HARMOGS: ROBUST 3D GAUSSIAN SPLATTING IN THE WILD VIA CONFLICT-AWARE GRADIENT HARMONIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

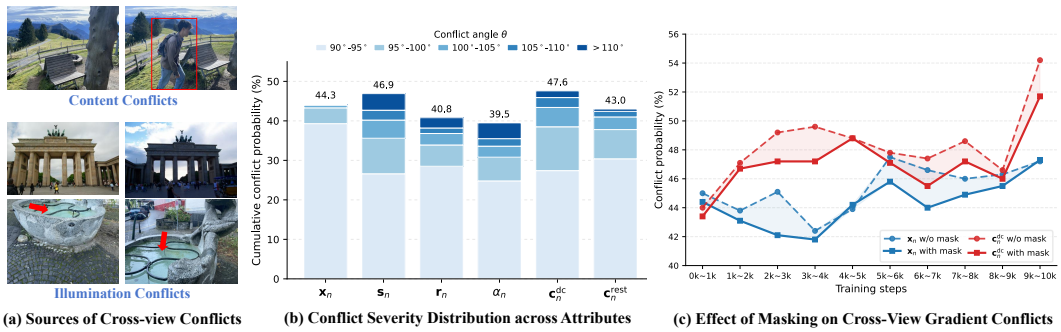
In-the-wild 3D Gaussian Splatting remains challenging due to transient distractors and illumination-induced cross-view appearance inconsistencies. Existing methods mainly rely on image-level masking to suppress unreliable supervision, but masking alone cannot fully eliminate residual occlusions or resolve illumination-induced inconsistencies, both of which can introduce conflicting cross-view gradients. These unresolved conflicts may destabilize Gaussian optimization and lead to visible reconstruction artifacts. We propose a conflict-aware 3DGS framework that addresses this problem from both image-space supervision and gradient-level optimization. Semantic Consistency-Guided Masking learns pixel-wise consistency scores to adaptively refine prior masks and suppress unreliable supervision before gradient formation. A dual-view Conflict-Aware Gradient Harmonization strategy further reconciles view-specific gradients by mutually rotating them into an orthogonal configuration, reducing negative directional interference across views. We also introduce conflict-aware densification and pruning to stabilize Gaussian growth and suppress persistently conflicting primitives. Extensive experiments on standard in-the-wild benchmarks demonstrate that our method achieves state-of-the-art rendering quality under complex transient distractors and cross-view inconsistencies.

1 INTRODUCTION

Reconstructing a reliable 3D representation from a collection of images is a fundamental problem in computer vision, with broad applications in VR/AR, film production, and game development. Recent advances, including Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), have achieved remarkable success in photorealistic scene modeling. However, these methods typically assume that training images are captured under consistent illumination and contain stable scene content, which rarely holds in unconstrained real-world environments.

In in-the-wild scenes, captured images are often affected by transient distractors and appearance variations, such as moving pedestrians, vehicles, and illumination changes. Existing methods mainly address this issue by suppressing unreliable image-level supervision. Some approaches (Zhang et al., 2024; Ren et al., 2024; Ungermann et al., 2024; Sabour et al., 2023) learn to down-weight transient regions in a self-supervised manner, but they often struggle with large or semantically complex occlusions. Recent methods (Tang et al., 2025; Li et al., 2025) further use pretrained segmentation models to generate object-level mask candidates, and then identify dynamic distractors with hand-crafted criteria, such as reconstruction residual thresholds or temporal tracking. However, these rule-based pipelines remain sensitive to segmentation quality and heuristic design, limiting their robustness in complex real-world scenes. To address these limitations, we introduce Semantic Consistency-Guided Masking, which learns a pixel-wise consistency score during training and uses it to adaptively refine a precomputed SAM-based (Kirillov et al., 2023) binary prior mask. This design combines adaptive reliability estimation with object-level semantic priors, making the mask more robust to complex transient occlusions and initial segmentation errors. By suppressing unreliable image evidence before gradient formation, it reduces noisy gradients in Gaussian optimization.

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.



(a) Sources of Cross-view Conflicts (b) Conflict Severity Distribution across Attributes (c) Effect of Masking on Cross-View Gradient Conflicts

Figure 1: (a) In-the-wild image collections often contain content conflicts caused by transient objects and illumination conflicts across views, which introduce unreliable supervision. (b) When two views are jointly used to optimize a shared Gaussian representation, their induced gradients often conflict across different Gaussian attributes. (c) Applying masks to unreliable image regions only marginally reduces the conflict probability during training, while a substantial level of cross-view conflict still persists. Statistics are reported on the NeRF On-the-go dataset (Ren et al., 2024).

Although this masking strategy alleviates noisy supervision, image-space masking alone cannot fully resolve optimization conflicts in in-the-wild 3DGS. As shown in Figure 1(b) and (c), cross-view gradient conflicts are common during Gaussian optimization and remain significant even after unreliable regions are masked. This is because masks may fail to fully exclude complex occlusions, while illumination inconsistencies introduce view-dependent supervision that cannot be handled by masking. Consequently, inconsistent cross-view supervision can still produce incompatible Gaussian updates, potentially destabilizing optimization and leading to visible reconstruction artifacts. Motivated by this observation, we introduce a dual-view 3DGS training framework with Conflict-Aware Gradient Harmonization. At each iteration, two training views are randomly sampled to compute view-specific gradients for Gaussian attributes. We then harmonize the two gradients by mutually rotating their directions into an orthogonal configuration. This explicitly eliminates negative directional interference between the two views while largely preserving their original descent directions, yielding a more compatible update for Gaussian optimization.

Building on this framework, we further introduce conflict-aware densification and pruning. Our densification strategy aggregates harmonization-aware view-space gradients from both views to guide Gaussian growth with more reliable two-view signals. The pruning strategy tracks per-Gaussian cross-view conflicts and periodically decays the opacity of highly conflicting primitives, allowing persistently unstable Gaussians to be naturally removed by the standard pruning process.

Our contributions are summarized as follows: 1) We propose a conflict-aware gradient harmonization framework for robust in-the-wild 3D Gaussian Splatting. By explicitly reconciling cross-view gradient conflicts, our method reduces incompatible Gaussian updates caused by transient distractors and illumination-induced inconsistencies. 2) We introduce Semantic Consistency-Guided Masking, which adaptively refines the prior mask and effectively suppresses transient distractors. We further develop conflict-aware densification and pruning strategies to stabilize Gaussian optimization under dual-view training, leading to more reliable Gaussian growth and pruning. 3) Extensive experiments on standard in-the-wild benchmarks demonstrate that our method effectively handles complex transient distractors and cross-view inconsistencies, achieving state-of-the-art rendering quality and reconstruction fidelity.

2 Related Works

Novel view synthesis. Synthesizing photorealistic images from novel viewpoints remains a fundamental yet challenging problem in machine learning and graphics. A seminal work, Neural Radiance Fields (NeRF) (Mildenhall et al., 2021), represents scenes implicitly by parameterizing color and density with multilayer perceptrons (MLPs), achieving remarkable rendering quality. Building upon this paradigm, subsequent efforts have primarily focused on improving efficiency, either through compact scene representations (Chen et al., 2022a; Sun et al., 2022; Hu et al., 2022; Müller et al., 2022; Fridovich-Keil et al., 2022; Takikawa et al., 2021; Xu et al., 2022) or via network compression techniques (Gordon et al., 2023; Reiser et al., 2021). More recently, 3D Gaussian Splatting (3DGS)

(Kerbl et al., 2023) has emerged as a powerful explicit representation for novel view synthesis, offering real-time rendering while maintaining high reconstruction quality. This has motivated a series of follow-up studies that improve 3DGS in terms of rendering fidelity, optimization robustness, and scalability to complex scenes (Guédon & Lepetit, 2024; Liu et al., 2025; Lin et al., 2024).

In-the-Wild Novel View Synthesis. Novel view synthesis in the wild is challenging due to transient distractors and lighting variations across views. NeRF-W (Martin-Brualla et al., 2021) extended NeRF to handle unconstrained internet image collections, learning image-specific appearance embeddings to model lighting variations and using a transient field to address dynamic occlusions. Subsequent works (Yang et al., 2023; Chen et al., 2024; 2022b; Ren et al., 2024) built on this, enhancing modeling performance and transient object removal. Ha-NeRF (Chen et al., 2022b) replaced the 3D transient volume with 2D visibility maps, using CNN-based global appearance encoding to improve appearance modeling. NeRF On-the-go (Ren et al., 2024) applied uncertainty-guided transient filtering using DINOv2 features, enabling transient removal on video sequences without semantic priors. RobustNeRF (Sabour et al., 2023) treated dynamic interference as outliers in a robust optimization framework, using iteratively reweighted least-squares (IRLS) to down-weight transient pixels. Despite these advances, these methods still suffer from NeRF's inherent issues, such as long training times and slow rendering. 3DGS-based methods (Zhang et al., 2024; Kulhanek et al., 2024; Li et al., 2025) provide an effective solution to these limitations. Wild-GS (Xu et al., 2024) aligns appearance information from reference images to 3D Gaussians via triplane cropping, modeling high-frequency details while maintaining 3DGS's real-time performance. To better handle transient interference, SpotLessSplats (Sabour et al., 2025) combines semantic segmentation features with iterative reconstruction residuals to adaptively identify transient regions. HybridGS (Lin et al., 2025) introduces a dual-branch architecture for static and transient elements, improving the distinction between transient objects and static backgrounds. Robust3DGS (Ungermann et al., 2024) leverages the IRLS technique from RobustNeRF (Sabour et al., 2023), further incorporating image residual clustering for self-supervised transient masking.

3 METHOD

3.1 PRELIMINARIES

3D Gaussian Splatting (Kerbl et al., 2023) represents a scene with a set of anisotropic Gaussian primitives. Each Gaussian is parameterized by learnable attributes $\Theta_n = \{\mathbf{x}_n, \mathbf{s}_n, \mathbf{r}_n, \alpha_n, \mathbf{c}_n\}$, where $\mathbf{x}_n \in \mathbb{R}^3$ is the center position, $\mathbf{s}_n \in \mathbb{R}^3$ is the scale, $\mathbf{r}_n \in \mathbb{R}^4$ is the rotation quaternion, $\alpha_n \in \mathbb{R}$ is the opacity, and $\mathbf{c}_n$ denotes the spherical-harmonic color coefficients. Given a camera view, Gaussians are projected onto the image plane and rendered by alpha compositing:

$$\mathbf{C} = \sum_{n=1}^{N} \mathbf{c}_n w_n \prod_{m=1}^{n-1} (1 - w_m), \tag{1}$$

where w_n denotes the per-pixel contribution weight of the n-th projected Gaussian, determined by its learnable opacity and its projected 2D Gaussian response at the pixel.

