

Full Length Article

Depth-consistent 3D Gaussian Splatting via physical defocus modeling and multi-view geometric supervision

Yu Deng¹, Baozhu Zhao¹, Junyan Su¹, Xiaohan Zhang¹, Qi Liu^{*}

Department of Future Technology, South China University of Technology, Guangzhou, 511400, China

ARTICLE INFO

Keywords:

3D scene reconstruction
Depth-of-field
Differentiable rendering
Novel view synthesis

ABSTRACT

Three-dimensional reconstruction in scenes with extreme depth variations remains challenging due to inconsistent supervisory signals between near-field and far-field regions. Existing methods fail to simultaneously address inaccurate depth estimation in distant areas and structural degradation in close-range regions. This paper proposes a novel computational framework that integrates depth-of-field supervision and multi-view consistency supervision to advance 3D Gaussian Splatting. Our approach comprises two core components: (1) Depth-of-field Supervision employs a scale-recovered monocular depth estimator (e.g., Metric3D) to generate depth priors, leverages defocus convolution to synthesize physically accurate defocused images, and enforces geometric consistency through a novel depth-of-field loss, thereby enhancing depth fidelity in both far-field and near-field regions; (2) Multi-View Consistency Supervision employing LoFTR-based semi-dense feature matching to minimize cross-view geometric errors and enforce depth consistency via least squares optimization of reliable matched points. By unifying defocus physics with multi-view geometric constraints, our method achieves superior depth fidelity, demonstrating a 0.8 dB PSNR improvement over the state-of-the-art method on the Waymo Open Dataset. This framework bridges physical imaging principles and learning-based depth regularization, offering a scalable solution for complex depth stratification in urban environments.

1. Introduction

Three-dimensional scene reconstruction from multi-view images remains a cornerstone capability for applications ranging from autonomous driving to immersive virtual reality. While Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) revolutionized photorealistic novel view synthesis, subsequent advances in 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) have achieved unprecedented real-time rendering speeds through differentiable Gaussian primitives. However, these methods face critical limitations when reconstructing scenes with substantial depth variations, as distant structures often exhibit positional inaccuracies due to insufficient supervision signals and overfitting to training viewpoints.

Current approaches for large-scale scene reconstruction rely predominantly on multi-view stereo (MVS) techniques (Furukawa et al., 2015) or volumetric neural representations (Barron et al., 2022). The former establishes geometric consistency through hand-crafted features (Xu & Tao, 2019), while the latter optimizes implicit fields

through photometric loss (Wang et al., 2023). Recent extensions such as GaussianPro (Cheng et al., 2024a) attempt to mitigate extreme distant interference through sky segmentation masks, yet introduce new artifacts from imperfect matting and sparse depth supervision. This limitation stems from a fundamental challenge: when objects appear at varying distances across frames, conventional supervision struggles to resolve scale ambiguities, particularly for distant regions receiving insufficient pixel-level gradients.

Our work addresses these limitations through two synergistic innovations. First, we leverage depth-of-field effects as implicit geometric supervision. By modeling the physical correlation between defocus blur and scene depth via adaptive kernel convolution, we derive gradient signals that guide Gaussians towards their geometrically consistent positions. Second, we introduce a hybrid depth estimation framework that integrates multi-view feature matching (LoFTR) (Wang et al., 2024c) with monocular depth completion (Metric3D) (Hu et al., 2024a), resolving scale ambiguities while preserving structural details. This dual strategy effectively constrains Gaussian distributions

* Corresponding author.

E-mail addresses: 202311093429@mail.scut.edu.cn (Y. Deng), 202320163293@mail.scut.edu.cn (B. Zhao), ft_su.junyan@mail.scut.edu.cn (J. Su), ftxiaohanzhang@mail.scut.edu.cn (X. Zhang), drliuqi@scut.edu.cn (Q. Liu).

URL: <https://drliuqi.github.io/> (Q. Liu)

¹ These authors contributed equally to this work.

across varying depth layers, particularly benefiting distant regions reconstruction where traditional methods fail.

The technical contributions of this work are threefold:

- A physics-aware defocus convolution model that translates optical principles into geometric constraints, using adaptive kernel designs (Gaussian, Polygonal, or SmoothStep) to adapt to different types of cameras and dynamic focus optimization to enhance depth consistency.
- A multiscale depth alignment framework combining global monocular depth recovery with local grid-based correction, achieving view-consistent depth estimation without manual masking.
- A gradient-aware density control mechanism that prioritizes structurally critical regions through depth-gradient statistics.

Extensive validation on urban (Waymo) and unbounded (Mip-NeRF 360) scenes demonstrates the efficacy of our approach. Quantitative results show that our model achieves 35.17 PSNR on Waymo, outperforming SOTA methods. Qualitative analyses reveal significant improvements in near region structure recovery, particularly for vehicles and buildings. These advancements establish new state-of-the-art performance for depth-aware scene reconstruction while preserving the computational efficiency central to 3DGS frameworks.

2. Related work

2.1. Multi-view stereo

MVS represents a fundamental computer vision task that aims to reconstruct high-fidelity 3D models from a collection of calibrated images. The existing MVS methods can be broadly categorized into traditional geometry-based approaches and contemporary learning-based approaches.

Traditional MVS approaches typically derive their camera parameters predominantly from Structure-from-Motion (SfM) methods (Schonberger & Frahm, 2016; Snavely et al., 2006) or Simultaneous Localization and Mapping (SLAM) frameworks (Engel et al., 2014; Mur-Artal et al., 2015). Within this paradigm, seminal methods proposed by Campbell et al. (2008), Furukawa et al. (2015), and Xu and Tao (2019) establish explicit pixel correspondences through hand-crafted features and rigorous geometric constraints.

Learning-based approaches (Ji et al., 2017; Kar et al., 2017; Zhou et al., 2023) have revolutionized MVS, initiated by the pioneering end-to-end architecture introduced by Yao et al. (2018). Contemporary methods (Feng et al., 2023; Ma et al., 2022b) leverage learned representations for robust depth regression, while advanced techniques such as cascade cost volumes (Gu et al., 2020) and feature matching networks (Giang et al., 2021) have substantially enhanced both performance and computational efficiency.

2.2. Neural radiance field

A radiance field establishes a mapping from a 3D spatial coordinate (x, y, z) and viewing directions parameterized by polar angle θ and azimuthal angle ϕ to a nonnegative radiance value, characterizing the light-matter interaction within the environment (Chen & Wang, 2024). NeRF (Mildenhall et al., 2021) pioneered the paradigm of representing a scene as an emissive volumetric function, implemented via a position-encoded neural network that enables differentiable rendering through volumetric quadrature.

Volumetric representations integrated with deep-learning techniques and volumetric ray-marching were initially proposed by Sitzmann et al. (2019) and Henzler et al. (2019). Significant advancements in this domain include Instant-NGP (Müller et al., 2022), which employs multi-resolution hash grids, and Plenoxels (Fridovich-Keil et al.,

2022), which utilizes sparse voxel grids for efficient optimization. Several approaches enhance rendering efficiency through scene reparameterization to generate more compact representations, notably Mip-NeRF 360 (Barron et al., 2022), Zip-NeRF (Barron et al., 2023), and F2-NeRF (Wang et al., 2023).

Recent research has increasingly focused on addressing defocus blur in neural rendering, with seminal works including RawNeRF (Mildenhall et al., 2022), AR-NeRF (Kaneko, 2022), and NeRFocus (Wang et al., 2022), all capable of synthesizing depth-of-field effects. Subsequently, Deblur-NeRF (Ma et al., 2022a) mitigates image blurriness resulting from defocus by implementing a Deformable Sparse Kernel module, while DP-NeRF (Lee et al., 2023) specifically addresses geometric and appearance consistency challenges in defocused scenarios. Nevertheless, as noted by (Wang et al., 2024b), these approaches continue to encounter substantial limitations regarding computational efficiency and real-time rendering capabilities.

2.3. 3D Gaussian splatting

Novel view synthesis has evolved significantly with NeRF (Mildenhall et al., 2021) setting a milestone for photorealistic rendering. Building upon this foundation, 3DGS (Kerbl et al., 2023) introduces a paradigm shift that models scenes as a collection of 3D Gaussian primitives rendered via differentiable rasterization, simultaneously achieving high-quality reconstruction and real-time rendering capabilities. This approach extends traditional splatting-based rasterization (Zwicker et al., 2002) by optimizing Gaussian primitives with explicit geometry and appearance attributes (Cheng et al., 2024a; Li et al., 2024).

Recent advances in 3DGS research have predominantly focused on efficiency and quality enhancements. Innovative methods have been proposed to optimize Gaussian contributions through scale-based evaluation strategies (Lee et al., 2024) and sophisticated visibility assessment techniques (Fan et al., 2023). Concurrent developments have yielded substantial improvements in rendering fidelity (Blanc et al., 2024; Huang et al., 2024; Yu et al., 2024), computational efficiency (Girish et al., 2024), and the capacity to handle large-scale scenes with complex geometry (Kerbl et al., 2024; Liu et al., 2024b).

Despite these significant advancements, fundamental limitations persist in the 3DGS representation framework. The inherent absence of true volumetric density fields manifests as view-dependent consistency issues and rendering artifacts (Mai et al., 2024; Radl et al., 2024). Furthermore, the conventional pinhole camera model employed in standard 3DGS implementations inherently restricts its application domain to All-in-Focus (AiF) scenarios (Wang et al., 2024b).

Recent research has addressed these fundamental challenges through innovative depth-of-field rendering approaches. DOF-GS (Wang et al., 2024b) introduces a finite aperture camera model coupled with explicit, differentiable defocus rendering guided by the Circle-of-Confusion (CoC), thereby enabling both adjustable depth-of-field effects and the generation of AiF images from defocused training data. Similarly, Cinematic Gaussians (Wang et al., 2024a) leverages multi-view LDR images with varying exposure times, apertures, and focus distances to reconstruct high-dynamic-range (HDR) radiance fields, incorporating analytical convolutions of Gaussians based on a thin-lens camera model. Unlike these prior works that primarily focus on either depth-of-field effects rendering or HDR reconstruction, our approach uniquely leverages depth-of-field information as an additional structural supervision signal to achieve more geometrically accurate 3D reconstruction.

3. Method

Our computational framework integrates dual supervisory paradigms for geometrically consistent 3D Gaussian Splatting, as depicted in Fig. 2. The architecture operates on multi-view inputs through dual-branch processing, establishing metric-scale depth priors while enforcing cross-view consistency constraints. Guided by this

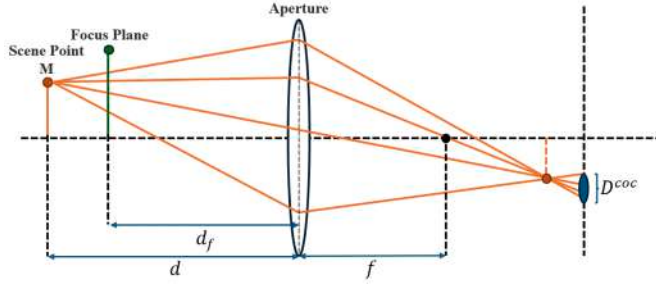


Fig. 1. A schematic illustrating the principle of depth of field blur. When a scene point M at a distance d does not lie on the focus plane (at distance d_f), it creates a blurred spot on the image plane known as the circle of confusion (with a diameter of D^{coc}), causing the image to be out of focus. f represents the focal length of the lens.

computational paradigm, we first introduce our depth-aware defocus modeling that formulates physical optics principles as differentiable geometric constraints. Subsequent sections systematically elaborate our hybrid depth estimation framework comprising global scale calibration and local depth refinement modules, culminating in multi-view geometric consistency enforcement.

