et al., 2025b) rely on camera poses obtained from SfM and are sensitive to pose inaccuracies. To mitigate this, prior methods (Sun et al., 2026; Duan et al., 2026; Lin et al., 2025a; Ji & Yao, 2025; Huang et al., 2025) jointly optimize poses and 3D Gaussians. For unordered images, I3DGS (Meuleman et al., 2026) and TrackGS (Shi et al., 2026) leverage multi-stage training and global track constraints. However, since they assume that input images do not contain distractors, simply leveraging their methods is challenging in unconstrained scenarios. To address this, although several methods (Gupta et al., 2026; Fujimura et al., 2026; Wang & Lee, 2026) have been proposed, they either assume sequential inputs or estimate poses from unconstrained images, resulting in imprecise camera poses. Unlike these methods, PAGauss precisely rectifies camera poses for unordered image sequences via source refinement using distractor-free images in unconstrained scenarios.

3 Preliminaries

3D Gaussian splatting. More recently, 3DGS has emerged as a method for representing scenes as collections of differentiable Gaussians $\mathcal{G} = \{g_k\}_{k=1}^{N}$. Each Gaussian g_k is parameterized by its position $\mu_k \in \mathbb{R}^3$, covariance matrix $\Sigma_k \in \mathbb{R}^{3\times3}$, which is decomposed into a rotation matrix $R_k \in \mathrm{SO}(3)$ and a scaling matrix $S_k \in \mathbb{R}^{3\times3}$, opacity $\alpha_k \in [0,1]$, and view-dependent color $c_k \in \mathcal{C}^{N_{\text{sh}}}$ represented via spherical harmonics (SH) coefficients N_{sh}. For novel view synthesis, 3D Gaussians $\mathcal{G}$ are projected onto the 2D image plane via the viewing transformation $W \in \mathbb{R}^{3\times3}$ and covariance matrix $\Sigma_k' = JW\Sigma_k W^T J^T$ where $J \in \mathbb{R}^{2\times3}$ is the Jacobian of the Taylor approximation of the projective transformation. Given a camera viewpoint, each pixel color $\hat{C}$ is rendered as:

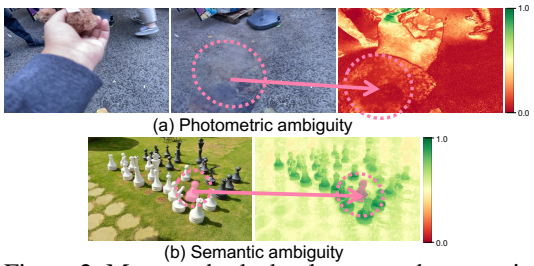


Figure 2: Most methods that leverage photometric errors or semantic similarity struggle to identify ambiguous distractors that have similar colors or semantic levels to static regions.

$$\hat{C} = \sum_{k=1}^{N} c_k \alpha_k \prod_{j=1}^{k-1} (1 - \alpha_j)\,. \tag{1}$$

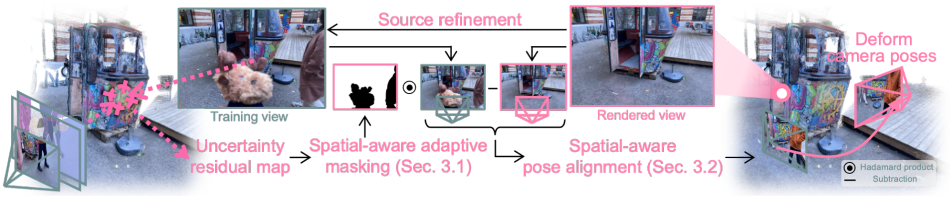


Figure 3: Given unconstrained images, PAGauss considers 3D spatial relationships across viewpoints to construct uncertainty residual maps. Then, we robustly identify ambiguous distractors regardless of photometric errors or semantic levels. Furthermore, we rectify imprecise camera poses with a source refinement strategy using distractor-free images, improving 3D Gaussian alignment.

In unconstrained scenarios, a transient mask $\mathcal{M}$ identifies pixels corresponding to distractors or occlusion regions that should be excluded from the optimization of 3D Gaussians. Therefore, mask annotations are crucial not only for excluding distractors and occlusion regions during optimization but also for evaluating the robustness against transient elements across methods. All 3D Gaussians $\mathcal{G}$ are optimized by minimizing the $\mathcal{L}_{\text{GS}}$ between the rendered image $\hat{I} \in \mathbb{R}^{3\times H\times W}$ and the ground truth image $I^{gt} \in \mathbb{R}^{3\times H\times W}$ using a transient mask $\mathcal{M}$ to focus on static regions, as follows:

$$\mathcal{L}_{\text{GS}} = (1-\lambda)\mathcal{L}_1(\hat{I}, I^{gt}) \odot \mathcal{M} + \lambda\mathcal{L}_{\text{D-SSIM}}(\hat{I}, I^{gt}) \odot \mathcal{M}, \tag{2}$$

where $\mathcal{L}_1$ and $\mathcal{L}_{\text{D-SSIM}}$ are the L_1 and SSIM losses, λ is a weighting factor, and $\odot$ denotes the Hadamard product. All 3D Gaussian anchors and camera poses are respectively initialized from the 3D point cloud and camera poses produced by SfM.

3D foundation model. Given a set of images $\mathcal{I} = \{I_i\}_{i=1}^N$, a 3D foundation model first encodes input images into semantic features to estimate 3D point maps, depth maps, and camera parameters for all views in a single forward pass:

$$\{K_i, E_i, D_i, U_i\}_{i=1}^N = \mathcal{H}_{\text{3D}}(\mathcal{E}_{\text{3D}}(\mathcal{I})) = \mathcal{F}_{\text{3D}}(\mathcal{I}), \tag{3}$$

where $\mathcal{E}_{\text{3D}}$ denotes the image encoder that extracts multi-view semantic features, $\mathcal{H}_{\text{3D}}$ denotes the geometry prediction head that estimates dense 3D geometry and camera parameters, $K_i \in \mathbb{R}^{3\times3}$ and $E_i \in \mathbb{R}^{3\times3}$ denote the camera intrinsic and extrinsic matrices, $D_i \in \mathbb{R}^{H\times W}$ is the pixel-aligned depth map which records the 3D position of each pixel, and U_i represents the per-pixel 3D confidence score that captures the 3D spatial multi-view consistency across unconstrained images.

4 Method

We present the novel PAGauss framework, as depicted in Fig. 3. Firstly, we propose the spatial-aware adaptive mask that leverages 3D multi-view consistency to identify ambiguous distractors (Sec. 4.1). Secondly, we introduce the spatial-aware camera pose alignment that refines camera poses using distractor-free images, improving 3D Gaussian alignment (Sec. 4.2). Finally, we introduce the simple but effective multi-stage training, avoiding a chicken-and-egg problem (Sec. 4.3).

4.1 Spatial-aware adaptive masking

Spatial-aware local adaptive masking. Previous methods (Li et al., 2026a; Wang et al., 2026c; Hu et al., 2026; Zheng et al., 2025; Ren et al., 2024) rely on photometric errors or semantic features extracted using a 2D VFM to identify distractors. However, they still struggle to capture ambiguous distractors that have similar colors or semantic similarity with static elements, as shown in Fig. 2. This limitation stems from the fact that photometric errors and semantic features are limited to 2D representations, neglecting 3D spatial understanding across viewpoints, which makes it difficult to identify ambiguous distractors in real-world



Figure 4: We visualize cross-view attention maps, showing that 3D foundation models inherently perceive 3D multi-view consistent regions.

scenarios. To overcome this, we consider 3D spatial relationships across viewpoints to robustly discern ambiguous distractors through 3D multi-view consistency using DA3 (Lin et al., 2025b), which inherently exhibits outlier filtering behavior, as presented in Fig. 4. Specifically, we first extract per-pixel 3D confidence scores across viewpoints as:

$$(\hat{U}_i)_{i=1}^{N} = \mathcal{F}_{\text{3D}}(\hat{\mathcal{I}}),\ \hat{\mathcal{I}} = \{I_i \mid i = 1, \ldots, N,\ i \neq k\} \cup \{\hat{I}_k\}. \tag{4}$$

By leveraging them, we obtain the uncertainty residual map $\beta = \|U - \hat{U}\|$, which reflects 3D multi-view consistency across viewpoints. Although the uncertainty residual map β can be used directly as in prior works (Li et al., 2026a; Zheng et al., 2026; Park et al., 2025; Kulhanek et al., 2024), it remains difficult to construct precise transient masks and identify subtle distractors.

To address these, unlike prior works, we exploit 3D multi-view consistency to derive an instance-level uncertainty map by aggregating uncertainty residuals within each instance. Following recent works (Bao et al., 2026; Pan et al., 2026b), we obtain instance masks from a segmentation model S (Kirillov et al., 2023), as $m_i^j = S(I_i^{\text{gt}})$, where j denotes the index of each object instance. Given the instance masks, we compute the average uncertainty scores for each instance as:

$$\mathcal{B}_i^j(t) = \frac{\sum_{p \in m_i^j} \tilde{\beta}_i(p,t)}{|m_i^j|}, \quad \tilde{\beta}_i = \frac{\beta_i}{\sum_j \beta_j}, \tag{5}$$

where t is the current iteration. At each iteration, utilizing the average uncertainty scores for each instance, a spatial-aware threshold $\mathcal{T}_{\text{geo}}$ is adaptively defined with a learnable weighting value $\ell^{(w)}$ to capture ambiguous distractors, as:

$$\mathcal{T}_{\text{geo}} = \mathbb{E}\left[\mathcal{B}_i\right] + \mathbb{V}\left[\mathcal{B}_i\right](1 + \ell^{(w)}), \tag{6}$$

where $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote the mean and variance of the instance-level uncertainty values. Finally, spatial-aware mask $\mathcal{M}_{\text{geo}}$ is obtained by selecting instances that exceed the spatial-aware adaptive threshold, as follows:

$$\mathcal{M}_{\text{geo}} = \{m_i^j \mid \mathcal{B}_i^j \geq \mathcal{T}_{\text{geo}}, j \in 1, \ldots, N\}, \tag{7}$$

where N is the number of instances. By leveraging them, compared to others (Park et al., 2026; Wang et al., 2026c; Li et al., 2026a), PAGauss precisely discerns ambiguous distractors from unconstrained images without auxiliary elements (*e.g.*, Dual 3DGS, MLPs) and additional pre-training.

Context-aware mask complement. However, it remains challenging to detect subtle distractors or handle cases where the instance masks are inaccurate. To address these issues, although DroneSplat (Tang et al., 2025) utilizes the SAM2 tracking head, it remains highly sensitive to the quality of the instance masks. To tackle this, we consider masked regions as contextual priors to identify unexplored diverse distractors. Specifically, we first consider regions in $1 - \mathcal{M}_{\text{geo}}$ with uncertainty exceeding $\mathbb{E}\left[\mathcal{B}_i\right]$. Then, given extracted semantic features from the DA3 within the masking regions, we construct the context-aware mask $\mathcal{M}_{\text{context}}$ by masking candidate regions which have high cosine similarity with distractors as:

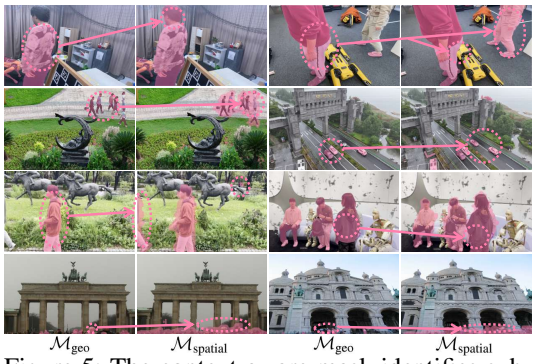


Figure 5: The context-aware mask identifies subtle distractors that are not captured by $\mathcal{M}_{\text{geo}}$.

$$\mathcal{M}_{\text{context}} = \{m_i^j \mid \mathcal{B}_i^j \geq \mathbb{E}[\mathcal{B}_i], \text{sim}((1 - \mathcal{M}_{\text{geo}}) \odot \mathcal{E}_{\text{3D}}(I_i), m_i^k \odot \mathcal{E}_{\text{3D}}(I_i)) \geq 0.9\}, \tag{8}$$

where m_i^k is defined as an element of $\mathcal{M}_{\text{geo}}$. $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. Finally, by combining both cues as $\mathcal{M}_{\text{spatial}} = 1 - (\mathcal{M}_{\text{geo}} \cup \mathcal{M}_{\text{context}})$, we robustly identify distractors, even when they are subtle or their instance masks are imperfect in Fig. 5, enabling artifact-free 3D reconstruction.

4.2 Spatial-aware camera pose alignment

Camera pose refinement. Prior methods rely on camera poses obtained from SfM. However, applying SfM to unconstrained images produces misaligned camera poses due to incorrect matching

Methods	Train hrs / FPS	Activist			Gazebo			Lagoon			Miniature			Pagoda			Terrace		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o camera pose optimization																			
WildGaussians	0.65 / 106	19.43	0.680	0.214	22.58	0.771	0.198	20.23	0.728	0.228	16.35	0.662	0.256	22.26	0.751	0.260	22.65	0.686	0.193
DeGauss	1.91 / 101	17.17	0.595	0.216	21.04	0.765	0.289	19.52	0.622	0.224	18.43	0.798	0.208	20.79	0.796	0.238	18.96	0.508	0.304
RobustSplat	0.77 / 103	20.91	0.779	0.161	24.52	0.754	0.293	23.72	0.831	0.147	18.32	0.786	0.213	22.54	0.788	0.249	16.52	0.388	0.592
DualSplat	3.23 / 102	20.68	0.732	0.158	23.94	0.800	0.234	23.95	0.795	0.109	18.65	0.750	0.206	22.79	0.809	0.225	22.87	0.735	0.177
AsymGS	2.64 / 106	20.71	0.774	0.162	22.99	0.816	0.172	20.35	0.747	0.241	18.33	0.764	0.241	21.67	0.771	0.268	22.18	0.791	0.170
PAGauss (Ours)	0.87 / 105	21.03	0.786	0.154	24.75	0.832	0.143	24.09	0.824	0.120	22.01	0.828	0.239	23.04	0.821	0.218	23.05	0.790	0.168
w/ camera pose optimization																			
CF3DGS + RobustSplat	1.03 / 107	20.99	0.776	0.162	24.85	0.837	0.154	24.07	0.835	0.144	18.28	0.777	0.219	23.07	0.821	0.204	21.59	0.616	0.378
LongSplat + RobustSplat	1.42 / 101	20.35	0.723	0.169	24.21	0.808	0.157	22.62	0.766	0.164	19.45	0.767	0.204	21.94	0.783	0.218	16.37	0.330	0.602
I3DGS + RobustSplat	1.74 / 103	20.09	0.569	0.476	23.32	0.701	0.399	22.84	0.663	0.425	20.26	0.785	0.294	23.07	0.824	0.300	22.21	0.535	0.532
PAGauss (Ours)	1.32 / 108	21.23	0.803	0.132	25.32	0.872	0.129	24.18	0.845	0.098	22.05	0.829	0.197	23.13	0.829	0.206	23.77	0.801	0.161

Table 2: Quantitative results on the WildArena dataset. Best and second are highlighted.

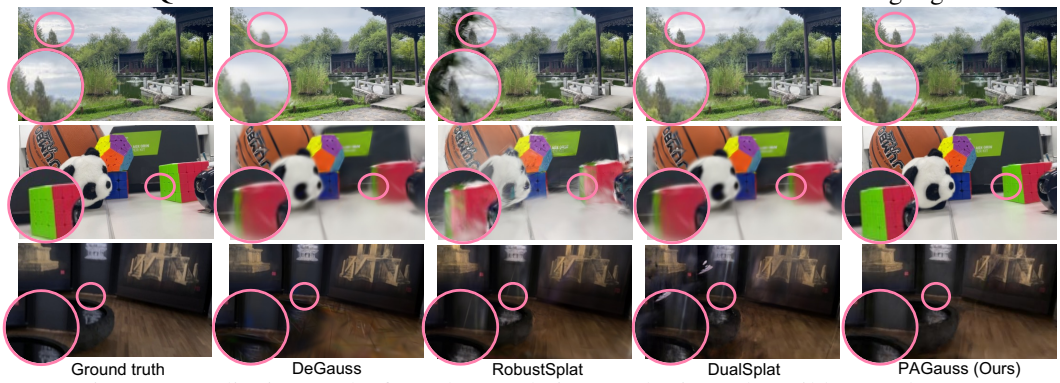


Figure 7: Qualitative results from the novel view synthesis on the WildArena dataset.

points, as shown in Fig. 6. This issue is largely overlooked, remaining a major bottleneck and leading to 3D Gaussian misalignment. To solve this, we refine the imprecise camera extrinsics, improving 3D Gaussian alignment across viewpoints. Given rotation $l_i^{(r)} \in \mathbb{R}^{1\times3}$ and translation $l_i^{(t)} \in \mathbb{R}^{1\times3}$ embeddings per image, we refine the camera extrinsics by applying the Lie algebra to the rotation and adding translation embedding as follows:

$$\hat{\mathbf{R}}_i = \exp(l_i^{(r)})\mathbf{R}_i, \quad \hat{\mathbf{T}}_i = \mathbf{T}_i + l_i^{(t)}, \tag{9}$$

where $\mathbf{R}_i$ and $\mathbf{T}_i$ are the camera rotation and translation of the image I_i. However, merely refining camera extrinsics using photometric errors misaligns camera poses in ambiguous scenarios where distractors have similar colors to static regions, leading to 3D Gaussian misalignment.