3.2 JOINT OPTIMIZATION OF CROSS-VIEW GAUSSIAN PARAMETERS

Our key insight is that in-the-wild training images often contain cross-view inconsistencies, including illumination conflicts across views and content conflicts caused by transient distractors. When optimizing the model with such inconsistent observations, the gradients derived from different views may point to conflicting update directions, leading to unstable optimization and reconstruction artifacts. To mitigate this, we jointly consider two different training views at each iteration and introduce a gradient harmonization strategy, detailed in Section 3.3, to reconcile their update directions.

Specifically, at each training iteration, we independently sample two views $\mathbf{V}_1$ and $\mathbf{V}_2$ with their corresponding ground-truth images $\mathbf{I}_1$ and $\mathbf{I}_2$. Let $\tilde{\mathbf{I}}_i$ denote the rendered image for view $i \in \{1, 2\}$, and $\mathbf{M}_i$ be the view-specific mask designed to filter out transient distractors, which will be detailed in Section 3.4. The masked images are given by $\mathbf{I}_i^{\mathbf{M}} = \mathbf{M}_i \odot \mathbf{I}_i$ and $\tilde{\mathbf{I}}_i^{\mathbf{M}} = \mathbf{M}_i \odot \tilde{\mathbf{I}}_i$. The reconstruction objective for each view is formulated as:

$$\mathcal{L}_i = (1 - \lambda_{rec}) \|\tilde{\mathbf{I}}_i^{\mathbf{M}} - \mathbf{I}_i^{\mathbf{M}}\|_1 + \lambda_{rec} \, \mathrm{DSSIM}(\tilde{\mathbf{I}}_i^{\mathbf{M}}, \mathbf{I}_i^{\mathbf{M}}), \quad i \in \{1, 2\}, \tag{2}$$

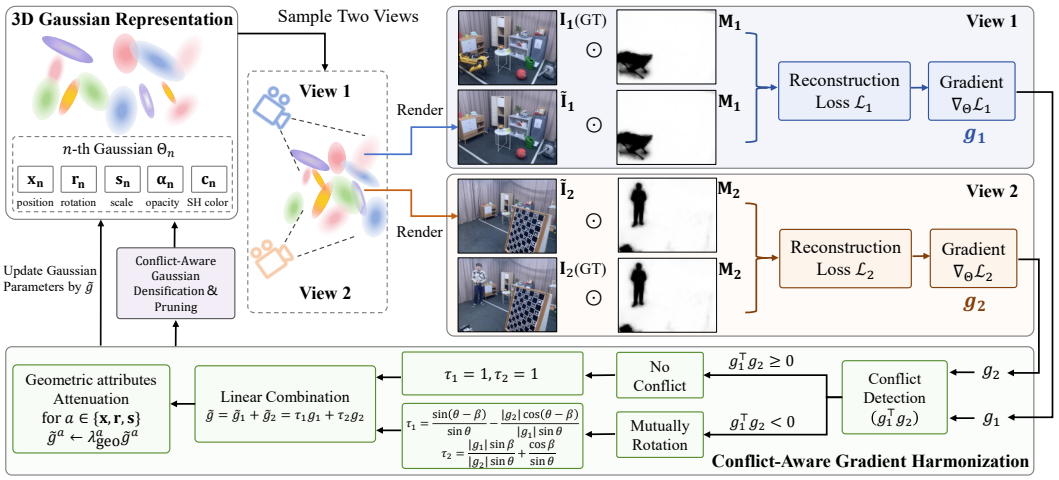


Figure 2: Overview of the proposed framework. At each training iteration, two views are sampled and rendered from the current 3D Gaussians. Semantic Consistency-Guided Masking generates view-specific masks to suppress transient regions, and the masked reconstruction losses $\mathcal{L}_1$ and $\mathcal{L}_2$ produce the corresponding gradients g_1 and g_2. When a cross-view conflict is detected, Conflict-Aware Gradient Harmonization reconciles the two gradients by mutually rotating them toward orthogonal directions, which is implemented by computing conflict-aware combination coefficients. The harmonized gradients from the two views are then summed to update the Gaussian parameters.

where λ_{rec} is a weighting hyperparameter, DSSIM is the structural dissimilarity index measure (Wang et al., 2004).

3.3 Conflict-Aware Gradient Harmonization

Let Θ denote the set of optimizable Gaussian parameters. Given the two view-specific losses $\mathcal{L}_1$ and $\mathcal{L}_2$, we compute their corresponding gradients as:

$$g_1 = \nabla_\Theta \mathcal{L}_1, \quad g_2 = \nabla_\Theta \mathcal{L}_2. \tag{3}$$

A gradient conflict arises when the two gradients form an obtuse angle, i.e., $g_1^\top g_2 < 0$. To mitigate such interference, we harmonize g_1 and g_2 by mutually rotating them within the subspace $\mathcal{V} = \mathrm{span}(g_1, g_2)$ until the orthogonality constraint $\tilde{g}_1^\top \tilde{g}_2 = 0$ is satisfied, as shown in Figure 3. Given the initial angle $\theta = \arccos\left(\frac{g_1^\top g_2}{\|g_1\|\|g_2\|}\right) \in (\frac{\pi}{2}, \pi)$, the total angular correction required to reach orthogonality is $\theta - \frac{\pi}{2}$. We introduce an allocation parameter $\rho \in [0, 1]$ to distribute this correction between the two gradients. Specifically, g_1 is rotated by $\beta = \rho(\theta - \frac{\pi}{2})$, while g_2 is rotated by $\bar{\beta} = (1 - \rho)(\theta - \frac{\pi}{2})$. To describe this rotation, we construct an orthonormal basis $\{u_1, u_2\}$ for $\mathcal{V}$, where u_1 is aligned with the direction of g_1:

$$u_1 = \frac{g_1}{\|g_1\|}, \quad u_2 = \frac{g_2 - (g_2^\top u_1)u_1}{\|g_2 - (g_2^\top u_1)u_1\|}. \tag{4}$$

Given that $g_2^\top u_1 = \|g_2\| \cos\theta$ and the magnitude of the orthogonal component is $\|g_2\| \sin\theta$, the expression can be written as:

$$u_2 = \frac{g_2 - \|g_2\| \cos\theta \cdot u_1}{\|g_2\| \sin\theta}. \tag{5}$$

Under this basis, the original gradients can be represented as $g_1 = \|g_1\| u_1$ and $g_2 = \|g_2\|(\cos\theta \cdot u_1 + \sin\theta \cdot u_2)$. We then rotate the two gradients while preserving their original magnitudes, yielding the harmonized gradients:

$$\tilde{g}_1 = \|g_1\|(\cos\beta \cdot u_1 + \sin\beta \cdot u_2), \quad \tilde{g}_2 = \|g_2\|(-\sin\beta \cdot u_1 + \cos\beta \cdot u_2). \tag{6}$$

Since both harmonized gradients remain in the subspace $\mathcal{V} = \mathrm{span}(g_1, g_2)$, their sum can be expressed as a linear combination of the original gradients:

$$\tilde{g} = \tilde{g}_1 + \tilde{g}_2 = \tau_1 g_1 + \tau_2 g_2. \tag{7}$$

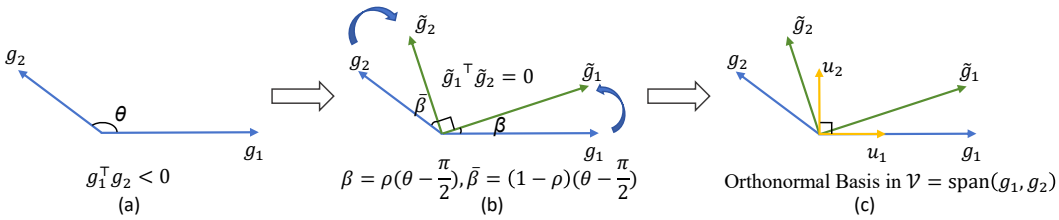


Figure 3: Illustration of the Cross-View Conflict-Aware Gradient Harmonization process. (a) A gradient conflict is detected when the angle θ between g_1 and g_2 is obtuse ($g_1^\top g_2 < 0$). (b) The gradients are mutually rotated within the subspace $\mathcal{V}$ by angles β and $\bar{\beta}$ respectively, until they satisfy the orthogonality constraint $\tilde{g}_1^\top \tilde{g}_2 = 0$. (c) Construction of the orthonormal basis $\{u_1, u_2\}$ to facilitate the analytical decomposition and recombination of gradients within the 2D plane.

By substituting the expressions of $\tilde{g}_1$ and $\tilde{g}_2$, the coefficients are derived as:

$$\tau_1 = \frac{\sin(\theta-\beta)}{\sin\theta} - \frac{\|g_2\|\cos(\theta-\beta)}{\|g_1\|\sin\theta}, \quad \tau_2 = \frac{\|g_1\|\sin\beta}{\|g_2\|\sin\theta} + \frac{\cos\beta}{\sin\theta}. \tag{8}$$

When no conflict exists, i.e., $g_1^\top g_2 \geq 0$, we set $\tau_1 = \tau_2 = 1$. In practice, we apply the above harmonization independently to each Gaussian attribute. For an attribute a, we use its view-specific gradients $g_{1,a} = \nabla_a \mathcal{L}_1$ and $g_{2,a} = \nabla_a \mathcal{L}_2$ to compute the corresponding coefficients τ_1^a and τ_2^a, where τ_i^a weights the gradient of attribute a from view i. For geometric attributes $a \in \{\mathbf{x}, \mathbf{r}, \mathbf{s}\}$, since geometry-level conflicts are more likely to induce reconstruction artifacts, we further introduce an attenuation factor when such conflicts are detected:

$$\lambda_{\text{geo}}^a = \exp\left(-k \cdot \max(0, -\cos\theta_a)\right), \tag{9}$$

where θ_a is the angle between $g_{1,a}$ and $g_{2,a}$, and k controls the attenuation strength. The final harmonized gradient for geometric attributes is then updated by $\tilde{g}^a \leftarrow \lambda_{\text{geo}}^a \tilde{g}^a$.

