Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DCE-GS: Sparse-View Gaussian Splatting with Depth-Prior-Guided Warm-up and Cross-View Confidence-Guided EMA Supervision

Mingrong Ren

renmingrong@bjut.edu.cn

Beijing University of Technology

Wenkang liu

Beijing University of Technology

Yan Yang

Beijing University of Technology

Yanxi Li

Beijing University of Technology

Research Article

Keywords: 3D Gaussian Splatting, Sparse-view novel view synthesis, Monocular depth prior, Exponential moving average

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11285307/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DCE-GS: Sparse-View Gaussian Splatting with Depth-Prior-Guided Warm-up and Cross-View Confidence-Guided EMA Supervision

Mingrong Ren[1*], Wenkang Liu[1], Yan Yang[1], Yanxi Li[1]

[1]School of Information Science and Technology, Beijing University of Technology, Beijing, 100124, Beijing, China.

*Corresponding author(s). E-mail(s): renmingrong@bjut.edu.cn;
Contributing authors: wenkang2024@emails.bjut.edu.cn; yangyan123@emails.bjut.edu.cn; lyx182@emails.bjut.edu.cn;

Abstract

Sparse-view 3D Gaussian Splatting (3DGS) reconstruction poses a severely underconstrained problem, often yielding inaccurate geometry and degraded novel-view synthesis. Depth priors provide useful geometric guidance but may also inject noise that is more detrimental under sparse views. We present DCE-GS, a two-stage framework that progressively alleviates sparse-view overfitting. In the first stage, a monocular depth prior is used only during warm-up to establish reliable initial geometry without propagating depth errors into later optimization. In the second stage, an online model and an exponential moving average (EMA) model are maintained as complementary representations: Gaussian dropout diversifies the online model, and a topology-aware EMA aggregates these diverse configurations without an extra optimizer. Instead of assuming EMA smoothing implies geometric correctness, we construct pseudo-stereo views and use cross-view consistency to select reliable regions that supervise the online model, suppressing erroneous geometry and floating Gaussians. Only the EMA model is retained at inference. Experiments on LLFF and Mip-NeRF 360 demonstrate competitive sparse novel-view synthesis, with ablations validating both stages.

Keywords: 3D Gaussian Splatting, Sparse-view novel view synthesis, Monocular depth prior, Exponential moving average

1 Introduction

3D Gaussian Splatting (3DGS) [1] has recently emerged as an effective scene representation for a wide range of applications, including novel view synthesis, immersive visualization, robotic perception and navigation, and 3D scene reconstruction. Novel view synthesis aims to learn a scene representation from posed images and render photorealistic images from unseen viewpoints. 3DGS represents a scene using explicit anisotropic Gaussian primitives and combines efficient optimization with real-time rendering, making it particularly attractive for high-quality and efficient novel-view rendering. However, under sparse-view supervision, the highly flexible Gaussian representation is severely underconstrained. Different Gaussian configurations can reproduce the same limited observations, yet they may correspond to substantially different scene geometries. This

ambiguity often leads to overfitting, manifested as floaters, distorted structures, and degraded rendering quality at novel viewpoints.

Existing studies address this ambiguity from several complementary directions. Depth-regularized methods such as DNGaussian [2] and FSGS [3] demonstrate that monocular priors can establish useful scene geometry, but persistent supervision may also propagate structural depth errors. DropoutGS [4], DropGaussian [5], and PairDropGS [6] show that stochastic Gaussian suppression is an effective and lightweight defense against overfitting; however, perturbation alone does not provide additional geometric evidence or explicitly assess the reliability of predictions at unseen views. Model-ensemble approaches provide another valuable cue. CoR-GS [7] and SE-GS [8] exploit diversity or consistency across Gaussian states to improve generalization. Maintaining multiple gradient-optimized Gaussian fields and their optimizer states, however, increases training cost. AsymGS [9] shows that a topology-aware exponential moving average (EMA) model offers a more efficient temporal alternative. However, because the EMA model incorporates information from previous iterations, it may inherit and accumulate geometric errors that persist across training iterations.