To address this issue, we refine imprecise camera poses using multi-view images randomly sampled from the training views as follows:

$$\mathcal{L}_{\text{pose}} = \frac{1}{N}\sum_{i=1}^{N}(||\log(\bar{\mathbf{R}}_i\hat{\mathbf{R}}_i^{\top})||_2^2 + ||\bar{\mathbf{T}}_i - \hat{\mathbf{T}}_i||_2^2), \tag{10}$$

where $\bar{\mathbf{R}}_i$ and $\bar{\mathbf{T}}_i$ denote the camera rotation and translation estimated by DA3. Note that, to avoid invalid rotation optimization, we apply the exponential map to SO(3). In this way, the spatial-aware camera pose refinement benefits from a bootstrapping effect, improving the 3D Gaussian alignment and enhancing depth consistency.



Figure 6: Since SfM estimates camera poses based on feature correspondences, distractors in unconstrained images often lead to inaccurate feature matching, resulting in imprecise pose alignment.

Adaptive source refinement. However, leveraging unconstrained images with DA3 inaccurately aligns camera poses due to distractors. To mitigate this and precisely refine camera poses, we leverage distractor-free rendered images to estimate camera poses. Notably, we avoid additional computational overhead from repeatedly sampling rendered images by iteratively caching them.

4.3 Multi-stage training scheme

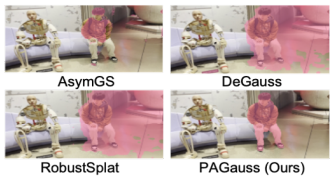
Initial training phase. Since jointly optimizing 3D Gaussians and camera poses at the initial stage leads to a chicken-and-egg problem, balancing their optimization is crucial. To this end, we propose the simple yet effective multi-stage training strategy. Initially, we first focus on optimizing the 3D Gaussians to identify ambiguous distractors and construct the 3D structures using Eq. 2.

Intermediate training phase. After the initial phase, to mitigate the 3D Gaussian misalignment and refine imprecise camera poses, we refine initialized camera poses as follows:

$$\mathcal{L}_{\text{mid}} = \mathcal{L}_{GS} + \lambda_1\mathcal{L}_{pose}, \tag{11}$$

Methods	NeRF On-the-go			Drone Imagery			Photo Tourism			Ambiguous Wild		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o camera pose optimization												
RobustSplat	22.72	0.784	0.174	21.17	0.575	0.468	18.41	0.820	0.179	20.34	0.672	0.360
AsymGS	23.40	0.801	0.135	21.36	0.618	0.369	25.14	0.864	0.169	20.30	0.738	0.236
DualSplat	23.37	0.820	0.150	21.57	0.646	0.260	25.01	0.870	0.158	20.60	0.748	0.223
RefineSplat	23.50	0.823	0.142	21.63	0.647	0.256	25.10	0.861	0.152	21.76	0.778	0.195
PAGauss (Ours)	23.73	0.828	0.143	21.78	0.686	0.223	25.14	0.883	0.172	21.52	0.783	0.201
w/ camera pose optimization												
CF3DGS + RobustSplat	22.63	0.772	0.183	21.13	0.657	0.247	18.70	0.806	0.218	21.62	0.786	0.213
LongSplat + RobustSplat	21.49	0.800	0.221	19.19	0.448	0.291	15.76	0.700	0.338	17.62	0.718	0.257
I3DGS + RobustSplat	20.39	0.767	0.248	20.83	0.595	0.349	17.11	0.694	0.331	18.64	0.767	0.243
PAGauss (Ours)	23.95	0.836	0.132	21.97	0.698	0.200	25.27	0.880	0.159	21.85	0.796	0.190

Table 3: Quantitative results on the in-the-wild datasets. Best and second are highlighted.



(a) Comparison of transient masks.

Method	Accuracy↑	Precision↑	Recall↑	IoU↑	Average↑
RobustSplat	0.906	0.833	0.925	0.847	0.886
AsymGS	0.939	0.828	0.947	0.831	0.878
DualSplat	0.964	0.863	0.950	0.858	0.915
LongSplat + RobustSplat	0.830	0.882	0.872	0.936	0.881
I3DGS + RobustSplat	0.903	0.724	0.703	0.555	0.721
PAGauss (Ours)	0.972	0.918	0.957	0.883	0.937

(b) Quantitative segmentation results on the WildArena dataset.

Figure 8: Robustness of the spatial-aware mask.

Scene	SfM	VGGT	DA3	Ours	PSNR↑	SSIM↑	LPIPS↓
Ambiguous Wild	✓	-	-	-	18.64	0.502	0.256
	-	✓	-	-	17.58	0.412	0.288
	-	-	✓	-	18.47	0.498	0.279
	-	-	-	✓	19.83	0.599	0.227
NeRF On-the-go	✓	-	-	-	24.03	0.870	0.132
	-	✓	-	-	22.14	0.853	0.184
	-	-	✓	-	23.65	0.862	0.160
	-	-	-	✓	24.39	0.878	0.126

(a) Ablation of the camera pose initialization.

Scene	Methods	ATE↓	RPE_{trans} ↓	RPE_{rot}↓	Average↓
Ambiguous Wild	VGGT	0.1956	0.2634	2.4526	0.9705
	SfM	0.1031	0.1874	1.3412	0.5439
	I3DGS + RobustSplat	0.1431	0.2102	1.9891	0.7808
	LongSplat + RobustSplat	0.1643	0.2329	2.0107	0.8026
	PAGauss (Ours)	0.0920	0.1523	1.1774	0.4739
Wild Arena	VGGT	0.1837	0.2541	2.3259	0.9809
	SfM	0.0739	0.1324	1.0843	0.4302
	I3DGS + RobustSplat	0.1643	0.2018	1.9765	0.7941
	LongSplat + RobustSplat	0.1832	0.2275	2.0219	0.8120
	PAGauss (Ours)	0.0598	0.0985	0.9240	0.3608

(b) Ablation of the camera pose estimation.

Table 4: Analysis of the camera pose initialization and estimation.

where λ_1 is 0.2. During this phase, we do not perform densification to stabilize pose refinement.

Final training phase. At the final stage, we optimize only the 3D Gaussians to improve rendering quality and represent detailed static structures, following initial phase.

5 Experimental Setup and Empirical Analysis

Implementation details, datasets, and baselines. Our method is built upon RobustSplat (Fu et al., 2025), following its default hyperparameters. We set the number of reference images $N = 4$ to construct $\mathcal{M}_{\text{geo}}$ and $\mathcal{M}_{\text{context}}$. We activate $\mathcal{M}_{\text{spatial}}$ after 500 iterations. For the adaptive source refinement, we randomly sample views from cached images. The initial training phase is performed for the first 30K iterations, followed by the intermediate training phase until 35K iterations, and the final training phase until 40K iterations. We obtain ground-truth camera poses from SfM using distractor-free images, enabling SfM to accurately estimate the camera poses. To evaluate whether PAGauss can identify ambiguous distractors and rectify camera poses, we utilize five benchmark datasets: WildArena, Ambiguous Wild (Park et al., 2026), Photo Tourism (Snavely et al., 2006), Drone Imagery (Tang et al., 2025), and NeRF On-the-go (Ren et al., 2024). We compare our method against prior works (Wang et al., 2026c; Li et al., 2026a; Jiang et al., 2024; Wang et al., 2025c; Kulhanek et al., 2024; Zhang et al., 2024; Sabour et al., 2024). We thoroughly report runtime, memory usage, and hyperparameters. More details are included in the Appendix.

Novel view synthesis results. In Tab. 2 and Tab. 3, PAGauss consistently shows impressive results. While prior works that leverage photometric errors or semantic similarity show competitive results, they still struggle to identify ambiguous distractors that share similar colors or semantics with static regions. However, PAGauss leverages 3D multi-view consistency to robustly discern ambiguous distractors regardless of colors or semantic similarity, showing notable results in Fig. 7.

Camera pose estimation results. In Tab. 4-(a) and Tab. 4-(b), we observe that inaccurate camera pose initialization caused by distractors and unreliable image matching during SfM leads to 3D Gaussian misalignment, degrading rendering quality. While prior methods (Meuleman et al., 2026;

Method	PSNR↑	SSIM↑	LPIPS↓
DUSt3R	22.84	0.803	0.250
MASt3R	22.91	0.810	0.247
π^3	22.97	0.816	0.242
VGGT	23.06	0.820	0.238
DA3	23.13	0.829	0.206

(a) Sensitivity of the 3D VFM.

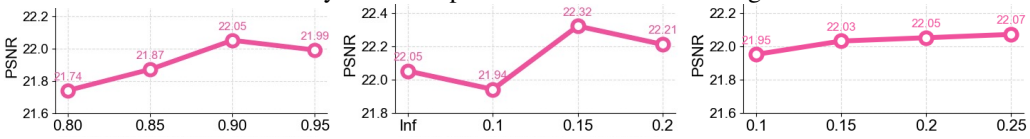
Method	PSNR↑	SSIM↑	LPIPS↓
WildGaussians	22.26	0.751	0.260
$+\mathcal{M}_{\text{spatial}}$	22.79	0.780	0.246
RobustSplat	22.54	0.788	0.249
$+\mathcal{M}_{\text{spatial}}$	23.02	0.803	0.235

(b) Adaptability of the $\mathcal{M}_{\text{spatial}}$.

Method		PSNR↑	SSIM↑	LPIPS↓
$\mathcal{M}_{\text{geo}}$	$\mathcal{M}_{\text{context}}$			
✓	-	22.39	0.798	0.229
✓	✓	23.13	0.829	0.206
SEED	Ours			
✓	-	22.49	0.814	0.217
-	✓	23.13	0.829	0.206

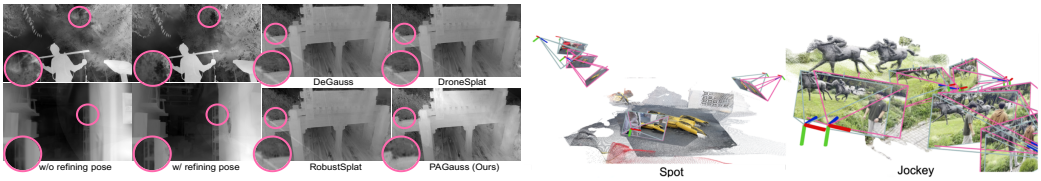
(c) Sensitivity of the $\mathcal{M}_{\text{spatial}}$.

Table 5: Analysis of the spatial-aware mask on the Pagoda scene.



(a) Threshold of the $\mathcal{M}_{\text{context}}$. (b) Value of the camera distance. (c) Value of the λ_1 for $\mathcal{L}_{\text{pose}}$.

Figure 9: Ablation of the hyperparameters across each module on the Miniature scene.



(a) Ablation of the depth consistency. (b) Visualization of w/ and w/o the source refinement.

Figure 10: Analysis of the depth consistency and the source refinement.

Lin et al., 2025a; Fu et al., 2024b) show promising results, since they estimate camera poses using unconstrained images, they imprecisely estimate camera poses, which becomes a bottleneck to achieving high quality results. In contrast, PAGauss precisely refines camera poses with distractor-free images through the adaptive source refinement strategy, improving 3D Gaussian alignment.

Segmentation results. We evaluate robustness to distractors using newly introduced mask annotations, enabling analysis unavailable in prior works. In Fig. 8-(a), we observe that prior works struggle to construct accurate transient masks due to their reliance on semantic similarity or photometric errors, particularly when static regions exhibit similar colors or semantics to distractors. We also quantitatively evaluate how robust PAGauss is. Compared to prior works, PAGauss achieves high precision, demonstrating robustness. The results imply that leveraging 3D spatial relationships enables robust mask construction, highlighting the importance of 3D multi-view consistency.

6 Technical Analysis

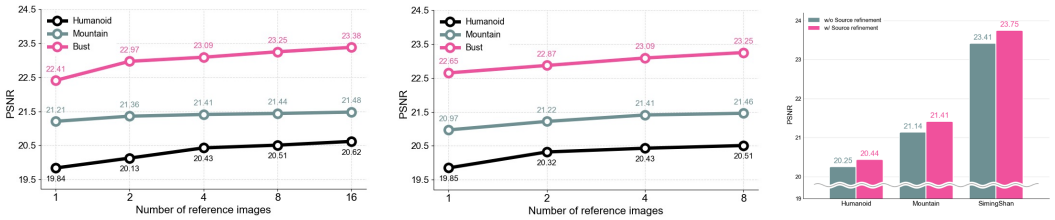
Robustness of the spatial-aware mask. In Tab. 5-(a), the results indicate that PAGauss robustly identifies ambiguous distractors, demonstrating its insensitivity to the choice of foundation model. $\mathcal{M}_{\text{spatial}}$ can be integrated as a plug-and-play module in Tab. 5-(b). This indicates that 3D multi-view consistency improves robustness regardless of semantic similarity or colors.

Method	PSNR↑	SSIM↑	LPIPS↓
Transient mask signal			
+ Photometric error	21.09	0.770	0.285
+ Semantic similarity	21.34	0.803	0.261
+ 3D uncertainty value	22.05	0.829	0.197

Table 6: Analysis of the spatial-aware mask signal on the Miniature scene.

We also confirm that leveraging 3D uncertainty maps as a signal in constructing transient masks shows robustness, yielding better results in Tab. 6. Moreover, PAGauss maintains strong results even when replacing the instance masks extracted from SAM with SEED (Van den Bergh et al., 2015), confirming its low dependence on the SAM in Tab. 5-(c). We evaluate threshold values for $\mathcal{M}_{\text{context}}$ in Fig. 9-(a). We observe that distractors share similar characteristics across scenes, highlighting the importance of contextual cues in identifying distractors. Furthermore, in Fig. 11-(a), we find that leveraging more images helps identify multi-view consistent regions, indicating its scalability. This suggests that harnessing 3D spatial relationships leads to improved generalization.

Effect of the spatial-aware camera pose alignment. PAGauss confirms its insensitivity and robustness in refining camera poses and selecting reference images for source refinement in Fig. 9-(b) and Fig. 9-(c). The results indicate our strategy accurately rectifies camera poses, improving 3D Gaussian alignment in Fig. 10-(a). We also demonstrate the scalability of PAGauss by regularizing camera poses using estimated poses from SfM with a larger number of images in Fig. 11-(b). Fur-

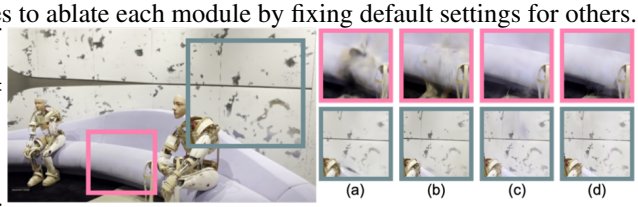


(a) Number of images for $\mathcal{M}_{\text{spatial}}$. (b) Number of images for $\mathcal{L}_{\text{pose}}$. (c) Source refinement.

Figure 11: Ablation of the number of reference images for each module. For each module, we only change the number of reference images to ablate each module by fixing default settings for others.

Method	Humanoid PSNR↑	SSIM↑	LPIPS↓
Baseline	19.10	0.812	0.309
(a) $+\mathcal{M}_{\text{spatial}}$	19.72	0.843	0.248
(b) + Camera pose alignment	19.93	0.855	0.236
(c) + Source refinement	20.25	0.862	0.213
(d) + Multi-stage training	20.43	0.871	0.200

(a) Effectiveness of each module.



(b) Visualization of the qualitative results.

Figure 12: Impact of each module. We show quantitative and visual results on the Humanoid scene.

thermore, our source refinement strategy achieves a bootstrapping effect by refining camera poses with distractor-free rendered images, improving overall performance in Fig. 11-(c). We also visualize how camera poses are aligned with and without adaptive source refinement in Fig. 10-(b).

Design and module analysis. We explore alternative designs in Tab. 7. We observe that using deformable MLPs to refine camera poses sacrifices training time. We also find that the joint training strategy suffers from the chicken-and-egg problem. However, PAGauss avoids these issues using the multi-stage training. We reveal how each module affects rendering results in Fig. 12-(a). The results indicate that the spatial-aware mask and refining poses mitigate artifacts in Fig. 12-(b).

Method	Train hrs	PSNR ↑	SSIM ↑	LPIPS ↓
Refining camera pose				
Deformable MLP layers	0.94	18.35	0.584	0.301
Learnable params	0.55	18.69	0.598	0.298
Training strategy				
Joint-training	-	16.53	0.527	0.359
Multi-stage training	-	18.69	0.598	0.298

Table 7: Ablation of different approach usages.

Discussion. We examine PAGauss in sparse-view scenarios using the Brandenburg Gate scene in Fig. 13. The results highlight that the more images we use to refine camera poses the more visual quality is enhanced, demonstrating its scalability. We also observe that leveraging distractor-free images enables more precise camera pose alignment, improving overall performance compared to others that estimate camera poses from unconstrained images.