3.4 Semantic Consistency-Guided Masking

Given a ground-truth image $\mathbf{I}_i$, we use a frozen DINOv2 (Oquab et al., 2023) backbone to extract intermediate features $\mathbf{F}_i \in \mathbb{R}^{C \times H' \times W'}$. A lightweight predictor f_ψ maps these semantic priors to a pixel-wise reliability scale σ_i, where a larger value indicates lower supervision reliability:

$$\sigma_i = \text{softplus}(f_\psi(\mathbf{F}_i) + \delta_0), \tag{10}$$

where $\delta_0 = \ln(e-1)$ is an initialization offset ensuring σ_i starts near unity. The predicted σ_i is defined on the DINOv2 feature grid. To guide the learning of σ_i, we construct a semantic consistency-aware residual target from the rendered image $\tilde{\mathbf{I}}_i$ and the ground-truth image $\mathbf{I}_i$:

$$\mathbf{E}_i = \min\left(1, \frac{d_{\cos}(\Phi_{\tilde{I}_i}, \Phi_{I_i})}{s_{\text{sem}}}\right) \odot \left\| \mathcal{D}(\tilde{\mathbf{I}}_i) - \mathcal{D}(\mathbf{I}_i) \right\|_1, \tag{11}$$

where $\Phi_{\tilde{I}_i}$ and Φ_{I_i} are DINOv2 features of the rendered and ground-truth images, respectively, and $\mathcal{D}(\cdot)$ denotes downsampling to the DINOv2 feature resolution. The term $d_{\cos}(\cdot,\cdot) = 1 - \frac{\Phi_{\tilde{I}_i} \cdot \Phi_{I_i}}{\|\Phi_{\tilde{I}_i}\| \|\Phi_{I_i}\|}$ computes the cosine distance at each spatial location in the feature space. We set $s_{\text{sem}} = 0.5$. This formulation increases the response of $\mathbf{E}_i$ only when both the appearance residual and semantic discrepancy are significant, thereby providing a more reliable optimization target. Based on this, the predictor f_ψ is optimized by:

$$\mathcal{L}_{\text{inc}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left(\frac{\mathbf{E}_i(p)}{2\sigma_i^2(p) + \epsilon} + \lambda_{\text{inc}} \log(\sigma_i(p) + \epsilon) \right), \tag{12}$$

where Ω denotes the spatial domain of the feature map and λ_{inc} controls the regularization strength. In practice, $\mathbf{E}_i$ is treated as a detached target when optimizing $\mathcal{L}_{\text{inc}}$, so this loss only updates the predictor f_ψ and does not propagate gradients to the Gaussian parameters. This formulation encourages the predictor to assign larger σ_i to regions characterized by both high photometric residuals and semantic

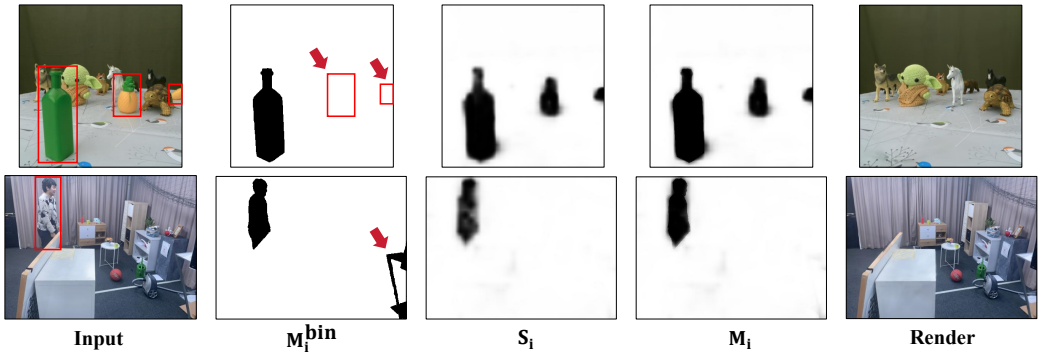


Figure 4: Visualization of Semantic Consistency-Guided Masking. The SAM-based binary mask $\mathbf{M}_i^{\text{bin}}$ provides a coarse region-level reliability prior, but may miss transient regions or incorrectly suppress reliable areas. The learned consistency score $\mathbf{S}_i$ captures pixel-wise inconsistency during training and adaptively refines this binary prior. As shown in the highlighted regions, the final mask $\mathbf{M}_i$ suppresses missed transient distractors while correcting erroneous binary-mask decisions, leading to more robust rendering results.

discrepancies. At each training iteration, σ_i is first bilinearly upsampled to the image resolution and then converted into a consistency score $\mathbf{S}_i$:

$$\mathbf{S}_i = \exp\left(-\frac{\sigma_i^2}{c_\sigma}\right), \tag{13}$$

where c_σ is a scaling hyperparameter. We further construct a SAM-based (Kirillov et al., 2023; Li et al., 2024) binary mask $\mathbf{M}_i^{\text{bin}}$ following AsymGS (Li et al., 2025), where $\mathbf{M}_i^{\text{bin}} = 1$ indicates stable regions and $\mathbf{M}_i^{\text{bin}} = 0$ indicates potential transient regions. The learned consistency score $\mathbf{S}_i$ is then combined with this prior to obtain the final reconstruction mask used in Eq. 2:

$$\mathbf{M}_i = \mathbf{M}_i^{\text{bin}} \odot (\mathbf{S}_i)^{\eta_{\text{s}}} + (1 - \mathbf{M}_i^{\text{bin}}) \odot (\mathbf{S}_i)^{\eta_{\text{t}}}, \tag{14}$$

where η_{s} and η_{t} control the correction intensity for stable and transient regions, respectively. When $\mathbf{M}_i$ is used in the reconstruction objective, it is detached from the predictor branch, so the reconstruction loss updates the Gaussian parameters without affecting f_ψ.

3.5 Conflict-Aware Gaussian Densification and Pruning

Conflict-aware joint densification. Under our dual-view optimization setting, the single-view densification statistics used in standard 3DGS are insufficient for reliable Gaussian growth. Moreover, directly accumulating the magnitudes of raw view-space gradients may cause conflicting yet individually large signals to trigger unnecessary densification. We therefore aggregate densification signals from both views, and modulate their view-space gradients using the position-specific conflict-aware coefficients from Section 3.3. Let $\mathcal{V}_i$, $r^{(i)}$, and $\nabla_{\mathbf{p}_{\text{view}}^{(i)}} \mathcal{L}_i$ denote the visible Gaussian set, projected 2D radius, and raw view-space gradient in view $i \in \{1, 2\}$, respectively. Given the coefficient $\tau_i^{\mathbf{x}}$ computed for the Gaussian position attribute $\mathbf{x}$, we use its reciprocal as an attenuation factor to suppress conflicting gradients. We define the conflict-modulated view-space gradient as:

$$\widetilde{\nabla}_{\mathbf{p}_{\text{view}}^{(i)}} \mathcal{L}_i = \frac{1}{\tau_i^x} \nabla_{\mathbf{p}_{\text{view}}^{(i)}} \mathcal{L}_i. \tag{15}$$

We then update the joint densification statistics by

$$r_{\max} \leftarrow \max\left(r_{\max}, \mathbb{I}_{\mathcal{V}_1} r^{(1)}, \mathbb{I}_{\mathcal{V}_2} r^{(2)}\right), \qquad \nabla_{\text{accum}} \leftarrow \nabla_{\text{accum}} + \sum_{i=1}^{2} \mathbb{I}_{\mathcal{V}_i} \left\| \widetilde{\nabla}_{\mathbf{p}_{\text{view}}^{(i)}} \mathcal{L}_i \right\|. \tag{16}$$

The densification criterion is then applied to these joint statistics: the average accumulated gradient determines whether a Gaussian is densified, while its current 3D scale determines whether it is cloned or split. The maximum projected radius $r_{\max}$ is retained for screen-space size-based pruning.

Table 1: Quantitative results on the NeRF On-the-go dataset. The **best** and second-best values in each column are highlighted in bold and underlined, respectively.

Scene	High Occlusion			Medium Occlusion			Low Occlusion		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RobustNeRF (Sabour et al., 2023)	20.60	0.602	0.379	21.72	0.741	0.248	16.60	0.407	0.480
NeRF On-the-go (Ren et al., 2024)	22.37	0.753	0.212	22.50	0.780	0.205	20.13	0.627	0.287
3DGS (Kerbl et al., 2023)	19.03	0.649	0.340	19.19	0.709	0.220	19.68	0.649	0.199
Mip-Splatting (Yu et al., 2024)	19.25	0.664	0.333	19.73	0.684	0.279	20.03	0.661	0.195
GS-W (Zhang et al., 2024)	18.52	0.645	0.335	21.04	0.737	0.208	19.75	0.660	0.287
WildGaussian (Kulhanek et al., 2024)	23.03	0.771	0.172	22.80	0.811	0.092	20.62	0.658	0.235
SpotLessSplats (Sabour et al., 2025)	21.92	0.710	0.222	22.79	0.817	0.162	20.02	0.596	0.276
HybridGS (Lin et al., 2025)	23.05	0.768	0.204	23.51	0.830	0.160	21.42	0.684	0.268
DroneSplat (Tang et al., 2025)	23.52	0.810	<u>0.136</u>	23.27	0.813	0.102	21.53	0.697	**0.162**
AsymGS (GS-GS) (Li et al., 2025)	<u>24.34</u>	<u>0.825</u>	0.150	<u>24.56</u>	<u>0.872</u>	<u>0.090</u>	<u>21.91</u>	<u>0.728</u>	0.189
AsymGS (EMA-GS) (Li et al., 2025)	24.12	0.818	0.154	24.32	0.864	<u>0.090</u>	21.77	0.722	**0.162**
Ours	**24.67**	**0.844**	**0.123**	**25.09**	**0.879**	**0.084**	**22.17**	**0.730**	<u>0.164</u>

Conflict-guided pruning. We consider persistent gradient conflicts across views as a potential indicator of unstable Gaussians, which may correspond to floating artifacts. Based on this intuition, we introduce a conflict-guided pruning strategy to suppress such Gaussians during optimization. For the n-th Gaussian at iteration t, we compute its conflict from the gradients of center position $\mathbf{x}_n$ and opacity α_n in the two training views. Let $g_{i,n,a}^{(t)} = \nabla_{a_n} \mathcal{L}_i^{(t)}$ denote the gradient of attribute $a \in \{\mathbf{x}, \alpha\}$ for Gaussian n from view i. The per-Gaussian instantaneous conflict is defined as:

$$C_n^{(t)} = \max_{a \in \{\mathbf{x}, \alpha\}} \left[\max \left(0, -\cos \left(g_{1,n,a}^{(t)}, g_{2,n,a}^{(t)} \right) \right) \right]. \tag{17}$$

A larger $C_n^{(t)}$ indicates stronger disagreement between the two views for Gaussian n. To reduce the effect of noisy per-iteration gradients, we maintain a per-Gaussian EMA conflict score and periodically decay its opacity:

$$H_n^{(t)} = \gamma H_n^{(t-1)} + (1-\gamma) C_n^{(t)}, \qquad \alpha_n \leftarrow \alpha_n \cdot \exp \left(-\lambda_{\text{prune}} H_n^{(t)} \right). \tag{18}$$

The opacity decay is applied every 100 iterations. We set $\gamma = 0.99$ and $\lambda_{\text{prune}} = 0.3$ by default. Gaussians with persistent conflicts gradually lose opacity and are eventually removed by the standard Gaussian pruning step, thereby suppressing floating artifacts while preserving optimization stability.