3.1. Defocus convolution

Accurate modeling of defocus effects constitutes an essential component in establishing comprehensive geometric and radiometric supervision signals (Cui & Knoll, 2024) for 3D Gaussian Splatting-based scene reconstruction. This section presents a physics-driven defocus convolution framework that emulates physical imaging characteristics to achieve high-fidelity depth-of-field reconstruction through optical principle formulation.

3.1.1. Physical imaging model

As shown in Fig. 1, in optical imaging systems, scene points at distinct depth layers generate varying Circle of Confusion (CoC) dimensions through lens propagation. Considering a 3D scene point with object distance d , focal length f , and focus distance d_f , the CoC diameter in optical space is derived as:

$$D^{(coc)} = \frac{f^2 |d - d_f|}{F \cdot d \cdot (d_f - f)} \quad (1)$$

where F represents the f-number (defined as f/A , with A denoting the physical aperture diameter). For digital imaging applications, we convert the optical CoC diameter to pixel space through sensor-image scaling:

$$D^{(pixel)} = D^{(coc)} \cdot \frac{w_i}{w_s} \quad (2)$$

where $D^{(pixel)}$ corresponds to the effective defocus diameter in digital coordinates, w_i indicates the image resolution width (pixels), and w_s specifies the sensor's physical width (mm).

3.1.2. Adaptive kernel design

Our framework incorporates three distinct convolution kernels to address diverse defocus characteristics found in optical systems. This multi-kernel approach provides the flexibility to balance physical realism with computational efficiency.

Gaussian Blur Kernel For baseline defocus simulation, we employ a standard Gaussian kernel. Its softness provides a natural-looking blur. The kernel is defined as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (3)$$

where σ relates to the circle of confusion (CoC) radius R through the formulation:

$$\sigma_G = \frac{D^{(pixel)}}{k_s} \quad (4)$$

Here, k_s represents a normalization coefficient (default: 20) that scales the physical CoC diameter to kernel space. To ensure energy conservation, we enforce unitary integral constraint via kernel normalization:

$$K_{\text{norm}} = \frac{G}{\sum_{x,y} G(x, y)} \quad (5)$$

SmoothStep Blur Kernel To better preserve sharp edges in regions with high depth discontinuity, a known limitation of Gaussian blur, we implement a hyperbolic tangent-based kernel. Its S-shaped transition profile offers a superior trade-off between blurring and edge preservation. The formulation is:

$$K(x, y) = 0.5 + 0.5 \tanh(0.25(r^2 - x^2 - y^2) + 0.5) \quad (6)$$

where r denotes the kernel radius, and (x, y) represent the coordinates of each position within the kernel. This formulation achieves controlled edge preservation through its S-shaped transition profile, particularly effective for depth discontinuity regions.

Polygonal Blur Kernel To achieve the highest degree of physical realism, especially for simulating the characteristic 'bokeh' from a lens's aperture blades, we introduce a parametric polygonal kernel. As shown in Fig. 3, this model explicitly encodes the aperture geometry. The formulation is as follows:

$$a_i^{\text{Poly}}(p) = o_i \cdot \beta_i \cdot K(p) \quad (7)$$

where N (default: 8) represents the number of aperture blades, $\beta_i = 1 / \sum_{p \in \Omega} K(p)$ ensures normalization, and $K(p)$ integrates radial attenuation $W(r)$ with geometric containment $H(p)$:

$$K(p) = H(p) \cdot W(r(p)) \quad (8)$$

where $r(p) = \sqrt{x^2 + y^2}$ represents the Euclidean distance from point p to the center. The radial weight function employs cosine attenuation:

$$W(r) = \cos\left(\frac{\pi}{2} \cdot \frac{r}{R}\right) \cdot \mathbf{1}[r \leq R] \quad (9)$$

where R denotes the predefined kernel radius, $\mathbf{1}[r \leq R]$ is the indicator function. The vertices of the polygon are given by $v_i = \left(R \cos\left(\frac{2\pi i}{N}\right), R \sin\left(\frac{2\pi i}{N}\right)\right)$ for $i = 1, 2, \dots, N$, where $v_{N+1} = v_1$ ensures the polygon closure.

The cross-product function $C(p, v_i, v_{i+1})$, which determines the relative position of point p with respect to the edge formed by vertices v_i and v_{i+1} :

$$C(p, v_i, v_{i+1}) = (v_{i+1,x} - v_{i,x})(p_y - v_{i,y}) - (v_{i+1,y} - v_{i,y})(p_x - v_{i,x}). \quad (10)$$

Using this cross-product criterion, the indicator function $H(p)$, which determines whether a point p resides inside the N -sided polygon based on the cross-product criterion:

$$H(p) = \begin{cases} 1, & \text{if } \bigwedge_{i=1}^N C(p, v_i, v_{i+1}) < 0 \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

This parametric modeling accurately reproduces the optical characteristics observed in real-world Polygonal apertures, as evidenced by the qualitative comparison in Fig. 3(b).

The distinct visual effects of these three kernels are compared in Fig. 4, guiding the choice of kernel for different application requirements.

3.1.3. Dynamic focus optimization

We optimize focus distance d_f using depth distribution statistics $\{d_{1/3}, d_{1/2}, d_{2/3}, \mu_d\}$ (terciles and mean depth). The optimal focus distance minimizes:

$$d_f^* = \arg \min_{d_f} \sum_{p \in \Omega} w(p) \|d(p) - d_f\|^2 \quad (12)$$

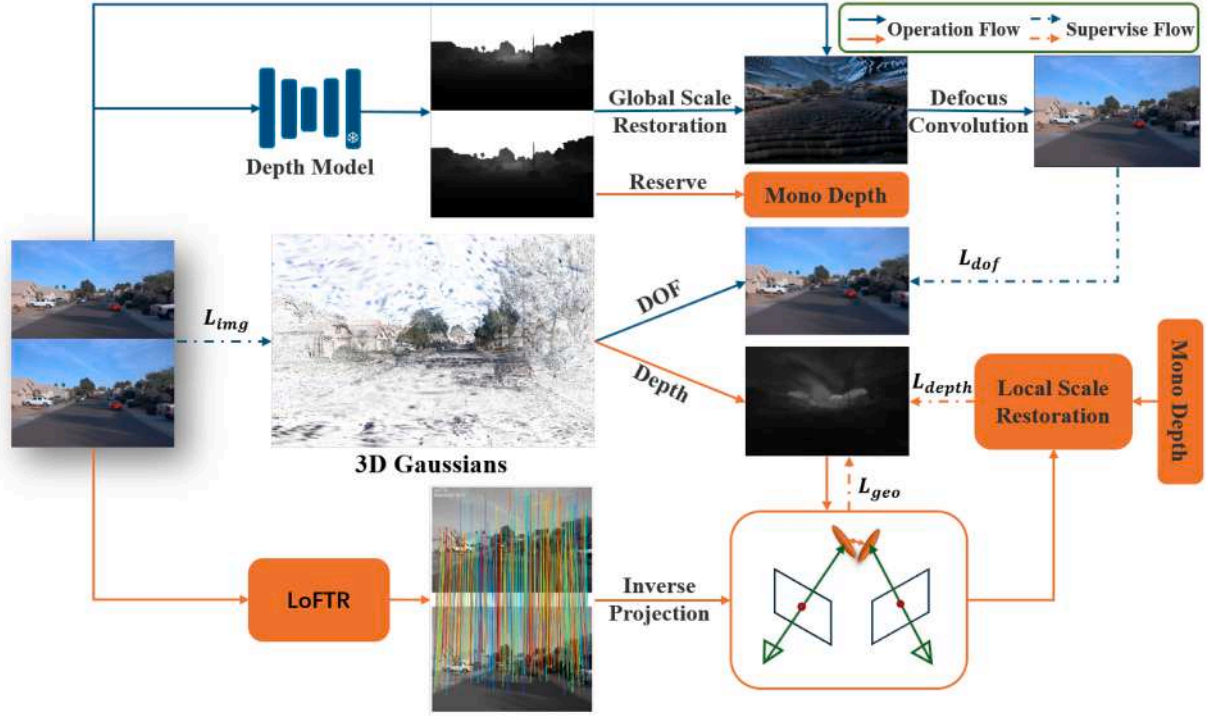


Fig. 2. Our framework consists of two core technical components: (a) Depth-of-Field Supervision (Blue Flow) addressing inaccuracies in distant scenes and difficulties in recovering structures in near-field scenes. The pipeline takes multi-view images as input, obtains scale-ambiguous depth predictions through a monocular depth estimator (e.g., Metric3D), and calculates true depth maps via a multi-view depth scale recovery algorithm. Defocus convolution is then utilized to generate defocused images from both rendered and ground truth images, with the final \mathcal{L}_{dof} loss between these defocused images supervising the 3DGS training. (b) Multi-View Consistency Supervision (Orange Flow) resolving cross-view geometric alignment issues. Initially, semi-dense feature matching is performed across multi-view images using LoFTR, minimizing the error \mathcal{L}_{geo} between 3D points corresponding to matched pixels to enhance cross-view geometric consistency. Simultaneously, a depth consistency loss \mathcal{L}_{depth} employs local depth maps recovered through least squares optimization from accurately matched points with reliable depth information to optimize the depth rendered by 3DGS.

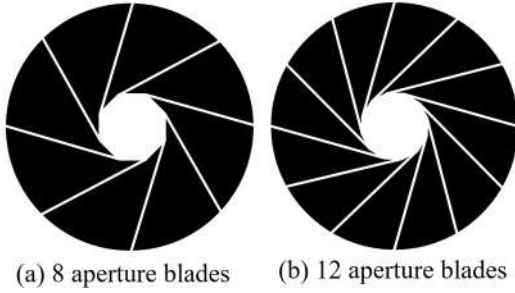


Fig. 3. Illustration of polygonal aperture mechanisms in a camera lens: (a) octagon aperture blades and (b) dodecagon aperture blades.

3.1.4. Defocus synthesis

Given an input image I , the final defocused image I_{out} is synthesized as follows:

$$I_{out}(p) = \sum_{q \in \mathcal{N}(p)} I(q) K(R(d_p))(p - q) \quad (13)$$

where $\mathcal{N}(p)$ denotes the neighborhood of pixel p , $K(R)$ represents the convolution kernel associated with the CoC size R , which is determined by the depth d_p of the scene point corresponding to pixel p .

To enhance computational efficiency while maintaining physical accuracy, we implement a separable convolution approach. Specifically, we decompose the 2D convolution into two sequential 1D convolutions along the horizontal and vertical directions:

$$I_{out} = (I * k_x) * k_y \quad (14)$$

where k_x and k_y are 1D kernel functions in the horizontal and vertical directions, respectively. This separable implementation reduces the computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(2n)$ for an $n \times n$ kernel, significantly accelerating the processing while preserving the physical fidelity of the defocus effects.

3.1.5. Depth-of-field loss

Our loss design addresses two critical and distinct aspects of the reconstruction: ensuring photometric accuracy in sharp, in-focus regions, and enforcing physical plausibility of blur in out-of-focus regions. This dual supervision is key to achieving both geometric accuracy and optical realism.

