CasDeblurGS: Cascaded 2D-to-3D Multi-View Consistency for 3D Gaussian Splatting from Two Blurry Images

Haeyun Choi[1*†], Minhyuk Jang[2*], and I-Gil Kim[2]

[1] University of Virginia, Charlottesville, VA, USA
phh3ps@virginia.edu
[2] KT R&D Center, Seoul, Republic of Korea
{minhyuk.jang,i-gil.kim}@kt.com

https://haeyun-choi.github.io/Cascaded2D3D_page/

Abstract. Free-viewpoint 3D scene media is increasingly important for immersive applications, yet practical capture often suffers from severe view sparsity and motion blur. Although neural rendering has advanced sparse-view synthesis, existing blur-aware methods typically require substantial multi-view redundancy, accurate camera poses, or costly per-scene optimization. We address a stringent yet practical setting: reconstructing a coherent 3D scene from only two motion-blurred images with known intrinsics, without input-view poses, auxiliary sharp images, or per-scene test-time optimization. To this end, we propose **CasDeblurGS**, a cascaded framework that progressively recovers reliable cross-view information from local 2D correspondences to global 3D guidance. Stage 1 constructs locally reliable guidance through occlusion-aware correspondence filtering, while Stage 2 aggregates the intermediate restorations into a provisional pose-free 3D Gaussian representation whose input-view re-renders provide dense global guidance for final restoration. The resulting views enable a more coherent 3D representation and higher-quality novel-view synthesis. Experiments on real-world and synthetic Deblur-NeRF scenes show consistent gains over strong baselines, improving PSNR by 1.19 dB and 2.11 dB, respectively. Progressive ablations, cross-view correspondence visualization, and camera reprojection analysis further demonstrate improvements in both rendering quality and multi-view geometric consistency.

Keywords: Gaussian Splatting · Deblurring · Sparse Reconstruction

1 Introduction

Recent advances in neural rendering, from neural radiance fields (NeRF) [24] to 3D Gaussian Splatting (3DGS) [12], have substantially improved the fidelity and efficiency of novel-view synthesis. Together with recent feed-forward reconstruction methods, which directly recover 3D representations from sparse images

[*] Equal contribution.
[†] This work was done at KT R&D Center.

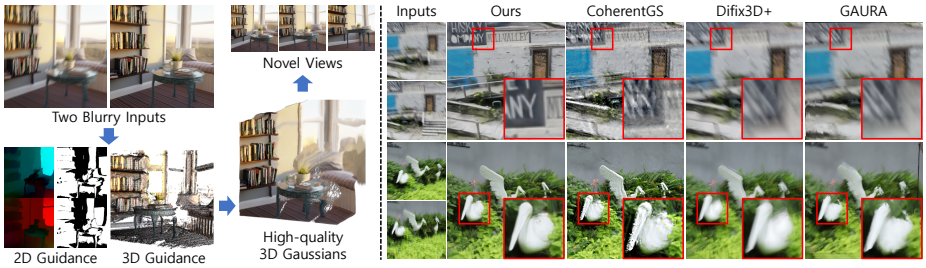

Fig. 1: (*Left*) Given two blurry images, CasDeblurGS progressively combines local 2D guidance and global 3D guidance for coherent 3D Gaussian reconstruction. (*Right*) Compared to CoherentGS [39], Difix3D+ [36], and GAURA [8], ours produces sharper details and more coherent novel views in the challenging two-view blurry setting.

[2, 5, 9, 10, 38, 41], these advances are making free-viewpoint 3D media increasingly practical for applications such as XR, immersive telepresence, and digital content creation [1, 7, 11, 13, 44].

Despite this progress, practical capture remains challenging. Dense multi-view acquisition may be infeasible because of limited capture time, sensing conditions, or platform stability, forcing reconstruction to rely on only a few observations. When these sparse inputs are further degraded by motion blur from handheld imaging, platform vibration, or low-light exposure, image details and cross-view correspondences are simultaneously corrupted, making reliable scene recovery substantially more difficult.

Recent studies have tackled this challenge by incorporating blur-aware modeling into neural rendering or jointly optimizing image restoration and scene representation [6, 14, 22, 26, 46, 48]. However, these approaches remain difficult to deploy in practice. First, they typically rely on per-scene test-time optimization and substantial multi-view redundancy to stabilize the joint estimation of blur and scene structure, often requiring 20–40 input images. Such assumptions are difficult to satisfy when only a few blurry views are available and rapid reconstruction is desired. Second, they depend on reliable geometric initialization under blur, typically in the form of accurate camera poses or coarse scene geometry. Such initialization is already difficult to recover from blurry frames using standard structure-from-motion pipelines [27]. When both view redundancy and geometric initialization are weak, the problem becomes ill-posed, often leading to unstable and spurious geometry.

These limitations become even more severe in sparse-view settings, where blur further weakens the already limited cross-view constraints. Consequently, even methods designed for sparse-view blurry reconstruction remain fragile in the extreme two-view regime [15, 39]. Under such limited observations, the joint estimation of blur and scene structure becomes severely ill-conditioned, often yielding blurry renderings, noisy geometry, and floating artifacts.

In this work, we consider a stringent yet practical setting: reconstructing a coherent 3D scene from only two motion-blurred images with known camera in-

trinsics, without input-view camera poses, auxiliary sharp images, or per-scene test-time optimization. The central difficulty is that severe blur corrupts the already limited cross-view evidence required for both local correspondence estimation and global 3D aggregation. Directly transferring unreliable 2D correspondences can introduce misaligned structures, while aggregating inconsistent observations in 3D can propagate these errors into the reconstructed scene.