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate our method on three widely used in-the-wild benchmarks: PhotoTourism (Martin-Brualla et al., 2021), NeRF On-the-go (Ren et al., 2024), and RobustNeRF (Sabour et al., 2023). PhotoTourism consists of internet landmark photo collections, featuring substantial appearance variations across time, weather, and seasons, as well as transient occluders such as pedestrians, vehicles, and scaffolding. Following prior works (Chen et al., 2022b; Martin-Brualla et al., 2021; Kulhanek et al., 2024), we evaluate on three scenes: Brandenburg Gate, Sacre Coeur, and Trevi Fountain. NeRF On-the-go contains handheld video sequences with dynamic occlusions, and we evaluate on all six public scenes spanning low-, medium-, and high-occlusion settings. RobustNeRF assesses reconstruction robustness under severe transient distractions, including challenging occlusions and inconsistent observations across views. We evaluate on all four scenes: Statue, Android, Crab, and Yoda. Following previous methods (Sabour et al., 2023; Ren et al., 2024; Kerbl et al., 2023; Yu et al., 2024), we evaluate rendering quality using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018).

4.2 Comparison with State-of-the-Art Methods

As shown in Tables 1, 2, and 3, our method consistently achieves superior reconstruction quality over both NeRF-based and Gaussian-based baselines across all evaluated in-the-wild datasets, including NeRF On-the-go, RobustNeRF, and PhotoTourism. These results demonstrate the robustness and

Table 2: Quantitative comparison on the RobustNeRF dataset. The **best** and second-best results in each column are highlighted in bold and underlined, respectively.

Scene	Statue			Android			Yoda			Crab		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RobustNeRF (Sabour et al., 2023)	20.60	0.760	0.150	23.28	0.750	0.130	29.78	0.820	0.150	-	-	-
NeRF On-the-go (Ren et al., 2024)	21.58	0.770	0.240	23.50	0.750	0.210	29.96	0.830	0.240	-	-	-
3DGS (Kerbl et al., 2023)	21.02	0.810	0.160	23.11	0.810	0.130	26.33	0.910	0.140	29.74	-	-
Mip-Splatting (Yu et al., 2024)	22.08	0.860	0.135	23.45	0.801	0.106	27.96	0.933	0.136	29.18	0.929	0.129
GS-W (Zhang et al., 2024)	21.99	0.862	0.102	24.23	0.824	0.090	32.74	0.957	0.084	33.22	0.952	0.088
WildGaussian (Kulhanek et al., 2024)	23.25	0.886	0.105	24.57	0.827	0.085	32.84	0.956	0.091	32.81	0.952	0.092
SpotLessSplats (Sabour et al., 2025)	22.54	0.840	0.130	25.05	0.850	0.090	33.66	0.960	0.100	34.43	-	-
HybridGS (Lin et al., 2025)	22.93	0.870	0.100	25.15	0.850	0.070	35.32	0.960	**0.070**	35.17	0.960	0.080
AsymGS (GS-GS) (Li et al., 2025)	23.47	**0.894**	0.097	25.61	**0.857**	0.071	37.18	**0.969**	0.074	36.18	**0.964**	0.078
AsymGS (EMA-GS) (Li et al., 2025)	23.49	0.890	0.096	25.47	0.849	0.068	36.50	0.967	0.077	35.60	0.961	0.080
Ours	**23.95**	0.893	**0.095**	**26.09**	**0.857**	**0.067**	**37.29**	**0.969**	0.074	**36.50**	**0.964**	**0.077**

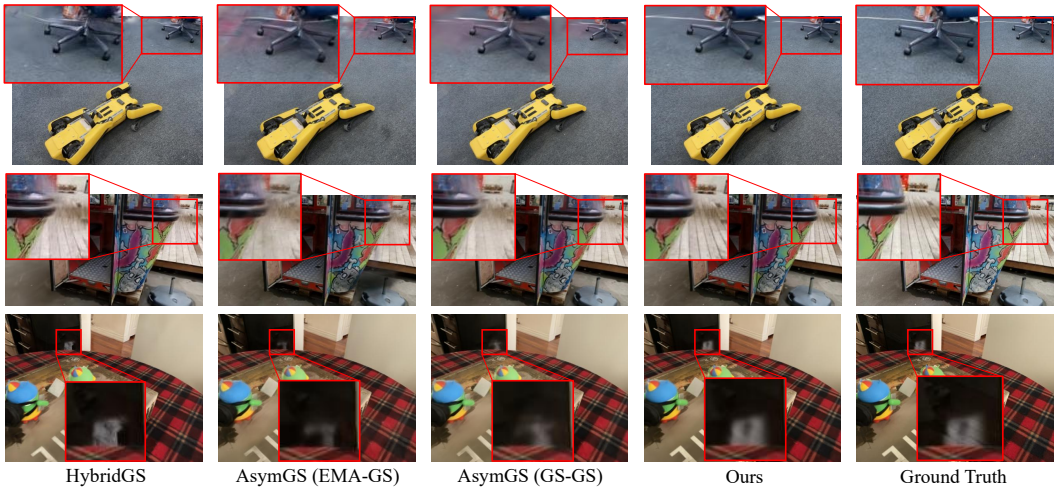


Figure 5: Qualitative comparisons on challenging in-the-wild scenes. Our method suppresses transient artifacts more effectively than current state-of-the-art methods, yielding cleaner renderings and more faithful scene details.

generality of our approach under challenging transient distractors and cross-view inconsistencies. Figure 5 further provides visual comparisons with representative state-of-the-art methods. The qualitative results show that our method effectively suppresses floating artifacts, leading to cleaner renderings and more faithful reconstructions.

4.3 Ablation Study

Gradient harmonization and Gaussian structure refinement. As shown in Table 4, our full dual-view framework achieves the best reconstruction quality among all variants. Compared with their single-view counterparts, the dual-view variants generally improve reconstruction performance under different masking configurations, especially in terms of PSNR, including w/o $\mathbf{M}_i$, w/o $\mathbf{S}_i$ (only $\mathbf{M}_i^{\text{bin}}$), and w/o $\mathbf{M}_i^{\text{bin}}$ (only $\mathbf{S}_i$). Specifically, dual-view optimization improves PSNR from 22.27 to 22.43 without $\mathbf{M}_i$, from 23.25 to 23.46 when only $\mathbf{M}_i^{\text{bin}}$ is used, and from 23.43 to 23.52 when only $\mathbf{S}_i$ is used. These results indicate that Gradient Harmonization helps exploit two-view optimization signals while mitigating their potential conflicts.

We further ablate the two conflict-aware components. Removing conflict-aware Gradient Harmonization reduces PSNR from 23.98 to 23.50 and SSIM from 0.818 to 0.802, indicating that directly combining two-view gradients without conflict reconciliation leads to less stable optimization. Removing conflict-aware densification and pruning also degrades performance, reducing PSNR to 23.84. Although this variant remains competitive, it uses more Gaussians than the full model (413K vs. 344K). This suggests that conflict-aware structure refinement, i.e., densification and pruning, not only improves reconstruction quality but also produces a more compact Gaussian representation by suppressing unstable primitives and avoiding redundant growth.

Table 3: Quantitative comparison on the PhotoTourism dataset. The **best** and second-best results in each column are highlighted in bold and underlined, respectively.

Scene	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF (Mildenhall et al., 2021)	18.90	0.882	0.138	15.60	0.846	0.163	16.14	0.696	0.282
3DGS (Kerbl et al., 2023)	19.37	0.880	0.141	17.44	0.835	0.204	17.58	0.709	0.266
Mip-Splatting (Yu et al., 2024)	20.01	0.877	0.166	17.54	0.831	0.203	17.36	0.684	0.319
NeRF-W (Martin-Brualla et al., 2021)	24.17	0.891	0.152	19.20	0.803	0.192	18.97	0.698	0.265
Ha-NeRF (Chen et al., 2022b)	24.04	0.887	0.139	20.02	0.801	0.171	20.18	0.691	0.223
K-Planes (Fridovich-Keil et al., 2023)	25.49	0.879	0.224	20.61	0.774	0.265	22.67	0.714	0.317
RefinedFields (Kassab et al., 2023)	26.64	0.886	-	22.26	0.817	-	23.42	0.737	-
GS-W (Zhang et al., 2024)	23.51	0.897	0.166	19.39	0.825	0.211	20.06	0.723	0.274
WildGaussian (Kulhanek et al., 2024)	27.77	0.927	0.133	22.56	0.859	0.177	23.63	0.766	0.228
AsymGS (GS-GS) (Li et al., 2025)	28.56	0.938	0.109	23.78	0.887	**0.139**	24.52	0.790	0.202
AsymGS (EMA-GS) (Li et al., 2025)	28.50	0.937	0.115	23.37	0.882	0.150	23.85	0.775	0.242
Ours	**29.08**	**0.945**	**0.102**	**24.58**	**0.892**	0.143	**24.97**	**0.807**	**0.201**

Table 4: Ablation study on the NeRF On-the-go dataset. We compare the full model with recent state-of-the-art methods, examine the contribution of each component in the proposed dual-view framework, and further report single-view variants to isolate the effect of dual-view optimization.