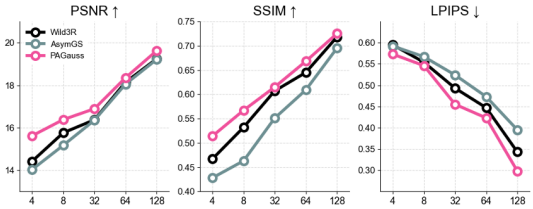


Figure 13: Ablation of the sparse-view scenarios.

WildArena dataset. We capture 82 natural scenes, including challenging unconstrained scenarios with diverse viewpoints (*e.g.*, bird's-eye, frontal, 360°), appearance variations, and ambiguous distractors, along with mask annotations for distractors. Specifically, unlike previous datasets, which lack mask annotations for distractors, the WildArena dataset is, to the best of our knowledge, the first to provide mask annotations that enable robustness analysis against distractors across baselines. We also provide the leaderboard for a fair comparison. More details are included in the Appendix.

7 Conclusion

We propose PAGauss, a novel framework that considers 3D spatial relationships across viewpoints to identify ambiguous distractors and refine imprecise camera poses in unconstrained scenarios. Leveraging 3D multi-view consistency, our spatial-aware adaptive mask robustly identifies ambiguous distractors regardless of colors or semantic similarity, while spatial-aware camera pose alignment precisely refines camera poses using distractor-free images through the source refinement strategy. We also release the WildArena dataset with challenging scenarios and mask annotations that enable robustness analysis against distractors across baselines. Extensive results across in-the-wild datasets demonstrate that PAGauss consistently shows superior robustness and high-quality results.

REFERENCES

Honggyu An, Jaewoo Jung, Mungyeom Kim, Chaehyun Kim, Minkyeong Jeon, Jisang Han, Kazumi Fukuda, Takuya Narihira, Hyunah Ko, Junsu Kim, et al. Learning compact 3d representations from feed-forward novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 363–373, 2026.

Yanqi Bao, Jing Liao, Jing Huo, and Yang Gao. Distractor-free generalizable 3d gaussian splatting. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=G33Iemmj3Z.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19457–19467, 2024.

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, et al. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026.

Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12952, 2022.

Woong Oh Cho, In Cho, Seoha Kim, Jeongmin Bae, Youngjung Uh, and Seon Joo Kim. 4d scaffold gaussian splatting with dynamic-aware anchor growing for efficient and high-fidelity dynamic scene reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3363–3371, 2026.

Hiba Dahmani, Moussab Bennehar, Nathan Piasco, Luis Roldao, and Dzmitry Tsishkou. Swag: Splatting in the wild images with appearance-conditioned gaussians. In *Proceedings of the European Conference on Computer Vision*, pp. 325–340. Springer, 2024.

Hongbo Duan, Pengting Luo, Chengzhi Zhao, Yuanhao Chiang, Fangming Liu, and Xueqian Wang. Dynactivegs: Active gaussian splatting for dynamic scene reconstruction. *arXiv preprint arXiv:2608.01178*, 2026.

Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Sparse-view gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024.

Chuanyu Fu, Yuqi Zhang, Kunbin Yao, Guanying Chen, Yuan Xiong, Chuan Huang, Shuguang Cui, and Xiaochun Cao. Robustsplat: Decoupling densification and dynamics for transient-free 3dgs. *arXiv preprint arXiv:2506.02751*, 2025.

Stephanie Fu, Mark Hamilton, Laura E Brandt, Axel Feldmann, Zhoutong Zhang, and William Freeman. Featup: A model-agnostic framework for features at any resolution. In *Proceedings of the International Conference on Learning Representations*, volume 2024, pp. 45324–45350, 2024a.

Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20796–20805, 2024b.

Yuki Fujimura, Takahiro Kushida, Kazuya Kitano, Takuya Funatomi, and Yasuhiro Mukaigawa. Wildsplatter: Feed-forward 3d gaussian splatting with appearance control from unconstrained images. *arXiv preprint arXiv:2604.21182*, 2026.

Yuto Furutani, Takashi Otonari, Kaede Shiohara, and Toshihiko Yamasaki. Wild3r: Feed-forward 3d gaussian splatting from unconstrained sparse photo collection. *arXiv preprint arXiv:2606.11894*, 2026.

Meiying Gu, Jiawei Zhang, Jiahe Li, Xiaohan Yu, Haonan Luo, Jin Zheng, and Xiao Bai. Sparsesurf: Sparse-view 3d gaussian splatting for surface reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4311–4319, 2026.

Vinayak Gupta, Chih-Hao Lin, Shenlong Wang, Anand Bhattad, and Jia-Bin Huang. Generalizable sparse-view 3d reconstruction from unconstrained images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33217–33226, 2026.

Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. *arXiv preprint arXiv:2512.04012*, 2025.

Liang Han, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Binocular-guided 3d gaussian splatting with view consistency for sparse view synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pp. 68595–68621, 2024.

Jianwei Hu, Tingxuan Huang, Hengyu Zhou, Ningna Wang, Xiaohu Guo Jinshan Lai, and Bin Wang. Mvfusion-gs: Motion-variance guided temporal attention for high-quality dynamic gaussian splatting. *arXiv preprint arXiv:2607.01578*, 2026.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4220–4230, 2024.

Hwasik Jeong, Seungryong Lee, Gyeongjin Kang, Seungkwon Yang, Xiangyu Sun, Seungtae Nam, and Eunbyung Park. 2xplat: Two experts are better than one generalist. *arXiv preprint arXiv:2603.21064*, 2026.

Bo Ji and Angela Yao. Sfm-free 3d gaussian splatting via hierarchical training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21654–21663, 2025.

Yuheng Jiang, Zhehao Shen, Yu Hong, Chengcheng Guo, Yize Wu, Yingliang Zhang, Jingyi Yu, and Lan Xu. Robust dual gaussian splatting for immersive human-centric volumetric videos. *ACM Transactions on Graphics*, 43(6):1–15, 2024.

Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T Barron, Noah Snavely, and Aleksander Hołyński. Zipmap: Linear-time stateful 3d reconstruction via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21748–21759, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

Injae Kim, Chaehyeon Kim, Minseong Bae, Minseok Joo, and Hyunwoo J Kim. F4splat: Feed-forward predictive densification for feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2603.21304*, 2026.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Dmytro Kotovenko, Olga Grebenkova, and Björn Ommer. Edgs: Eliminating densification for efficient convergence of 3dgs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41065–41076, 2026.

Jonas Kulhanek and Torsten Sattler. Nerfbaselines: Consistent and reproducible evaluation of novel view synthesis methods. *arXiv preprint arXiv:2406.17345*, 2024.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024.

JongMin Lee and Sungjoo Yoo. Dense-sfm: Structure from motion with dense consistent matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6404–6414, 2025.

Joo Chan Lee, Jong Hwan Ko, and Eunbyung Park. Optimized minimal 3d gaussian splatting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 38, pp. 135864–135888, 2026.

Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic gaussian splatting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pp. 5384–5409, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Proceedings of the European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

Chengqi Li, Zhihao Shi, Yangdi Lu, Wenbo He, and Xiangyu Xu. Robust neural rendering in the wild with asymmetric dual 3d gaussian splatting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 38, pp. 50166–50191, 2026a.

Hao Li, Sicheng Li, Xiang Gao, Abudouaihati Batuer, Lu Yu, and Yiyi Liao. Gifstream: 4d gaussian-based immersive video with feature stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21761–21770, 2025a.

Moyang Li, Zihan Zhu, Marc Pollefeys, and Daniel Barath. Droid-slam in the wild. *arXiv preprint arXiv:2603.19076*, 2026b.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5521–5531, 2022.

Xinze Li, Pengxu Chen, Yiyuan Wang, Weifeng Su, and Wentao Cheng. S-vggt: Structure-aware subscene decomposition for scalable 3d foundation models. *arXiv preprint arXiv:2603.17625*, 2026c.

Yanyan Li, Youxu Fang, Zunjie Zhu, Kunyi Li, Yong Ding, and Federico Tombari. 4d gaussian splatting slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25019–25028, 2025b.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8508–8520, 2024.

Chin-Yang Lin, Cheng Sun, Fu-En Yang, Min-Hung Chen, Yen-Yu Lin, and Yu-Lun Liu. Longsplat: Robust unposed 3d gaussian splatting for casual long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27412–27422, 2025a.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025b.

Wei Long, Haifeng Wu, Shiyin Jiang, Jinhua Zhang, Xinchun Ji, and Shuhang Gu. Idesplat: Iterative depth probability estimation for generalizable 3d gaussian splatting. *arXiv preprint arXiv:2601.03824*, 2026.

Cheng-You Lu, Yi-Shan Hung, Wei-Ling Chi, Hao-Ping Wang, Charlie Li-Ting Tsai, Yu-Cheng Chang, Yu-Lun Liu, Thomas Do, and Chin-Teng Lin. Df3dv-1k: A large-scale dataset and benchmark for distractor-free novel view synthesis. *arXiv preprint arXiv:2604.13416*, 2026.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

Alexander Markin, Vadim Pryadilshchikov, Artem Komarichev, Ruslan Rakhimov, Peter Wonka, and Evgeny Burnaev. T-3dgs: Removing transient objects for 3d scene reconstruction. *arXiv preprint arXiv:2412.00155*, 2024.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 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.

Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18039–18048, 2024.

Andreas Meuleman, Linus Franke, Boris Zhestiankin, Camille Montemagni, and George Drettakis. Immediate 3d gaussian splat reconstruction of unordered input with global consistency. *arXiv preprint arXiv:2607.14481*, 2026.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *Proceedings of the European Conference on Computer Vision*, 2024.

Shanshan Pan, Jiale Chen, Yilin Liu, and Hui Huang. Topogs: Planar reconstruction via topology-aware 3d gaussian splatting. *arXiv preprint arXiv:2607.16838*, 2026a.

Tianbo Pan, Xingyi Yang, Shizun Wang, and Xinchao Wang. Visual geometry transformer in the wild: Distractor-free 3d reconstruction. *arXiv preprint arXiv:2606.22787*, 2026b.

Wongi Park, Myeongseok Nam, Siwon Kim, Sangwoo Jo, and Soomok Lee. Forestsplats: Deformable transient field for gaussian splatting in the wild. *arXiv preprint arXiv:2503.06179*, 2025.

Wongi Park, Jiyeon Lim, Minjae Lee, Myeongseok Nam, Seongjun Choi, Jungwoo Kim, Soomok Lee, William J Beksi, and SangHyun Lee. Rectifying mask via entropy for distractor-free 3dgs in ambiguous scenarios. *arXiv preprint arXiv:2606.29496*, 2026.

Mia Gaia Polansky, George Kopanas, Stephan Garbin, Todd Zickler, and Dor Verbin. Eulerian gaussian splatting using hashed probability pyramids. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19045–19053, 2026.

Shiwei Ren, Tianci Wen, Yongchun Fang, and Biao Lu. Fastgs: Training 3d gaussian splatting in 100 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26094–26103, 2026.

Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. 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.

Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20626–20636, 2023.

Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David J Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *arXiv preprint arXiv:2406.20055*, 2024.

Vikram Sandu, Mayurdeep Pathak, and Rajiv Soundararajan. Temporally aware densification for dynamic 3d gaussian splatting. *arXiv preprint arXiv:2606.23212*, 2026.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Proceedings of the European Conference on Computer Vision*, 2016.

Kangmin Seo and Jae-Pil Heo. Per-view gaussian predictions enable training-free distractor filtering in feed-forward 3dgs. *arXiv preprint arXiv:2608.26951*, 2026.

Dongbo Shi, Shen Cao, Lubin Fan, Bojian Wu, Jinhui Guo, Ligang Liu, and Renjie Chen. Trackgs: Optimizing colmap-free 3d gaussian splatting with global track constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8960–8968, 2026.

Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *Proceedings of the ACM SIGGRAPH Conference*, pp. 835–846. ACM, 2006.

Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20675–20685, 2024.

Shuo Sun, Unal Artan, Malcolm Mielle, Achim J Lilienthaland, and Martin Magnusson. Dense dynamic scene reconstruction and camera pose estimation from multi-view videos. *arXiv preprint arXiv:2603.12064*, 2026.

Jiadong Tang, Yu Gao, Dianyi Yang, Liqi Yan, Yufeng Yue, and Yi Yang. Dronesplat: 3d gaussian splatting for robust 3d reconstruction from in-the-wild drone imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 833–843, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 1363–1389, 2023.

Yuzhou Tang, Dejun Xu, Yongjie Hou, Zhenzhong Wang, and Min Jiang. Nexussplats: Efficient 3d gaussian splatting in the wild. *arXiv preprint arXiv:2411.14514*, 2024.

Atakan Topaloğlu, Kunyi Li, Michael Niemeyer, Nassir Navab, A Murat Tekalp, and Federico Tombari. Oraclegs: Grounding generative priors for sparse-view gaussian splatting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 77–87, 2026.

Michael Van den Bergh, Xavier Boix, Gemma Roig, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. *International Journal of Computer Vision*, 111:298–314, 2015.

Diwen Wan, Ruijie Lu, and Gang Zeng. Superpoint gaussian splatting for real-time high-fidelity dynamic scene reconstruction. *arXiv preprint arXiv:2406.03697*, 2024.

Fusang Wang, Nathan Piasco, Moussab Bennehar, Luis Roldão, Dzmitry Tsishkou, and Fabien Moutarde. Lesv: Language embedded sparse voxel fusion for open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2604.01388*, 2026a.

Haoran Wang, Guoxi Huang, Fan Zhang, David Bull, and Nantheera Anantrasirichai. Prune wisely, reconstruct sharply: Compact 3d gaussian splatting via adaptive pruning and difference-of-gaussian primitives. *arXiv preprint arXiv:2602.24136*, 2026b.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a.

Peihao Wang, Yuehao Wang, Dilin Wang, Sreyas Mohan, Zhiwen Fan, Lemeng Wu, Ruisi Cai, Yu-Ying Yeh, Zhangyang Wang, Qiang Liu, et al. Steepest descent density control for compact 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26663–26672, 2025b.

Rui Wang, Quentin Lohmeyer, Mirko Meboldt, and Siyu Tang. 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, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024a.

Weijie Wang, Yeqing Chen, Zeyu Zhang, Hengyu Liu, Haoxiao Wang, Zhiyuan Feng, Wenkang Qin, Feng Chen, Zheng Zhu, Donny Y Chen, et al. Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. *arXiv preprint arXiv:2509.19297*, 2025d.

Xu Wang, Zhiru Wang, Shiyun Xie, Chengwei Pan, and Yisong Chen. Dualsplat: Robust 3d gaussian splatting via pseudo-mask bootstrapping from reconstruction failures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4912–4921, 2026c.

Yihao Wang, Marcus Klasson, Matias Turkulainen, Shuzhe Wang, Juho Kannala, and Arno Solin. Desplat: Decomposed gaussian splatting for distractor-free rendering. *arXiv preprint arXiv:2411.19756*, 2024b.

Yunsong Wang and Gim Hee Lee. Flow4dgs-slam: Optical flow-guided 4d gaussian splatting slam. *arXiv preprint arXiv:2604.22339*, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 3502–3516, 2022.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024a.

Mingxuan Wu, Huang Huang, Justin Kerr, Chung Min Kim, Anthony Zhang, Brent Yi, and Angjoo Kanazawa. Predict-optimize-distill: A self-improving cycle for 4d object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6575–6584, 2025.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21551–21561, 2024b.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, et al. Scal3r: Scalable test-time training for large-scale 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21760–21771, 2026.

Haolin Xiong. *SparseGS: Real-time 360 sparse view synthesis using Gaussian splatting*. University of California, Los Angeles, 2024.

Xinmiao Xiong, Bangya Liu, Hao Wang, Dayou Li, Nuo Chen, Andrew Feng, Mingyu Ding, Suman Banerjee, Yang Zhou, and Zhiwen Fan. Accelerating transformer-based monocular slam via geometric utility scoring. *arXiv preprint arXiv:2604.08718*, 2026.

Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections. *arXiv preprint arXiv:2407.12306*, 2024a.

Jiacong Xu, Yiqun Mei, and Vishal M Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *arXiv preprint arXiv:2406.10373*, 2024b.

Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19595–19604, 2024.

Qi Yang, Le Yang, Geert Van Der Auwera, and Zhu Li. Hybridgs: High-efficiency gaussian splatting data compression using dual-channel sparse representation and point cloud encoder. *arXiv preprint arXiv:2505.01938*, 2025.

Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, and Mingkui Tan. 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.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *Proceedings of the International Conference on Learning Representations*, volume 2024, pp. 9142–9159, 2024.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456, 2024.

Wenhao Yuan, Yiyuan Ge, and Deli Cai. Stereogs: Sparse-view 3d gaussian splatting via stereo priors. *arXiv preprint arXiv:2606.30545*, 2026.

Youngsik Yun, Jeongmin Bae, Hyunseung Son, Seoha Kim, Hahyun Lee, Gun Bang, and Youngjung Uh. Compensating spatiotemporally inconsistent observations for online dynamic 3d gaussian splatting. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference*, pp. 1–9, 2025.

Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *Proceedings of the European Conference on Computer Vision*, pp. 341–359. Springer, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21936–21947, 2025.

Xiaoxue Zhang, Xiaoxu Zheng, Yixuan Yin, Tiao Zhao, Kaihua Tang, Michael Bi Mi, Zhan Xu, and Dave Zhenyu Chen. Anchorsplat: Feed-forward 3d gaussian splatting with 3d geometric priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18924–18933, 2026a.