Our key idea is to recover reliable cross-view information *progressively*, from local 2D correspondences to global 3D guidance. We propose **CasDeblurGS**, a cascaded 2D-to-3D multi-view consistency framework. In Stage 1, a frozen stabilizer first produces alignment-friendly observations, after which our occlusion-aware cross-view guidance module (OCGM) filters unreliable optical-flow correspondences through forward–backward consistency and constructs masked reference warps for intermediate restoration. In Stage 2, these restorations are aggregated by a frozen pose-free 3DGS backbone into a provisional 3D representation whose input-view re-renders provide dense global guidance for final restoration. The final restored views are then passed through the same frozen backbone to construct the output 3D representation for novel-view synthesis. At inference time, CasDeblurGS requires only the two blurry images and their intrinsics, with no camera poses, depth maps, external generative priors, or scene-specific optimization.

This progressive design combines complementary strengths of 2D and 3D guidance. Local warping selectively transfers fine cross-view evidence where correspondences are reliable, whereas the provisional 3D representation aggregates information from both views to provide dense scene-level feedback. Consequently, the cascade improves both the quality of the restored observations and their consistency for downstream 3D reconstruction.

Experiments on synthetic and real-world Deblur-NeRF scenes demonstrate consistent gains over strong blur-aware and restoration-based baselines. CasDeblurGS improves PSNR over the strongest competing results by 1.19 dB on real-world scenes and 2.11 dB on synthetic scenes. Progressive ablations, cross-view correspondence visualization, and camera reprojection analysis further demonstrate improvements in both novel-view synthesis quality and multi-view geometric consistency.

Our contributions are summarized as follows:

- We introduce CasDeblurGS, a pose-free feed-forward framework for 3D reconstruction from only two motion-blurred images with known intrinsics, without auxiliary sharp images or per-scene test-time optimization.
- We propose a cascaded 2D-to-3D guidance strategy that establishes locally reliable cross-view correspondences through OCGM and then exploits pose-free 3D re-rendering to provide dense global guidance for final restoration.
- We demonstrate consistent improvements on real-world and synthetic Deblur-NeRF scenes, supported by progressive ablations, cross-view correspondence visualization, and camera reprojection analysis.

2 Related Work

2.1 Radiance Fields from Blurry Images

Recent advances in NeRF and 3D Gaussian Splatting (3DGS) have spurred research on modeling camera motion blur for sharp 3D scene reconstruction. A common strategy is to simulate the blur formation process by modeling camera motion during exposure as a continuous trajectory. Existing methods parameterize this trajectory using SE(3) interpolation for linear motion [32, 37, 48], Bézier curves for complex nonlinear paths [17], or Neural ODEs for flexible temporal transformations [18, 19]. However, such trajectory optimization typically requires substantial multi-view redundancy to jointly stabilize blur and geometry estimation. In sparse-view settings, insufficient multi-view constraints make this estimation unreliable; in the extreme two-view blurry regime, standard SfM initialization becomes infeasible and cross-view correspondences are severely weakened, leading to unstable geometry estimation.

Alternatively, some methods address sparse blurry inputs using 2D structural or semantic priors, or external generative models. For instance, HQGS [20] leverages 2D edges and semantic cues, S2Gaussian [31] resolves feature-space cross-view inconsistencies during 2D super-resolution, and CoherentGS [39] augments virtual views with deblurring networks and diffusion priors. However, these methods rely heavily on local 2D cues or external priors. In extreme two-view settings with limited geometric constraints, they struggle to ensure global 3D coherence, often producing structural distortions and cross-view inconsistencies in the final renderings.

2.2 Feed-forward 3D Gaussian Splatting

While traditional 3DGS relies on per-scene optimization, recent generalizable feed-forward methods [45] directly predict 3D Gaussians from sparse views. Building on earlier generalizable NeRFs [3,33,42], 3DGS-based models now dominate this direction, using epipolar geometry, cost volumes, depth-aware matching, or hybrid volume-pixel representations for cross-view aggregation [2, 5, 29, 35, 38, 43]. Pose-free variants further remove explicit pose dependence through coarse-to-fine alignment or canonical-space prediction [9, 10, 41]. However, these methods largely assume sharp inputs and remain vulnerable to severe motion blur.

GAURA [8] jointly addresses feed-forward reconstruction and deblurring, but it assumes known camera poses, uses kernel-based synthetic blur, and remains unverified in the extreme two-view regime. More fundamentally, feed-forward aggregation depends on reliable cross-view correspondence cues, such as photometric consistency or epipolar features, which degrade severely under strong blur. This problem is amplified in two-view settings, where no additional observations are available to compensate for the lost constraints.

2.3 Multi-view Image Deblurring

Multi-view deblurring exploits geometric cues such as disparity and depth for restoration [25, 40, 49], but its reliance on correspondence estimation limits its effectiveness in extreme two-view blurry settings where high-frequency details are severely degraded and reliable camera poses and depth are unavailable.

Fundamentally, improving 2D image quality alone does not guarantee 3D consistency for novel view synthesis [16, 17, 22, 32]. Recent diffusion-based approaches aim at 3D-consistent restoration [21,23,36], yet remain vulnerable when severe blur collapses geometric matching cues. Motivated by these limitations, we propose a cascaded design that progresses from flow-based local 2D guidance to global 3D constraints via pose-free 3DGS re-rendering, enabling stable reconstruction without external generative priors at inference time.

3 Preliminaries

3.1 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) represents a scene using a set of anisotropic 3D Gaussians:

$$\mathcal{G} = \{g_m\}_{m=1}^{M}, \qquad g_m = (\boldsymbol{\mu}_m, \mathbf{q}_m, \mathbf{s}_m, \alpha_m, \mathbf{c}_m), \tag{1}$$

where $\boldsymbol{\mu}_m \in \mathbb{R}^3$ denotes the Gaussian center, $\mathbf{q}_m$ the rotation, $\mathbf{s}_m$ the scale, α_m the opacity, and $\mathbf{c}_m$ the appearance parameters, e.g., spherical harmonics coefficients [12, 41]. Given a target viewpoint v_t, the differentiable 3DGS rasterizer renders an image as:

$$\hat{I}_t = \mathrm{Render}(\mathcal{G}, v_t). \tag{2}$$

This formulation supports efficient differentiable rendering and serves as the scene representation used throughout our pipeline.