We therefore propose DCE-GS, a two-stage framework that progressively improves sparse-view reconstruction through depth-prior-guided warm-up and dual-model confidence-guided self-supervision. In the first stage, we stabilize a single online model using global-local monocular depth guidance, establishing a reliable initial geometric structure before introducing self-supervision. After warm-up, the depth prior is removed, and a topology-aware EMA model is initialized from the online model. In the second stage, the EMA model provides a stable reference for the online model without requiring an additional gradient-optimized Gaussian field. Instead of maintaining two independently optimized Gaussian fields as in CoR-GS [7], we construct the EMA model by exponentially averaging the Gaussian parameters of the online model across training iterations. This EMA model suppresses short-term stochastic fluctuations and provides a more stable reference for self-supervision. However, Gaussian parameters from successive training iterations are highly correlated, limiting the diversity of information available to the EMA model. To increase the diversity of these parameters, we introduce Gaussian dropout into the online model during training on observed views. Different subsets of Gaussian primitives are randomly dropped across iterations, requiring the remaining primitives to repeatedly explain the input images. This perturbation encourages the online model to explore different parameter configurations, thereby providing more diverse updates for the EMA model. Even so, a stable EMA model does not necessarily imply geometrically correct predictions, as errors shared across multiple training iterations may also be preserved. We therefore identify reliable EMA predictions through pseudo-stereo projection validity and patch-level consistency checks, and use them to supervise the online model at the pseudo-left views.

We evaluate DCE-GS on LLFF and Mip-NeRF 360 under sparse-view settings. Quantitative and qualitative comparisons demonstrate competitive novel-view synthesis quality, while ablations examine the contributions of depth-guided warm-up and confidence-guided EMA supervision. Our main contributions are as follows.

- We introduce DCE-GS, an efficient two-stage framework for sparse-view 3D Gaussian Splatting. The first stage is a monocular depth-prior-guided warm-up that uses a scale- and shift-invariant global-local Pearson objective to establish reliable geometry, after which the prior is removed to avoid propagating its uncertainty.
- The second stage maintains a gradient-optimized online model and a topology-aware EMA model. Gaussian dropout diversifies the online model as a representation-level regularizer, and the EMA model is updated from the online model rather than optimized independently.
- Within this second stage, we propose EMA-guided confidence-aware self-supervision, in which the EMA model generates pseudo-stereo views and only regions validated by projection validity and patch-level agreement supervise the online model.

2 Related Work

2.1 Novel View Synthesis

Novel view synthesis renders photorealistic images from unseen viewpoints given posed observations. Neural Radiance Fields (NeRF) [10] represent scenes as continuous volumetric radiance fields with remarkable quality, and subsequent work improves anti-aliasing and unbounded-scene modeling through multi-scale representations [11]. A complementary line learns cross-scene priors by conditioning on image features [12], aggregating source-view evidence at rendering time [13], or incorporating multi-view geometry [14, 15]. These methods reduce target-scene observations but require cross-scene pre-training, while per-scene NeRF optimization remains expensive due to dense ray sampling.

More recently, 3D Gaussian Splatting (3DGS) [1] represents a scene with anisotropic Gaussian primitives rendered through differentiable rasterization, enabling fast optimization and real-time rendering. Follow-up work addresses aliasing [16], deblurring [17], learned anchors [18], and surface-oriented representations [19, 20]. These methods still rely on adequate multi-view coverage; under sparse observations the primitives can overfit training cameras and produce inaccurate reconstructions at unseen viewpoints.

2.2 Sparse Novel View Synthesis

Sparse novel view synthesis is ill-posed because limited observations under-constrain scene geometry and appearance. Early NeRF-based methods inject priors or regularization, e.g., semantic consistency [21], patch-based regularization from unobserved viewpoints [22], entropy minimization [23], frequency control [24], sparse SfM supervision [25], and monocular depth ranking [26]. With 3DGS, similar ideas extend to explicit Gaussians: depth-regularized methods include DNGaussian [2], FSGS [3], CoherentGS [27], SparseGS [28], and LoopSparseGS [29], while reliability-aware supervision is explored by CDGS [30], RDG-GS [31], and NexusGS [32]. Semantic guidance is likewise adopted [33, 34]. These methods demonstrate the value of geometric priors but also motivate limiting inaccurate monocular estimates.