Method	PSNR↑	SSIM↑	LPIPS↓	Training Time (h)	GPU Memory (GB)	Points (K)
Baselines						
AsymGS (GS-GS) (Li et al., 2025)	23.61	0.809	0.143	0.41	3.73	546
AsymGS (EMA-GS) (Li et al., 2025)	23.40	0.801	0.135	0.27	2.82	531
Dual-view Setting						
Ours	23.98	0.818	0.123	0.33	3.97	344
Ours w/o $\mathbf{M}_i$	22.43	0.770	0.169	0.15	3.21	444
Ours w/o $\mathbf{S}_i$ (only $\mathbf{M}_i^{\text{bin}}$)	23.46	0.806	0.134	0.16	3.38	430
Ours w/o $\mathbf{M}_i^{\text{bin}}$ (only $\mathbf{S}_i$)	23.52	0.799	0.139	0.30	3.92	396
Ours w/o Gradient Harmonization	23.50	0.802	0.131	0.33	3.96	471
Ours w/o Conf.-Aware Dens. & Prun.	23.84	0.814	0.128	0.30	3.98	413
Single-view Setting						
Ours (single-view)	23.49	0.801	0.139	0.21	3.29	393
Ours (single-view) w/o $\mathbf{M}_i$	22.27	0.776	0.166	0.14	3.08	467
Ours (single-view) w/o $\mathbf{S}_i$ (only $\mathbf{M}_i^{\text{bin}}$)	23.25	0.803	0.133	0.15	3.08	462
Ours (single-view) w/o $\mathbf{M}_i^{\text{bin}}$ (only $\mathbf{S}_i$)	23.43	0.804	0.136	0.20	3.26	399

Design for masking. Removing the final mask $\mathbf{M}_i$ clearly degrades performance, dropping PSNR from 23.98 to 22.43 in the dual-view setting and from 23.49 to 22.27 in the single-view setting. These results show that using masks to suppress unreliable transient regions is crucial for robust in-the-wild reconstruction. We further evaluate the two components of $\mathbf{M}_i$ separately. In the dual-view setting, using only the SAM-based binary mask $\mathbf{M}_i^{\text{bin}}$ achieves 23.46 PSNR and 0.134 LPIPS, while using only the learned consistency score $\mathbf{S}_i$ achieves 23.52 PSNR and 0.139 LPIPS. Similar results are observed in the single-view setting, where the two variants obtain 23.25 and 23.43 PSNR, respectively, both outperforming the no-mask variant. These results show that both components are effective, while their combination consistently achieves the best performance, indicating that the binary mask and learned consistency score provide complementary cues for suppressing unreliable regions.

5 Conclusion

In this work, we present a novel conflict-aware 3D Gaussian Splatting framework for robust in-the-wild scene reconstruction. We first introduce Semantic Consistency-Guided Masking to adaptively refine prior masks and effectively suppress unreliable supervision. Since masking alone cannot fully resolve cross-view optimization conflicts, we further propose dual-view Conflict-Aware Gradient Harmonization to reconcile view-specific gradients into an orthogonal configuration. Conflict-aware densification and pruning are also developed to stabilize Gaussian structure optimization by guiding reliable growth and removing persistently conflicting primitives. Extensive experiments on three challenging in-the-wild benchmarks demonstrate the effectiveness and robustness of our approach.

REFERENCES

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pp. 333–350. Springer, 2022a.

Jiahao Chen, Yipeng Qin, Lingjie Liu, Jiangbo Lu, and Guanbin Li. Nerf-hugs: Improved neural radiance fields in non-static scenes using heuristics-guided segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19436–19446, 2024.

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, 2022b.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5501–5510, 2022.

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12479–12488, 2023.

Cameron Gordon, Shin-Fang Chng, Lachlan MacDonald, and Simon Lucey. On quantizing implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 341–350, 2023.

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5354–5363, 2024.

Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. Efficientnerf efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12902–12911, 2022.

Karim Kassab, Antoine Schnepf, Jean-Yves Franceschi, Laurent Caraffa, Jeremie Mary, and Valérie Gouet-Brunet. Refinedfields: Radiance fields refinement for unconstrained scenes. *arXiv preprint arXiv:2312.00639*, 3, 2023.

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

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.

Jonas Kulhanek and Torsten Sattler. NerfBaselines: Consistent and reproducible evaluation of novel view synthesis methods. In *Proceedings of the 39th International Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, pp. 467–484. Springer, 2024.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5166–5175, 2024.

Jingyu Lin, Jiaqi Gu, Lubin Fan, Bojian Wu, Yujing Lou, Renjie Chen, Ligang Liu, and Jieping Ye. Hybridgs: Decoupling transients and statics with 2d and 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 788–797, 2025.

Yifei Liu, Zhihang Zhong, Yifan Zhan, Sheng Xu, and Xiao Sun. Maskgaussian: Adaptive 3d gaussian representation from probabilistic masks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 681–690, 2025.

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.

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.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

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.

Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14335–14345, 2021.

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 Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *ACM Transactions on Graphics*, 44(2):1–11, 2025.

Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5459–5469, 2022.

Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11358–11367, 2021.

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.

Paul Ungermann, Armin Ettenhofer, Matthias Nießner, and Barbara Roessle. Robust 3d gaussian splatting for novel view synthesis in presence of distractors. In *DAGM German Conference on Pattern Recognition*, pp. 153–167. Springer, 2024.

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 (CVPR)*, pp. 4912–4921, June 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.

Jiacong Xu, Yiqun Mei, and Vishal M Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *Advances in Neural Information Processing Systems*, 37:103334–103355, 2024.

Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5438–5448, 2022.

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.

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.

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 *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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

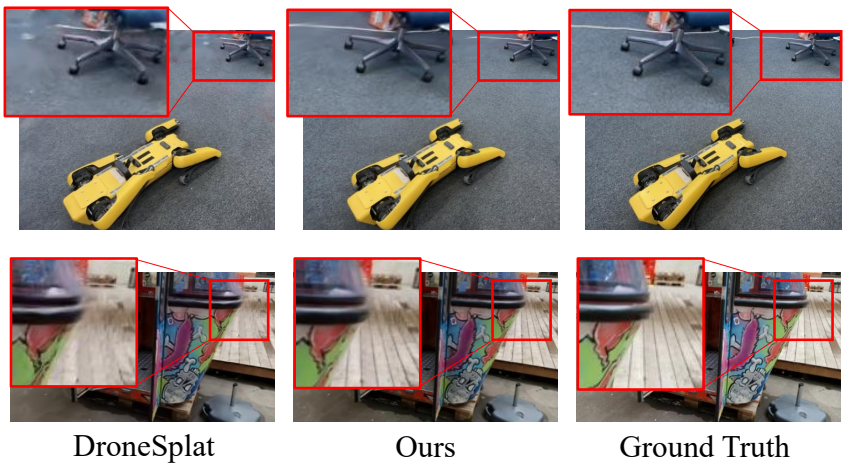


Figure 6: Additional qualitative comparisons with DroneSplat (Tang et al., 2025).

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability and clarity of the manuscript. Specifically, after completing the initial draft, we used AI assistance to refine sentence structures, improve the overall flow of the paper, enhance the clarity of explanations across different sections, and check grammatical correctness. For mathematical derivations and proofs, we used AI tools to provide additional perspectives and suggestions, as well as to assist in reviewing the derivation process and identifying potential errors or inconsistencies. All AI-assisted suggestions and modifications were carefully reviewed, verified, and refined by the authors. The authors take full responsibility for the final manuscript and all contents presented therein.

We emphasize that AI tools were not used for discovering or selecting references, developing research methodologies, designing experiments, or generating the core research ideas. All research concepts, methodologies, and contributions were independently proposed and developed by the authors. The authors take full responsibility for the content of this manuscript, including any text refined with the assistance of LLMs. We have carefully reviewed all AI-assisted modifications to ensure their accuracy, originality, and compliance with ethical research standards, and to prevent plagiarism or any form of scientific misconduct.

A IMPLEMENTATION DETAILS

Our Gaussian model is built upon Mip-Splatting (Yu et al., 2024). We follow AsymGS (Li et al., 2025) for the overall densification and pruning schedule, while adapting the optimization settings to our framework. Relative to the default settings, the learning rates of Gaussian attributes are uniformly scaled by 0.7 on NeRF On-the-go and RobustNeRF, and by 0.5 on PhotoTourism. The reconstruction loss weight λ_{rec} is set to 0.25. For Semantic Consistency-Guided Masking, we adopt a $5{,}000$-iteration warm-up, during which only the prior binary mask is used for reconstruction. The learning rate of the σ_i predictor is set to 0.001, the regularization weight λ_{inc} is set to 0.5, and the scaling hyperparameter c_σ for the consistency score $\mathbf{S}_i$ is set to 0.2. The mask correction parameters are set to $\eta_{\text{s}} = 1.2$ and $\eta_{\text{t}} = 3$ for stable and transient regions, respectively. For Conflict-Aware Gradient Harmonization, we set the geometric attenuation strength to $k = 0.5$ and use $\rho = 0.5$ as the default gradient-correction allocation parameter for all experiments unless otherwise stated.

B ADDITIONAL RESULTS

Complete per-scene results on the NeRF On-the-go dataset are reported in Table 5, and additional qualitative comparisons are provided in Figures 6 and 7.