3.2 Pose-free Feed-forward 3DGS

Pose-free feed-forward 3DGS reconstructs a scene directly from sparse images and their camera intrinsics without requiring input-view camera extrinsics. In particular, NoPoSplat [41] learns a mapping:

$$\mathcal{G} = h_\eta\big(\{(I_i, K_i)\}_{i=1}^{N}\big), \tag{3}$$

where I_i and K_i denote the i-th input image and its camera intrinsics, respectively. NoPoSplat uses the first input view to establish a canonical coordinate frame and directly predicts the Gaussians associated with all input views in this shared space. This enables direct multi-view fusion without externally supplied input-view camera poses. The camera intrinsics are incorporated into the input representation to account for camera geometry and alleviate scene-scale ambiguity. The resulting Gaussian representation can then be rendered from a target viewpoint expressed in the same canonical frame using Eq. (2).

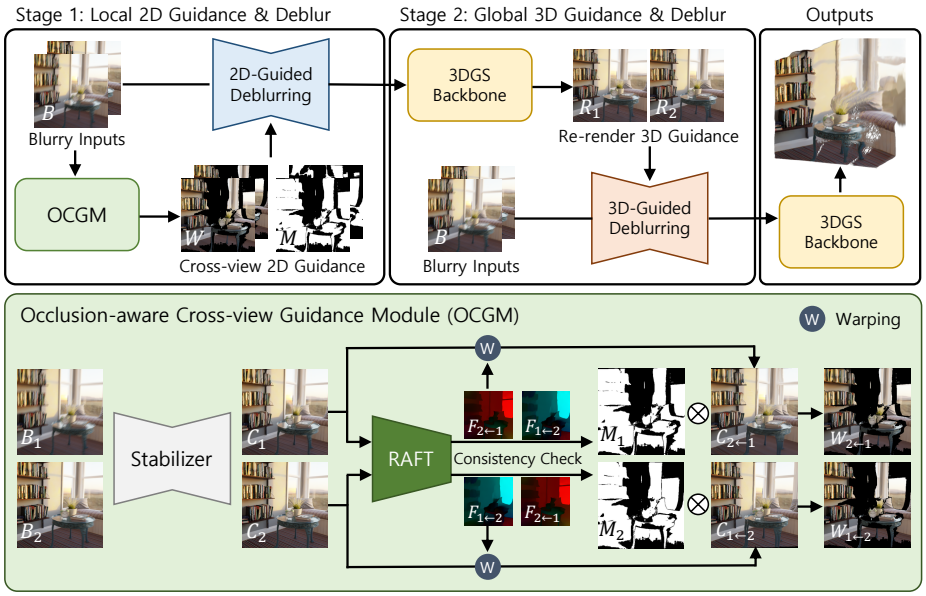

Fig. 2: Overview of CasDeblurGS. (Top) The cascade constructs local 2D guidance, then uses a frozen pose-free 3DGS backbone to provide global 3D guidance for final restoration. Final restorations feed the same backbone for novel-view synthesis. (Bottom) OCGM stabilizes inputs, estimates bidirectional RAFT flow, derives forward–backward consistency masks, and produces masked cross-view warps for Stage 1.

In this work, we instantiate the backbone in the two-view setting ($N = 2$) and keep h_η frozen throughout the pipeline. As detailed in Sec. 4, its canonical-space reconstruction and re-rendering capabilities are used for global 3D guidance and final novel-view synthesis, without scene-specific test-time optimization.

4 Method

Given two motion-blurred images and their known camera intrinsics, $\{(B_i, K_i)\}_{i=1}^{2}$, CasDeblurGS progressively recovers reliable cross-view guidance to restore the input views and construct a pose-free 3D Gaussian representation for novel-view synthesis. No ground-truth or pre-estimated input-view camera poses are provided to the framework. As illustrated in Fig. 2, the pipeline comprises two cascaded restoration stages followed by final pose-free 3D reconstruction.

Stage 1 establishes locally reliable correspondence-based guidance and uses it to produce sharper intermediate restorations with improved cross-view consistency. Stage 2 aggregates these intermediate views into a provisional 3D representation and re-renders it at the input viewpoints, providing dense global guidance for final restoration. The final restored views are then processed by the same frozen 3DGS backbone to construct the output representation for novel-view synthesis.

The stabilizer $\mathcal{S}_\phi$ [4], the RAFT-based optical-flow estimator [30], and the pose-free 3DGS backbone h_η [41] remain frozen throughout the pipeline. The 2D- and 3D-guided restoration networks, $\mathcal{D}_\theta^{2D}$ and $\mathcal{D}_\psi^{3D}$, are trained offline in a stage-wise manner. All modules are fixed at inference, and CasDeblurGS therefore requires no scene-specific test-time optimization.

The key design of CasDeblurGS is the progressive construction of reliable cross-view guidance. Stage 1 suppresses unreliable local correspondences through occlusion-aware masking, whereas Stage 2 lifts the intermediate restorations into a shared 3D representation and projects the aggregated scene information back to the image plane. Algorithm 1 summarizes the complete inference pipeline. Training and implementation details are provided in Sec. 5.1 and the supplementary material.

4.1 Stage 1: Local 2D Guidance and Deblurring

Stage 1 constructs locally reliable correspondence-based guidance from the two input views. It transfers cross-view information only where the estimated correspondence is sufficiently trustworthy. The stage consists of alignment preconditioning, occlusion-aware cross-view guidance generation, and 2D-guided restoration.