Beyond geometric priors, another line regularizes optimization dynamics. CoR-GS [7] and MMGS [35] exploit disagreement across multiple Gaussian fields, while DropoutGS [4], DropGaussian [5], and PairDropGS [6] randomly suppress Gaussian primitives to alleviate training-view overfitting. Relatedly, consistency regularization encourages stable predictions under perturbations: Temporal Ensembling [36] and Mean Teacher [37] average predictions or model weights to provide smoother supervision without optimizing a second model, and SE-GS [8] exploits uncertainty from perturbed predictions.

Closest to ours, AsymGS [9] introduces a dynamic EMA Proxy to replace a second independently optimized 3DGS model, reducing cost while preserving consistency. In contrast, our EMA model provides pseudo signals for unseen views rather than acting as a proxy for another model. Moreover, these signals are not used directly; their reliability is first evaluated through projection validity and patch-level agreement, and only validated regions supervise the online model.

3 Method

The pipeline of the proposed DCE-GS is illustrated in Fig. 1. Our framework progressively alleviates overfitting and inaccurate geometry in sparse-view 3D Gaussian Splatting through a two-stage strategy. In the first stage, a depth-prior-guided warm-up is introduced to establish a reliable initial geometric structure for subsequent sparse-view optimization. In the second stage, the online model and the EMA model are maintained as two complementary Gaussian models. The online model is optimized via gradient descent, and Gaussian dropout is employed to introduce diversity into the optimization. The EMA model aggregates these diverse states into a more stable representation without requiring an additional optimizer. To provide reliable self-supervision for the online model, we construct pseudo-stereo views and perform cross-view consistency validation with the EMA model. Only geometrically and photometrically consistent regions are retained as reliable pseudo-view supervision, which helps suppress erroneous geometry and floating Gaussians caused by sparse-view overfitting. We describe the first stage in Sec. 3.1 and divide the second

stage into topology-aware EMA model construction (Sec. 3.2) and cross-view confidence-guided supervision (Sec. 3.3).

3.1 Depth-Prior-Guided Warm-up

Existing depth priors can provide useful geometric guidance, but their prediction errors and uncertainty may also impose inaccurate geometric constraints, particularly under sparse-view supervision. To exploit their geometric benefits without allowing such errors to affect later optimization, we introduce a Depth-Prior-Guided Warm-up for a single online model $\mathcal{G}^O$ to establish a reliable initial geometric structure, after which the depth prior is removed and the reconstruction is further refined in the second stage using pseudo-view stereo supervision.

For a training camera C, let I be its ground-truth image and D^m the monocular depth prior pre-computed with Marigold [38]. Let $(\hat{I}, \hat{D}, A)$ denote the color, depth, and accumulated alpha rendered by $\mathcal{G}^O$. We compute the alpha-normalized expected depth $D(p) = \hat{D}(p)/(A(p)+\epsilon)$, where p is a pixel and ϵ is a small constant for numerical stability. This normalization prevents variations in accumulated opacity from being interpreted as depth changes.