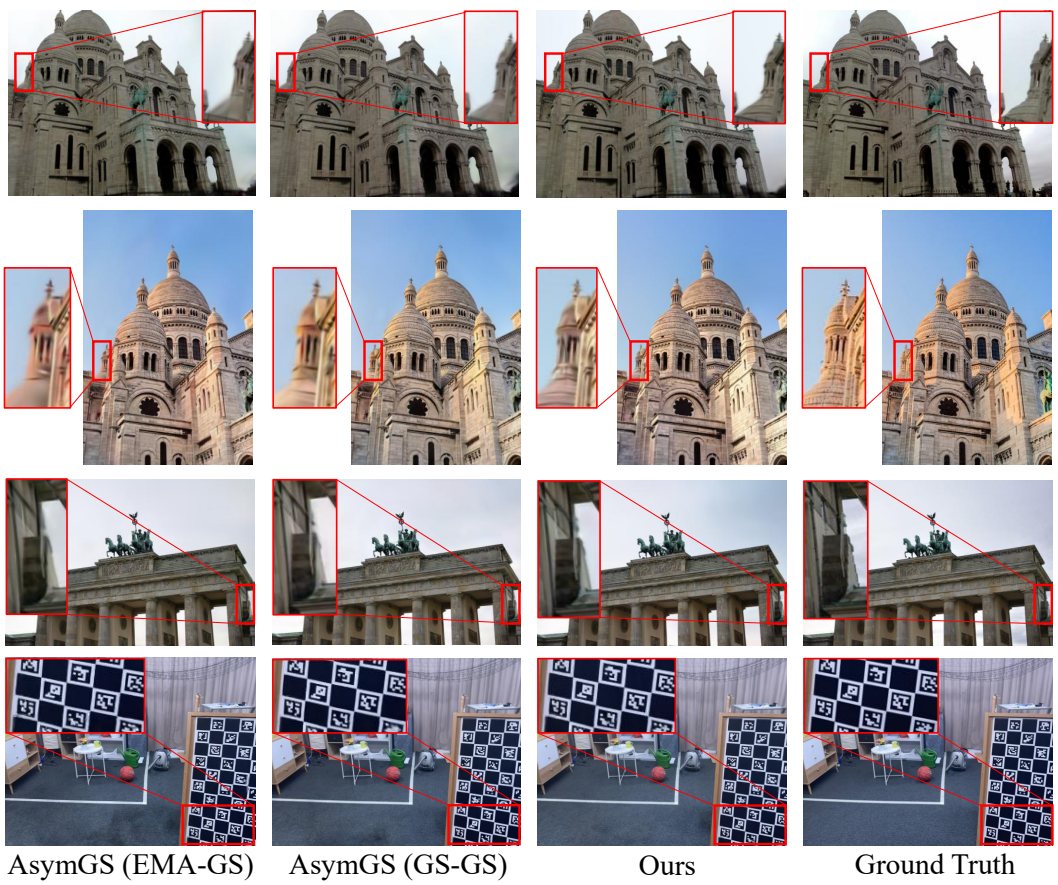


Figure 7: Additional qualitative comparisons with AsymGS (Li et al., 2025).

Table 5: Per-scene quantitative results on On-the-go dataset. The **best** and second-best values in each column are highlighted in bold and underlined, respectively.

Scene	Mountain			Fountain			Corner			Patio			Spot			Patio-High		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RobustNeRF (Sabour et al., 2023)	17.54	0.496	0.383	15.65	0.318	0.576	23.04	0.764	0.244	20.39	0.718	0.251	20.65	0.625	0.391	20.54	0.578	0.366
NeRF On-the-go (Ren et al., 2024)	20.15	0.644	0.259	20.11	0.609	0.314	24.22	0.806	0.190	20.78	0.754	0.219	23.33	0.787	0.189	21.41	0.718	0.235
3DGS (Kerbl et al., 2023)	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 (Yu et al., 2024)	19.86	0.649	0.200	20.19	0.672	0.189	21.15	0.728	0.230	18.31	0.639	0.328	20.18	0.689	0.338	18.31	0.639	0.328
WildGaussian (Kulhanek et al., 2024)	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 (Sabour et al., 2025)	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
HybridGS (Lin et al., 2025)	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 (Tang et al., 2025)	21.45	0.694	**0.158**	21.60	0.699	0.165	24.65	0.814	0.098	21.88	0.812	0.106	24.44	0.827	**0.094**	22.60	0.792	0.177
AsymGS (GS-GS) (Li et al., 2025)	22.00	0.740	0.199	21.83	0.717	0.180	26.15	0.885	0.085	22.97	0.860	0.096	25.52	0.854	0.135	23.17	0.796	0.164
AsymGS (EMA-GS) (Li et al., 2025)	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
Ours	**22.21**	**0.741**	0.177	**22.12**	**0.718**	**0.151**	**26.58**	**0.886**	0.081	**23.59**	**0.871**	0.087	**25.91**	**0.879**	0.096	**23.43**	**0.809**	**0.149**

We further compare our method with DualSplat (Wang et al., 2026). We follow the evaluation protocol adopted by AsymGS (Li et al., 2025) and NeRF-Baselines (Kulhanek & Sattler, 2025), and reproduce DualSplat using the same metric implementations to ensure a consistent comparison. Specifically, the released evaluation code of DualSplat computes LPIPS on images normalized to $[0, 1]$, whereas NeRF-Baselines evaluation protocol normalizes inputs to $[-1, 1]$ before LPIPS computation. For SSIM evaluation, DualSplat uses zero-padded convolution and averages over all spatial locations, while the NeRF-Baselines protocol adopts valid convolution and excludes window centers within five pixels of the image boundary. Therefore, we report DualSplat results reproduced under the unified NeRF-Baselines evaluation protocol rather than directly using the originally reported values. To further ensure a fair comparison, we follow the dataset preprocessing pipeline used by AsymGS, as the two implementations differ in image undistortion and downsampling resolution. The quantitative comparison with DualSplat is reported in Table 6.

Table 6: Additional quantitative comparison with DualSplat (Wang et al., 2026) on the NeRF On-the-go dataset under the unified AsymGS/NeRF-Baselines evaluation protocol. Average results across different occlusion levels are also reported. The **best** values in each column are highlighted in bold.

Scene	High Occlusion			Medium Occlusion			Low Occlusion			Average		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DualSplat	23.91	0.831	**0.122**	23.94	0.859	0.088	21.48	0.701	0.170	23.11	0.797	0.127
Ours	**24.67**	**0.844**	0.123	**25.09**	**0.879**	**0.084**	**22.17**	**0.730**	**0.164**	**23.98**	**0.818**	**0.123**

Table 7: Additional quantitative comparison with DualSplat (Wang et al., 2026) on the NeRF On-the-go dataset using the DualSplat preprocessing pipeline and evaluation settings. Average results across different occlusion levels are also reported. The **best** values in each column are highlighted in bold.

Scene	High Occlusion			Medium Occlusion			Low Occlusion			Average		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DualSplat	24.61	0.872	**0.071**	24.09	0.863	0.067	21.57	0.727	**0.129**	23.42	0.820	**0.088**
Ours	**24.65**	**0.873**	0.079	**24.83**	**0.897**	**0.047**	**21.82**	**0.736**	0.143	**23.77**	**0.835**	0.090

To further examine the influence of different preprocessing and evaluation pipelines, we additionally evaluate both methods using the dataset preprocessing pipeline and evaluation settings adopted by DualSplat. The corresponding results are reported in Table 7.

Across both evaluation settings, our method consistently achieves higher average PSNR and SSIM compared with DualSplat, demonstrating improved reconstruction fidelity under different preprocessing and metric protocols. Under the unified AsymGS/NeRF-Baselines protocol (Table 6), HarmoGS improves the average PSNR from 23.11 to 23.98 and the average SSIM from 0.797 to 0.818. Similar improvements are observed under the DualSplat preprocessing and evaluation settings (Table 7), where our method achieves higher average PSNR and SSIM while maintaining competitive LPIPS performance. These results suggest that the performance gains of HarmoGS are robust to different evaluation pipelines.

C Limitations

We observe that Conflict-Aware Gradient Harmonization can be sensitive to mask quality. As shown in Table 4, when the mask is less reliable, e.g., w/o $\mathbf{M}_i$ and w/o $\mathbf{M}_i^{\text{bin}}$ (only $\mathbf{S}_i$), the dual-view setting slightly underperforms its single-view counterpart on SSIM or LPIPS in a few cases. We hypothesize that inaccurate masks may introduce unreliable regions into the two-view optimization process. In this case, harmonizing gradients from two views can lead to a compromise between inconsistent signals, which may slightly weaken fine structural details. Nevertheless, the dual-view setting still brings clear PSNR improvements, indicating that it improves the overall reconstruction fidelity even when local perceptual details are mildly affected. This also highlights the importance of accurate inconsistency masking for fully exploiting cross-view gradient harmonization.

D Additional Analysis of Gradient Harmonization Design Choices

D.1 Gradient Harmonization Granularity

Our gradient harmonization is performed in a **scene-wide per-attribute** manner rather than a per-primitive manner. Specifically, for each Gaussian attribute group $a \in \{\mathbf{x}, \mathbf{r}, \mathbf{s}, \alpha, \mathbf{c}\}$, we flatten the gradients of all N Gaussians into a vector:

$$g_{i,a} \in \mathbb{R}^{Nd_a}, \tag{19}$$

where d_a denotes the dimensionality of attribute a. Given two sampled views, we compute the conflict measure and harmonization coefficients (τ_1^a, τ_2^a) based on the corresponding scene-level attribute gradients. The same coefficients are then applied to all Gaussian primitives within the attribute group.

This design avoids unstable conflict estimation caused by independently computing conflicts for individual primitives, where weakly visible or noisy Gaussians may introduce unreliable gradient

directions. In contrast, our conflict-aware pruning strategy operates at the primitive level, as it aims to identify individual Gaussians with persistent cross-view conflicts and remove unstable primitives.

To further investigate the impact of harmonization granularity, we compare different gradient aggregation strategies on the NeRF On-the-go dataset. As shown in Table 8, scene-wide per-attribute harmonization achieves the best overall performance, demonstrating that aggregating gradients at the attribute level provides a more reliable estimation of cross-view conflicts while preserving effective optimization signals.

Table 8: Ablation study on different gradient harmonization granularities. Results are evaluated on the NeRF On-the-go dataset.

Granularity	PSNR↑	SSIM↑	LPIPS↓
Primitive-wise, all attributes	23.75	0.814	0.128
Primitive-wise per attribute	23.85	0.817	0.125
Scene-wide, all attributes	23.78	0.815	0.130
Scene-wide per attribute (Ours)	**23.98**	**0.818**	**0.123**

D.2 Gradient Conflict and Multi-view Averaging

Negative inner products between view-specific gradients naturally arise during multi-view optimization. In in-the-wild reconstruction, such conflicts can be caused by inconsistent observations, including transient occlusions and cross-view appearance variations. Direct aggregation of conflicting gradients may introduce incompatible update directions, which can destabilize Gaussian optimization and lead to reconstruction artifacts.