Supervision for In-Focus Regions. To optimize sharp regions, we employ a standard reconstruction loss, \mathcal{L}_{rgb} , which combines a pixel-wise L1 loss and a structural similarity (SSIM) term:

$$\mathcal{L}_{rgb} = (1 - \lambda_{dssim}) \mathcal{L}_{L1}^{(rgb)} + \lambda_{dssim} \mathcal{L}_{SSIM}^{(rgb)} \quad (15)$$

The constituent losses are defined as:

$$\mathcal{L}_{L1}^{(rgb)} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|I_{rend}(p) - I_{gt}(p)\|_1 \quad (16)$$

$$\mathcal{L}_{SSIM}^{(rgb)} = 1 - \frac{(2\mu_{rend}\mu_{gt} + C_1)(2\sigma_{rend}\sigma_{gt} + C_2)}{(\mu_{rend}^2 + \mu_{gt}^2 + C_1)(\sigma_{rend}^2 + \sigma_{gt}^2 + C_2)} \quad (17)$$

where μ and σ represent local means and standard deviations computed over 11×11 windows, with constants $C_1 = 0.01^2$ and $C_2 = 0.03^2$ preventing numerical instability during division. This component primarily supervises in-focus regions through direct pixel-wise comparison (\mathcal{L}_{L1}) and structural similarity preservation (\mathcal{L}_{SSIM}).

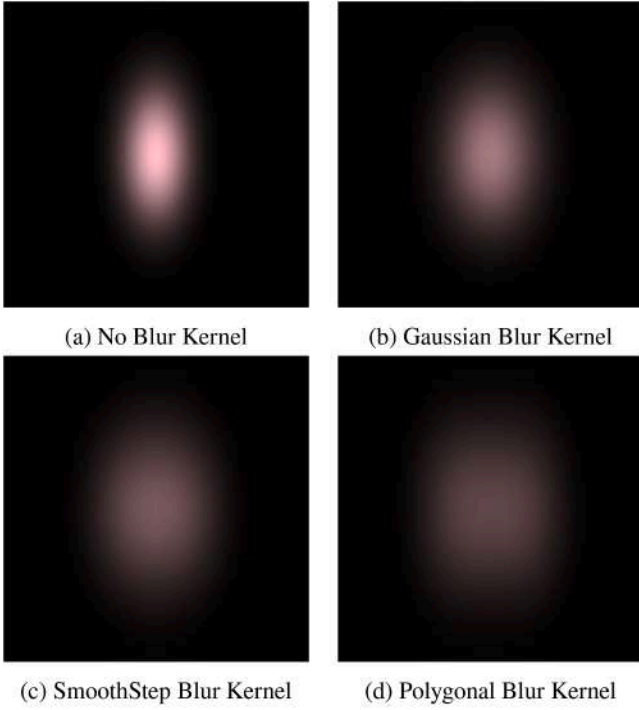


Fig. 4. Comparative analysis of defocus convolution techniques: (a) Original (no blur) provides baseline sharpness; (b) implementing radially symmetric blur via bell-shaped intensity profiles to simulate natural defocus; (c) preserving edge structures through S-curve transitions using hyperbolic tangent functions; (d) emulating optical apertures with geometric containment and radial attenuation for realistic bokeh effects.

Supervision for Out-of-Focus Regions. To specifically supervise the physically-based defocus effects, we introduce a dedicated loss term, \mathcal{L}_{dof} . This loss is computed between the ground-truth image convolved with our physics-based kernel and the rendered image similarly convolved. It shares the same structure as the in-focus loss but operates on the defocused images:

$$\mathcal{L}_{\text{L1}}^{(\text{dof})} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|I_{\text{rend}}^{(\text{dof})}(p) - I_{\text{gt}}^{(\text{dof})}(p)\|_1 \quad (18)$$

$$\mathcal{L}_{\text{SSIM}}^{(\text{dof})} = 1 - \text{SSIM}(I_{\text{rend}}^{(\text{dof})}, I_{\text{gt}}^{(\text{dof})}) \quad (19)$$

Composite Defocus Loss. These components are combined into the final depth-of-field loss \mathcal{L}_{dof} :

$$\mathcal{L}_{\text{dof}} = (1 - \lambda_{\text{dssim}}^{(\text{dof})}) \mathcal{L}_{\text{L1}}^{(\text{dof})} + \lambda_{\text{dssim}}^{(\text{dof})} \mathcal{L}_{\text{SSIM}}^{(\text{dof})} \quad (20)$$

This hierarchical structure allows for targeted optimization of focus accuracy and defocus physics, which is crucial for high-quality synthesis.

3.1.6. Depth-aware Gaussian density control

Traditional 3D Gaussian reconstruction struggles in defocused regions due to uniform density allocation, which limits the effectiveness of Gaussian points for downstream tasks (Wu et al., 2024). Our gradient-aware strategy adaptively modulates point density using optical gradients derived from the CoC physics. The proposed approach achieves three fundamental improvements over conventional methods: (1) Edge preservation through gradient-sensitive prioritization of points in depth discontinuity regions, (2) Adaptive efficiency via automatic density balancing based on quantile-based gradient statistics, and (3) Physical consistency inherited from the optical imaging model through CoC constraints. The control parameter $\tau \in [0, 1]$ governs the quality-efficiency trade-off, where higher τ values enhance reconstruction fidelity at the cost of increased computational resources, as formalized in the preservation criterion.

Gradient Computation The CoC gradient magnitude $\nabla \mathcal{L}_{\text{Dof}}$ is computed through differentiable rendering:

$$\nabla \mathcal{L}_{\text{Dof}} = \frac{\partial \mathcal{L}_{\text{dof}}}{\partial \mathbf{x}} \quad (21)$$

the definition of \mathcal{L}_{dof} refers to Section 3.1.5.

Adaptive Density Modulation A quantile-based preservation criterion prioritizes structurally critical regions:

$$\mathcal{M}_{\text{keep}} = \mathbb{I}[\|\nabla \mathcal{L}_{\text{Dof}}\| \geq Q_{\tau}(\{\|\nabla \mathcal{L}_{\text{Dof}}\|\})] \quad (22)$$

where $\tau = 0.2$ preserves the top 20% of points with the highest gradients, and Q_{τ} denotes the τ th quantile.

Pruning Criterion The pruning logic combines opacity thresholding with gradient-aware selection:

$$\mathcal{M}_{\text{prune}} = (\alpha < \alpha_{\min}) \vee [(\|\nabla \mathbf{x}\| < g_{\min}) \wedge \neg \mathcal{M}_{\text{keep}}] \quad (23)$$

where α_{\min} and g_{\min} are empirically determined thresholds for opacity and spatial gradients, respectively. The logical operators \vee (OR), \wedge (AND), and \neg (NOT) implement the compound pruning condition.

The base values for α_{\min} and g_{\min} are identical to those in the original 3DGS framework (Kerbl et al., 2023), ensuring a fair comparison. Our key contribution is the adaptive criterion $\mathcal{M}_{\text{keep}}$, which introduces a principled, physics-aware mechanism that modulates the pruning process based on geometric significance. It is important to note that our method acts as a protective supplement to the original densification logic, rather than adaptively lowering the global threshold. As detailed in Appendix A.1, our experiments confirm that simply tuning the global threshold is a fragile strategy that can lead to training instability, whereas our approach provides a stable performance improvement.

Gradient Statistics Interquartile-range weighting prevents outlier dominance in gradient accumulation:

$$w_i = \exp\left(-\frac{\|\nabla \mathcal{L}_{\text{Dof},i}\| - Q_{0.25}}{Q_{0.75} - Q_{0.25} + \epsilon}\right) \quad (24)$$

$$\hat{G}_i = \hat{G}_i + w_i \|\nabla \mathcal{L}_{\text{Dof},i}\| \quad (25)$$

with $Q_{0.25}$ and $Q_{0.75}$ denoting the 25th and 75th percentiles of gradient magnitudes.

3.2. Global scale recovery for monocular depth maps

3.2.1. Motivation for scale recovery

Monocular depth estimation methods (Almalioglu et al., 2022; Liu et al., 2024a) estimate depth values up to an unknown scale factor for each image, resulting in metric inconsistencies across multi-view observations. These scale ambiguities lead to misaligned geometries when integrating depth maps into 3DGS. Our scale recovery algorithm jointly optimizes per-image parameters to ensure view-consistent depth relationships, which are critical for realistic depth-of-field synthesis.

3.2.2. Scale recovery via geometric consistency

Given N images with camera parameters $\{(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i)\}$ and their corresponding raw monocular depth maps $\{D_m\}$, we model the scaled depth at a pixel coordinate \mathbf{p} as:

$$\tilde{D}_m(\mathbf{p}) = s_i \cdot D_m(\mathbf{p}) + b_i \quad (26)$$

where $s_i > 0$ and b_i are learnable scale and shift parameters, respectively.

Feature Matching For image pairs (I_i, I_j) , we employ the LoFTR descriptor to establish semi-dense local feature correspondences.

Theoretical Depth Ratio For valid matches $(\mathbf{p}_i^k \leftrightarrow \mathbf{p}_j^k)$, we compute:

$$\gamma_{ij}^k = \mathbf{e}_z^T \left(\mathbf{R}_{ji} \mathbf{x}_i^k + \frac{\mathbf{t}_{ji}}{Z_i^k} \right) \quad (27)$$

where $\mathbf{x}_i^k = \mathbf{K}_i^{-1} \mathbf{p}_i^k$ and $Z_i^k \approx \tilde{D}_m(\mathbf{p}_i^k)$.

Joint Optimization We define an objective function that is minimized over the set of all scale and shift parameters $\{s_i, b_i\}_{i=1}^N$ across all images. The function combines a reprojection term and a depth ratio consistency term:

$$\begin{aligned} loss = & \sum_{(i,j,k)} \underbrace{\left\| \pi(\mathbf{R}_j \tilde{\mathbf{x}}_i^k + \mathbf{t}_j) - \mathbf{p}_j^k \right\|^2}_{\text{Reprojection}} \\ & + \lambda \sum_{(i,j,k)} \underbrace{\left(\frac{D_m(\mathbf{p}_j^k)}{D_m(\mathbf{p}_i^k)} - \gamma_{ij}^k \right)^2}_{\text{Ratio Consistency}} \end{aligned} \quad (28)$$

where $\tilde{\mathbf{x}}_i^k = \tilde{D}_m^k \mathbf{x}_i^k$ and $\lambda = 0.5$. This optimization yields the optimal parameters $\{s_i^*, b_i^*\}$, which represent the globally consistent scale and shift for each view.

Depth Alignment We apply the optimized parameters $\{s_i^*, b_i^*\}$ to the raw monocular depths to obtain the final, globally aligned depth maps $\{D_a\}$:

$$D_a(\mathbf{p}) = s_i^* D_m(\mathbf{p}) + b_i^* \quad (29)$$

The aligned depths $\{D_a\}$ from Eq. (29) enable consistent blur synthesis through 3DGS as described in Section 3.1.4.

3.3. Geometric consistency supervision via feature matching

3.3.1. Depth rendering with 3DGS

For a given viewpoint, the rendered depth value $D_r(\mathbf{x})$ at pixel coordinate \mathbf{x} is computed through alpha compositing of K ordered Gaussians:

$$D_r(\mathbf{x}) = \sum_{k=1}^K \alpha_k T_k d_k \quad (30)$$

where $T_k = \prod_{l=1}^{k-1} (1 - \alpha_l)$ denotes the cumulative transmittance up to the k th Gaussian, α_k is the opacity of the k th Gaussian, and d_k represents the depth value of the Gaussian center.

Note that this formulation is the standard for depth rendering via alpha compositing, analogous to the color rendering equation in the original 3DGS framework (Kerbl et al., 2023). The sum of weights, $\sum \alpha_k T_k$, represents the total accumulated alpha. For the opaque surfaces targeted by our reconstruction, the optimization process naturally drives this sum to converge to a value near 1. Deviations from 1 are physically meaningful, representing rays viewing empty space or semi-transparent geometry.