Alignment preconditioning with a Stabilizer. Directly estimating optical flow between severely motion-blurred inputs is unreliable because blur suppresses and distorts structures shared across views. We therefore first process each input using a frozen pretrained stabilizer:

$$C_1 = \mathcal{S}_\phi(B_1), \qquad C_2 = \mathcal{S}_\phi(B_2). \tag{4}$$

The stabilized views (C_1, C_2) are not treated as final restorations. Instead, they serve as alignment-friendly observations from which more reliable cross-view correspondences can be estimated.

Occlusion-aware cross-view guidance. From the stabilized views, we estimate bidirectional optical flow using a frozen RAFT model [30]:

$$F_{1\leftarrow 2} = \mathrm{RAFT}(C_2, C_1),\ F_{2\leftarrow 1} = \mathrm{RAFT}(C_1, C_2). \tag{5}$$

Here, $F_{b\leftarrow r}$ denotes a flow field defined on the base-view coordinates of b, used to map content from the reference view r into view b. Not every estimated correspondence provides reliable restoration guidance. Occlusions, motion boundaries, out-of-bounds projections, and blur-induced flow errors can introduce misaligned or duplicated structures. We therefore validate each correspondence using forward–backward consistency. For a pixel location x in the base view, its mapped coordinate in the reference view is

$$x' = x + F_{b\leftarrow r}(x). \tag{6}$$

The corresponding forward–backward residual is

$$e(x) = \|F_{b\leftarrow r}(x) + F_{r\leftarrow b}(x')\|_2 . \tag{7}$$

To avoid over-rejecting valid correspondences under large displacements, we use the motion-adaptive threshold

$$T(x) = \tau + \alpha \left(\|F_{b\leftarrow r}(x)\|_2 + \|F_{r\leftarrow b}(x')\|_2\right), \tag{8}$$

where τ provides a base tolerance for small estimation and resampling errors, and α adjusts the tolerance according to the motion magnitude. The validity mask is defined as

$$M_{b\leftarrow r}(x) = \mathbb{1}[\text{in-bounds}(x')] \cdot \mathbb{1}[e(x) \le T(x)]. \tag{9}$$

Only correspondences satisfying both the in-bounds and forward–backward consistency conditions are retained. We then construct the masked cross-view warp

$$W_{b\leftarrow r} = M_{b\leftarrow r} \odot \mathcal{W}\big(C_r, F_{b\leftarrow r}\big), \tag{10}$$

where $\mathcal{W}$ denotes backward warping. We refer to this combination of bidirectional flow estimation, forward–backward validation, and masked warping as the *occlusion-aware cross-view guidance module* (OCGM). Applying OCGM in both directions produces the guidance–mask pairs $(W_{1\leftarrow 2}, M_{1\leftarrow 2})$ and $(W_{2\leftarrow 1}, M_{2\leftarrow 1})$.

2D-guided deblurring. For each view, we concatenate the original blurry input, the corresponding masked cross-view warp, and its validity mask:

$$\begin{aligned} \tilde{S}_1 &= \mathcal{D}_\theta^{2D}\big(\text{concat}(B_1, W_{1\leftarrow 2}, M_{1\leftarrow 2})\big), \\ \tilde{S}_2 &= \mathcal{D}_\theta^{2D}\big(\text{concat}(B_2, W_{2\leftarrow 1}, M_{2\leftarrow 1})\big). \end{aligned} \tag{11}$$

Using the original blurry image as the base preserves image evidence that may be altered by the stabilizer, while the masked warp transfers complementary information from the other view. The explicit validity mask additionally indicates where the cross-view guidance can be trusted. The resulting intermediate restorations $(\tilde{S}_1, \tilde{S}_2)$ provide sharper observations with improved cross-view consistency for constructing the provisional 3D representation in Stage 2.

4.2 Stage 2: Global 3D Guidance and Deblurring

Stage 2 aggregates the intermediate restorations into a provisional 3D representation and projects the resulting scene-level information back to the image plane for final restoration. While Stage 1 transfers locally reliable information through correspondence-based masked warping, Stage 2 provides dense global guidance induced by a shared 3D representation.

Provisional 3D representation and re-render guidance. We feed the intermediate restorations $(\tilde{S}_1, \tilde{S}_2)$ and their camera intrinsics into the frozen pose-free 3DGS backbone:

$$\mathcal{G} = h_\eta\big((\tilde{S}_1, K_1), (\tilde{S}_2, K_2)\big). \tag{12}$$

We then re-render the resulting 3D Gaussian representation at the two input viewpoints:

$$R_1 = \mathrm{Render}(\mathcal{G}, v_1), \qquad R_2 = \mathrm{Render}(\mathcal{G}, v_2), \tag{13}$$

where v_1 denotes the first input viewpoint in the canonical camera frame established by the pose-free backbone, and v_2 is determined by the relative camera geometry inferred for the second view. Both viewpoints are internal to the backbone's canonical-space formulation; no ground-truth or pre-estimated input-view poses are provided to CasDeblurGS. Because $\mathcal{G}$ is jointly constructed from both intermediate restorations, the re-rendered images (R_1, R_2) encode scene information aggregated across the two views and provide dense 3D guidance over the full image plane.

3D-guided deblurring. For each view, we combine the original blurry input with its corresponding re-render guidance:

$$\begin{aligned} \hat{S}_1 &= \mathcal{D}^{3D}_\psi\big(\mathrm{concat}(B_1, R_1)\big), \\ \hat{S}_2 &= \mathcal{D}^{3D}_\psi\big(\mathrm{concat}(B_2, R_2)\big). \end{aligned} \tag{14}$$

The original blurry input preserves view-specific image evidence, while the re-rendered view supplies complementary scene-level structure aggregated through the shared 3D representation. Unlike the spatially selective correspondence guidance used in Stage 1, the re-rendered images provide dense guidance across the full image plane. Stage 2 thereby produces the final restored views $(\hat{S}_1, \hat{S}_2)$ using information jointly derived from both intermediate restorations.