Xiyu Zhang, Jingyu Zhuang, Hongjia Zhai, Zizheng Yan, Jinwei Chen, Guofeng Zhang, and Qingnan Fan. Wildsplat: Feedforward gaussian splatting from unposed in-the-wild images. *arXiv preprint arXiv:2607.05347*, 2026b.

Yangfan Zhao, Hanwei Zhang, Ke Huang, Qiufeng Wang, Zhenzhou Shao, and Dengyu Wu. Ru4d-slam: Reweighting uncertainty in gaussian splatting slam for 4d scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1492–1502, 2026.

Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni. Wildgs-slam: Monocular gaussian splatting slam in dynamic environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2025.

Jianhao Zheng, Liyuan Zhu, Zihan Zhu, and Iro Armeni. Wildpose: A unified framework for robust pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28903–28913, 2026.

Xingguang Zhong, Liren Jin, Marija Popovic, Jens Behley, and Cyrill Stachniss. Dynamic visual slam using a general 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21805–21815, 2026.

Junsheng Zhou, Zhifan Yang, Liang Han, Wenyuan Zhang, Kanle Shi, Shenkun Xu, and Yu-Shen Liu. 4c4d: 4 camera 4d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11829–11839, 2026a.

Qinzheng Zhou, Zaychik Liu, Lijing Lu, and Zhihang Li. Confidence-guided multi-scale aggregation for sparse-view high-resolution 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19054–19064, 2026b.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024.

Siting Zhu, Renjie Qin, Guangming Wang, Jiuming Liu, and Hesheng Wang. Semgauss-slam: Dense semantic gaussian splatting slam. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 21174–21181, 2025.

Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Shengyin Jiang, Long Chen, Zhi-Xin Yang, and Jiwen Lu. Dvgt: Driving visual geometry transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14658–14668, 2026.

APPENDIX

A Related work **2**

Feed-forward methods. . . . 2

Advanced visual SLAM. . . . 2

Sparse-view scenarios. . . . 2

Dynamic 3D reconstruction. . . . 2

B Descriptions of the WildArena dataset **2**

Details of dataset acquisition. . . . 2

Evaluation of the WildArena dataset. . . . 3

WildArena leaderboard. . . . 3

Author statement of the WildArena dataset license. . . . 4

C Experimental setup and empirical analysis **4**

Implementation details, datasets, and metrics. . . . 4

Detailed qualitative results. . . . 4

Detailed quantitative results. . . . 5

Additional segmentation results. . . . 5

Comparison of the memory usage. . . . 5

D Additional technical analysis **6**

Robustness of the spatial-aware mask. . . . 6

Effect of the 3D spatial bootstrapping. . . . 6

Robustness of camera pose alignment. . . . 6

Efficacy of using multiple images. . . . 7

Failure case analysis. . . . 7

Additional design analysis. . . . 7

E Discussion **7**

Practical applications. . . . 7

Limitations. . . . 7

Future work. . . . 8

F Large language models **8**

Use of large language models. . . . 8

G Video demonstration **8**

Video demonstration of novel view synthesis. . . . 8

A Related work

Feed-forward methods. Feed-forward methods (Jeong et al., 2026; Charatan et al., 2024; Wang et al., 2025d; Zhang et al., 2026a) have emerged as an efficient alternative to traditional pipelines (Yu et al., 2024; Kerbl et al., 2023), enabling single-pass inference. Recent approaches (Fujimura et al., 2026; Zhang et al., 2026b; Furutani et al., 2026; Gupta et al., 2026) combine these methods with 3D foundation models (Wang et al., 2025a; Lin et al., 2025b; Wang et al., 2024a). However, they still utilize unconstrained images containing distractors to estimate camera poses, resulting in inaccurate poses that become a bottleneck for improving rendering quality. In contrast, PAGauss rectifies camera poses using distractor-free images through the source refinement strategy, improving 3D Gaussian alignment. We visualize key distinctions in estimating camera poses, as shown in Fig. 14.

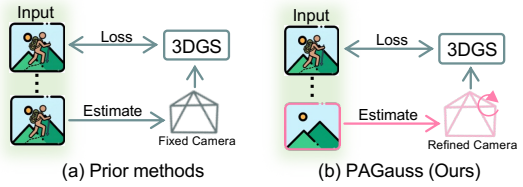


Figure 14: Key distinctions of the camera pose estimation in unconstrained scenarios.

Advanced visual SLAM. Visual SLAM methods (Xiong et al., 2026; Zhu et al., 2025; Yan et al., 2024; Matsuki et al., 2024) estimate camera poses and reconstruct 3D geometry from RGB image sequences. However, they are primarily designed for static scenes and struggle in the presence of dynamic objects. To address this, prior works (Zheng et al., 2025; 2026; Li et al., 2026b; Zhao et al., 2026; Li et al., 2025b) leverage semantic priors or predefined object categories to remove distractors and estimate camera poses using unconstrained images. Although promising, they fail to identify ambiguous distractors that share similar colors or semantic similarity with static regions, resulting in inaccurate camera pose estimation. Orthogonal to these that rely on 2D representations, PAGauss robustly captures ambiguous distractors using 3D spatial relationships across viewpoints and precisely estimates camera poses using distractor-free images through the source refinement.

Sparse-view scenarios. While previous works (Wu et al., 2024b; Pan et al., 2026a; Ren et al., 2026; Kotovenko et al., 2026; Xiong, 2024) typically utilize dense image collections, several approaches (Gu et al., 2026; Yuan et al., 2026; Han et al., 2024; Wu et al., 2024b; Zhou et al., 2026b) have explored 3D reconstruction from sparse constrained image sets. However, extending these methods to unconstrained scenarios remains challenging due to distractors. Inspired by this, recent approaches (Furutani et al., 2026; Gupta et al., 2026; Fujimura et al., 2026) have been proposed. Although they show competitive results, they overlook the imprecise camera poses obtained from unconstrained images. In contrast, PAGauss addresses this overlooked issue by refining imprecise camera poses.

Dynamic 3D reconstruction. Recent works (Li et al., 2024; 2025a; Lee et al., 2026; Cho et al., 2026; Yun et al., 2025) extend 3DGS to dynamic 3D reconstruction using deformation fields. Early approaches (Sun et al., 2024; Wu et al., 2024a; Yang et al., 2024) primarily deform canonical-space Gaussians over time, which introduces redundant computations due to updating an excessive number of Gaussians. To address this issue, modern methods (Lee et al., 2024; Wan et al., 2024; Huang et al., 2024; Zhou et al., 2026a; Sandu et al., 2026) interpolate sparse control points to drive Gaussian motion, enabling high-quality rendering and motion editing. However, they mainly focus on rendering dynamic elements that persist throughout the scene using consistently captured image sequences. In contrast, PAGauss leverages casually captured images to identify distractors that appear transiently across multiple viewpoints, showing high-quality results in real-world scenarios.

B Descriptions of the WildArena dataset

Details of dataset acquisition. The WildArena dataset consists of 82 scenes manually captured using an iPhone 12 Pro, a Samsung Galaxy S25, and several other mobile devices. The acquisition process follows previous methodologies (Sabour et al., 2023; Ren et al., 2024; Lu et al., 2026; Park et al., 2026) to ensure consistency. During each capture, exposure, white balance, and ISO were kept fixed. This dataset contains ambiguous scenarios and diverse camera viewpoints (*e.g.*, bird's-eye, frontal,

Dataset	Scene	Indoor	Outdoor	Ambiguity	Mask GT	360°
WildArena (Ours)	82	42	40	✓	✓	✓
Ambiguous Wild	18	10	8	✓	✗	✗
DF3DV-1K	1K	726	322	▲	✗	✓
NeRF On-the-go	12	2	10	✗	✗	✓
Drone Imagery	3	0	3	✗	✗	✗
Photo Tourism	25	0	25	✗	✗	✗
RobustNeRF	5	5	0	✗	✗	✗
DeGauss	5	4	0	✗	✗	✗

Table 8: Comparison of the unconstrained benchmark datasets.

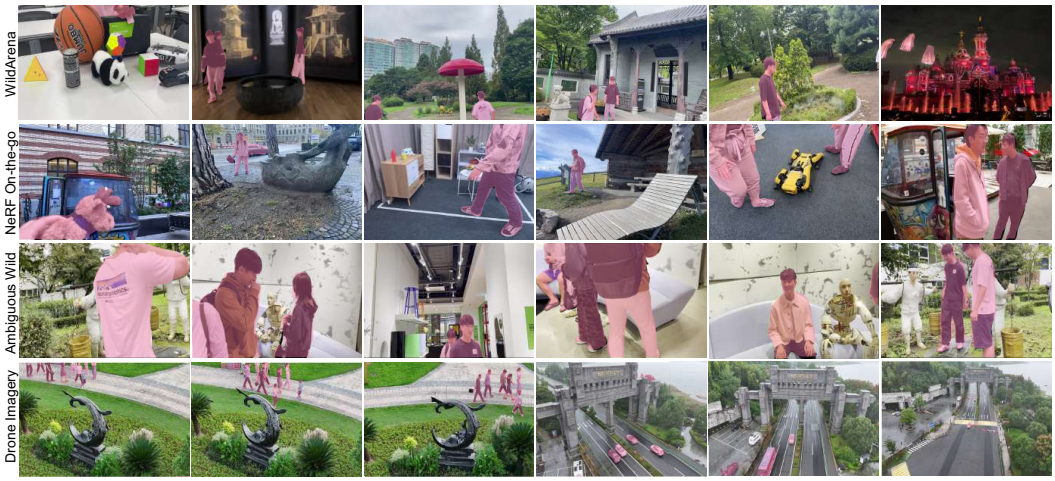


Figure 15: Examples of mask annotations for in-the-wild datasets.

360°, ground-level) that are underexplored in previous datasets, along with mask annotations for distractors. For each scene, approximately 10–12 images are held out as test views, following previous works. Furthermore, the scenes cover a wide range of occlusion ratios, from 5% to 80%, defined as the proportion of pixels occupied by distractors relative to the total number of pixels in each image. Images were captured using multiple mobile devices at resolutions of 1920×1080 and 1920×1200.

We also observe that existing datasets (Sabour et al., 2023; Ren et al., 2024; Snavely et al., 2006; Lu et al., 2026; Wang et al., 2025c; Li et al., 2022; Tang et al., 2025; Xu et al., 2024b) do not provide mask annotations for distractors. However, mask annotations for distractors are important for evaluating whether existing methods can robustly identify distractors and occlusions in unconstrained scenarios and assessing how such distractors affect the 3D novel view synthesis results. To facilitate robustness analysis in these challenging scenarios, we provide distractor mask annotations for WildArena and augment existing datasets with complementary annotations, as shown in Fig. 15. To ensure high-quality mask annotations for in-the-wild datasets, we leverage SAM and SAM2 to automatically generate initial masks. We then manually refine inaccurate masks and further validate mask annotations using a professional data annotation service. We summarize the key distinctions between the WildArena dataset and prior datasets in Tab. 8. We also show samples from each scene in the WildArena dataset in Fig. 31.

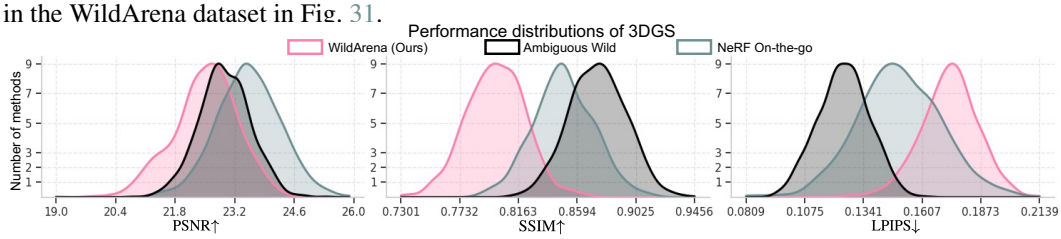


Figure 16: Difficulty analysis of performance for unconstrained scenarios.

Evaluation of the WildArena dataset. To establish the WildArena dataset as a challenging benchmark for unconstrained 3D reconstruction, we evaluate the WildArena dataset from two complementary perspectives: the ambiguity of distractors and the quality of the mask annotations. To assess the ambiguity of distractors, we adopt the evaluation protocol introduced in RefineSplat (Park et al., 2026), which characterizes ambiguity in terms of color similarity and semantic similarity between distractors and static scenes. To assess the quality of the mask annotations, we recruit multiple human annotators to manually verify the annotations and conduct a user study, as depicted in Fig. 36. These evaluations comprehensively validate the ambiguity and mask annotation quality of WildArena as a benchmark for unconstrained 3D reconstruction. Moreover, we observe that leveraging SfM (Schönberger & Frahm, 2016) on the WildArena dataset tends to produce incorrect initialization of 3D Gaussians, as shown in Fig. 32. We further analyze the difficulty of the WildArena dataset based on the performance distributions of 3DGS methods, as shown in Fig. 16. Compared to prior datasets (Park et al., 2026; Ren et al., 2024), WildArena exhibits lower PSNR and SSIM and higher LPIPS, indicating that it contains challenging scenarios for 3D reconstruction.

WildArena leaderboard. Prior methods often adopt different evaluation protocols, hindering fair comparison. These protocols vary in several aspects, including image resolution, downscaling strategies, image distortion, and the choice of LPIPS variant. Such differences can lead to substantial variations in performance across evaluation protocols, as shown in Fig. 17. To facilitate consistent and reproducible evaluation, following previous works (Chen et al., 2026; Kulhanek & Sattler, 2024; Myrzakhan et al., 2024), we establish a unified protocol by splitting the data into train and test sets across datasets and providing a public leaderboard, enabling fair comparisons across prior and future work, as illustrated in Fig. 37. Furthermore, we release unified reproduction results, including rendered views and distractor masks, to facilitate transparent and reproducible comparisons. We hope that the leaderboard will promote fair evaluation and facilitate further progress for real-world scenarios.

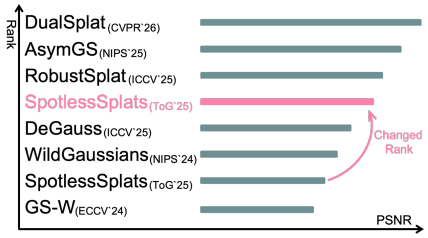


Figure 17: Impact of changing evaluation protocol.

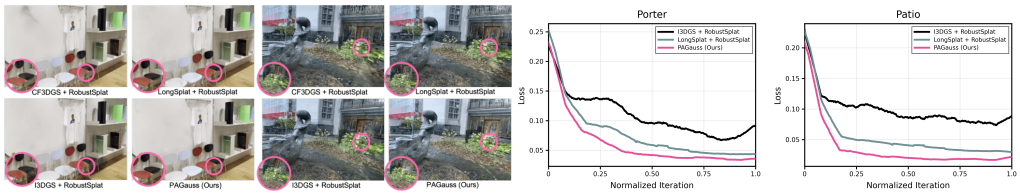
Author statement of the WildArena dataset license. The authors of the WildArena dataset confirm that the dataset was collected, processed, and released in accordance with applicable ethical guidelines and regulations. The authors take full responsibility for any potential violations of rights or ethical standards associated with its use. The dataset is released under the CC BY-NC-ND 4.0 license, permitting free sharing and downloading for non-commercial and research purposes, provided that appropriate credit is given to the authors and no derivative works are created.

C Experimental setup and empirical analysis

Dataset	Indoor	Appear. chng	360°	Ambiguity	BEV	Clean
WildArena (Ours)	✓	✓	✓	✓	✓	✓
Ambiguous Wild	✓	✗	✗	✗	✗	✗
NeRF On-the-go	✓	✗	✓	✗	✗	✗
Drone Imagery	✗	✗	✗	✗	✓	✗
Photo Tourism	✗	✓	✗	✗	✗	✗

Table 9: Characteristics of the unconstrained in-the-wild datasets.

Implementation details, datasets, and metrics. We use Adam (Kingma, 2014) to optimize 3D Gaussians and refine imprecise camera poses. We follow the default training hyperparameters of RobustSplat (Fu et al., 2025), making minor specific hyperparameter modifications, which are summarized in Tab. 17. Furthermore, to reduce time complexity in calculating the uncertainty residual map, we cache rendered images. Moreover, to address appearance variations in the Photo Tourism dataset (Snavely et al., 2006), following prior approaches (Kulhanek et al., 2024; Yang et al., 2023; Park et al., 2025), we employ per-view appearance embeddings of size 32, processed by an appearance MLP with 2 layers of size 128 and ReLU activations. We use instance masks extracted with SAM (Kirillov et al., 2023), following prior works (Tang et al., 2025; Bao et al., 2026); this process takes approximately 20 minutes for the WildArena, NeRF On-the-go, Drone Imagery, and Ambiguous Wild datasets, and about 53 minutes for the Photo Tourism dataset. For the context-aware mask, following recent works (Kulhanek et al., 2024; Ren et al., 2024; Bao et al., 2026; Zhou et al., 2024; An et al., 2026; Wang et al., 2026a; Wu et al., 2025), we upsample the semantic features extracted from the DA3 encoder $\mathcal{E}_{3D}$ to the original image resolution using bilinear interpolation. We utilize five benchmark datasets to evaluate the robustness of PAGauss in diverse real-world scenarios, as indicated in Tab. 9. In the following subsections, we report training time (Train hrs) and frames-per-second (FPS), computed on an RTX PRO 5000 GPU unless stated otherwise. We also reproduce the baselines following the official implementations in Tab. 16. Note that we exclude several methods (Hu et al., 2026; Zhang et al., 2026b) from comparison due to lack of public code. For the sparse-view scenarios, we randomly sample training views, following prior approaches (Seo & Heo, 2026; Furutani et al., 2026; Gupta et al., 2026). We adopt the widely used PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) to evaluate rendering quality. For camera pose estimation, we evaluate camera pose accuracy using three metrics, ATE, $\text{RPE}_{\text{trans}}$, and RPE_{rot}, which measure the absolute trajectory error, relative translation error, and relative rotation error, respectively. In particular, to evaluate the accuracy of camera poses estimated by PAGauss, we derive ground truth camera poses from SfM with distractor-free rendered images produced by several prior methods (Yang et al., 2025; Jiang et al., 2024; Li et al., 2026a; Wang et al., 2026c), ensuring the ground truth poses are obtained independently of our approach. We also evaluate robustness against distractors based on segmentation results across baselines using accuracy, precision, recall, and IoU metrics.