3.3.2. Cross-view feature matching

For image pairs (I_i, I_j) , we extract semi-dense correspondences using a pre-trained LoFTR model:

$$\mathcal{M}_{ij} = \{(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})\}_{m=1}^M \quad (31)$$

where matches are filtered by a confidence threshold $\tau = 0.5$.

3.3.3. Geometric consistency loss

As shown in Fig. 5, for each matched pair $(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}) \in \mathcal{M}_{ij}$:

- 1) Project 2D matches to 3D space:

$$\mathbf{p}_i^{(m)} = \pi_i^{-1}(\mathbf{x}_i^{(m)}, D_r(\mathbf{x}_i^{(m)})) \quad (32)$$

$$\mathbf{p}_j^{(m)} = \pi_j^{-1}(\mathbf{x}_j^{(m)}, D_r(\mathbf{x}_j^{(m)})) \quad (33)$$

- 2) Compute position discrepancy:

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{M}|} \sum_{m=1}^M \|\mathbf{p}_i^{(m)} - \mathbf{p}_j^{(m)}\|_1 \quad (34)$$

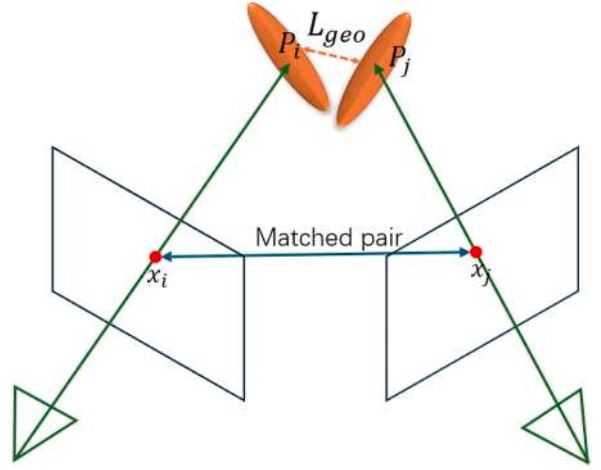


Fig. 5. Schematic diagram of geometric consistency loss calculation: Feature points matched in two views are projected into 3D space using rendered depth, and geometric constraints are imposed by minimizing the distance between these projected points.

3.4. Local scale restoration

Accurate depth estimation constitutes a fundamental requirement for effectively harnessing depth-of-field effects to strengthen structural supervision within 3D Gaussian Splatting. While monocular depth estimation offers initial depth cues, its inherent scale ambiguity prevents direct application. To address this limitation, we propose a local scale restoration framework through adaptive regional analysis.

3.4.1. Grid-based local regions

Given a rendered depth map $D_r \in \mathbb{R}^{H \times W}$ and monocular depth map $D_m \in \mathbb{R}^{H \times W}$, we decompose the image domain into a grid of local regions. To ensure our method is robust to varying image resolutions, we employ an adaptive strategy. Instead of fixing the number of grid cells (h, w) , we constrain the size of each cell, (g_h, g_w) , to be within an empirically established range $[g_{\min}, g_{\max}]$ (15 to 60 pixels). This range balances two factors: cells must be large enough for stable parameter estimation (g_{\min}) but small enough to capture local depth variations (g_{\max}). The number of cells (h, w) is then derived from the image dimensions (H, W) to meet these size constraints. Integer division is used to tile the grid, ensuring all boundary pixels are included in the final row and column of cells without clipping.

To guarantee robust parameter estimation, each grid cell must maintain a minimum of 5 valid feature points. This density constraint prevents numerical instability during local linear transformation computation, particularly addressing challenges from sparse or non-uniform feature distributions.

3.4.2. Local linear transformation

Within each grid cell $G_{i,j}$, we establish a parametric mapping between rendered depth D_r and monocular depth D_m through:

$$D_r^{(i,j)} = s_{i,j} D_m^{(i,j)} + t_{i,j}, \quad (35)$$

where $s_{i,j}$ (scale factor) and $t_{i,j}$ (translation offset) denote grid-specific transformation parameters. These parameters are optimized via **Tikhonov-regularized least squares** minimization:

$$\min_{s_{i,j}, t_{i,j}} \sum_{p \in G_{i,j}} \|D_r(p) - (s_{i,j} D_m(p) + t_{i,j})\|_2^2 + \lambda \left\| \begin{bmatrix} s_{i,j} \\ t_{i,j} \end{bmatrix} \right\|_2^2, \quad (36)$$

where p indexes valid feature points within the grid cell, and $\lambda = 10^{-6}$ serves as the regularization coefficient to ensure numerical stability in ill-posed conditions.

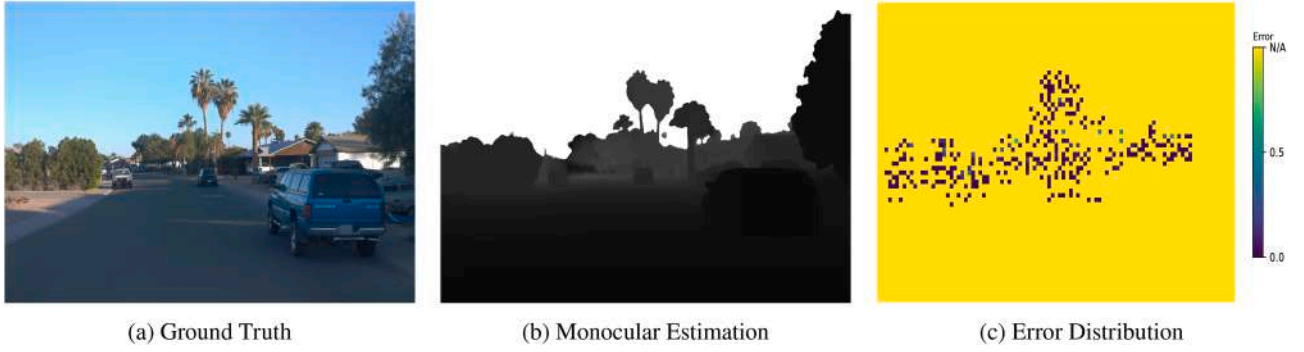


Fig. 6. Comparative visualization of depth estimation components: (a) Ground truth, (b) Monocular depth prediction from our proposed method, (c) Spatial error mapping. Error is computed only in regions with sufficient feature matches (non-yellow areas) and normalized to $[0, 1]$, where cooler colors indicate lower error. The 'N/A' label on the color bar denotes regions where error was not computed.

The closed-form solution derives from the normal equations formulation:

$$\begin{bmatrix} s_{i,j} \\ t_{i,j} \end{bmatrix} = (X^T X + \lambda I)^{-1} X^T y, \quad (37)$$

with design matrix $X = [D_m(p) \quad 1]$ containing monocular depth measurements and an intercept term, and observation vector $y = D_r(p)$ comprising rendered depth values.

3.4.3. Depth error map generation and visualization

The depth error $E_{i,j}$ for each grid cell is computed as the mean absolute difference between the rendered depth values and their corresponding transformed monocular depth estimates:

$$E_{i,j} = \frac{1}{|G_{i,j}|} \sum_{p \in G_{i,j}} |D_r(p) - (s_{i,j} D_m(p) + t_{i,j})|, \quad (38)$$

where $|G_{i,j}|$ represents the number of valid points in grid cell $G_{i,j}$, $D_r(p)$ denotes the rendered depth at point p , and $s_{i,j} D_m(p) + t_{i,j}$ is the transformed monocular depth using the optimized local linear transformation parameters.

These grid-level errors are then interpolated to the original image resolution to form a comprehensive depth error map $E \in \mathbb{R}^{H \times W}$. To enhance visualization contrast and handle grid boundaries and regions with insufficient feature points, we apply the following min-max normalization:

$$\hat{E}(p) = \begin{cases} \frac{E(p) - E_{\min}}{E_{\max} - E_{\min}}, & \text{if } E(p) \neq 1 \text{ and } E(p) \text{ is finite,} \\ 1, & \text{otherwise.} \end{cases} \quad (39)$$

where E_{\min} and E_{\max} represent the minimum and maximum error values among all valid pixels, respectively. The default value of 1 is preserved for pixels in invalid regions or grid cells with insufficient feature points.

To quantitatively analyze the spatial distribution characteristics of depth estimation errors, Fig. 6 provides a multi-modal comparison comprising three aligned representations: (a) Ground truth RGB image, (b) Monocular depth estimation results, and (c) Corresponding disparity error heatmap generated by our method.

This multi-view visualization quantitatively reveals spatial error concentration patterns and pinpoints challenging regions characterized by complex geometric configurations, occlusion boundaries, and fine-scale structural details.

3.4.4. Depth consistency loss

We design a depth consistency loss function to optimize the alignment between rendered and monocular depths:

$$\mathcal{L}_{\text{depth}} = \mathcal{L}_{\text{abs}} + \alpha \mathcal{L}_{\text{corr}}, \quad (40)$$

where α is a balancing coefficient empirically determined to regulate the influence of correlation error.

where the **absolute error term** \mathcal{L}_{abs} penalizes deviations in valid regions:

$$\mathcal{L}_{\text{abs}} = \frac{1}{|\Omega_{\text{valid}}|} \sum_{p \in \Omega_{\text{valid}}} \hat{E}(p), \quad (41)$$

with valid regions Ω_{valid} defined as pixels where both error map values are non-zero and both rendered and monocular depths are positive.

The **correlation error term** $\mathcal{L}_{\text{corr}}$ enforces consistency in invalid regions using min-max normalized depth maps:

$$\mathcal{L}_{\text{corr}} = \left| 1 - \frac{1}{|\Omega_{\text{invalid}}|} \sum_{p \in \Omega_{\text{invalid}}} \hat{D}_r(p) \hat{D}_m(p) \right|, \quad (42)$$

where $\Omega_{\text{invalid}} = \{p \in \Omega | p \notin \Omega_{\text{valid}}\}$ is the complement of the valid region, and \hat{D}_r and \hat{D}_m are min-max normalized to $[0, 1]$ as follows:

$$\hat{D}(p) = \frac{D(p) - D_{\min}}{D_{\max} - D_{\min} + \epsilon}, \quad (43)$$

with $\epsilon = 10^{-8}$ added for numerical stability.

The proposed depth consistency loss addresses three critical challenges in depth alignment. First, the explicit min-max normalization of rendered and monocular depth maps (\hat{D}_r and \hat{D}_m) to the unit interval inherently handles scale variations between different regions, eliminating the need for ad-hoc calibration. Second, the dual-term design (\mathcal{L}_{abs} for valid regions and $\mathcal{L}_{\text{corr}}$ for invalid regions) ensures robustness against uneven feature point distributions by decoupling the optimization constraints. Finally, the integration of normalized correlation in $\mathcal{L}_{\text{corr}}$ and the previously computed error map $\hat{E}(p)$ in \mathcal{L}_{abs} jointly improve numerical stability, effectively handling regions with insufficient valid matches while maintaining consistent depth relationships. These attributes are achieved with only a single hyperparameter α to balance the two loss terms.

3.5. Total loss formulation

The overall optimization objective is formulated as a weighted combination of four key components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{dof}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}} \quad (44)$$

where \mathcal{L}_{rgb} ensures photometric accuracy through RGB reconstruction error minimization, \mathcal{L}_{dof} enforces optical consistency by aligning depth-of-field effects, \mathcal{L}_{geo} maintains multi-view geometric consistency across adjacent frames through feature matching, and $\mathcal{L}_{\text{depth}}$ aligns rendered depths with monocular depth priors using our proposed depth consistency loss.