4.3 Final 3D Representation for NVS

We apply the same frozen pose-free 3DGS backbone once more to the final restored views:

$$\mathcal{G}^\star = h_\eta\big((\hat{S}_1, K_1), (\hat{S}_2, K_2)\big). \tag{15}$$

The provisional representation $\mathcal{G}$ is constructed from the intermediate restorations solely to generate the re-render guidance used in Stage 2. In contrast, $\mathcal{G}^\star$ is constructed from the final restored views and serves as the output 3D representation for novel-view synthesis.

Given $\mathcal{G}^\star$, we render novel viewpoints using the differentiable 3DGS rasterizer. CasDeblurGS performs no per-scene or test-time optimization of $\mathcal{G}^\star$. Instead, the proposed cascade supplies the frozen pose-free backbone with sharper and more mutually consistent input observations, enabling it to construct a more coherent final 3D representation.

Algorithm 1 Inference Pipeline of CasDeblurGS

Require: Two blurry images and intrinsics $\{(B_i, K_i)\}_{i=1}^{2}$
Ensure: Restored images $(\hat{S}_1, \hat{S}_2)$ and 3D representation $\mathcal{G}^\star$
Note. All modules are frozen at inference.
Stage 1: Local 2D Guidance and Deblurring
$C_1 \leftarrow \mathcal{S}_\phi(B_1)$ ▷ stabilization
$C_2 \leftarrow \mathcal{S}_\phi(B_2)$
$F_{1\leftarrow 2} \leftarrow \mathrm{RAFT}(C_2, C_1)$ ▷ optical flow
$F_{2\leftarrow 1} \leftarrow \mathrm{RAFT}(C_1, C_2)$
$M_{1\leftarrow 2} \leftarrow \mathrm{FBMask}(F_{1\leftarrow 2}, F_{2\leftarrow 1})$ ▷ validity mask
$M_{2\leftarrow 1} \leftarrow \mathrm{FBMask}(F_{2\leftarrow 1}, F_{1\leftarrow 2})$
$W_{1\leftarrow 2} \leftarrow M_{1\leftarrow 2} \odot \mathcal{W}(C_2, F_{1\leftarrow 2})$ ▷ masked warp
$W_{2\leftarrow 1} \leftarrow M_{2\leftarrow 1} \odot \mathcal{W}(C_1, F_{2\leftarrow 1})$
$\tilde{S}_1 \leftarrow \mathcal{D}_\theta^{2D}(\mathrm{concat}(B_1, W_{1\leftarrow 2}, M_{1\leftarrow 2}))$ ▷ 2D guidance
$\tilde{S}_2 \leftarrow \mathcal{D}_\theta^{2D}(\mathrm{concat}(B_2, W_{2\leftarrow 1}, M_{2\leftarrow 1}))$
Stage 2: Global 3D Guidance and Deblurring
$\mathcal{G} \leftarrow h_\eta\big((\tilde{S}_1, K_1), (\tilde{S}_2, K_2)\big)$ ▷ pose-free 3DGS backbone
$R_1 \leftarrow \mathrm{Render}(\mathcal{G}, v_1)$ ▷ re-render
$R_2 \leftarrow \mathrm{Render}(\mathcal{G}, v_2)$
$\hat{S}_1 \leftarrow \mathcal{D}_\psi^{3D}(\mathrm{concat}(B_1, R_1))$ ▷ 3D guidance
$\hat{S}_2 \leftarrow \mathcal{D}_\psi^{3D}(\mathrm{concat}(B_2, R_2))$
Final 3D Representation for Novel View Synthesis
$\mathcal{G}^\star \leftarrow h_\eta\big((\hat{S}_1, K_1), (\hat{S}_2, K_2)\big)$
return $(\hat{S}_1, \hat{S}_2, \mathcal{G}^\star)$

5 Experiments

5.1 Experimental Setup

Datasets and two-view evaluation protocol. We train our restoration networks on the synthetic camera-motion-blur dataset introduced in DeepDeblurRF [6] and evaluate on the camera-motion-blur subset of Deblur-NeRF [22], which comprises five synthetic scenes and ten real-world scenes. For a fair comparison, we evaluate all methods on 85 fixed tuples across 15 scenes: 25 synthetic and 60 real-world. Each tuple contains two blurry inputs and one held-out target view for novel-view synthesis. The complete input–target mappings are provided in the supplementary material. All images are resized with preserved aspect ratio and center-cropped to 256×256 to match the pose-free 3DGS backbone [41].

Compared methods. We compare CasDeblurGS with representative sparse-view 3DGS methods. SE-GS [47] performs per-scene optimization for few-shot novel-view synthesis, while GAURA [8] is a generalizable feed-forward method for sparse blurry inputs. CoherentGS [39] targets sparse motion blur using video diffusion priors and alternating per-scene optimization.

To test whether stronger 2D restoration before reconstruction is sufficient, we construct two deblurring-then-reconstruction baselines. Specifically, we pair the

Table 1: Novel-view synthesis results on real-world and synthetic Deblur-NeRF scenes. DAVANet and Difix3D+ use the same frozen pose-free 3DGS backbone as ours. Best and second-best are highlighted in **bold** and underlined, respectively.

			Real-world			Synthetic		
Method	Pose-Free	Generalizable	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SE-GS [47]	×	×	15.90	0.398	0.443	18.76	0.560	0.338
GAURA [8]	×	✓	17.67	0.508	0.444	17.91	0.524	0.414
CoherentGS [39]	×	×	19.90	<u>0.660</u>	<u>0.292</u>	21.11	<u>0.687</u>	<u>0.259</u>
DAVANet [49]	✓	✓	19.34	0.573	0.360	19.61	0.625	0.292
Difix3D+ [36]	✓	✓	<u>19.94</u>	0.611	0.311	<u>21.30</u>	0.673	0.262
Ours	✓	✓	**21.13**	**0.700**	**0.228**	**23.41**	**0.796**	**0.165**

same frozen pose-free 3DGS backbone used in CasDeblurGS with DAVANet [49], a stereo deblurring network, and Difix3D+ [36], a diffusion-based 3D enhancement method. For these controls, the "Pose-Free" and "Generalizable" refer to the integrated pipelines rather than to the restoration models alone. All methods use the same two-view protocol and input–target mappings (see Supplementary Material), with official implementations adapted where necessary.