Monocular depth is generally defined only up to an affine transformation. We therefore measure its agreement with the rendered depth using the Pearson correlation coefficient. For two vectors x and y sampled over the same valid pixels, we define

$$\rho(x,y) = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2}\sqrt{\sum_j (y_j - \bar{y})^2} + \epsilon}, \quad (1)$$

In Eq. 1, j indexes the samples and $\bar{x}$ and $\bar{y}$ are their respective means. Let Ω denote the set of valid image pixels and let a subscript indicate restriction to the corresponding set. The global term in Eq. 2 preserves the overall scene layout,

$$\mathcal{L}^g_{\text{dep}} = 1 - \rho(D_\Omega, D^m_\Omega)\,, \quad (2)$$

whereas the local term refocuses the optimization on fine-scale depth changes. Given a collection $\mathcal{B}$ of valid image patches, it is defined in Eq. 3:

$$\mathcal{L}^l_{\text{dep}} = \frac{1}{|\mathcal{B}|}\sum_{B\in\mathcal{B}} [1 - \rho(D_B, D^m_B)]\,. \quad (3)$$

This global-local formulation is invariant to the unknown scale and shift of the monocular prior while retaining local geometric details, following the motivation of global-local depth regularization [2].

The photometric objective $\mathcal{L}_{\text{rgb}}$ follows standard 3DGS and combines an L_1 term with D-SSIM. The complete warm-up objective is given in Eq. 4:

$$\mathcal{L}_{\text{warm}} = \mathcal{L}_{\text{rgb}} + \lambda_d \left(\mathcal{L}^g_{\text{dep}} + \lambda_{\text{loc}}\mathcal{L}^l_{\text{dep}}\right). \quad (4)$$

Here λ_d controls the overall depth regularization and λ_{loc} balances its local component. We use T_w to denote the iteration at which warm-up ends. At T_w, the depth loss is disabled and the EMA model is initialized by copying the online model. This avoids incorporating highly unreliable early updates into the EMA model and provides a reliable basis for subsequent EMA updates.

3.2 Topology-Aware EMA Model Construction

After the warm-up stage, the Gaussian representation is further optimized from the observed images. Under sparse-view supervision, however, the limited observations may not sufficiently constrain the reconstruction, making the online model sensitive to optimization fluctuations. Although an independently optimized second Gaussian field could provide complementary predictions, it would require an additional set of trainable parameters. We therefore maintain an online model $\mathcal{G}^O$ and an EMA model $\mathcal{G}^E$. $\mathcal{G}^O$ is optimized through gradient-based updates, while $\mathcal{G}^E$ is updated from the online model to provide a more stable representation without introducing another gradient-optimized Gaussian field. Specifically, at T_w, the EMA model is initialized by copying the online model, giving $\mathcal{G}^E_{T_w} = \mathcal{G}^O_{T_w}$. The EMA model has no optimizer and does not receive gradients. After each online-model optimization step, its parameters are updated as in Eq. 5:

$$\theta^E_t = \alpha\theta^E_{t-1} + (1-\alpha)\theta^O_t, \quad (5)$$

where θ^O_t and θ^E_t denote the parameters of the online and EMA models, respectively, and α is the EMA decay coefficient [9].