The balancing weights λ_{geo} and λ_{depth} control the relative influence of geometric and depth constraints, respectively. This formulation enables joint optimization of appearance, geometry, and optical properties within a unified framework, effectively leveraging both learned monocular depth priors and multi-view geometric constraints.

4. Experiment

4.1. Datasets

Our experiments were conducted on four datasets: Waymo (Sun et al., 2020), Mip-NeRF360 (Barron et al., 2022), SS3DM (Hu et al., 2024b), and the YouTube (Cheng et al., 2024b) dataset. Waymo dataset is a large-scale urban dataset. Mip-NeRF360 is a common NeRF benchmark. The detailed descriptions and experimental results for both the SS3DM and YouTube datasets will be provided in the Section 4.3.3.

4.2. Implementation details

Our implementation extends the original 3DGS framework (Kerbl et al., 2023) by incorporating enhanced depth-of-field controls while maintaining compatibility for fair comparisons. All experiments were conducted using the original 3DGS configuration over 30,000 iterations on NVIDIA GeForce RTX 4090 GPUs.

Key optical parameters adhere to standard photographic configurations: a 50mm focal length (equivalent to a standard lens), an $f/5.6$ aperture (providing a moderate depth of field), and a 36mm full-frame sensor. To balance defocus realism with rendering performance, we limited the maximum blur kernel size to 7×7 pixels.

When the Dynamic Focus Strategy is activated, the focus distance d_f is optimized through statistical analysis of depth distributions. Empirical evaluation demonstrates that both median depth (d_{med}) and first-tercile depth ($d_{1/3}$) exhibit comparable performance metrics. The median depth configuration is adopted as the default selection to maximize computational efficiency.

Our depth-aware density control preserves the top 20 % of Gaussians by applying a $\tau = 0.2$ quantile threshold to depth gradients, thereby prioritizing structural regions while maintaining computational efficiency.

4.3. Results

To rigorously validate our approach, we established comparative benchmarks against five state-of-the-art novel view synthesis methods: **GaussianPro** (Cheng et al., 2024a), **Instant-NGP** (Müller et al., 2022), **Mip-NeRF 360** (Barron et al., 2022), **Zip-NeRF** (Barron et al., 2023), and the baseline **3DGS** (Kerbl et al., 2023). The quantitative performance comparison between our method and these approaches is systematically presented in Table 1.

4.3.1. Quantitative comparisons

Table 1 presents the comprehensive evaluation of our method against state-of-the-art approaches on Waymo and Mip-NeRF 360

Table 1

Revised comparison with state-of-the-art methods on Waymo and Mip-NeRF 360 datasets. Our method shows significant gains over the baseline through principled adaptation. **Ours (Default)** refers to our default configuration from the original manuscript. **Ours (All)** refers to the full framework with tuned supervision weights. **Ours (Best)** reports the per-scene best PSNR achieved across all experimental configurations, demonstrating the peak performance potential of our framework under this metric.

Method	Waymo			Mip-NeRF 360		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Instant-NGP (Müller et al., 2022)	30.98	0.886	0.281	25.59	0.699	0.331
Mip-NeRF 360 (Barron et al., 2022)	30.09	0.909	0.262	27.69	0.792	0.237
Zip-NeRF (Barron et al., 2023)	34.22	0.939	0.205	28.54	0.828	0.189
3DGS (Kerbl et al., 2023)	34.04	0.942	0.224	27.21	0.815	0.214
GaussianPro (Cheng et al., 2024a)	34.37	0.945	0.210	27.92	0.825	0.208
ConsistentGaussian (Ours, Default)	35.17	0.950	0.205	27.71	0.824	0.197
ConsistentGaussian (Ours, All)	–	–	–	27.92	0.827	0.189
ConsistentGaussian (Ours, Best)	–	–	–	27.95	0.826	0.195

datasets, analyzed through three key metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) (Wang et al., 2004), and the learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). The experimental results for GaussianPro are sourced directly from its original implementation.

Analysis on Unbounded Scenes and Adaptive Supervision. The Mip-NeRF 360 dataset, comprising nine distinct scenes (five outdoor and four indoor) that present diverse and challenging unbounded environments, serves as an ideal testbed for analyzing the adaptability of supervision strategies. Unlike the more structurally homogeneous forward-facing scenes in Waymo, its diversity (e.g., from the structured ‘room’ to the texture-rich ‘treehill’) motivates a deeper, scene-specific investigation into the interplay between our framework’s components and scene characteristics. Our initial results with default parameters, while robust on Waymo, were less conclusive on this dataset, prompting this focused analysis.

As shown in the revised Table 1, our full framework with tuned weights (**Ours, All**) improves the average PSNR to 27.92 dB. Furthermore, by selecting the configuration that yielded the best PSNR for each scene (**Ours, Best**), the average PSNR is further lifted to a peak of **27.95 dB**. While this does not surpass the specialized Zip-NeRF model, it highlights a key insight: for scenes with unreliable geometric priors due to repetitive textures (e.g., ‘treehill’), a targeted application of our supervision signals is most effective. Conversely, for well-structured scenes (e.g., ‘room’), the full suite of losses yields substantial improvements. This demonstrates that our framework’s value lies not only in its strong baseline performance but also in its flexibility, providing a toolkit and principled guidelines for practitioners to achieve optimal results by adapting the supervision strategy to the scene at hand.)

Cross-Dataset Analysis The 7.25 dB PSNR disparity between Waymo (35.17 dB) and Mip-NeRF 360 (27.92 dB) reflects fundamental challenges in outdoor driving scenarios with extensive depth ranges compared to indoor scenes with constrained depth variation. Our methodology specifically targets structural inconsistencies in distant views, which contrasts with the Mip-NeRF 360 dataset’s characteristics where most sub-scenes exhibit limited depth variation. Furthermore, our analysis identifies implementation discrepancies between the Mip-NeRF 360 variants used in GaussianPro and our baseline experiments.

4.3.2. Qualitative comparisons

Our method demonstrates significant visual improvements in complex outdoor scenarios, as evidenced by comparative reconstructions on the Waymo dataset (Fig. 7). The key enhancement manifests in resolving the inherent near-far supervision conflict through depth-aware physical modeling, particularly evident in scenes containing both close-range structural elements and distant environmental features.

Geometric Consistency Enhancement As shown in Fig. 7 (red boxes), our approach maintains coherent geometric patterns across various depth layers where baseline methods exhibit structural fragmentation or spatial artifacts. This improvement stems from two synergistic mechanisms: 1) The depth-sensitive density control prioritizes representation capacity allocation to geometrically critical regions, and 2) The hybrid depth estimation framework ensures metric consistency across multi-view observations through Eq. (29). These technical components jointly address the dual challenges of preserving near-field structural details while maintaining far-field depth coherence.

Technical Implementation All visualizations strictly adhere to the rendering configurations defined in Section 4.2, with kernel parameterization following the specifications in Section 3.1.2. Depth-dependent blur synthesis is governed by the physical imaging model in Eq. (1).

4.3.3. Additional experiments

We performed additional experiments on diverse datasets to validate the generalizability of the ConsistentGaussian model (ours). Beyond the widely adopted real-world Mip-Nerf 360 and Waymo datasets

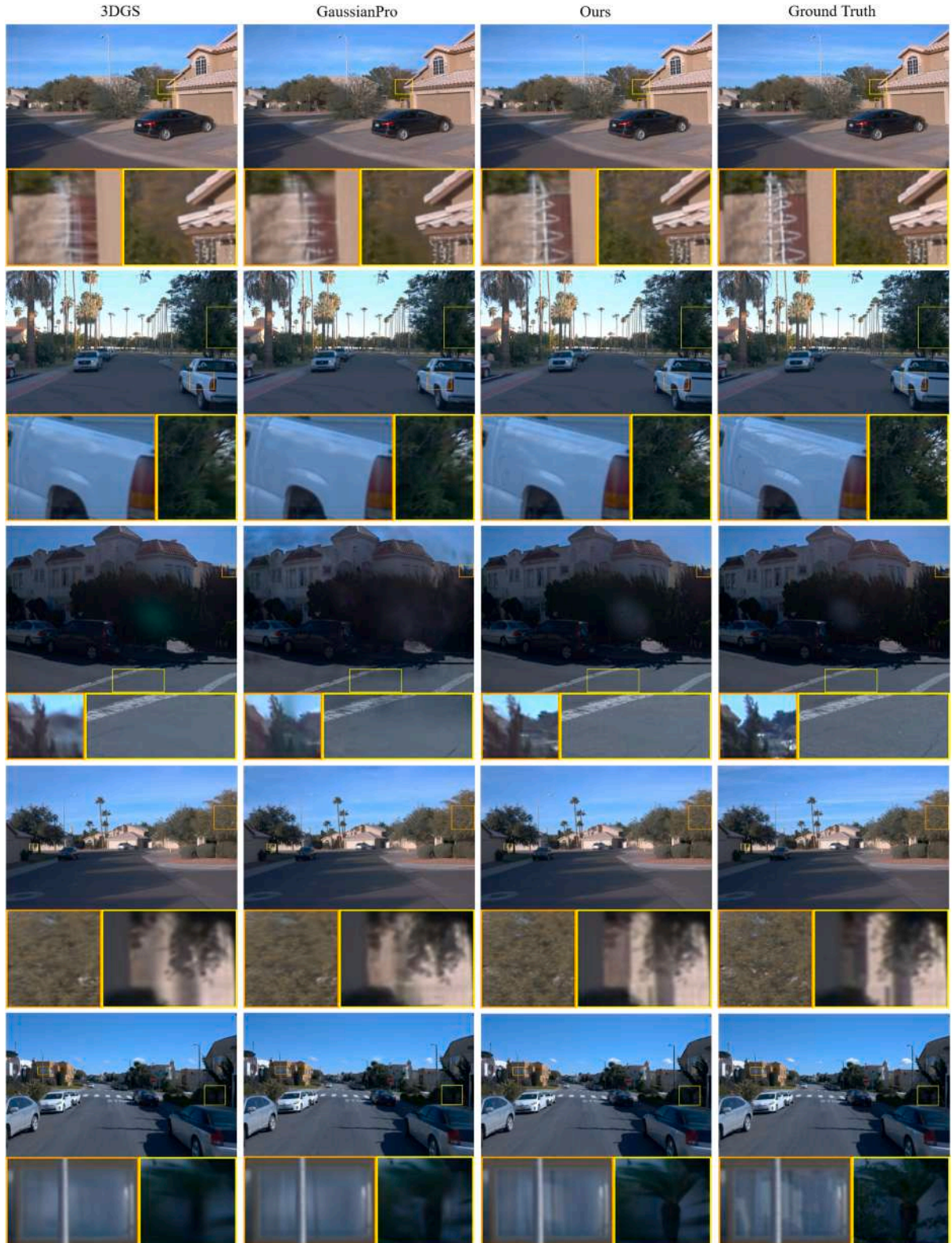


Fig. 7. Visual comparison on Waymo dataset. Our method achieves superior detail preservation in both mid-range and distant scenes compared to 3DGS and GaussianPro.

in the 3D reconstruction field, we specifically introduced three distinct data sources: the SS3DM synthetic autonomous driving dataset (Hu et al., 2024b), YouTube video sequences (Cheng et al., 2024a), and our self-collected LibraryDoF dataset captured using consumer-grade cameras. Through experimental validation of these three types of

differentiated datasets, we further corroborated the robustness advantages of our proposed method in handling multi-source heterogeneous data.