Implementation details. We use a pretrained NAFNet [4] as the frozen stabilizer $\mathcal{S}_\phi$ and RAFT-Large [30] as the frozen optical-flow estimator. For h_η, we use the NoPoSplat checkpoint [41] trained on RealEstate10K with two 256×256 inputs. Both restoration networks, $\mathcal{D}_\theta^{2D}$ and $\mathcal{D}_\psi^{3D}$, follow the NAFNet architecture with a base width of 64.

Stage 1 takes a 7-channel concatenation of the blurry base view, masked cross-view warp, and validity mask, whereas Stage 2 takes a 6-channel concatenation of the blurry base view and its 3D re-render guidance. We train the two restoration networks stage-wise. The Stage 1 network is trained first, after which it is fixed while training the Stage 2 network.

Each stage is optimized for 400K iterations using AdamW with an initial learning rate of 10^{-3}, weight decay of 10^{-3}, and $\beta_1 = \beta_2 = 0.9$. We use cosine learning-rate decay with a minimum learning rate of 10^{-7} and a batch size of 32 per GPU on two NVIDIA A100 GPUs. The training objective combines an ℓ_1 reconstruction loss and a VGG-based perceptual loss [28,34]. For OCGM, we set the forward–backward consistency parameters to $\tau = 1.0$ and $\alpha = 0.01$ in all experiments. Sensitivity analysis for these parameters, runtime comparisons, and additional implementation details are provided in the supplementary material.

Evaluation metrics. We evaluate novel-view renderings against the corresponding sharp reference images using PSNR, SSIM, and LPIPS. We report the average over all fixed evaluation tuples separately for the synthetic and real-world subsets. Higher PSNR and SSIM indicate better reconstruction quality, whereas lower LPIPS indicates greater perceptual similarity.

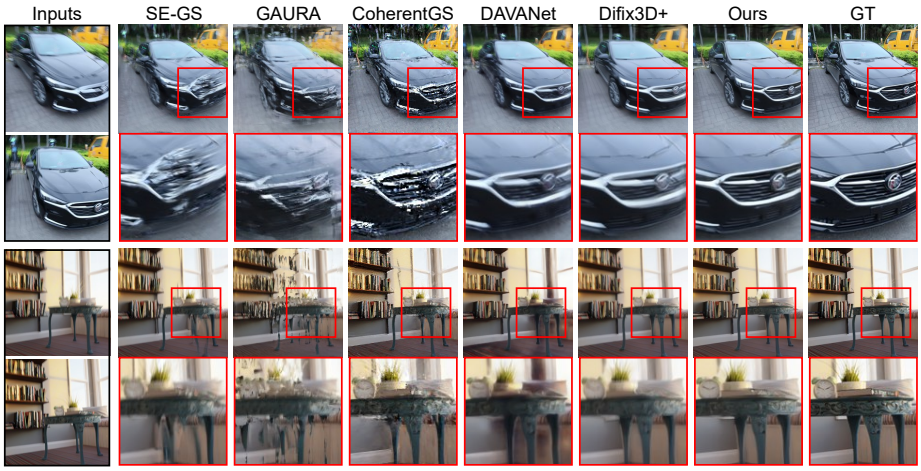

Fig. 3: Qualitative novel-view synthesis comparisons on real-world and synthetic Deblur-NeRF scenes. The upper vehicle scene is real-world and the lower indoor scene synthetic. The Inputs column shows the two blurry views. For each method, the upper image shows the rendered target view, while the lower enlarges the red-boxed region.

5.2 Comparison with Prior Methods

Quantitative results. Table 1 compares CasDeblurGS with prior methods on the real-world and synthetic Deblur-NeRF subsets. Our method achieves the best performance across all metrics on both subsets. On real-world scenes, CasDeblurGS obtains 21.13 dB PSNR, 0.700 SSIM, and 0.228 LPIPS, outperforming the best competing results by 1.19 dB, 0.040, and 0.064, respectively. The gains are larger on synthetic scenes, where our method improves PSNR by 2.11 dB, SSIM by 0.109, and LPIPS by 0.094.

The gaps also reflect differences between the baselines' native settings and our extreme two-view regime. SE-GS addresses few-shot reconstruction from sparse posed views but does not explicitly handle blur-corrupted correspondences. GAURA is trained with 8–12 source views and uses 10 views at inference, leaving substantially less multi-view redundancy for epipolar aggregation when restricted to two inputs. CoherentGS directly targets sparse motion blur but reports configurations with 3, 6, or 9 views, leaving its alternating reconstruction and generative expansion more weakly constrained under only two inputs.

The consistent gains over DAVANet and Difix3D+, which use the same frozen NoPoSplat backbone, show that applying a strong 2D restoration model before reconstruction is insufficient. Instead, the results demonstrate the benefit of progressively incorporating reliable local correspondence cues and global 3D guidance. Moreover, CasDeblurGS remains pose-free and generalizable without per-scene optimization.

Qualitative results. As shown in Fig. 3, SE-GS and GAURA retain substantial blur, while CoherentGS and the restoration-based baselines partially reduce blur but oversmooth details or introduce structural artifacts. These limitations are particularly visible around the vehicle grille in the real-world example and the table edges and legs in the synthetic example. In contrast, CasDeblurGS recovers sharper boundaries and structures that more closely match the ground-truth views. These results show that the cascade improves both sharpness and structural fidelity in novel-view renderings. Additional qualitative results are provided in the supplementary material.