(a) Ablation of novel view synthesis results. (b) Ablation of the multi-stage training strategy.

Figure 18: Analysis of the 3D novel-view synthesis and the training convergence.

Detailed qualitative results. We provide detailed qualitative results, as presented in Fig. 26, Fig. 27, Fig. 28, Fig. 29, and Fig. 30. We observe that prior methods (Li et al., 2026a; Kulhanek et al., 2024; Wang et al., 2025c), which primarily rely on photometric errors or semantic similarity, struggle to discern ambiguous distractors, leading to artifact-prone results. Among these methods, DualSplat (Wang et al., 2026c) and AsymGS (Li et al., 2026a) perform relatively better, as they leverage instance masks obtained from SAM to precisely identify distractors. However, since they also rely on 2D representations as a condition to construct transient masks, they still struggle to identify subtle and ambiguous distractors, especially when the masks extracted from SAM are incorrect. In contrast, since PAGauss leverages 3D multi-view consistency, PAGauss robustly identifies ambiguous distractors regardless of colors or semantic similarity. Moreover, considering context to identify distractors, PAGauss can capture subtle and fine-grained distractors, showing artifact-free 3D reconstruction results. PAGauss also performs favorably on Photo Tourism with significant domain shifts (*e.g.*, appearance variations, lighting), supporting its robustness in unconstrained scenarios.

Detailed quantitative results. We also show detailed quantitative results, as shown in Tab. 18, Tab. 19, Tab. 20, and Tab. 21. Compared to prior approaches (Dahmani et al., 2024; Chen et al., 2022; Tang et al., 2024; Xu et al., 2024a; Wang et al., 2024b; Markin et al., 2024), since PAGauss rectifies imprecise camera poses and robustly identifies ambiguous distractors, PAGauss consistently shows impressive results. In particular, we combine RobustSplat with unposed methods (*e.g.*, CF3DGS (Fu et al., 2024b), LongSplat (Lin et al., 2025a), I3DGS (Meuleman et al., 2026)) and compare quantitative results. Although they show promising results, since they utilize unconstrained images containing distractors, they struggle to precisely estimate camera poses. Unlike them, PAGauss leverages distractor-free images through the source refinement strategy that accurately refines camera poses. Note that we exclude InstantSplat (Fan et al., 2024) from our comparisons, as it targets sparse-view scenarios and periodically invokes Structure-from-Motion (SfM), leading to excessive GPU and training overhead as the number of input images increases in real-world scenarios.

Method	Accuracy↑	Precision↑	Recall↑	loU↑	Average↑
WildGaussians	0.857	0.807	0.820	0.829	0.828
DroneSplat	0.877	0.812	0.831	0.848	0.842
RobustSplat	0.911	0.838	0.907	0.836	0.873
AsymGS	0.913	0.846	0.860	0.832	0.885
PAGauss	0.940	0.853	0.922	0.912	0.906

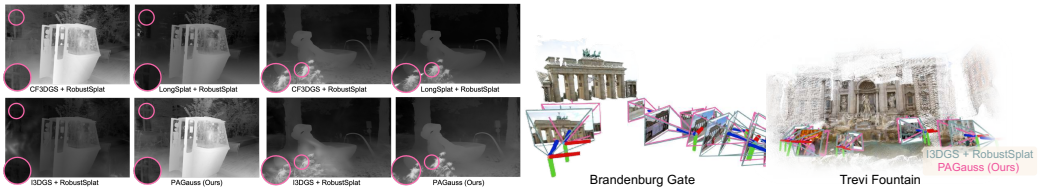
Table 10: Quantitative segmentation results for distractors on the NeRF On-the-go dataset.

Additional segmentation results. Additionally, we evaluate robustness against distractors across baselines on the NeRF On-the-go dataset. Although they (Tang et al., 2025; Kulhanek et al., 2024) show promising results, they still struggle to identify ambiguous distractors using photometric errors or semantic similarity. However, compared to prior works that focus on 2D representations, PAGauss leverages 3D multi-view consistency to capture ambiguous distractors regardless of colors or semantic similarity, demonstrating robustness, as presented in Tab. 10.

N	2	4	8	16
VRAM (GB)	10.32	13.71	16.80	21.83

Table 11: Analysis of the VRAM usage with numbers of images.

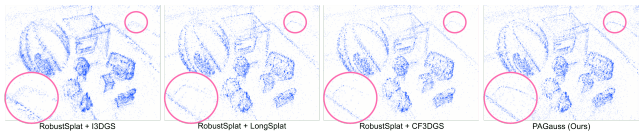
Comparison of the memory usage. We report memory usage, as shown in Tab. 22, Tab. 23, and Tab. 24. We observe that, by refining camera poses, PAGauss reduces misalignment of 3D Gaussians, achieving competitive memory usage compared to prior baselines (Li et al., 2026a; Kulhanek et al., 2024; Tang et al., 2025; Wang et al., 2025c). Note that we do not report the memory usage of DeGauss (Wang et al., 2025c) and AsymGS (Li et al., 2026a), as both incur substantial memory



(a) Ablation of the depth consistency. (b) Visualization of the camera pose alignment.

Figure 19: Analysis of the depth consistency and the camera pose alignment.

Scene	Method	PSNR↑	SSIM↑	LPIPS↓
Activist	DeGauss	17.17	0.595	0.216
	+ source refinement	18.01	0.623	0.198
Spot	RobustSplat	24.85	0.808	0.164
	+ source refinement	26.07	0.908	0.109



(a) Impact of the source refinement. (b) Visualization of the 3D point cloud.

Figure 20: Effectiveness of the spatial-aware pose alignment

overhead due to usage of auxiliary elements (*e.g.*, Dual 3DGS, MLPs), to ensure a fair comparison. We also report VRAM usage with varying numbers of images, as presented in Tab. 11.

D Additional Technical Analysis

Robustness of the spatial-aware mask. In Fig. 33, Fig. 34, and Fig. 35, compared to others that rely on 2D representations, PAGauss robustly identifies distractors across real-world scenarios. The results indicate that leveraging photometric errors or semantic similarity still struggles to identify ambiguous and subtle distractors. However, since PAGauss leverages 3D multi-view consistency and considers contexts in scenes, PAGauss robustly captures ambiguous and fine-grained distractors regardless of colors or semantic similarity. Furthermore, we examine whether learnable weighting factor $\ell^{(w)}$ is beneficial compared to fixed threshold and statistical measures (*e.g.*, mean, variance, percentile) on the Patio scene in Tab. 13. These results suggest that, since distractors vary across real-world scenarios, an adaptive threshold allows the weighting factor to be appropriately adjusted.

Effect of the 3D spatial bootstrapping. We also found that the outlier filtering behavior can be readily adapted to existing bootstrapping-based methods (Fu et al., 2025; Sabour et al., 2024). Specifically, bootstrapping-based methods estimate transient masks $\mathcal{M}_t$ as:

$$\mathcal{M}_t = \text{Sigmoid}(\text{MLP}_\theta(\mathcal{F}_{\text{3D}}(I_i))). \quad (12)$$

Method	PSNR↑	SSIM↑	LPIPS↓
RobustSplat	19.10	0.812	0.309
+ 3D spatial bootstrapping	19.49	0.841	0.285
SpotLessSplats	19.23	0.813	0.296
+ 3D spatial bootstrapping	19.56	0.849	0.267

Table 12: Analysis of the 3D spatial-aware bootstrapping on the Humanoid scene.

Instead of relying on 2D representations, we supervise MLP_θ with the 3D representations as:

$$\mathcal{L}_{conf} = \|\mathcal{M}_t - \beta\|. \quad (13)$$

Compared to prior bootstrapping approaches (Sabour et al., 2024; Fu et al., 2025) that primarily focus on photometric errors or semantic similarity and 2D representations, this leverages both 3D multi-view consistency and 2D representations, constructing transient masks regardless of colors or semantic similarity, enhancing robustness, as presented in Tab. 12.

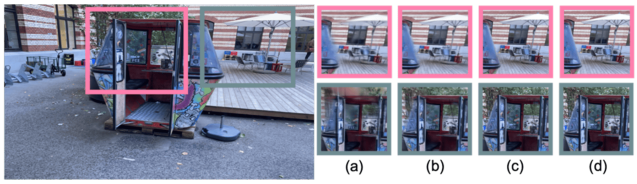
Robustness of camera pose alignment. We compare our method with unposed methods combined with RobustSplat in Fig. 18-(a). Although prior works show competitive results, since they utilize unconstrained images to estimate camera poses, they struggle to refine imprecise camera poses. Furthermore, we observe that jointly optimizing 3D Gaussians and camera poses creates an optimization dilemma, where it is unclear whether the 3D scene representation or the camera poses should be refined first, leading to a circular dependency just like chicken-and-egg problem. However, by adopting the multi-stage training strategy, PAGauss shows consistent improvements by achieving stable training convergence and mitigating the chicken-and-egg problem in Fig. 18-(b). Moreover, PAGauss achieves more consistent depth

$\mathcal{T}_{\text{geo}}$	PSNR↑	SSIM↑	LPIPS↓
0.5	20.37	0.813	0.156
$\mathbb{E}[\mathcal{B}_i] + \mathbb{V}[\mathcal{B}_i]$	21.45	0.827	0.099
$\mathbb{E}[\mathcal{B}_i] + \mathbb{V}[\mathcal{B}_i](1 + \ell^{(w)})$	22.31	0.861	0.092

Table 13: Ablation of the threshold.

Method	Patio-High PSNR↑	SSIM↑	LPIPS↓
Baseline	21.84	0.757	0.192
(a) $+\mathcal{M}_{\text{spatial}}$	23.16	0.834	0.163
(b) + Camera pose alignment	23.89	0.879	0.134
(c) + Source refinement	24.35	0.882	0.127
(d) + Multi-stage training	24.80	0.894	0.102

(a) Impact of each module.



(b) Visualization of the qualitative results.

Figure 22: Impact of each module. We show quantitative and visual results on the Patio-High scene.

through the spatial-aware pose refinement against other methods, as depicted in Fig. 19-(a) and Fig. 19-(b). Additionally, our source refinement strategy can be applied in a plug-and-play manner in improving 3D Gaussian alignment in Fig. 20-(a). We also visualize 3D point clouds, as shown in Fig. 20-(b). The results indicate that PAGauss effectively arranges 3D Gaussian anchors, showing artifact-free 3D reconstruction results. Note that since VGGSfM requires excessive GPU memory, we exclude it from our comparison.

Efficacy of using multiple images. For a fairer comparison with prior methods (Kulhanek et al., 2024; Li et al., 2026a; Wang et al., 2025c), we redesign the training strategy of these methods to use multi-view images as a batch, and evaluate their rendering quality, as presented in Tab. 14. The results indicate that they still struggle to discern ambiguous distractors due to relying on 2D representations, leading to artifact-prone 3D reconstruction. These findings indicate that leveraging 3D spatial relationships is important for identifying ambiguous distractors and refining camera poses in real-world scenarios.

Method	Humanoid PSNR↑	SSIM↑	LPIPS↓	Corner PSNR↑	SSIM↑	LPIPS↓
WildGaussians ($\mathcal{N}=4$)	19.07	0.802	0.353	23.59	0.811	0.154
AsymGS ($\mathcal{N}=4$)	19.53	0.820	0.272	25.90	0.881	0.109
DeGauss ($\mathcal{N}=4$)	17.11	0.734	0.284	25.74	0.867	0.134
RobustSplat ($\mathcal{N}=4$)	19.32	0.825	0.297	25.87	0.875	0.158
PAGauss (Ours)	20.43	0.871	0.200	26.31	0.897	0.103

Table 14: Ablation of multi-view images.

Failure case analysis. We observe that the spatial-aware adaptive masks often fail to capture subtle distractors, as illustrated in Fig. 21. However, such fine-grained transient elements do not significantly degrade rendering quality, allowing PAGauss to achieve competitive results across baselines. This issue has also been raised in previous approaches.



Figure 21: Failure cases.

Additional design analysis. We also reveal how each module affects overall performance in Fig. 22-(a) and 22-(b). The results indicate that leveraging the spatial-aware mask is important to capture ambiguous distractors. We further show that refining camera poses obtained from SfM with unconstrained images leads to improved performance, aligning 3D Gaussians. Moreover, we evaluate different semantic feature extractors for context-aware masks on the Statuette scene, as shown in Tab. 15. The results indicate that PAGauss remains insensitive across different semantic feature extractors.

Method	PSNR↑	SSIM↑	LPIPS↓
DINOv2	22.13	0.889	0.195
DIFT	22.19	0.890	0.183
ResNet50	20.04	0.753	0.203
VGG19	19.42	0.658	0.275
DA3	22.01	0.781	0.191

Table 15: Sensitivity of feature extractors.

E Discussion

Practical applications. Accurate 3D reconstruction benefits cultural heritage sites, historical landmarks, tourist attractions, and archaeological sites by enabling digital preservation, as shown in Fig. 25. PAGauss shows high-quality 3D reconstruction from unconstrained images despite transient distractors and varying illumination. This makes it practical to create digital twins for tourism, education, virtual exploration, and future restoration. These digital twins can provide immersive and interactive representations of real-world sites, enabling users to remotely explore, analyze, and interact with them from anywhere, without physical travel.



Figure 23: Limitations of PAGauss.

Limitations. Although PAGauss refines imprecise camera poses obtained from SfM and identifies ambiguous distractors, it still struggles to reconstruct 3D scenes from unconstrained images that

include scene-irrelevant views in Fig. 23. Furthermore, when distractors occupy more than 80% of the image area across most images, PAGauss finds it difficult to leverage 3D multi-view consistency to identify distractors and refine inaccurate camera poses.

Future work. To address the above issues, combining our approach with an outlier rejection strategy, such as RobustVGGT (Han et al., 2025), could effectively filter out scene-irrelevant views. Furthermore, to identify distractors that occupy a significant portion of an image, leveraging pre-defined object categories may also be beneficial. Developing a robust COLMAP pipeline that can reliably handle in-the-wild images is also an important direction for improving the robustness of 3D reconstruction in challenging real-world environments. Our method might be beneficial in autonomous driving scenarios for rectifying camera pose initialization and identifying ambiguous distractors. We hope PAGauss will inspire future research in challenging real-world environments.

F LARGE LANGUAGE MODELS

Use of large language models. Following ICLR 2027 guidelines, large language models were employed solely to assist with language editing, including refining phrasing, correcting minor typographical errors, and improving the overall clarity of the paper.

G VIDEO DEMONSTRATION

Video demonstration of novel view synthesis. We further provide novel-view synthesis animations rendered along camera trajectories across diverse scenes. These animations enable a more comprehensive assessment of 3D consistency in real-world scenarios. We encourage readers to view the animations using Adobe Acrobat, KDE Okular, or Firefox, as illustrated in Fig. 24.

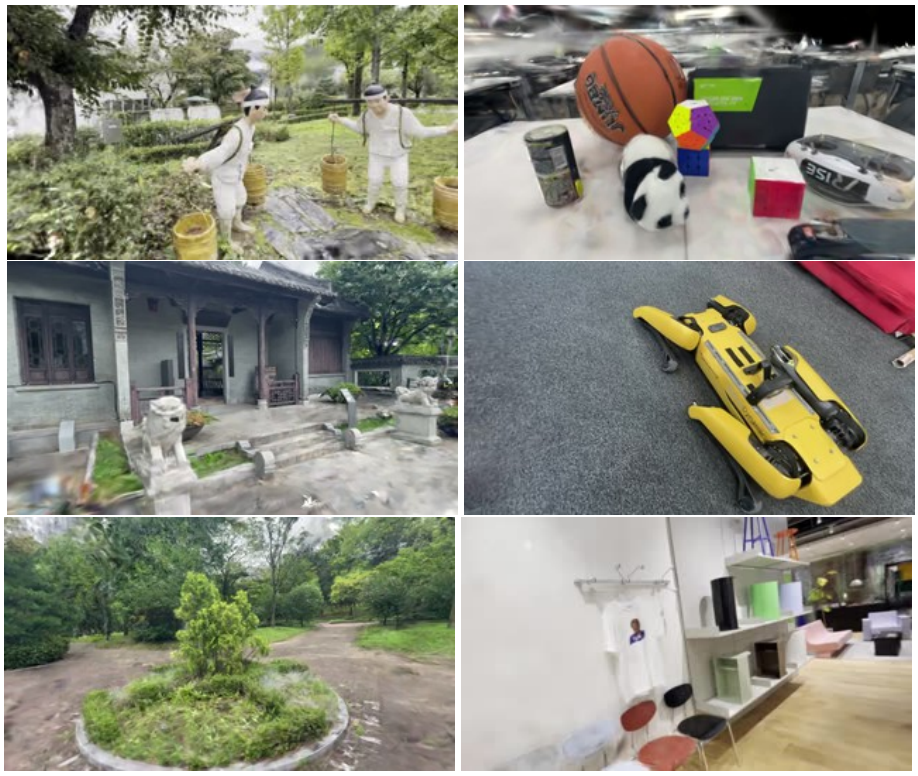


Figure 24: We visualize novel view synthesis results produced by PAGauss along a camera trajectory. Please view the animations in Adobe Acrobat, KDE Okular, or Firefox for the best experience.