The SS3DM benchmark provides controlled-environment evaluation for street view surface reconstruction, offering CARLA-simulator-

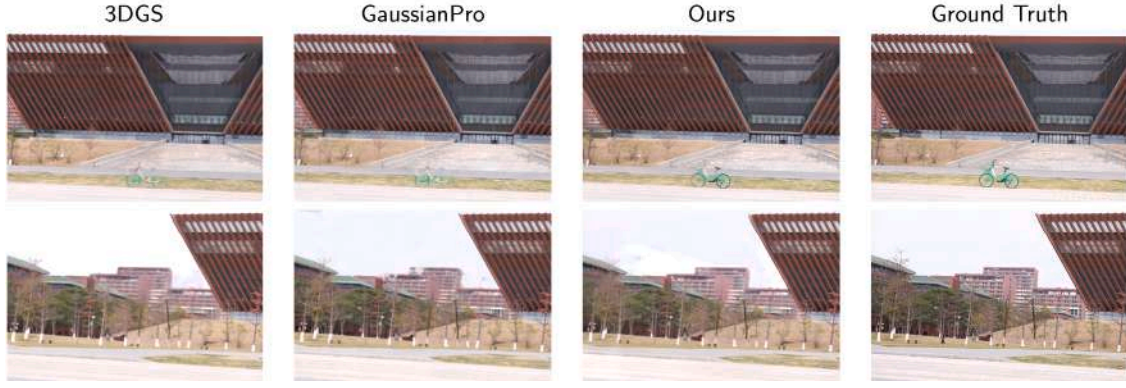


Fig. 8. Comparison of model renderings on the LibraryDoF dataset among our method, GaussianPro, and 3DGS. Our method shows distinct advantages, with more accurate detail restoration, and a more immersive visual experience.

Table 2

Comparison on SS3DM, YouTube, and LibraryDoF datasets.

Method	SS3DM			YouTube			LibraryDoF		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	30.47	0.891	0.253	34.98	0.960	0.081	23.13	0.729	0.340
GaussianPro	31.44	0.906	0.225	35.58	0.9645	0.071	22.62	0.731	0.330
Ours	33.83	0.928	0.192	36.50	0.972	0.057	24.81	0.810	0.241

generated 3D ground-truth meshes that enable precise geometric and photometric assessments. The YouTube dataset, curated by GaussianPro, contains four distinct Subscenes extracted from publicly available YouTube videos.

The SS3DM dataset contains eight virtual towns (Town01 – Town07, Town10), each featuring multiple sub-scenes. We standardize our evaluation by selecting the 150_streetsurf sub-scene in Town01 – Town07, while employing 200_streetsurf in Town10 which lacks the 150_streetsurf configuration. This strategy maintains the consistency of the evaluation while accommodating the inherent variations in the dataset.

For the YouTube dataset, each sub-scene corresponds to 360-degree aerial footage of iconic landmarks, such as the Eiffel Tower. We utilized the YouTube dataset to further validate the effectiveness of our proposed method.

To further evaluate the efficacy of our proposed method, we meticulously acquired a comprehensive dataset using a Canon R10 camera in front of the university library. The dataset comprises two distinct categories of images with approximately equal distribution and similar amounts: near-all-in-focus images and shallow depth-of-field images. The near-all-in-focus images were captured with the focus set at distant objects, simulating an all-in-focus effect, while the shallow depth-of-field images were obtained by shifting the focal plane forward under identical camera extrinsic parameters, thereby introducing defocus effects.

For model training implementation, we adhered to the splitting strategy used in the original 3DGS, with a crucial modification: we ensured that the training set contained a balanced mixture of both near-all-in-focus and shallow depth-of-field images, whereas the test set exclusively consisted of near-all-in-focus images. During the loss computation phase, the base reconstruction loss \mathcal{L}_{rgb} is only calculated when the depth-of-field strategy is disabled or the loaded image is near all-in-focus. In contrast, for shallow depth-of-field images, we computed \mathcal{L}_{dof} . When training comparative baseline models, we restricted the input to near-all-in-focus images exclusively, as these approaches lack the capability to process images with pronounced depth-of-field effects.

The scene-specific evaluation results of the synthetic SS3DM dataset are presented in Table 3, which provides a comprehensive overview

of performance metrics across its distinct scenarios. Table 2 further summarizes the mean metric comparisons among SS3DM (synthetic autonomous driving data), YouTube sequences (real-world dynamic scenes), and our LibraryDoF dataset (challenging handheld captures). Fig. 8 shows some rendering results from the LibraryDoF dataset. These cross-dataset comparisons demonstrate the consistent superiority of our method across diverse data modalities, highlighting its robust generalization capability from synthetic to real-world scenarios and across varying imaging conditions.

4.4. Ablation study

4.4.1. Experimental protocol

Our ablation analysis adopts two complementary strategies (as shown in Table 4): 1) Progressive integration of core components to isolate individual contributions, and 2) Comparative evaluation of blur kernel implementations. The first protocol sequentially adds depth-of-field supervision, gradient-aware density control, point matching loss, and monocular depth alignment to the 3DGS baseline. The second protocol substitutes the Gaussian blur kernel with Polygonal blur kernel or SmoothStep blur kernel in the full configuration. The “No Depth-of-field Loss” condition removes defocus supervision while retaining other components, quantifying its geometric regularization effect.

4.4.2. Component effectiveness

The Gaussian blur-based depth-of-field supervision establishes the foundational improvement (+0.57 dB PSNR), validating our physics-driven defocus modeling. Subsequent integration of gradient-aware density control contributes an additional +0.39 dB enhancement, demonstrating its efficacy in preserving geometrically critical regions. The marginal LPIPS improvement (0.206 vs 0.209) when adding point matching loss indicates enhanced structural consistency through multi-view correspondence constraints.

The full configuration achieves peak performance (35.17 dB PSNR), confirming the necessity of unified depth supervision. The 0.75 dB degradation in the “No Depth-of-field Loss” condition (34.42 dB vs 35.17 dB) quantitatively demonstrates defocus supervision’s critical role in geometric regularization, consistent with our theoretical analysis in Section 3.1.1.

4.4.3. Blur kernel analysis

All blur kernel implementations significantly outperform the baseline, with the Gaussian configuration achieving optimal PSNR (35.17 dB) and LPIPS (0.205). The Polygonal blur kernel generates physically accurate bokeh effects through parametric aperture modeling (Fig. 4(d)), effectively replicating real camera optics. The Polygonal and SmoothStep blur kernel implementations yield identical performance.

Table 3
Per-scene performance comparison across SS3DM towns.

Method	Town01			Town02			Town03			Town04		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	32.466	0.934	0.220	25.549	0.844	0.342	26.356	0.840	0.325	31.976	0.910	0.205
GaussianPro	32.425	0.929	0.220	31.588	0.918	0.233	31.101	0.880	0.250	31.713	0.913	0.202
Ours	34.400	0.953	0.178	34.000	0.947	0.184	35.903	0.944	0.168	34.199	0.934	0.172

Method	Town05			Town06			Town07			Town10		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	33.350	0.927	0.196	32.305	0.906	0.254	30.220	0.870	0.244	31.562	0.895	0.241
GaussianPro	31.959	0.918	0.204	32.210	0.908	0.246	29.372	0.886	0.213	31.145	0.895	0.232
Ours	34.432	0.940	0.174	34.082	0.918	0.231	30.944	0.877	0.225	32.710	0.913	0.202

Table 4
Ablation study with component integration and blur kernel comparisons on Waymo dataset.

Configuration	PSNR↑	SSIM↑	LPIPS↓	Δ PSNR	Δ SSIM	Δ LPIPS
3DGS (Baseline)	34.04	0.942	0.224	–	–	–
<i>Progressive Component Integration</i>						
+ Depth-of-field (Gaussian Blur Kernel)	34.61	0.946	0.217	+0.57	+0.004	–0.007
+ Gradient Control	35.00	0.948	0.209	+0.96	+0.006	–0.015
+ Point Match Loss	34.98	0.949	0.206	+0.94	+0.007	–0.018
Full (Gaussian Blur Kernel)	35.17	0.950	0.205	+1.13	+0.008	–0.019
<i>Blur Kernel Analysis</i>						
Full (Polygonal Blur Kernel)	35.03	0.948	0.209	+0.99	+0.006	–0.015
Full (SmoothStep Blur Kernel)	35.03	0.948	0.209	+0.99	+0.006	–0.015
No Depth-of-field Loss	34.42	0.944	0.221	+0.38	+0.002	–0.003

Implementation Guidelines The Gaussian blur kernel demonstrates superior overall performance across metrics, making it the default choice for general applications. The Polygonal blur kernel is recommended for scenarios requiring optical realism, particularly when synthesizing aperture-specific effects. The SmoothStep blur kernel provides enhanced edge preservation for high-frequency detail recovery and is suitable for post-processing applications. This parametric design framework maintains core reconstruction performance while accommodating diverse optical requirements.

4.4.4. Hyperparameter robustness

To address the practical usability of our method, we analyze the sensitivity of the key new hyperparameter introduced: the depth-gradient preservation quantile, τ (see Section 3.1.6). We evaluate the final reconstruction quality (PSNR) on three diverse scenes from the Mip-NeRF 360 dataset while varying τ over a wide range from 0.1 (preserving top 90% of gradients) to 0.9 (preserving top 10%). As shown in Table 5, the performance is highly stable across all scenes. The maximum PSNR deviation is less than 0.15 dB, demonstrating that our method is not sensitive to the precise choice of this parameter and that our default value ($\tau = 0.2$) is a robust choice for general use.

In summary, our framework is designed to be both principled and practical, minimizing the need for extensive hyperparameter tuning and ensuring its reproducibility.

5. Conclusion

We propose a physics-guided framework that enhances 3D Gaussian Splatting through depth-of-field-induced geometric supervision, addressing three fundamental challenges in neural scene reconstruction. First, our differentiable defocus convolution model physically emulates camera optics through parametric kernel design, achieving optically faithful bokeh effects while preserving computational efficiency through separable convolution operators. Second, our gradient-aware density control mechanism dynamically preserves geometrically critical structures through quantile-based pruning, particularly effective in maintaining urban scene integrity. Third, hierarchical depth alignment integrates

Table 5

Robustness analysis of the depth-gradient preservation quantile (τ) on the Mip-NeRF 360 dataset. PSNR↑ is stable across a wide range of τ values, indicating low sensitivity. The maximum performance deviation for each scene is shown in parentheses, confirming either “Highly Robust” (≤ 0.1 dB) or “Good Robustness” (0.1–0.2 dB) according to our protocol. Our default value is $\tau = 0.2$.

Scene	Preservation Quantile τ						
	0.1	0.2	0.4	0.5	0.6	0.8	0.9
bicycle ($\Delta 0.13$ dB)	25.31	25.35	25.39	25.37	25.35	25.33	25.26
garden ($\Delta 0.10$ dB)	27.50	27.58	27.53	27.56	27.55	27.58	27.60
room ($\Delta 0.14$ dB)	31.99	31.96	32.09	32.03	32.08	31.98	31.95

global monocular depth estimation with local grid-based corrections, significantly enhancing geometric consistency. Comprehensive evaluations demonstrate state-of-the-art performance in structured environments, showing enhanced rendering fidelity and improved depth estimation accuracy.

Current limitations in unbounded scene optimization suggest three research directions: 1) Temporal focus adaptation mechanisms for video-consistent dynamic scene reconstruction, 2) depth-discontinuous kernel blending strategies combining physically accurate polygonal blur kernel with detail-preserving SmoothStep blur kernel, and 3) adaptive optics simulations for extreme depth ranges using dynamic kernel scaling.