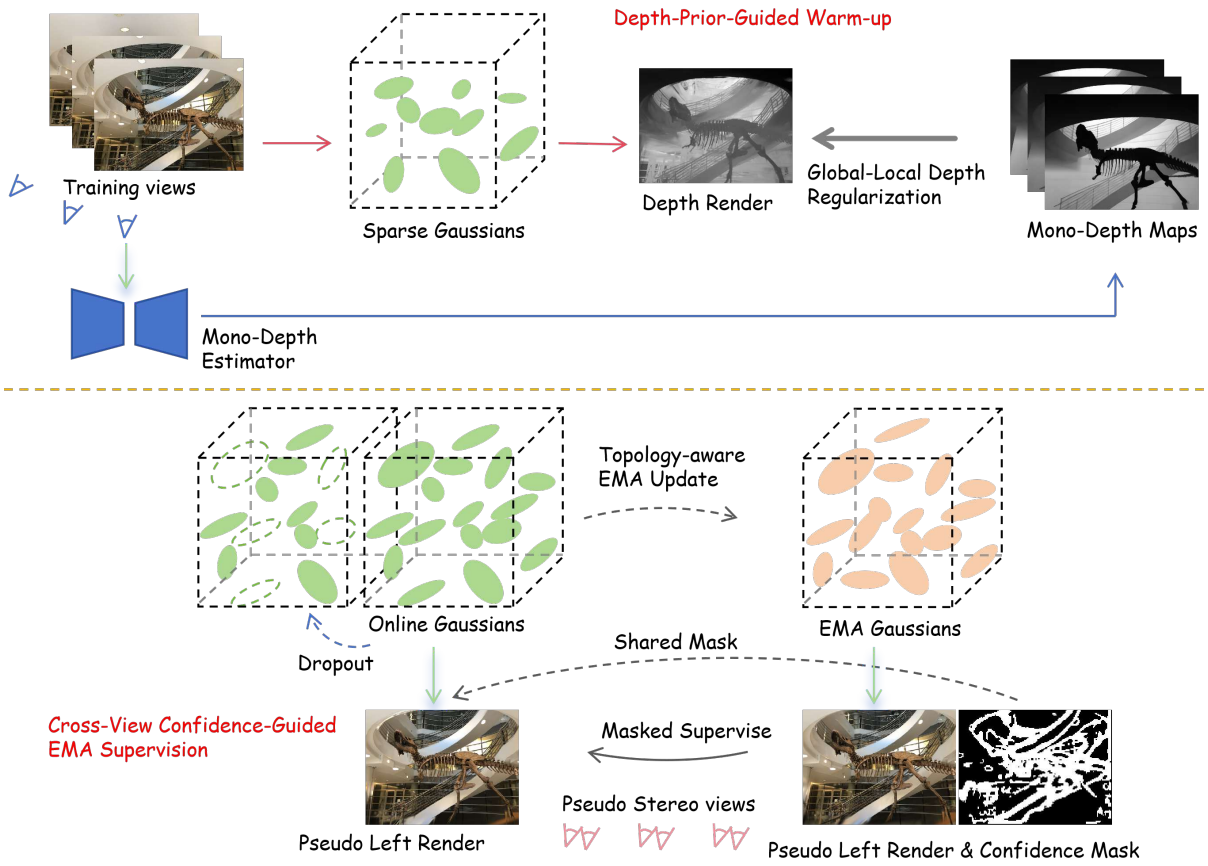

Fig. 1 Overview of the proposed two-stage framework. During the depth-prior-guided warm-up stage, a single online model is optimized with photometric supervision and global-local monocular depth regularization, where the monocular prior is pre-computed by a mono-depth estimator. When warm-up ends, the topology-aware EMA model is initialized by copying the online model and is subsequently synchronized with the online model through topology-aware EMA updates. In the second stage, Gaussian dropout is applied to the online model on observed views to diversify its optimization states. At pseudo views, only the EMA model renders the constructed stereo pair to identify projection-valid centers and cross-view-consistent patches, which yields the shared confidence mask. The online model is also rendered at the pseudo-left view, without dropout, and receives masked EMA supervision over the patches selected by this shared mask.

Since the EMA update is performed element-wise on corresponding Gaussians, the online and EMA models must maintain a consistent Gaussian correspondence throughout optimization. However, adaptive density control continuously changes the Gaussian topology through cloning, splitting, and pruning. To preserve the correspondence, we synchronize these topology changes between the two models. Specifically, every clone, split, or prune operation performed on the online model is mirrored in the EMA model. For newly created Gaussians, the corresponding EMA Gaussians are initialized from the online model and then updated through the EMA rule in subsequent iterations.

Though EMA alone may reduce short-term fluctuations in the online model, its effect is limited when consecutive updates are highly similar. To provide more diverse updates for the EMA model to accumulate, we introduce Gaussian dropout for the online model. At each iteration, a Bernoulli mask randomly removes a subset of Gaussians, and the remaining Gaussians are optimized to explain the observed image. As the dropout mask changes across iterations, the online model receives different optimization perturbations and is encouraged to rely on different subsets

of Gaussians rather than repeatedly adapting to the same configuration. After each optimization step, the EMA model is updated using the complete set of optimized Gaussian parameters. It therefore accumulates information from different dropout configurations over multiple iterations, rather than following any particular dropped subset. However, prediction stability alone is insufficient to establish geometric correctness, motivating the cross-view consistency test introduced next.