Method	Link	License
w/o camera pose optimization baseline		
3DGS	https://github.com/graphdeco-inria/gaussian-splatting	Gaussian-Splatting License
Mip-Splatting	https://github.com/autonomousvision/mip-splatting	Gaussian-Splatting License
Ha-NeRF	https://github.com/rover-xingyu/Ha-NeRF	
CR-NeRF	https://github.com/fcyycf/RobustSplat	Gaussian-Splatting License
Gaussian-W	https://github.com/EastbeanZhang/Gaussian-Wild	
NexusSplats	https://github.com/juliantang324/NexusSplats	MIT License
WildGaussians	https://github.com/jkulhanek/wild-gaussians/	MIT License
SpotLessSplats	https://github.com/lilygoli/SpotLessSplats	Apache 2.0 License
Splatfacto-W	https://github.com/KevinXu02/splatfacto-w	Apache 2.0 License
NerfBaselines	https://github.com/nerfbaselines/nerfbaselines	MIT License
DroneSplat	https://github.com/BITyia/DroneSplat	MIT License
T-3DGS	https://github.com/Vadim200116/T-3DGS	
ForestSplats	https://github.com/WongiPark0628/ForestSplats	
DeGauss	https://github.com/BatFaceWayne/DeGauss	GPL-3.0 License
RobustSplat	https://github.com/fcyycf/RobustSplat	Gaussian-Splatting License
RefineSplat	https://github.com/WongiPark0628/AmbiguousWild	MIT License
AsymGS	https://github.com/steveli88/AsymGS	Apache 2.0 License
DualSplat	https://github.com/Lanslot/DualSplat	Gaussian-Splatting License
Wild3R	https://github.com/furuschool/wild3r	
w/ camera pose optimization baseline		
SfM	https://github.com/colmap/colmap	BSD License
DUSt3R	https://github.com/naver/dust3r	CC BY-NC-SA 4.0 License
MASt3R	https://github.com/naver/mast3r	CC BY-NC-SA 4.0
π^3	https://github.com/yyfz/Pi3	BSD 3-Clause License
VGGT	https://github.com/facebookresearch/vggt	VGGT Non-Commercial License
DA3	https://github.com/ByteDance-Seed/Depth-Anything-3	Apache 2.0 License
CF3DGS	https://github.com/NVlabs/CF-3DGS	NVIDIA License
LongSplat	https://github.com/NVlabs/LongSplat	NVIDIA License
I3DGS	https://github.com/graphdeco-inria/i3dgs	Immediate3DGS License
In-the-Wild datasets		
Photo Tourism	https://github.com/ubc-vision/image-matching-benchmark	Apache 2.0 License
NeRF On-the-go	https://github.com/cvg/nerf-on-the-go	Apache 2.0 License
Drone Imagery	https://github.com/BITyia/DroneSplat	MIT License
Ambiguous Wild	https://github.com/WongiPark0628/AmbiguousWild	Apache 2.0 License

Table 16: The implementation repository and license.

Hyperparameters	WildArena	Ambiguous Wild	NeRF On-the-go	Drone Imagery	Photo Tourism
position_lr_init	0.00016	0.00016	0.00016	0.00016	0.00016
position_lr_final	0.0000016	0.0000016	0.0000016	0.0000016	0.00016
position_lr_delay_mult	0.01	0.01	0.01	0.01	0.01
feature_lr	0.0025	0.0025	0.0025	0.0025	0.0025
opacity_lr	0.1	0.1	0.1	0.1	0.1
scaling_lr	0.005	0.005	0.005	0.005	0.005
rotation_lr	0.001	0.001	0.001	0.001	0.001
percent_dense	0.01	0.01	0.01	0.01	0.01
λ	0.2	0.2	0.2	0.2	0.2
λ_1	0.2	0.2	0.2	0.2	0.2
densification_interval	1000	1000	1000	1000	1000
densify_from_iter	500	500	500	500	500
densify_grad_threshold	0.0002	0.0002	0.0002	0.0002	0.0002
total_iterations	40000	40000	40000	7000	100000

Table 17: Detailed hyperparameters for in-the-wild datasets.

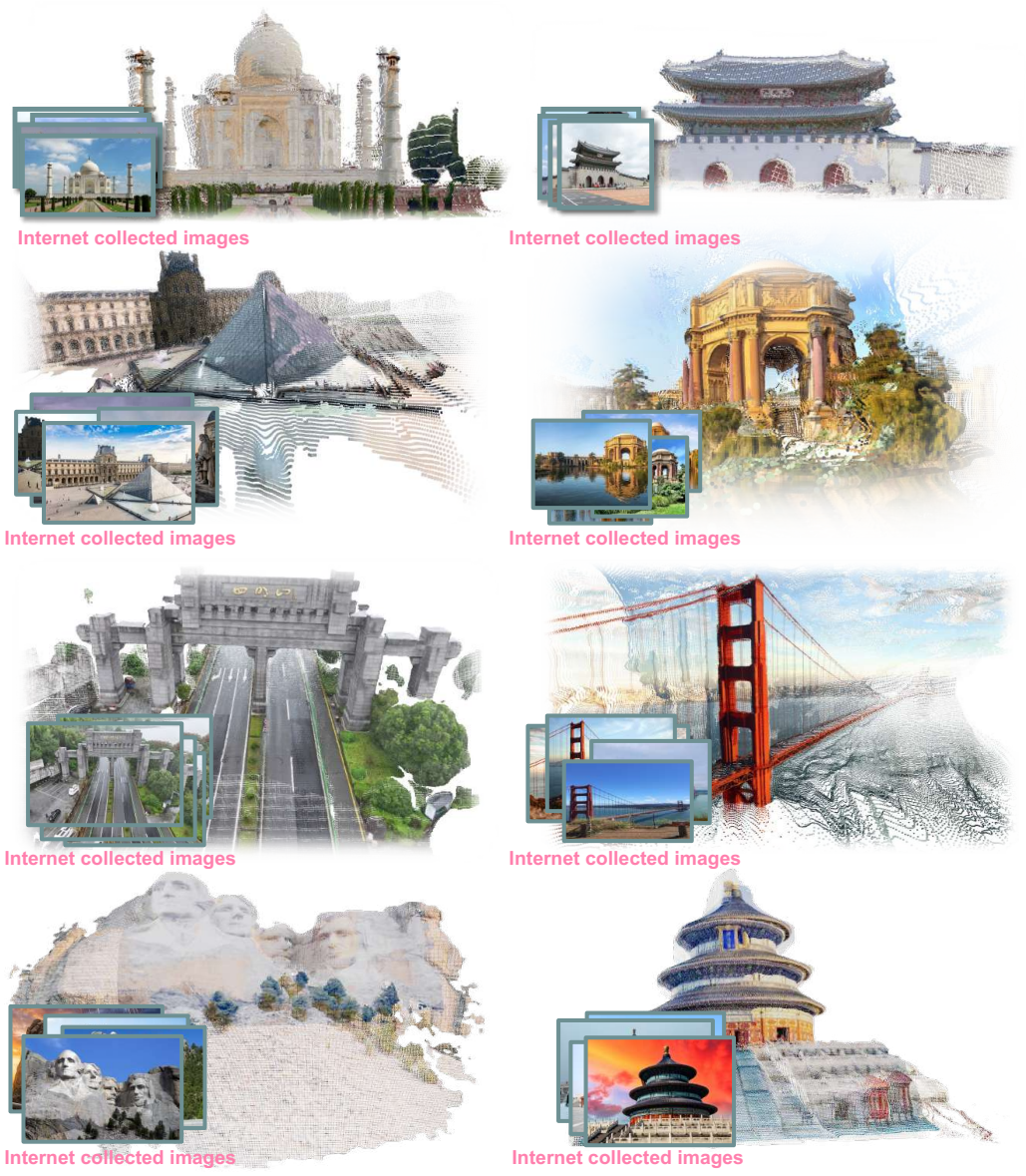


Figure 25: Social impact in real-world scenarios.

Method	Train hrs / FPS	Humanoid			Porter			Lounge			Bust			Jockey			Statuette		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o camera pose optimization																			
GS-W	1.01 / 106	18.31	0.759	0.251	19.65	0.685	0.260	21.21	0.885	0.206	20.40	0.552	0.461	13.33	0.451	0.481	19.08	0.625	0.308
NexusSplats	1.31 / 98	16.56	0.776	0.351	20.51	0.701	0.291	22.55	0.872	0.189	19.81	0.509	0.474	15.41	0.416	0.464	18.89	0.603	0.348
WildGaussians	1.47 / 101	19.02	0.801	0.355	18.54	0.664	0.272	18.52	0.825	0.266	19.59	0.457	0.585	13.46	0.395	0.578	18.03	0.549	0.441
DeGauss	2.92 / 95	16.34	0.745	0.302	16.07	0.748	0.241	16.74	0.420	0.443	17.13	0.538	0.346	17.13	0.538	0.346	18.04	0.598	0.331
RobustSplat	0.78 / 102	19.10	0.812	0.309	21.20	0.706	0.327	22.82	0.895	0.172	21.20	0.529	0.472	17.76	0.478	0.496	20.01	0.614	0.383
AsymGS	1.87 / 107	19.53	0.828	0.272	22.38	0.810	0.131	21.53	0.889	0.198	21.26	0.696	0.245	17.02	0.489	0.331	20.12	0.714	0.237
DualSplat	2.34 / 101	19.38	0.819	0.277	22.74	0.820	0.129	21.97	0.887	0.179	21.50	0.695	0.243	17.39	0.513	0.314	20.63	0.752	0.197
PAGauss (Ours)	0.82 / 104	19.41	0.847	0.225	23.90	0.875	0.125	23.19	0.899	0.154	22.43	0.704	0.219	18.36	0.592	0.298	21.88	0.782	0.184
w/ camera pose optimization																			
CF3DGS + RobustSplat	0.96 / 101	20.24	0.863	0.235	23.97	0.849	0.147	22.13	0.897	0.159	23.03	0.741	0.234	18.33	0.588	0.309	22.07	0.778	0.193
LongSplat + RobustSplat	1.19 / 104	15.06	0.743	0.306	18.14	0.213	0.766	16.75	0.807	0.204	18.69	0.703	0.270	19.81	0.546	0.289	17.32	0.745	0.257
I3DGS + RobustSplat	1.50 / 99	17.16	0.801	0.258	20.93	0.854	0.172	19.48	0.870	0.167	17.81	0.712	0.259	17.63	0.554	0.377	18.86	0.809	0.224
PAGauss (Ours)	0.98 / 103	20.43	0.871	0.200	24.04	0.873	0.130	22.72	0.902	0.143	23.09	0.749	0.221	18.85	0.598	0.255	22.01	0.781	0.191

Table 18: Quantitative results on the Ambiguous Wild dataset.

Method	Train hrs / FPS	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↓
w/o camera pose optimization																			
Ha-NeRF	- / <1	18.64	0.485	0.499	16.71	0.393	0.569	19.23	0.684	0.367	16.82	0.543	0.393	17.85	0.460	0.599	16.67	0.463	0.505
3DGS	0.35 / 116	19.40	0.638	0.213	19.96	0.659	0.185	20.90	0.713	0.241	17.48	0.704	0.199	20.77	0.693	0.316	17.29	0.604	0.363
Mip-Splatting	0.18 / 82	19.23	0.581	0.273	19.74	0.624	0.252	21.44	0.767	0.161	16.96	0.688	0.188	21.14	0.757	0.220	19.12	0.640	0.303
Splatfacto-W	1.03 / 43.1	20.70	0.661	0.169	20.37	0.662	0.187	21.53	0.739	0.241	15.58	0.491	0.536	20.03	0.683	0.324	15.58	0.491	0.536
GS-W	1.03 / 105	19.43	0.596	0.299	20.06	0.723	0.274	22.17	0.793	0.155	19.90	0.681	0.260	17.13	0.608	0.409	19.90	0.681	0.260
WildGaussians	1.56 / 94	20.43	0.653	0.255	20.81	0.662	0.215	24.16	0.822	0.045	21.44	0.800	0.138	23.82	0.816	0.138	22.23	0.725	0.206
SpotLessSplats	1.87 / 81	19.84	0.580	0.294	20.19	0.612	0.258	24.03	0.795	0.258	21.55	0.838	0.065	23.52	0.756	0.185	20.31	0.664	0.259
DeSplat	0.84 / 106	19.59	0.710	0.170	20.27	0.680	0.170	26.05	0.880	0.090	20.89	0.810	0.110	26.07	0.900	0.090	22.59	0.840	0.120
HybridGS	0.98 / 112	21.73	0.693	0.284	21.11	0.674	0.252	25.03	0.847	0.151	21.98	0.812	0.169	24.33	0.794	0.196	21.77	0.741	0.211
DroneSplat	1.35 / 101	21.23	0.687	0.162	21.54	0.705	0.168	24.77	0.823	0.106	21.85	0.816	0.107	24.37	0.821	0.095	22.53	0.778	0.181
RobustSplat	0.75 / 116	21.15	0.737	0.201	21.01	0.701	0.207	25.88	0.876	0.144	21.63	0.827	0.139	24.85	0.808	0.164	21.84	0.757	0.192
T-3DGS	1.72 / 103	20.62	0.703	0.223	20.83	0.681	0.218	26.14	0.890	0.114	20.96	0.819	0.154	25.84	0.893	0.127	22.76	0.831	0.170
AsymGS	2.9 / 110	21.93	0.735	0.162	21.61	0.709	0.162	25.77	0.876	0.089	22.87	0.853	0.091	25.09	0.839	0.152	23.14	0.797	0.156
DualSplat	1.8 / 107	21.86	0.747	0.189	21.28	0.706	0.192	26.27	0.895	0.115	21.64	0.828	0.142	25.96	0.908	0.108	23.26	0.836	0.157
RefineSplat	1.2 / 105	22.62	0.772	0.167	21.51	0.732	0.188	26.21	0.893	0.108	22.08	0.831	0.093	25.87	0.869	0.142	22.71	0.841	0.159
PAGauss (Ours)	0.93 / 112	22.51	0.748	0.179	21.53	0.734	0.187	26.25	0.894	0.110	22.23	0.844	0.103	25.98	0.870	0.147	23.89	0.879	0.134
w/ camera pose optimization																			
CF3DGS + RobustSplat	1.12 / 108	21.46	0.703	0.230	21.09	0.706	0.204	24.78	0.838	0.154	21.66	0.821	0.151	24.64	0.801	0.172	22.20	0.763	0.190
LongSplat + RobustSplat	1.70 / 105	22.17	0.799	0.214	23.44	0.831	0.173	24.74	0.876	0.147	21.29	0.836	0.183	18.98	0.730	0.326	18.35	0.731	0.283
I3DGS + RobustSplat	2.03 / 107	21.05	0.748	0.258	22.37	0.789	0.211	24.03	0.863	0.157	19.81	0.806	0.207	17.42	0.696	0.350	17.71	0.701	0.307
PAGauss (Ours)	1.43 / 110	22.70	0.761	0.164	21.60	0.733	0.187	26.31	0.897	0.103	22.31	0.861	0.092	26.01	0.872	0.144	24.80	0.894	0.102

Table 19: Quantitative results on the NeRF On-the-go dataset.

Method	Train hrs / FPS	Brandenburg Gate			Sacre Coeur			Trevi-Fountain		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o camera pose optimization										
Ha-NeRF	452 / <1	24.04	0.877	0.139	20.02	0.801	0.171	20.18	0.690	0.222
3DGS	2.2 / 57	19.33	0.884	0.132	17.70	0.845	0.176	17.08	0.714	0.241
Mip-Splatting	0.18 / 82	20.46	0.890	0.126	18.14	0.879	0.168	17.52	0.747	0.218
Splatfacto-W	1.05 / 40.2	26.82	0.911	0.126	22.56	0.876	0.157	22.28	0.766	0.235
GS-W	1.2 / 51	23.51	0.897	0.166	19.39	0.825	0.211	20.06	0.723	0.271
WildGaussians	11.5 / 86	27.77	0.927	0.133	22.56	0.859	0.177	23.63	0.766	0.228
SWAG	0.8 / 15	26.33	0.929	0.139	21.16	0.860	0.185	23.10	0.815	0.208
NexusSplats	6.81 / 88	27.12	0.924	0.143	22.61	0.857	0.180	23.92	0.761	0.253
DroneSplat	1.8 / 97	19.88	0.871	0.112	16.85	0.808	0.143	16.71	0.661	0.276
RobustSplat	1.01 / 103	19.74	0.887	0.146	17.87	0.847	0.178	17.63	0.213	0.727
ForestSplats	7.9 / 122	28.13	0.935	0.118	23.84	0.876	0.123	23.11	0.802	0.212
AsymGS	20.31 / 110	28.25	0.937	0.115	23.35	0.882	0.150	23.83	0.775	0.242
DualSplat	16.92 / 109	28.05	0.934	0.116	23.64	0.875	0.143	23.36	0.801	0.215
PAGauss (Ours)	8.31 / 106	28.31	0.933	0.115	23.25	0.885	0.169	23.86	0.833	0.234
w/ camera pose optimization										
CF3DGS + RobustSplat	7.96 / 104	20.71	0.887	0.173	17.70	0.831	0.210	17.69	0.701	0.272
LongSplat + RobustSplat	20.31 / 110	16.94	0.746	0.305	15.04	0.680	0.340	15.32	0.675	0.369
I3DGS + RobustSplat	10.79 / 104	18.31	0.798	0.272	16.56	0.669	0.371	16.48	0.617	0.350
PAGauss (Ours)	8.31 / 106	28.33	0.941	0.120	23.46	0.889	0.147	24.03	0.810	0.211

Table 20: Quantitative results on the Photo Tourism dataset.