CRedit authorship contribution statement

Yu Deng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation; **Baozhu Zhao:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Conceptualization; **Junyan Su:** Writing – review & editing, Validation, Formal analysis; **Xiaohan Zhang:** Writing – review & editing; **Qi Liu:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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

Appendix A.

A.1. analysis of densification strategy

To validate our adaptive preservation mechanism (see Section 3.1.6), we compared it against the simpler alternative of globally tuning the densification gradient threshold. As shown in Table A.1, this manual tuning is a fragile process. Lowering the threshold from the 3DGS default

Table A.1

Quantitative comparison of densification strategies on the Mip-NeRF 360 “bicycle” scene. Our method is compared against baselines with our adaptive mechanism disabled.

Method	densify grad threshold	PSNR↑
Ours (Adaptive)	0.0002	25.35
Baseline (Default Threshold)	0.0002	25.26
Baseline (Lowered Threshold)	≤ 0.00015	<i>Training Failed</i>

Table A.2

Robustness of geometric supervision weights (λ) on Mip-NeRF 360 scenes. PSNR↑ is stable for standard scenes, while pathological scenes benefit from principled attenuation.

Scene	Default Weights (1x)	Attenuated (10x)	Attenuated (100x)
bicycle	25.35	25.31	25.35
bonsai	32.44	32.64	32.80

Table A.3

Comparison of our adaptive grid strategy vs. a range of fixed grid sizes on Mip-NeRF 360 scenes. The optimal fixed grid size (bolded) is highly scene-dependent. Our adaptive method consistently achieves near-optimal performance automatically. Performance is measured in PSNR↑.

Grid Strategy	bicycle	garden	room
Ours (Adaptive)	25.35	27.59	32.09
8x8	25.33	27.58	32.10
16x16	25.34	27.50	32.01
32x32	25.37	27.57	32.08
64x64	25.38	27.53	31.99
128x128	25.34	27.54	31.97

of 0.0002 leads to training failure. In contrast, our adaptive method, which operates on the same stable default threshold, not only avoids this instability but also delivers a consistent performance gain over the baseline. While it is possible to achieve marginal gains by carefully tuning the threshold, this requires a delicate, scene-specific process. Our method provides a robust path to high performance without such manual intervention.

A.2. Robustness of empirical parameters

Our framework introduces several parameters that are empirically set. Here, we provide a detailed analysis to demonstrate their robustness and validate our default choices.

A.2.1. Geometric supervision weights

The weights for the geometric consistency loss (λ_{geo}) and depth consistency loss (λ_{depth}) are set to 0.05 and 0.005 by default. As shown in Table A.2, these defaults are robust for standard scenes like ‘bicycle’, where performance varies by only 0.04 dB across a 100x change in weights. For pathological scenes with unreliable geometric priors like ‘bonsai’, our analysis confirms the effectiveness of a key adaptive criterion: strategically reducing the weights leads to significant performance gains. This provides a clear, principled guideline for tuning in such specific cases, rather than requiring blind experimentation.

A.2.2. Adaptive grid strategy parameters

Our adaptive grid strategy constrains the cell size to an empirically established range of $[g_{\min}, g_{\max}] = [15, 60]$ pixels. The rationale for this effective range is rooted in the bias-variance tradeoff. We validate this choice with two analyses. First, Table A.3 shows that the optimal fixed grid size is highly scene-dependent, making manual tuning impractical. Our adaptive method consistently performs competitively against the best fixed grid in each case, automating this choice. Second, Table A.4 shows the method’s low sensitivity to the precise values of g_{\min} and g_{\max} , confirming our default range is a robust choice.

Table A.4

Sensitivity analysis of the adaptive grid boundaries $[g_{\min}, g_{\max}]$ on the ‘bicycle’ scene. PSNR↑ is highly stable around our default setting of [15, 60].

$[g_{\min}, g_{\max}]$	PSNR↑
[5, 50]	25.26
[5, 70]	25.32
[15, 50]	25.37
[15, 60]	25.35
[25, 50]	25.37
[25, 70]	25.30

Table A.5

Sensitivity analysis of maximum blur kernel size on four diverse Mip-NeRF 360 scenes. A smaller 3x3 kernel consistently provides either superior or statistically equivalent performance. Performance is measured in PSNR↑.

Scene	Kernel Size	PSNR↑
bicycle	3x3	25.38
	7x7	25.35
flowers	3x3	22.01
	7x7	22.02
garden	3x3	27.55
	7x7	27.53
room	3x3	32.10
	7x7	31.87

A.2.3. Maximum blur kernel size

Our analysis of the ‘maximum blur kernel size’, a key parameter in our defocus model, reveals a nuanced but important finding, as shown in Table A.5. We compared our default size of 7x7 against a smaller 3x3 kernel across four diverse scenes. The results indicate that a smaller kernel is a consistently strong choice. For scenes with prominent fine structures like ‘room’ and ‘bicycle’, the 3x3 kernel provides a modest but measurable performance benefit, improving PSNR by +0.23 dB and +0.03 dB, respectively. For other scenes such as ‘garden’ and ‘flowers’, the performance is highly robust to the kernel size, with negligible differences between the two settings. This comprehensive analysis validates our default 7x7 as a robust baseline, while also providing a clear, data-driven guideline for practitioners: for maximizing fidelity, especially in scenes with intricate geometry, selecting a smaller 3x3 kernel is a principled and effective optimization.

A.3. Additional qualitative comparisons on Mip-NeRF 360

To provide further visual evidence of our method’s performance on diverse and challenging unbounded scenes, Fig. A.1 presents qualitative comparisons against state-of-the-art methods on three scenes from the Mip-NeRF 360 dataset. These visualizations complement the quantitative results in Table 1 and demonstrate our method’s superior ability to reconstruct complex geometry and fine details.

A.4. Additional qualitative results on diverse datasets

To further demonstrate the generalization capabilities of our framework, this section provides qualitative comparisons on the synthetic SS3DM dataset. As illustrated in Fig. A.2, our method consistently produces reconstructions with higher fidelity and fewer artifacts compared to baseline methods.

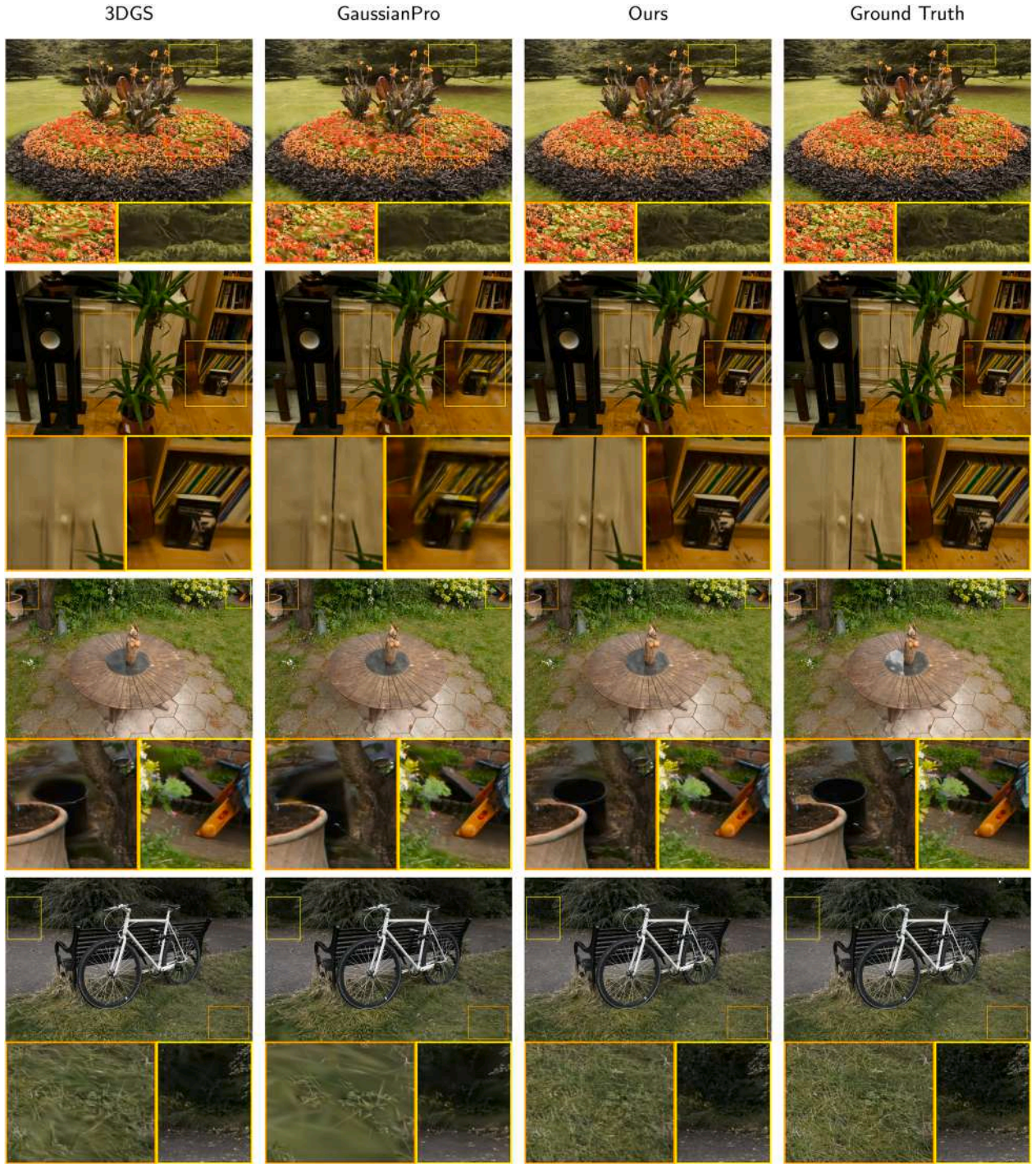


Fig. A.1. Qualitative comparison on diverse Mip-NeRF 360 scenes. (a) In the ‘flowers’ scene, our method reconstructs both intricate petal structures and distant foliage with significantly higher textural fidelity and fewer artifacts. (b) In the ‘room’ scene, our method excels at rendering fine geometric details, such as the cabinet handle, while producing sharper textures on the bookshelf and floor. (c) In the ‘garden’ scene, our method achieves a more complete reconstruction of complex objects (e.g., the bucket) and fine foliage, with a clear reduction in artifacts. (d) In the ‘bicycle’ scene, our method dramatically improves the realism of the ground plane, rendering grass and soil with superior detail and textural accuracy.