5.3 Ablation Studies

Progressive contribution of the cascade. Table 2 shows consistent gains as the stabilizer, Stage 1 2D guidance, and Stage 2 3D guidance are progressively introduced. The stabilizer provides an initial improvement by producing more alignment-friendly observations. Stage 1 yields the largest incremental PSNR gain (+0.58 dB), confirming the benefit of locally reliable cross-view correspondences. Stage 2 further improves PSNR by 0.25 dB and SSIM by 0.016, indicating that dense global guidance resolves residual inconsistencies that cannot be addressed by 2D warping alone. Additional component diagnostics are provided in the supplementary material.

Table 2: Ablation on real-world scenes.

	Blurry	+Stab.	+2D	+3D
PSNR ↑	20.07	20.30	<u>20.88</u>	**21.13**
SSIM ↑	0.612	0.655	<u>0.684</u>	**0.700**

Cross-view correspondence analysis. Figure 4 visualizes the progressive improvement in cross-view consistency. On BlurBasket, the number of matches increases from 69 for the blurry inputs to 77 after stabilization, 96 after Stage 1, and 112 after Stage 2. The limited gain from stabilization reflects its independent per-view processing, whereas Stage 1 establishes more reliable local correspondences through OCGM. Stage 2 further improves correspondence recovery in

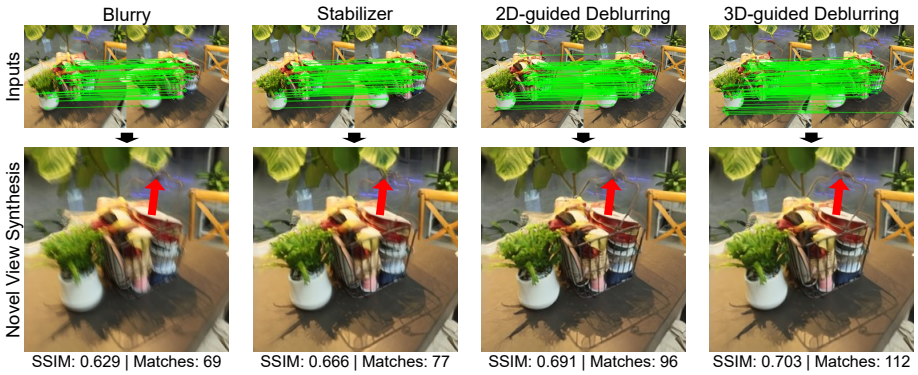

Fig. 4: Progressive ablation on real-world BlurBasket scene. Green lines denote cross-view 2D matches between input views, and red arrows highlight the basket handle.

Table 3: Camera reprojection analysis on real-world scenes.

Stage	Valid pts ↑	Reproj. err. ↓	Inliers@1px ↑	Inlier ratio ↑
Stage 1	95.1	0.2215	85.4	0.9168
Stage 2	**126.3**	**0.2001**	**113.5**	**0.9343**

regions that remain challenging for 2D warping alone, such as the basket handle highlighted by the red arrows. These results show that the cascade progressively improves both local correspondence reliability and global structural consistency.

Camera reprojection analysis. To examine whether Stage 2 improves cross-view geometric consistency rather than merely image sharpness, we conduct the reprojection analysis reported in Tab. 3. For each restored input pair, we extract mutual SIFT matches, triangulate them using the benchmark camera projection matrices, and compute symmetric reprojection errors in both views. We report averages across the evaluation pairs for valid triangulated points, median reprojection error, one-pixel inliers, and inlier ratio. The benchmark camera poses are used only for this evaluation and are never provided to our pose-free reconstruction pipeline. Compared with Stage 1, Stage 2 produces approximately 33% more valid triangulated points and one-pixel inliers while reducing reprojection error by 9.7%. The inlier ratio increases from 0.9168 to 0.9343, indicating that 3D re-render guidance improves the geometric consistency of the restored views.

6 Conclusion

We presented CasDeblurGS, a cascaded framework for pose-free 3D Gaussian Splatting from two motion-blurred images with known camera intrinsics. Our method progressively recovers reliable cross-view information: Stage 1 constructs locally trustworthy 2D guidance via occlusion-aware correspondence filtering, while Stage 2 uses pose-free 3D re-rendering to provide dense global guidance for final restoration. The resulting views enable coherent 3D reconstruction with a frozen feed-forward 3DGS backbone without input-view poses, auxiliary sharp images, or per-scene test-time optimization. Experiments on real-world and synthetic Deblur-NeRF scenes demonstrate consistent gains over sparse-view reconstruction and restoration baselines. Progressive ablations, reprojection analysis, and cross-view correspondence visualization further show that the proposed cascade improves both rendering quality and multi-view geometric consistency.

Limitations and future work. Our framework assumes known intrinsics and static scenes in a two-view setting. Its correspondence guidance may deteriorate under extreme blur, large occlusions, or limited overlap. Future work will address unknown intrinsics, dynamic scenes, and broader sparse-view settings.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25522885, Development of a World Foundation Model for Training and Deployment of Physical AI Systems).