Method	Train hrs / FPS	SimingShan			Sculpture		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o camera pose optimization							
3DGS	1.1 / 59	18.14	0.749	0.317	16.11	0.396	0.286
Mip-Splatting	0.8 / 63	19.78	0.742	0.295	17.22	0.404	0.249
Ha-NeRF	32 / < 1	22.26	0.767	0.298	18.35	0.412	0.274
Splatfacto-W	0.6 / 71	22.13	0.715	0.268	17.49	0.438	0.289
GS-W	0.6 / 104	22.41	0.706	0.243	17.34	0.414	0.271
WildGaussians	1.3 / 94	22.63	0.707	0.239	17.15	0.371	0.383
DroneSplat	1.5 / 104	23.16	0.759	0.152	18.48	0.489	0.225
RobustSplat	0.57 / 109	23.20	0.703	0.476	19.14	0.447	0.461
AsymGS	2.79 / 103	23.40	0.763	0.289	19.33	0.473	0.450
DualSplat	2.39 / 108	23.53	0.778	0.219	19.62	0.514	0.301
PAGauss (Ours)	0.70 / 102	23.71	0.783	0.187	19.85	0.589	0.259
w/ camera pose optimization							
CF3DGS + RobustSplat	0.70 / 105	23.66	0.800	0.211	18.60	0.514	0.283
LongSplat + RobustSplat	2.79 / 103	22.42	0.655	0.238	15.97	0.241	0.345
I3DGS + RobustSplat	2.39 / 108	22.78	0.734	0.280	18.89	0.456	0.419
PAGauss (Ours)	0.72 / 110	24.08	0.801	0.173	19.87	0.595	0.228

Table 21: Quantitative results on the Drone Imagery dataset.

Method	Activist	Gazebo	Lagoon	Miniature	Pagoda	Terrace
WildGaussians	207.24	232.56	180.88	60.32	42.10	265.92
RobustSplat	227.13	204.83	227.47	77.68	39.29	279.62
CF3DGS + RobustSplat	169.46	150.63	167.39	41.50	29.77	29.38
LongSplat + RobustSplat	184.97	154.84	190.52	58.75	30.64	215.34
PAGauss (Ours)	221.09	188.23	229.12	94.42	57.73	275.64

Table 22: Comparison of the memory usage on the WildArena dataset.

Method	Humanoid	Porter	Lounge	Bust	Jockey	Statuette
3DGS	413.04	447.08	247.98	472.09	396.69	370.94
Mip-Splatting	213.16	521.56	203.82	465.87	412.77	357.62
DroneSplat	110.01	389.84	114.13	294.58	288.72	266.31
RobustSplat	51.32	347.45	86.68	226.19	214.27	211.27
CF3DGS + RobustSplat	169.05	361.29	138.44	321.24	317.62	65.44
LongSplat + RobustSplat	53.25	341.76	85.51	233.98	239.26	222.62
PAGauss (Ours)	56.39	327.54	97.48	233.98	250.48	224.22

Table 23: Comparison of the memory usage on the Ambiguous Wild dataset.

Method	Mountain	Fountain	Corner	Patio	Spot	Patio-High
3DGS	447.08	413.04	396.69	247.98	277.27	103.08
Mip-Splatting	521.56	213.16	412.77	203.82	194.96	54.24
DroneSplat	389.84	110.01	288.72	114.13	79.24	22.12
RobustSplat	347.45	51.32	214.27	86.68	64.39	62.03
DualSplat	335.51	486.28	181.17	151.10	155.65	309.20
RefineSplat	306.91	46.85	218.63	86.61	25.80	20.31
CF3DGS + RobustSplat	299.52	459.37	180.58	75.07	142.64	286.54
LongSplat + RobustSplat	286.71	480.95	172.87	125.71	141.28	253.00
PAGauss (Ours)	304.05	458.22	184.50	75.50	142.68	286.32

Table 24: Comparison of the memory usage on the NeRF On-the-go dataset.



Figure 26: Qualitative results on the WildArena dataset.

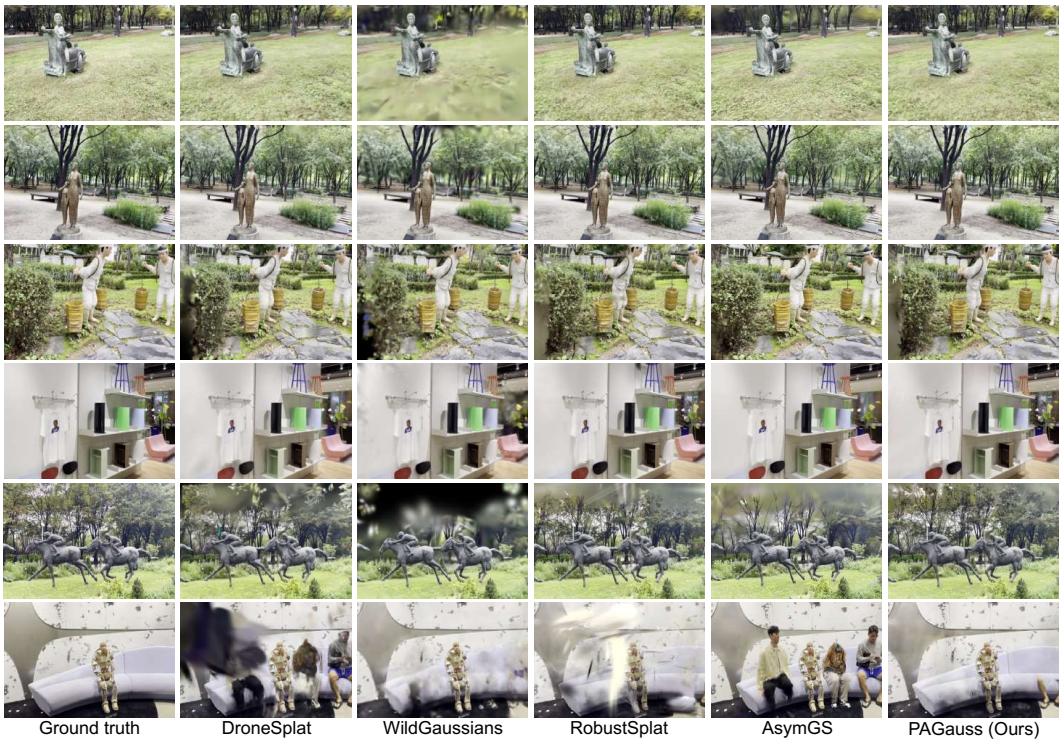


Figure 27: Qualitative results on the Ambiguous Wild dataset.

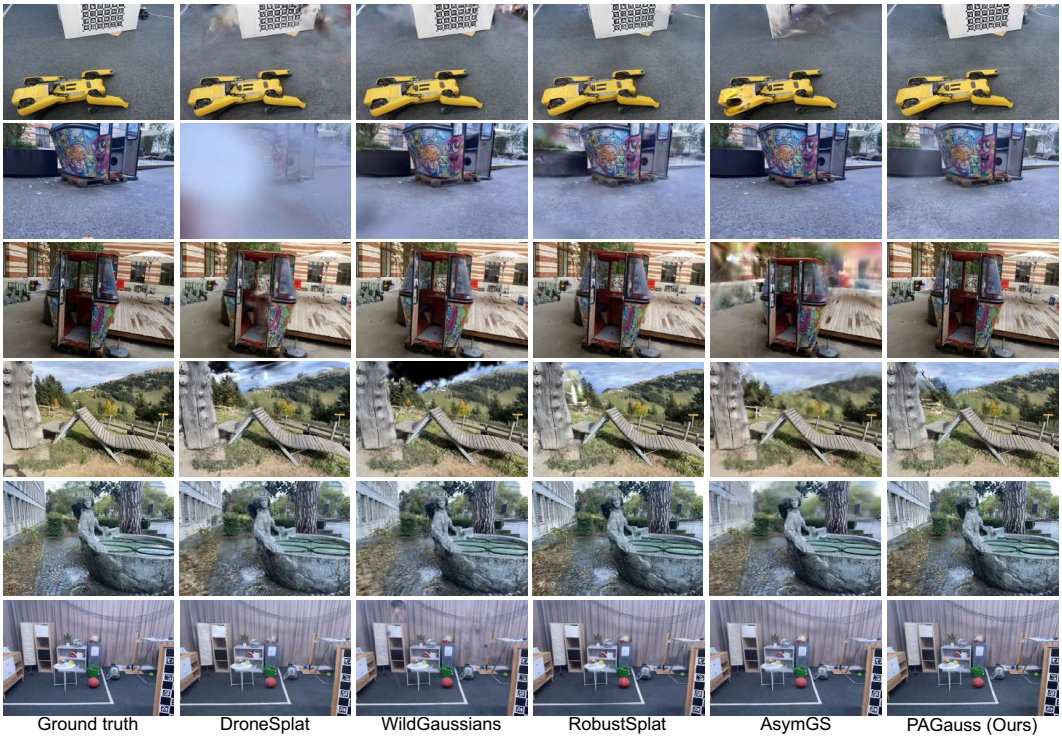


Figure 28: Qualitative results on the NeRF On-the-go dataset.



Figure 29: Qualitative results on the Photo Tourism dataset.

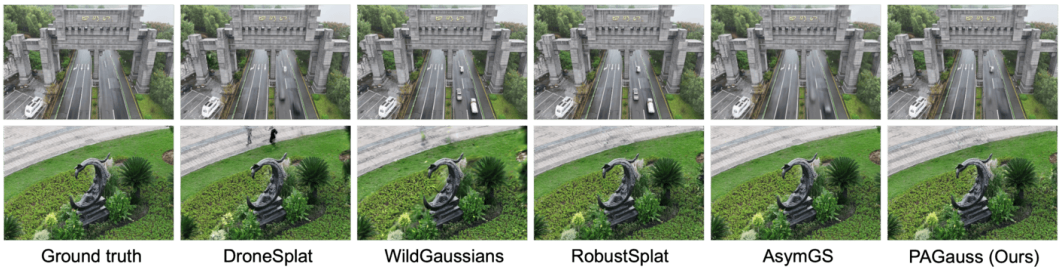


Figure 30: Qualitative results on the Drone Imagery dataset.



Figure 31: Additional sample images on the WildArena dataset.

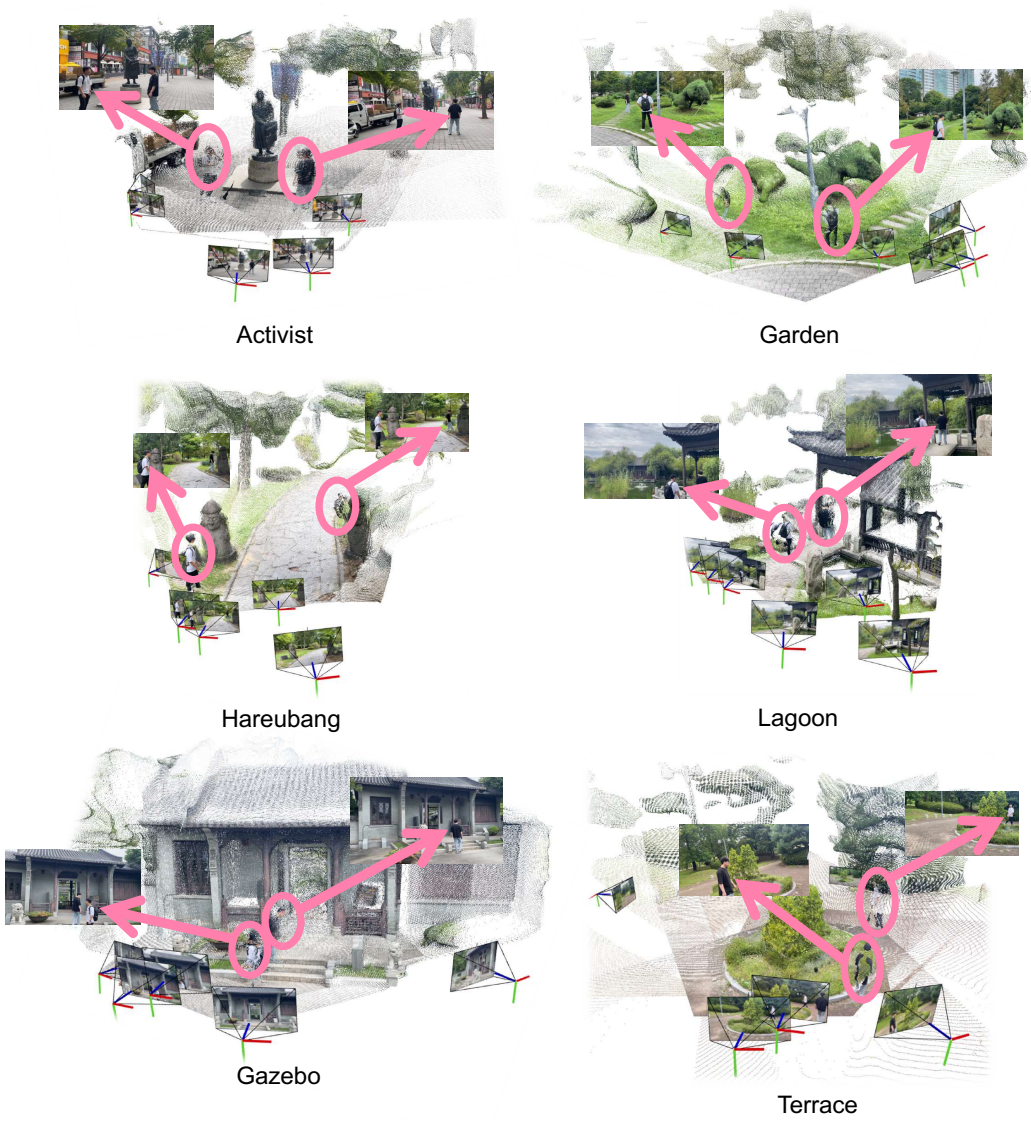


Figure 32: 3D Gaussians initialization with inaccurate 3D point clouds due to distractors.

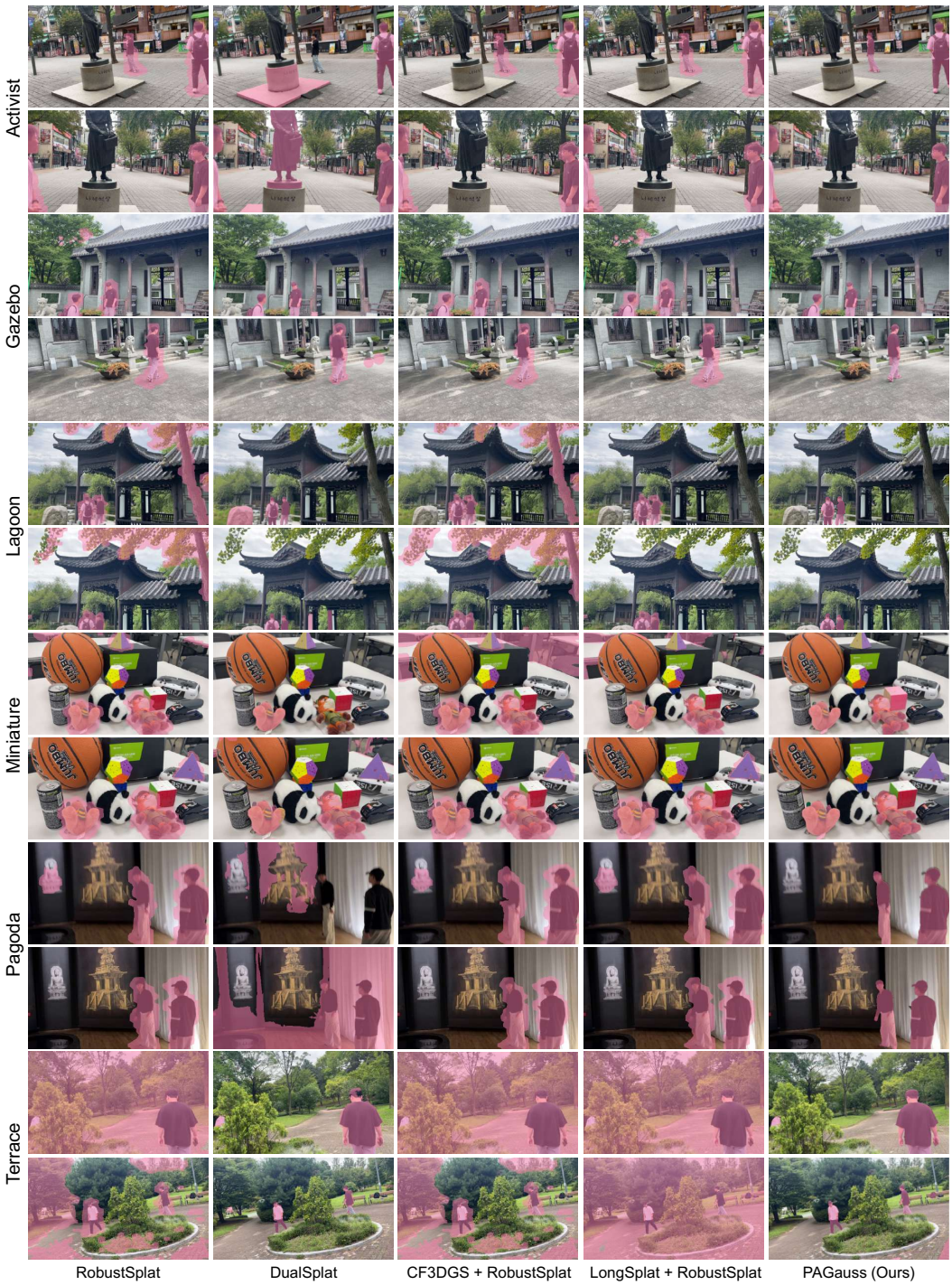


Figure 33: Comparison of transient masks in the WildArena dataset.