3.3 Cross-View Confidence-Guided EMA Supervision

In the EMA model, errors that persist across successive online-model updates may also be preserved by the EMA update. Therefore, instead of using all EMA predictions as supervision, we evaluate their reliability from nearby viewpoints. A reliable surface should produce consistent appearance and geometry when viewed from slightly different camera positions, whereas floaters and other erroneous structures are more likely to produce inconsistent projections. Based on this observation, we construct a pair of nearby pseudo views and use their cross-view consistency to select reliable EMA predictions.

We use each pair of adjacent training cameras to define an interpolation interval. Given their camera-to-world rotations (R_i, R_{i+1}) and centers (c_i, c_{i+1}), we obtain a pseudo-left camera C_l by spherical linear interpolation of rotation and linear interpolation of position [39],

$$\begin{aligned} R_l &= \mathrm{SLERP}(R_i, R_{i+1}, \beta), \\ c_l &= (1-\beta)c_i + \beta c_{i+1}, \qquad T_l = -R_l^{\mathsf{T}} c_l, \end{aligned} \tag{6}$$

In Eq. 6, $\mathrm{SLERP}(\cdot)$ denotes spherical linear interpolation of rotations, R_l and c_l are the camera-to-world rotation and center, T_l is the world-to-camera translation, and β is the interpolation coefficient. We translate C_l along its local horizontal axis by a signed baseline b to obtain the pseudo right camera C_r, while preserving rotation and intrinsics. Its magnitude follows a desired disparity and the median valid EMA depth, and its randomized sign avoids directional bias. This construction provides a rectified pseudo-stereo pair for cross-view consistency checking.

We only use the EMA model to evaluate pseudo-view reliability, so that the selection of reliable supervision is independent of the current online-model parameters. Without gradients, the EMA model renders color, alpha-normalized depth, and accumulated alpha as (I_l^E, D_l^E, A_l^E) at C_l and (I_r^E, D_r^E, A_r^E) at C_r. Superscript E denotes EMA predictions and subscripts l and r identify the views. For each pixel in the pseudo-left view, D_l^E is used to back-project the pixel into 3D and project it onto the pseudo-right view. The corresponding right-view color, depth, and alpha are then obtained by bilinear sampling:

$$\widetilde{I}^E_{r\to l} = \mathcal{W}\big(I_r^E, D_l^E, P_l, P_r\big)\,, \tag{7}$$

In Eq. 7, $\mathcal{W}$ is differentiable inverse warping and P_l and P_r are the two camera projection matrices. A correspondence is considered valid only when the projected location is within the image bounds, both depths are positive and finite, and both A_l^E and the warped right-view alpha exceed a predefined threshold. These conditions define the valid correspondence mask used for subsequent confidence estimation, while D_r^E and A_r^E are not involved in the photometric error.

We next validate local appearance around each candidate center rather than directly supervising every accepted pixel [40]. For a candidate q, let P_q be its 5×5 neighborhood and $S_q \subseteq P_q$ the pixels that pass the geometric checks above. We reject patches whose valid coverage $|S_q|/25$ is too low. For the valid pixels in S_q, we average the left RGB values and their right-to-left warped counterparts to form two pooled RGB descriptors. Their cosine discrepancy measures the relative RGB consistency between the two descriptors. Since cosine discrepancy is less sensitive to absolute RGB differences, we further compute the mean per-pixel photometric error, which directly measures absolute RGB discrepancies that may be hidden by descriptor pooling. The mean patch photometric error is defined in Eq. 8:

$$e_{\mathrm{pho}}(q) = \frac{1}{3|S_q|} \sum_{p\in S_q} \left\| I_l^E(p) - \widetilde{I}^E_{r\to l}(p) \right\|_1 . \tag{8}$$

The factor 3 averages over RGB channels. A patch is accepted only when all three criteria are satisfied. It has sufficient valid coverage, its cosine

difference is below the descriptor threshold, and $e_{\text{pho}}(q) < \tau_{\text{pho}}$, where τ_{pho} is the photometric threshold. This is a hard decision rather than a continuous loss weight.