References

1. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 6832–6852 (2025)
2. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19457–19467 (2024)
3. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14124–14133 (2021)
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European conference on computer vision. pp. 17–33. Springer (2022)
5. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European conference on computer vision. pp. 370–386. Springer (2024)
6. Choi, H., Yang, H., Han, J., Cho, S.: Exploiting deblurring networks for radiance fields. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6012–6021 (2025)
7. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
8. Gupta, V., Girish, R.S.V., Mukund Varma, T., Tewari, A., Mitra, K.: Gaura: Generalizable approach for unified restoration and rendering of arbitrary views. In: European Conference on Computer Vision. pp. 249–266. Springer (2024)
9. Hong, S., Jung, J., Shin, H., Han, J., Yang, J., Luo, C., Kim, S.: Pf3plat: Pose-free feed-forward 3d gaussian splatting. arXiv preprint arXiv:2410.22128 (2024)
10. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
11. Joshi, N., Carney, J., Kuo, N., Li, H., Peng, C., Brown, M.: Unconstrained large-scale 3d reconstruction and rendering across altitudes. arXiv preprint arXiv:2505.00734 (2025)
12. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023)
13. Khan, M., Fazlali, H., Sharma, D., Cao, T., Bai, D., Ren, Y., Liu, B.: Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8315–8321. IEEE (2025)

14. Lee, B., Lee, H., Sun, X., Ali, U., Park, E.: Deblurring 3d gaussian splatting. In: European Conference on Computer Vision. pp. 127–143. Springer (2024)
15. Lee, D., Kim, D., Lee, J., Lee, M., Lee, S., Lee, S.: Sparse-derf: Deblurred neural radiance fields from sparse view. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
16. Lee, D., Lee, M., Shin, C., Lee, S.: 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 (2023)
17. Lee, D., Oh, J., Rim, J., Cho, S., Lee, K.M.: Exblurf: Efficient radiance fields for extreme motion blurred images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17639–17648 (2023)
18. Lee, J., Kim, D., Lee, D., Cho, S., Lee, M., Lee, S.: Crim-gs: Continuous rigid motion-aware gaussian splatting from motion-blurred images. arXiv preprint arXiv:2407.03923 (2024)
19. Lee, J., Kim, D., Lee, D., Cho, S., Lee, M., Lee, W., Kim, T., Wee, D., Lee, S.: Comogaussian: Continuous motion-aware gaussian splatting from motion-blurred images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26415–26424 (2025)
20. Lin, X., Luo, S., Shan, X., Zhou, X., Ren, C., Qi, L., Yang, M.H., Vasconcelos, N.: Hqgs: High-quality novel view synthesis with gaussian splatting in degraded scenes. In: The Thirteenth International Conference on Learning Representations (2025)
21. Luo, Y., Zhou, S., Lan, Y., Pan, X., Loy, C.C.: 3denhancer: Consistent multi-view diffusion for 3d enhancement. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16430–16440 (2025)
22. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblur-nerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022)
23. Mao, Y., Wang, B., Kulkarni, N., Park, J.J.: Sir-diff: Sparse image sets restoration with multi-view diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21620–21630 (2025)
24. Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision (2020)
25. Pan, L., Dai, Y., Liu, M., Porikli, F.: Simultaneous stereo video deblurring and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4382–4391 (2017)
26. Peng, C., Tang, Y., Zhou, Y., Wang, N., Liu, X., Li, D., Chellappa, R.: Bags: Blur agnostic gaussian splatting through multi-scale kernel modeling. In: European Conference on Computer Vision. pp. 293–310. Springer (2024)
27. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Tang, S., Ye, W., Ye, P., Lin, W., Zhou, Y., Chen, T., Ouyang, W.: Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. arXiv preprint arXiv:2410.06245 (2024)
30. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)

31. Wan, Y., Shao, M., Cheng, Y., Zuo, W.: S2gaussian: Sparse-view super-resolution 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 711–721 (2025)
32. Wang, P., Zhao, L., Ma, R., Liu, P.: Bad-nerf: Bundle adjusted deblur neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4170–4179 (2023)
33. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2021)
34. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW) (2021)
35. Wei, D., Li, Z., Liu, P.: Omni-scene: Omni-gaussian representation for ego-centric sparse-view scene reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22317–22327 (2025)
36. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26024–26035 (2025)
37. Wu, R., Zhang, Z., Chen, M., Yan, Z., Zuo, W.: Deblur4dgs: 4d gaussian splatting from blurry monocular video. arXiv preprint arXiv:2412.06424 (2024)
38. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16453–16463 (2025)
39. Xu, Z., Feng, C., Li, Y., Zhao, J., Yang, J., Yu, W., Yuan, L., Tian, Y.: Breaking the vicious cycle: Coherent 3d gaussian splatting from sparse and motion-blurred views. arXiv preprint arXiv:2512.10369 (2025)
40. Yan, B., Ma, C., Bare, B., Tan, W., Hoi, S.C.: Disparity-aware domain adaptation in stereo image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13179–13187 (2020)
41. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In: International Conference on Learning Representations (2025)
42. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4578–4587 (2021)
43. Zhang, C., Xu, H., Wu, Q., Gambardella, C.C., Phung, D., Cai, J.: Pansplat: 4k panorama synthesis with feed-forward gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11437–11447 (2025)
44. Zhang, D., Li, G., Li, J., Bressieux, M., Hilliges, O., Pollefeys, M., Van Gool, L., Wang, X.: Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In: 2025 International Conference on 3D Vision (3DV). pp. 1091–1102. IEEE (2025)
45. Zhang, J., Li, Y., Chen, A., Xu, M., Liu, K., Wang, J., Long, X.X., Liang, H., Xu, Z., Su, H., et al.: Advances in feed-forward 3d reconstruction and view synthesis: A survey. arXiv preprint arXiv:2507.14501 (2025)
46. Zhao, A., Yu, P., Zhu, Z., Wei, M.: Bsgs: Bi-stage 3d gaussian splatting for camera motion deblurring. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8351–8359 (2025)

47. Zhao, C., Wang, X., Zhang, T., Javed, S., Salzmann, M.: Self-ensembling gaussian splatting for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4940–4950 (2025)
48. Zhao, L., Wang, P., Liu, P.: Bad-gaussians: Bundle adjusted deblur gaussian splatting. In: European Conference on Computer Vision. pp. 233–250. Springer (2024)
49. Zhou, S., Zhang, J., Zuo, W., Xie, H., Pan, J., Ren, J.S.: Davanet: Stereo deblurring with view aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10996–11005 (2019)