Increasing the number of views involved in each iteration provides additional multi-view information during optimization. However, simply aggregating more view-specific gradients does not explicitly address conflicts caused by inconsistent observations. To study the effect of view aggregation strategies, we compare standard multi-view gradient averaging with our gradient harmonization under different numbers of views per iteration. When increasing the number of views, the **learning rate** is scaled accordingly while keeping other training configurations unchanged to maintain stable training and effective convergence.

Table 9: Comparison between gradient harmonization and multi-view gradient averaging on the NeRF On-the-go dataset.

Method	Views per Iteration	Harmonization	PSNR↑	SSIM↑	LPIPS↓
Standard training	1	No	23.53	0.802	0.138
Multi-view averaging	2	No	23.50	0.802	0.131
Multi-view averaging	4	No	23.46	0.802	0.138
Multi-view averaging	8	No	23.41	0.800	0.141
HarmoGS (Ours)	2	Yes	**23.98**	**0.818**	**0.123**

The comparison indicates that simply increasing the number of views per iteration does not consistently improve reconstruction quality. Although multi-view averaging incorporates information from multiple observations, it does not directly reconcile conflicting update directions caused by inconsistent supervision. In contrast, HarmoGS explicitly harmonizes view-specific gradients and achieves improved reconstruction quality with only two sampled views.

D.3 Gradient-Correction Allocation Parameter

The gradient harmonization strategy introduces an allocation parameter ρ to distribute the required angular correction between two conflicting gradients. In all experiments, we set $\rho = 0.5$, which assigns equal correction to both gradients:

$$\beta = \bar{\beta} = \frac{1}{2}\left(\theta - \frac{\pi}{2}\right). \tag{20}$$

This symmetric choice treats the two view-specific gradients equally, avoiding preference toward either view and making the harmonization invariant to the ordering of gradient pairs. When $\rho \neq 0.5$, one gradient receives more correction while the other retains more of its original direction, introducing an asymmetric update.

To evaluate the influence of ρ, we conduct an ablation study on the NeRF On-the-go dataset with different values: $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$.

Table 10: Ablation study of the gradient-correction allocation parameter ρ on the NeRF On-the-go dataset.

ρ	PSNR↑	SSIM↑	LPIPS↓
0	23.72	0.814	0.129
0.25	23.89	0.816	0.127
0.5	**23.98**	**0.818**	**0.123**
0.75	23.87	0.817	0.127
1	23.72	0.817	0.128

The results show that the performance is approximately symmetric around $\rho = 0.5$, which is consistent with the exchangeability of the two gradients. The symmetric allocation achieves the best overall performance, indicating that balancing the correction between two conflicting gradients provides a more stable optimization direction.

D.4 Per-Gaussian Conflict for Pruning

Different from gradient harmonization, which operates on scene-level attribute gradients, conflict-aware pruning aims to identify individual unstable Gaussians. Therefore, we define the pruning statistic at the primitive level.

For the n-th Gaussian, we compute the conflict score from the gradients of its position and opacity attributes:

$$C_n^{(t)} = \max_{a \in \{\mathbf{x}, \alpha\}} \left[\max(0, -\cos(g_{1,n,a}^{(t)}, g_{2,n,a}^{(t)})) \right]. \tag{21}$$

The position gradient is a 3D vector and provides a continuous measure of directional disagreement between two views. In contrast, opacity is a scalar visibility-related parameter. For scalar gradients, cosine similarity reduces to sign agreement, where opposite signs indicate that the two views favor contradictory opacity updates. We therefore use scalar opacity conflicts as an additional indicator of inconsistent visibility contributions rather than interpreting them as continuous angular conflicts.

A single instantaneous conflict does not directly trigger pruning. Instead, we accumulate the conflict signal using an exponential moving average:

$$H_n^{(t)} = \gamma H_n^{(t-1)} + (1 - \gamma) C_n^{(t)}, \tag{22}$$

and decay the opacity of Gaussians with persistent conflicts:

$$\alpha_n \leftarrow \alpha_n \exp(-\lambda_{\text{prune}} H_n^{(t)}). \tag{23}$$

This design distinguishes transient optimization fluctuations from persistent cross-view inconsistencies, allowing unstable Gaussians to be gradually suppressed rather than removed based on a single conflicting observation.

D.5 Comparison with Gradient Projection Methods

Gradient conflicts commonly arise when multiple views provide inconsistent supervision for a shared Gaussian representation. Given two view-specific losses L_1 and L_2 with gradients g_1 and g_2, the standard joint update direction is:

$$d = -(g_1 + g_2). \tag{24}$$

The first-order change of L_1 induced by this update is:

$$\Delta L_1 \approx -\eta g_1^\top (g_1 + g_2) = -\eta(\|g_1\|^2 + g_1^\top g_2). \tag{25}$$

When $g_1^\top g_2 < 0$, the second gradient partially cancels the descent component of the first gradient. Under inconsistent cross-view supervision, such cancellation may weaken useful optimization signals and lead to compromised updates.

To address this issue, our Conflict-Aware Gradient Harmonization rotates conflicting gradients within their original subspace until they become orthogonal, while preserving their magnitudes. By distributing the angular correction between the two gradients, our method treats the two views symmetrically and avoids introducing preference toward either view.

In contrast, projection-based gradient conflict resolution methods such as PCGrad modify conflicting gradients by removing their interfering components, which changes both the gradient magnitude and direction. This behavior is suitable when conflicting objectives should be strictly decoupled, but may discard useful information when different views provide partially inconsistent yet complementary supervision.

We compare our gradient harmonization with naive gradient summation and PCGrad on the NeRF On-the-go dataset while keeping all other training configurations unchanged.

Table 11: Comparison between gradient harmonization and gradient projection methods on the NeRF On-the-go dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
Naive summation	23.50	0.802	0.131
PCGrad	23.75	0.809	0.126
HarmoGS (Ours)	**23.98**	**0.818**	**0.123**

The results show that both conflict-aware strategies improve over naive gradient summation. However, HarmoGS achieves better reconstruction quality than PCGrad across all metrics. This indicates that preserving the magnitude and symmetry of view-specific gradients provides a more effective way to reconcile cross-view inconsistencies in in-the-wild reconstruction.

E Analysis of View Pair Selection Strategy

Our framework adopts random pair sampling to select two views for gradient harmonization. Since all views optimize the same shared Gaussian representation, each sampled pair provides an additional stochastic constraint on the shared parameters rather than requiring independent optimization for every view pair. To analyze the influence of view pair selection, we compare random sampling with overlap-guided selection strategies based on geometric visibility.

For two views i and j, we measure their overlap using the co-visible point ratio:

$$O_{ij} = \frac{|P_i \cap P_j|}{\min(|P_i|, |P_j|)}, \tag{26}$$

where P_i denotes the set of points visible in view i. Based on this metric, we implement two overlap-guided strategies. For each randomly sampled anchor view, Top-1 selects the candidate view with the highest overlap, while Top-5 uniformly samples from the five candidates with the highest overlap.

We compare these strategies with uniform random sampling on the NeRF On-the-go dataset while keeping other training configurations unchanged.

The results indicate that overlap-guided selection does not consistently improve reconstruction quality compared with random sampling. Although highly overlapping views provide stronger geometric correspondence, repeatedly selecting similar viewpoints may reduce the diversity of cross-view supervision. In contrast, random sampling provides a broader distribution of view pairs while maintaining sufficient overlap, allowing the optimization process to encounter diverse cross-view inconsistencies.

Table 12: Comparison of different view pair selection strategies on the NeRF On-the-go dataset.

Pair Selection Strategy	PSNR↑	SSIM↑	LPIPS↓	Training Time (h)↓
Pose/Visibility-Guided Top-1	23.76	0.811	0.132	0.48
Pose/Visibility-Guided Top-5	23.84	0.813	0.133	0.51
Uniform Random Sampling (Ours)	**23.98**	**0.818**	**0.123**	**0.33**

F Additional Analysis on Cross-View Gradient Harmonization

This section provides additional analysis and derivation details for the proposed two-view gradient harmonization strategy. We first discuss the delayed nature of conflicts in single-view optimization in Section F.1. We then provide a detailed derivation of the harmonized gradients in Section F.2, followed by an orthogonality verification in Section F.3. Next, we derive how the harmonized update can be reformulated in the original gradient basis in Section F.4. Finally, we analyze the expected pairwise conflict coverage of random two-view sampling in Section F.5.

F.1 Implicit and Delayed Conflicts in Single-View Optimization

Assume the training set contains N views. For the i-th view, let its reconstruction objective be $\mathcal{L}_i(\Theta)$, where Θ denotes the shared optimizable Gaussian parameters. The corresponding view-specific gradient is

$$g_i = \nabla_\Theta \mathcal{L}_i(\Theta). \tag{27}$$

For two different views i and j, a pairwise gradient conflict occurs when

$$g_i^\top g_j < 0. \tag{28}$$

This condition indicates that the two views prefer incompatible update directions for the shared Gaussian parameters.

Conventional 3DGS optimization samples one view at each iteration. If view i is sampled, the parameter update is

$$\Theta^+ = \Theta - \eta g_i, \tag{29}$$

where η is the learning rate. Since only g_i is computed, the current iteration cannot observe whether this update is compatible with other views. However, the update may still increase the loss of another view j. Using a first-order Taylor expansion around Θ, we have

$$\mathcal{L}_j(\Theta^+) = \mathcal{L}_j(\Theta - \eta g_i) \approx \mathcal{L}_j(\Theta) - \eta \nabla_\Theta \mathcal{L}_j(\Theta)^\top g_i. \tag{30}$$

Since $\nabla_\Theta \mathcal{L}_j(\Theta) = g_j$, this becomes

$$\mathcal{L}_j(\Theta^+) \approx \mathcal{L}_j(\Theta) - \eta g_j^\top g_i. \tag{31}$$

If $g_i^\top g_j < 0$, then

$$-\eta g_j^\top g_i > 0, \tag{32}$$

and therefore

$$\mathcal{L}_j(\Theta^+) > \mathcal{L}_j(\Theta). \tag{33}$$

This means that the update induced by view i can increase the reconstruction loss of view j. Therefore, single-view optimization is not conflict-free. Rather, the conflict between view i and view j is unobservable at the current iteration because g_j is not computed. Such a conflict can only be implicitly revealed in later iterations when view j is sampled.