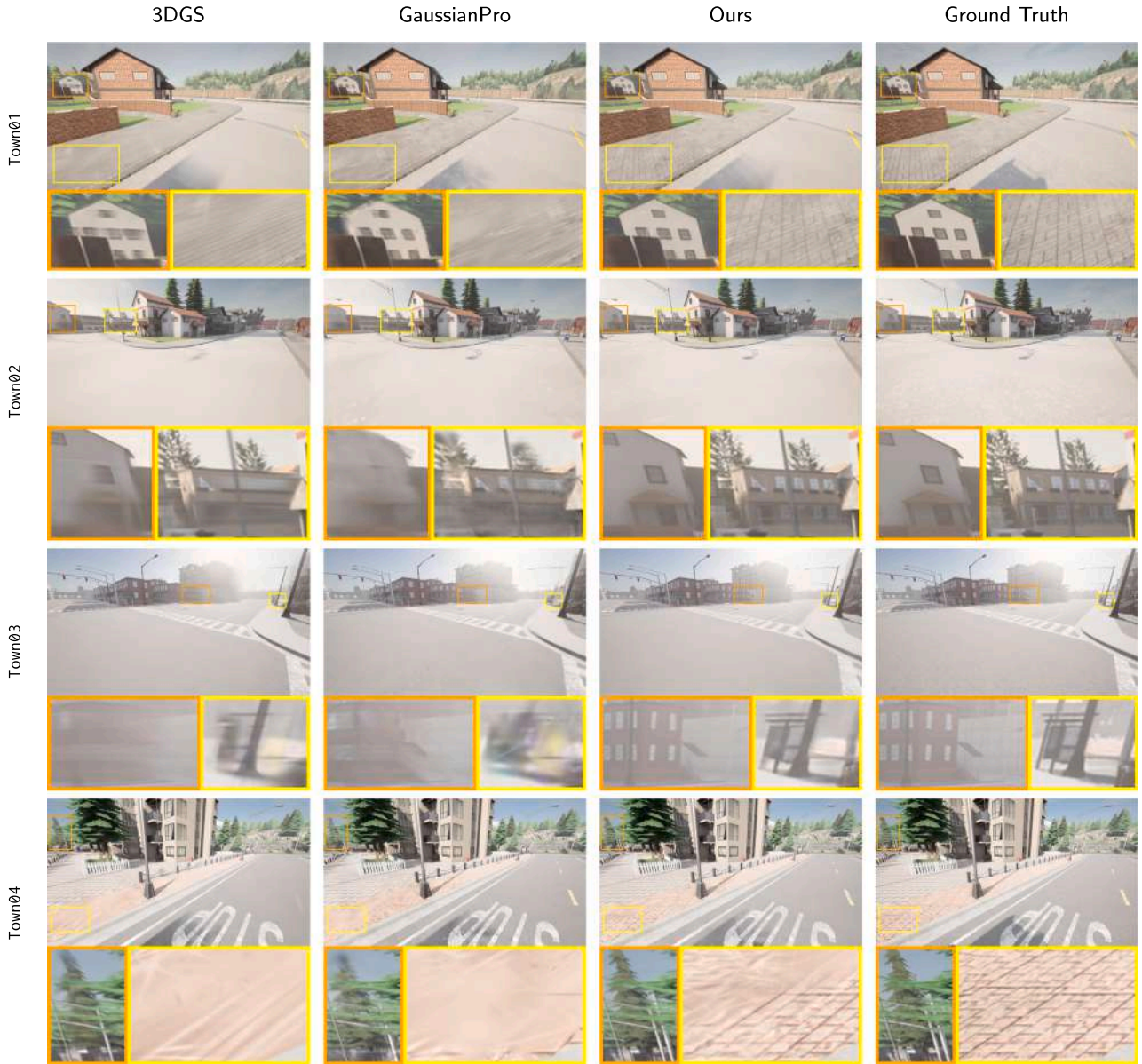


Fig. A.2. Qualitative comparison on the synthetic SS3DM dataset. Our method consistently achieves superior visual quality across various scenes. In Town01, the proposed method achieved higher fidelity in the reconstruction of building facades and road textures. In Town02, our method rendered distant architecture with enhanced realism and detail, particularly at road corners characterized by sparse views. In Town03, under complex lighting, our approach preserved sharp structural details on buildings and roadside signs, mitigating the blurring artifacts observed in baseline results. In Town04, our method generated more coherent geometry and clearer details for intricate elements, including road textures at sparsely-viewed corners and distant trees.

References

- Almalioglu, Y., Turan, M., Saputra, M. R. U., De Gusmão, P. P. B., Markham, A., & Trigoni, N. (2022). SelfVIO: Self-supervised deep monocular visual-inertial odometry and depth estimation. *Neural Networks*, 150, 119–136.
- Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2022). Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5470–5479).
- Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2023). Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19697–19705).
- Blanc, H., Deschaud, J.-E., & Paljic, A. (2024). RayGauss: Volumetric Gaussian-based ray casting for photorealistic novel view synthesis. <https://arxiv.org/abs/2408.03356>.
- Campbell, N. D. F., Vogiatzis, G., Hernández, C., & Cipolla, R. (2008). Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer vision—ECCV 2008: 10th European conference on computer vision, marseille, france, october 12–18, 2008, proceedings, part i 10* (pp. 766–779). Springer.
- Chen, G., & Wang, W. (2024). A survey on 3D Gaussian Splatting. arXiv:2401.03890.
- Cheng, K., Long, X., Yang, K., Yao, Y., Yin, W., Ma, Y., Wang, W., & Chen, X. (2024a). Gaussianpro: 3D Gaussian Splatting with progressive propagation. In *Forty-first international conference on machine learning*.
- Cheng, K., Long, X., Yang, K., Yao, Y., Yin, W., Ma, Y., Wang, W., & Chen, X. (2024b). Gaussianpro: 3D Gaussian Splatting with progressive propagation. In *Forty-first international conference on machine learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net.
- Cui, Y., & Knoll, A. (2024). Dual-domain strip attention for image restoration. *Neural Networks*, 171, 429–439.
- Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular slam. In *European conference on computer vision* (pp. 834–849). Springer.
- Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., & Wang, Z. (2023). LightGaussian: Unbounded 3D Gaussian compression with 15x reduction and 200+ FPS. arXiv:2311.17245.
- Feng, Z., Yang, L., Guo, P., & Li, B. (2023). CVRecon: Rethinking 3D geometric feature learning for neural reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 17750–17760).
- Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., & Kanazawa, A. (2022). Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5501–5510).
- Furukawa, Y., Hernández, C. et al. (2015). Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9 (1–2), 1–148.
- Giang, K. T., Song, S., & Jo, S. (2021). Curvature-guided dynamic scale networks for multi-view stereo. arXiv:2112.05999.
- Girish, S., Gupta, K., & Shrivastava, A. (2024). EAGLES: Efficient accelerated 3D Gaussians with lightweight encodings. In *European conference on computer vision* (pp. 54–71). Springer.
- Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2495–2504).
- Henzler, P., Mitra, N. J., & Ritschel, T. (2019). Escaping Plato's Cave: 3D shape from adversarial rendering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9984–9993).
- Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., & Shen, S. (2024a). Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. arXiv:2404.15506.
- Hu, Y., Wen, K., Zhou, H., Guo, X., & Liu, Y. (2024b). SS3DM: benchmarking street-view surface reconstruction with a synthetic 3D mesh dataset. In *Advances in neural information processing systems 38: Annual conference on neural information processing systems 2024, neurIPS 2024, vancouver, BC, Canada, December 10 - 15, 2024*.
- Huang, B., Yu, Z., Chen, A., Geiger, A., & Gao, S. (2024). 2D Gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference papers* (pp. 1–11).
- Ji, M., Gall, J., Zheng, H., Liu, Y., & Fang, L. (2017). Surfacenet: An end-to-end 3D neural network for multiview stereopsis. In *Proceedings of the IEEE international conference on computer vision* (pp. 2307–2315).
- Kaneko, T. (2022). AR-NeRF: Unsupervised learning of depth and defocus effects from natural images with aperture rendering neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18387–18397).
- Kar, A., Häne, C., & Malik, J. (2017). Learning a multi-view stereo machine. *Advances in Neural Information Processing Systems*, 30, 365–376.
- Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 139.
- Kerbl, B., Meuleman, A., Kopanas, G., Wimmer, M., Lanvin, A., & Drettakis, G. (2024). A hierarchical 3D Gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics (TOG)*, 43(4), 1–15.
- Lee, D., Lee, M., Shin, C., & Lee, S. (2023). DP-NeRF: Deblurred neural radiance field with physical scene priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12386–12396).
- Lee, J. C., Rho, D., Sun, X., Ko, J. H., & Park, E. (2024). Compact 3D Gaussian representation for radiance field. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21719–21728).
- Li, J., Shi, Y., Cao, J., Ni, B., Zhang, W., Zhang, K., & Van Gool, L. (2024). Mipmap-GS: Let Gaussians deform with scale-specific mipmap for anti-aliasing rendering. arXiv:2408.06286.
- Liu, X., Zhang, T., & Liu, M. (2024a). Joint estimation of pose, depth, and optical flow with a competition-cooperation transformer network. *Neural Networks*, 171, 263–275.
- Liu, Y., Luo, C., Fan, L., Wang, N., Peng, J., & Zhang, Z. (2024b). Citygaussian: Real-time high-quality large-scale scene rendering with Gaussians. In *European conference on computer vision* (pp. 265–282). Springer.
- Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., & Sander, P. V. (2022a). Deblur-NeRF: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12861–12870).
- Ma, Z., Teed, Z., & Deng, J. (2022b). Multiview stereo with cascaded epipolar raft. In *European conference on computer vision* (pp. 734–750). Springer.
- Mai, A., Hedman, P., Kopanas, G., Verbin, D., Futschik, D., Xu, Q., Kuester, F., Barron, J. T., & Zhang, Y. (2024). Ever: Exact volumetric ellipsoid rendering for real-time view synthesis. arXiv:2410.01804.
- Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P. P., & Barron, J. T. (2022). Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16190–16199).
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.
- Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4), 1–15.
- Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5), 1147–1163.
- Radl, L., Steiner, M., Parger, M., Weinrauch, A., Kerbl, B., & Steinberger, M. (2024). Stopthepop: Sorted Gaussian splatting for view-consistent real-time rendering. *ACM Transactions on Graphics (TOG)*, 43(4), 1–17.
- Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4104–4113).
- Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., & Zollhofer, M. (2019). Deepvoxels: Learning persistent 3D feature embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2437–2446).
- Snively, N., Seitz, S. M., & Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3D. In *ACM SIGGRAPH 2006 papers* (pp. 835–846).
- Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., & Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.
- Wang, C., Wolski, K., Kerbl, B., Serrano, A., Berman, M., Seidel, H.-P., Myszkowski, K., & Leimkühler, T. (2024a). Cinematic Gaussians: Real-time HDR radiance fields with depth of field. In *Computer graphics forum* (p. e15214). Wiley Online Library (vol. 43).
- Wang, P., Liu, Y., Chen, Z., Liu, L., Liu, Z., Komura, T., Theobalt, C., & Wang, W. (2023). F2-NeRF: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4150–4159).
- Wang, Y., Chakravarthula, P., & Chen, B. (2024b). DOF-GS: Adjustable depth-of-field 3D Gaussian Splatting for refocusing, defocus rendering and blur removal. arXiv:2405.17351.
- Wang, Y., He, X., Peng, S., Tan, D., & Zhou, X. (2024c). Efficient loFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21666–21675).
- Wang, Y., Yang, S., Hu, Y., & Zhang, J. (2022). Nerfocus: Neural radiance field for 3D synthetic defocus. arXiv:2203.05189.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.
- Wu, Y., Chen, X., Huang, X., Song, K., & Zhang, D. (2024). Unsupervised distribution-aware keypoints generation from 3D point clouds. *Neural Networks*, 173, 106158. <https://www.sciencedirect.com/science/article/pii/S0893608024000820>. <https://doi.org/10.1016/j.neunet.2024.106158>
- Xu, Q., & Tao, W. (2019). Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5483–5492).
- Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 767–783).
- Yu, Z., Chen, A., Huang, B., Sattler, T., & Geiger, A. (2024). Mip-splatting: Alias-free 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19447–19456).
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). 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).
- Zhou, H., Zhao, H., Wang, Q., Hao, G., & Lei, L. (2023). Miperv-MVS: Multi-scale iterative probability estimation with refinement for efficient multi-view stereo. *Neural Networks*, 162, 502–515.
- Zwicker, M., Pfister, H., Van Baar, J., & Gross, M. (2002). Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3), 223–238.