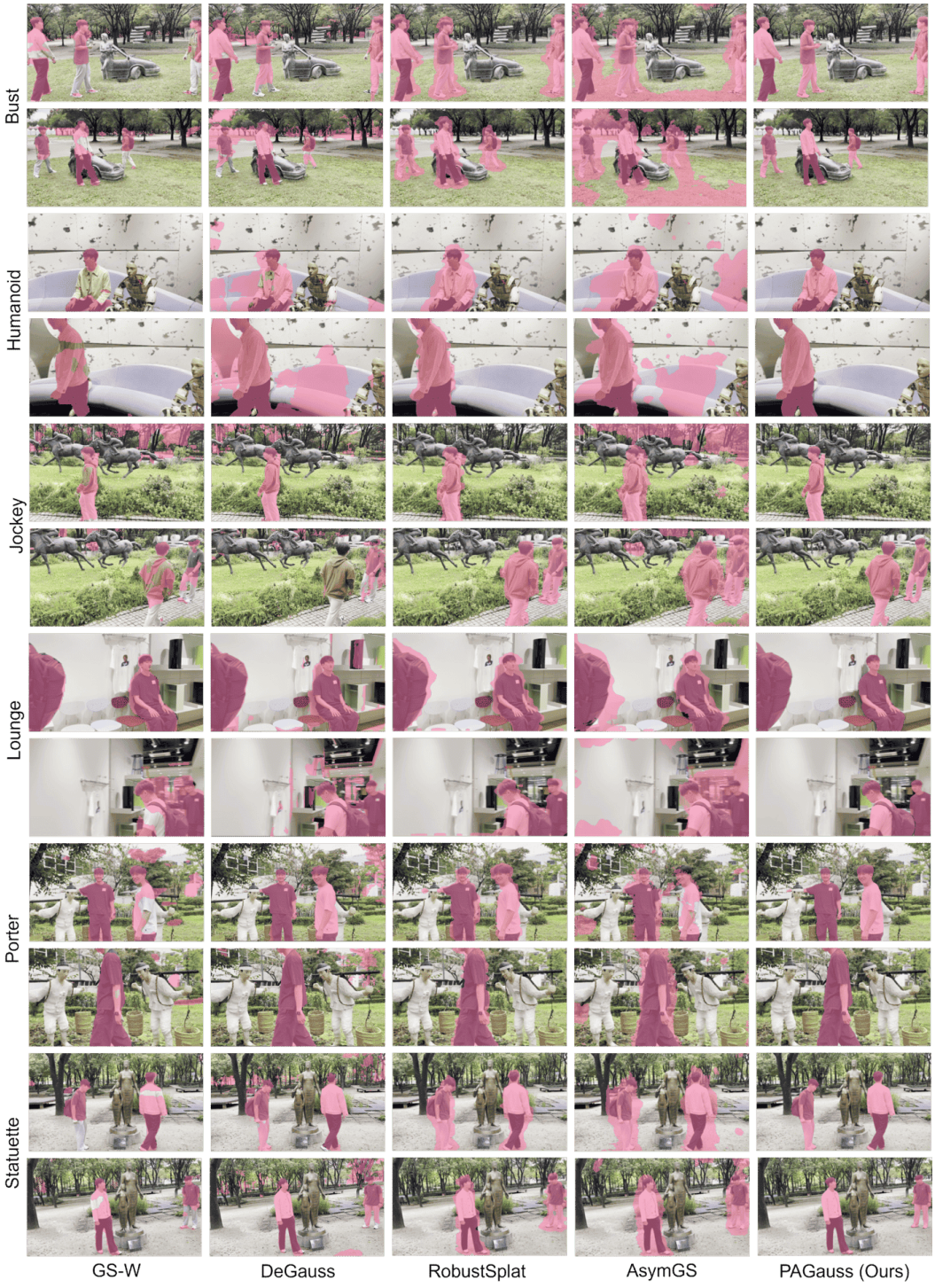


Figure 34: Comparison of transient masks in the Ambiguous Wild dataset.

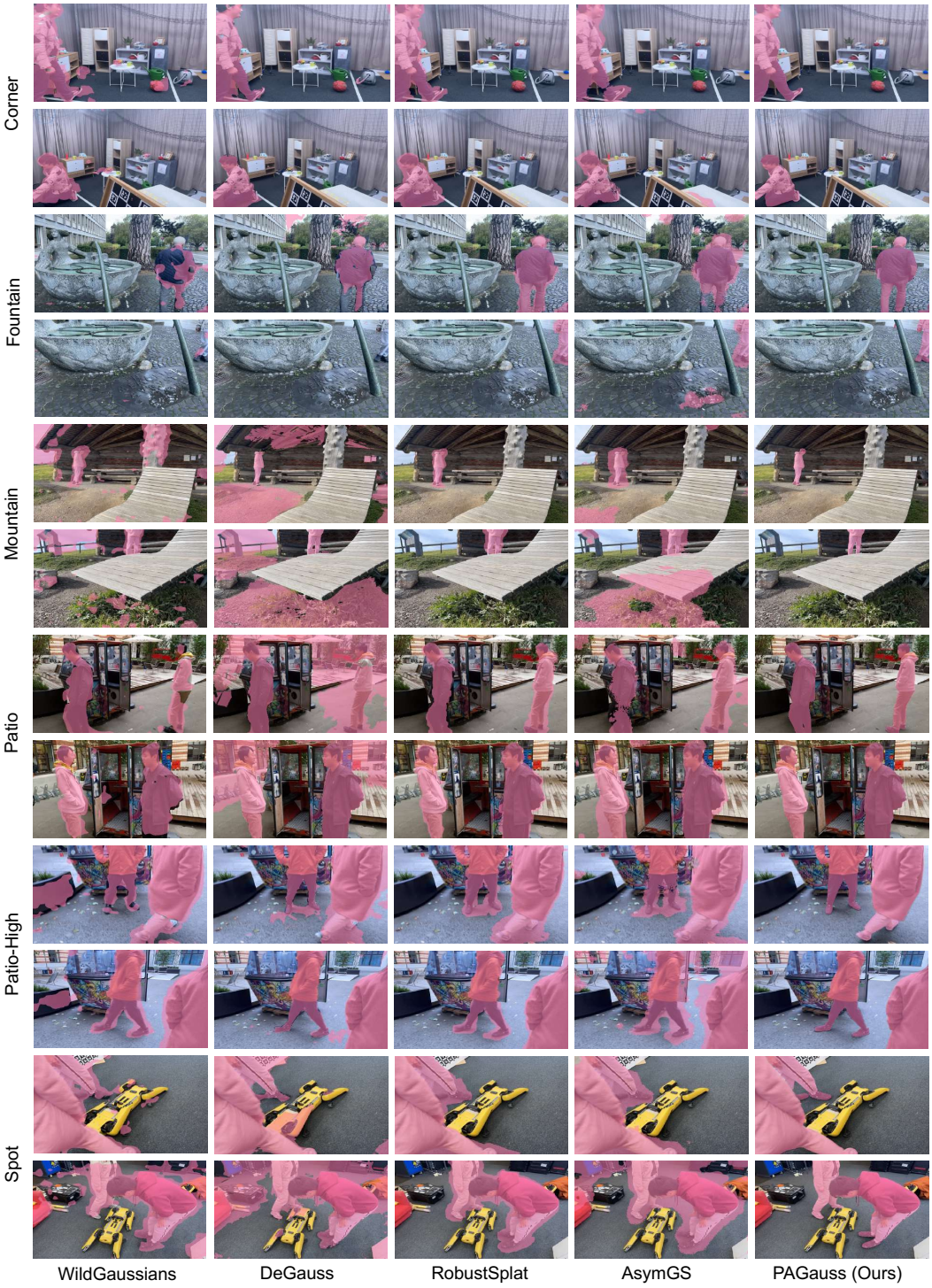


Figure 35: Comparison of transient masks in the NeRF On-the-go dataset.

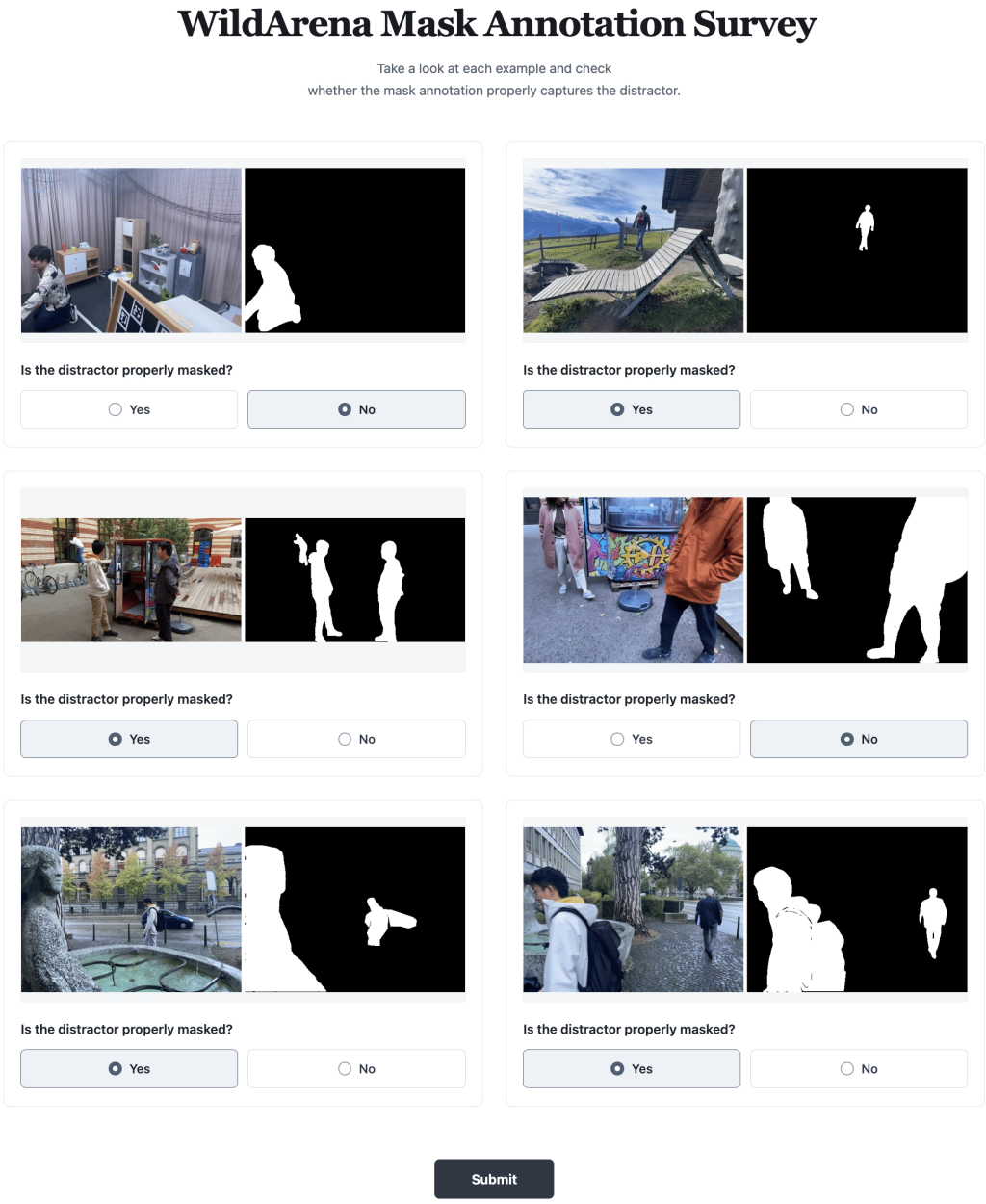


Figure 36: WildArena mask annotation survey. Each participant was shown an upper figure, which presents several distractors alongside their corresponding mask annotations. Participants were asked to verify whether each mask annotation correctly captures distractors.

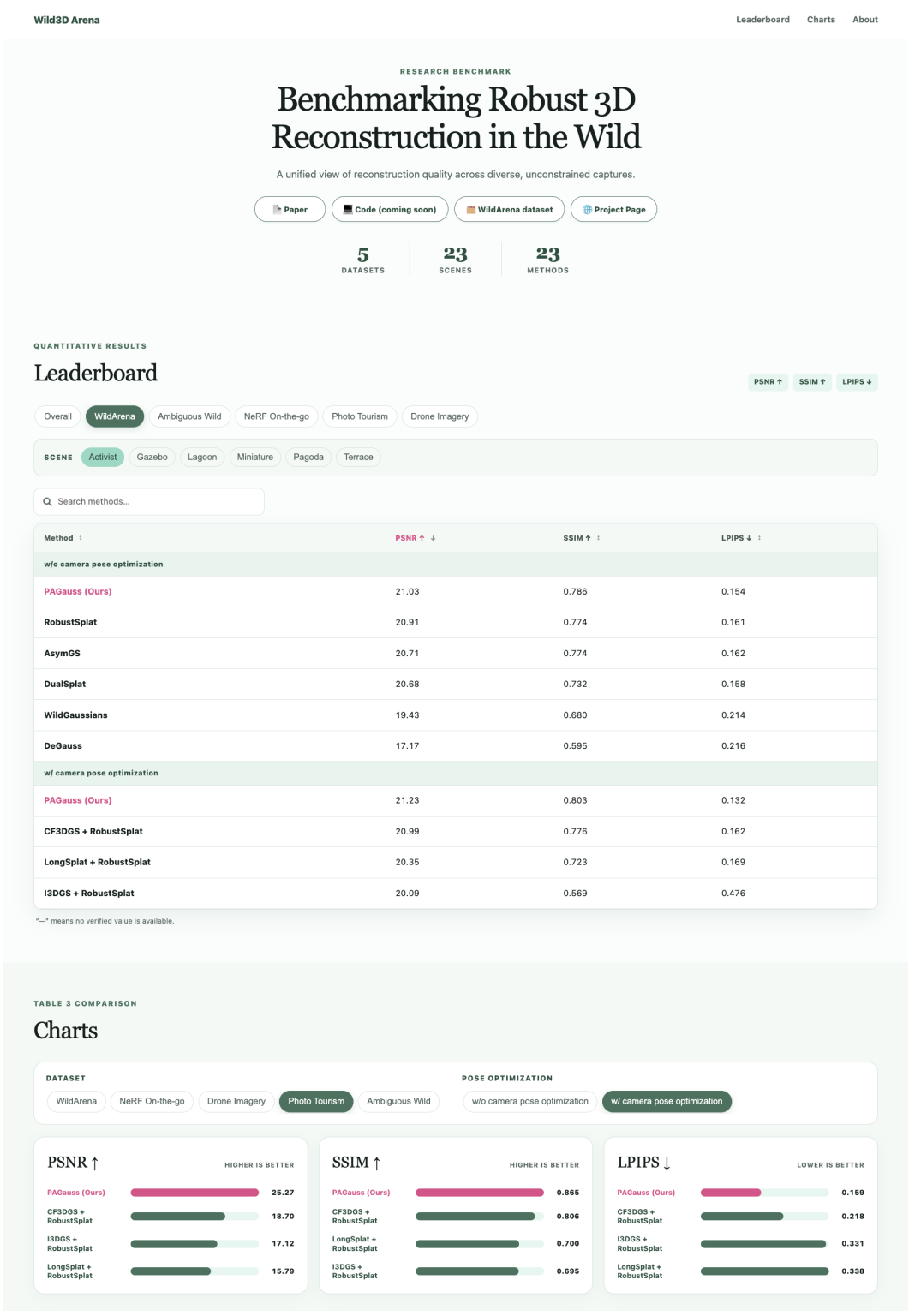


Figure 37: WildArena leaderboard for fair and reproducible evaluation of 3D reconstruction methods in real-world scenarios.