F.2 Detailed Derivation of Harmonized Gradients

Assume that g_1 and g_2 are nonzero and not exactly collinear. When $g_1^\top g_2 < 0$, the angle between them is

$$\theta = \arccos\left(\frac{g_1^\top g_2}{\|g_1\|\|g_2\|}\right) \in \left(\frac{\pi}{2}, \pi\right). \tag{34}$$

We construct an orthonormal basis $\{u_1, u_2\}$ for the two-dimensional subspace $\mathcal{V}$:

$$u_1 = \frac{g_1}{\|g_1\|}, \quad u_2 = \frac{g_2 - (g_2^\top u_1)u_1}{\|g_2 - (g_2^\top u_1)u_1\|}. \tag{35}$$

Since

$$g_2^\top u_1 = \|g_2\| \cos\theta, \tag{36}$$

the original gradients can be represented as

$$g_1 = \|g_1\| u_1, \quad g_2 = \|g_2\| \left(\cos\theta\, u_1 + \sin\theta\, u_2\right). \tag{37}$$

The original angular separation between g_1 and g_2 is $\theta > \frac{\pi}{2}$. To remove the negative inner-product conflict, their final angular separation should be $\frac{\pi}{2}$. Thus, the required angular correction is

$$\Delta\theta = \theta - \frac{\pi}{2}. \tag{38}$$

We introduce an allocation parameter $\rho \in [0, 1]$ to distribute this correction between the two gradients:

$$\beta = \rho\left(\theta - \frac{\pi}{2}\right), \quad \bar{\beta} = (1-\rho)\left(\theta - \frac{\pi}{2}\right). \tag{39}$$

Here, g_1 is rotated from angle 0 to angle β in the basis $\{u_1, u_2\}$, while g_2 is rotated from angle θ toward g_1 by $\bar{\beta}$. Therefore, the final angle of $\tilde{g}_2$ with respect to u_1 is

$$\theta - \bar{\beta} = \theta - (1-\rho)\left(\theta - \frac{\pi}{2}\right). \tag{40}$$

Since $\beta = \rho(\theta - \frac{\pi}{2})$, we have

$$\theta - \bar{\beta} = \frac{\pi}{2} + \beta. \tag{41}$$

Therefore, the harmonized gradients are

$$\tilde{g}_1 = \|g_1\| \left(\cos\beta\, u_1 + \sin\beta\, u_2\right), \tag{42}$$

and

$$\tilde{g}_2 = \|g_2\| \left(\cos\left(\frac{\pi}{2} + \beta\right) u_1 + \sin\left(\frac{\pi}{2} + \beta\right) u_2\right). \tag{43}$$

Using

$$\cos\left(\frac{\pi}{2} + \beta\right) = -\sin\beta, \quad \sin\left(\frac{\pi}{2} + \beta\right) = \cos\beta, \tag{44}$$

we obtain

$$\tilde{g}_2 = \|g_2\| \left(-\sin\beta\, u_1 + \cos\beta\, u_2\right). \tag{45}$$

This construction preserves the magnitudes of the two original gradients while changing only their directions within $\mathcal{V}$.

F.3 Orthogonality Verification

We now verify that the two harmonized gradients are orthogonal. Since $\{u_1, u_2\}$ is an orthonormal basis, we have

$$u_1^\top u_1 = 1, \quad u_2^\top u_2 = 1, \quad u_1^\top u_2 = 0. \tag{46}$$

Substituting the expressions of $\tilde{g}_1$ and $\tilde{g}_2$, we have

$$\begin{aligned}
\tilde{g}_1^\top \tilde{g}_2 &= \|g_1\|\|g_2\| \left(\cos\beta\, u_1 + \sin\beta\, u_2\right)^\top \left(-\sin\beta\, u_1 + \cos\beta\, u_2\right) && (47)\\
&= \|g_1\|\|g_2\| \Big(-\cos\beta\sin\beta\, u_1^\top u_1 + \cos^2\beta\, u_1^\top u_2 - \sin^2\beta\, u_2^\top u_1 + \sin\beta\cos\beta\, u_2^\top u_2 \Big) && (48)\\
&= \|g_1\|\|g_2\| \Big(-\cos\beta\sin\beta + \sin\beta\cos\beta \Big) && (49)\\
&= 0. && (50)
\end{aligned}$$

Therefore, the harmonized gradients satisfy

$$\tilde{g}_1^\top \tilde{g}_2 = 0. \tag{51}$$

This confirms that the proposed rotation removes the negative inner-product conflict between the sampled gradients.

F.4 Recombination in the Original Gradient Basis

The final update direction is defined as the sum of the two harmonized gradients:

$$\tilde{g} = \tilde{g}_1 + \tilde{g}_2. \tag{52}$$

Substituting the expressions of $\tilde{g}_1$ and $\tilde{g}_2$, we obtain

$$\tilde{g} = \|g_1\| (\cos\beta\, u_1 + \sin\beta\, u_2) + \|g_2\| (-\sin\beta\, u_1 + \cos\beta\, u_2) \tag{53}$$

$$= (\|g_1\| \cos\beta - \|g_2\| \sin\beta)\, u_1 + (\|g_1\| \sin\beta + \|g_2\| \cos\beta)\, u_2. \tag{54}$$

Since both $\tilde{g}_1$ and $\tilde{g}_2$ lie in $\mathrm{span}(g_1, g_2)$, their sum can be expressed as a linear combination of the original gradients:

$$\tilde{g} = \tau_1 g_1 + \tau_2 g_2. \tag{55}$$

Using

$$g_1 = \|g_1\| u_1, \quad g_2 = \|g_2\| (\cos\theta\, u_1 + \sin\theta\, u_2), \tag{56}$$

we have

$$\tau_1 g_1 + \tau_2 g_2 = (\tau_1 \|g_1\| + \tau_2 \|g_2\| \cos\theta)\, u_1 + \tau_2 \|g_2\| \sin\theta\, u_2. \tag{57}$$

By matching the coefficient of u_2, we obtain

$$\tau_2 \|g_2\| \sin\theta = \|g_1\| \sin\beta + \|g_2\| \cos\beta, \tag{58}$$

which gives

$$\tau_2 = \frac{\|g_1\| \sin\beta}{\|g_2\| \sin\theta} + \frac{\cos\beta}{\sin\theta}. \tag{59}$$

By matching the coefficient of u_1, we have

$$\tau_1 \|g_1\| + \tau_2 \|g_2\| \cos\theta = \|g_1\| \cos\beta - \|g_2\| \sin\beta. \tag{60}$$

Therefore,

$$\tau_1 = \frac{\|g_1\| \cos\beta - \|g_2\| \sin\beta - \tau_2 \|g_2\| \cos\theta}{\|g_1\|}. \tag{61}$$

Substituting the expression of τ_2 and simplifying, we obtain

$$\tau_1 = \frac{\sin(\theta - \beta)}{\sin\theta} - \frac{\|g_2\| \cos(\theta - \beta)}{\|g_1\| \sin\theta}. \tag{62}$$

Thus, the harmonized update can be implemented as

$$\tilde{g} = \tau_1 g_1 + \tau_2 g_2, \tag{63}$$

where

$$\tau_1 = \frac{\sin(\theta - \beta)}{\sin\theta} - \frac{\|g_2\| \cos(\theta - \beta)}{\|g_1\| \sin\theta}, \quad \tau_2 = \frac{\|g_1\| \sin\beta}{\|g_2\| \sin\theta} + \frac{\cos\beta}{\sin\theta}. \tag{64}$$

When no conflict exists, i.e., $g_1^\top g_2 \geq 0$, we do not perform harmonization and simply set $\tau_1 = \tau_2 = 1$.

F.5 Expected Conflict Coverage over the Training Set

We further analyze how random two-view sampling covers pairwise conflicts over the training set. Let $\mathcal{P} = \{(i, j) \mid 1 \leq i < j \leq N\}$ be the set of all unordered view pairs, and let $M = |\mathcal{P}| = \frac{N(N-1)}{2}$ be the number of all possible view pairs. Assume that one unordered view pair is sampled uniformly at random at each iteration. For a fixed pair (i, j), the probability that it is sampled in one iteration is

$$p = \frac{1}{M}. \tag{65}$$

Therefore, the probability that this pair is not sampled in one iteration is

$$1 - p = 1 - \frac{1}{M}. \tag{66}$$

After T independent training iterations, the probability that this pair has never been sampled is

$$\left(1 - \frac{1}{M}\right)^T. \tag{67}$$

Consequently, the probability that this pair has been sampled at least once after T iterations is

$$P_{\text{cover}}(T) = 1 - \left(1 - \frac{1}{M}\right)^T. \tag{68}$$

When M is large, this can be approximated by

$$P_{\text{cover}}(T) \approx 1 - \exp\left(-\frac{T}{M}\right). \tag{69}$$

This probability also gives the expected fraction of all view pairs that have been sampled at least once:

$$\mathbb{E}\left[\frac{\#\text{covered pairs}}{M}\right] = 1 - \left(1 - \frac{1}{M}\right)^T. \tag{70}$$

Therefore, random two-view sampling progressively covers the pairwise view space as training proceeds. However, exhaustive coverage of all possible view pairs is not required for optimization, since the sampled gradients are used to update the shared Gaussian representation and different sampled pairs provide complementary cross-view constraints during training. To reach a target pairwise coverage ratio q, the required number of iterations satisfies

$$T \geq \frac{\log(1-q)}{\log(1-\frac{1}{M})}. \tag{71}$$

For large M, this is approximately

$$T \gtrsim M \log \frac{1}{1-q}. \tag{72}$$

Therefore, reaching 50%, 80%, 95%, 99%, and 99.9% expected pair coverage requires approximately

$$T_{50} \approx 0.7M, \quad T_{80} \approx 1.6M, \quad T_{95} \approx 3M, \quad T_{99} \approx 4.6M, \quad T_{99.9} \approx 6.9M. \tag{73}$$

At this stage, most pairwise relationships have been exposed at least once in expectation.

G Broader Impacts

This work aims to improve the robustness of 3D Gaussian Splatting for in-the-wild scene reconstruction. By reducing the influence of transient distractors and cross-view inconsistencies, the proposed method can benefit real-world applications that require reliable novel view synthesis, such as virtual reality, digital content creation, cultural heritage preservation, and robotic perception in unconstrained environments.