For each candidate patch center q, we obtain a binary confidence indicator $M_{\text{conf}}(q)$ that determines whether the corresponding patch is used for pseudo-view supervision. At C_l, the online model produces a dropout-free rendering I_l^O, while I_l^E is the gradient-free reference. No online-model right-view rendering is needed. Both pseudo-left renderings use complete Gaussian sets, since dropout remains confined to observed-view optimization and would otherwise inject mask-dependent noise into pseudo-view supervision. The hard-masked confidence loss is defined in Eq. 9:

$$\mathcal{L}_{\text{conf}} = M_{\text{conf}} \cdot \Big[\left\| I_l^O - I_l^E \right\|_1 + \lambda_{\text{ssim}} \times \mathcal{L}_{\text{D-SSIM}}\left(I_l^O, I_l^E\right) \Big], \tag{9}$$

where M_{conf} assigns a binary weight to each candidate patch, with 1 for accepted patches and 0 otherwise, and λ_{ssim} controls the D-SSIM term. Rejected patches therefore contribute zero. The complete second-stage objective is given in Eq. 10:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{rgb}}^{\text{drop}} + \gamma(t)\mathcal{L}_{\text{conf}}, \tag{10}$$

where $\mathcal{L}_{\text{rgb}}^{\text{drop}}$ is the observed-view photometric loss evaluated with Gaussian dropout, t is the current iteration, and $\gamma(t)$ weights confidence supervision. We set $\gamma(t) = 0$ before T_c and progressively increase it thereafter.

4 Experiments

We evaluate the proposed framework on two real-world novel-view synthesis benchmarks under sparse input settings. We first describe the evaluation protocol and implementation details, then report quantitative comparisons with NeRF- and 3DGS-based methods. We subsequently present qualitative comparisons followed by ablation studies on the main components of our framework.

4.1 Experimental Setup

Datasets and evaluation protocol. We evaluate on LLFF with 3, 6, and 9 input views and Mip-NeRF 360 with 12 and 24 views, measuring PSNR, SSIM, and LPIPS.

Implementation details. Each run contains 10,000 iterations. Depth-prior regularization is active before iteration 2,000, at which point the EMA model is initialized from the online model and Gaussian dropout is enabled. Confidence-guided pseudo-view supervision is active from iteration 3,000 onward. The EMA decay coefficient is $\alpha = 0.8$, and the Gaussian dropout rate increases linearly from zero to 0.25. Monocular depth priors are pre-computed with Marigold [38], and 5×5 patches are used for cross-view confidence estimation. Unless otherwise stated, this configuration is used for all experiments.

4.2 Quantitative Comparison

Table 1 reports the results on LLFF. Our method performs consistently well across the 3-, 6-, and 9-view settings, obtaining favorable PSNR, SSIM, and LPIPS values relative to the compared methods. Compared with CoR-GS, our method improves all three metrics across the evaluated sparse-view settings.

Table 2 summarizes the results on Mip-NeRF 360. Our method remains competitive in both the 12- and 24-view settings, yielding favorable PSNR, SSIM, and LPIPS values relative to the compared methods. These results indicate that the proposed framework remains effective for more complex scenes and wider viewpoint variations.

4.3 Qualitative Comparison

Figures 2 and 3 provide qualitative comparisons under the sparsest setting of each benchmark. The red boxes highlight thin structures, high-frequency textures, and weakly observed regions in which sparse-view optimization is particularly prone to floaters, color bleeding, and over-smoothed geometry.

On Mip-NeRF 360 with 12 training views (Fig. 2), our results better preserve object silhouettes and foreground-background boundaries (e.g., the bicycle and garden scenes) and reduce color leakage in dark regions (the counter scene), compared with competing methods that tend to produce elongated primitives and blurred translucent regions. The proposed supervision is most helpful near depth discontinuities and in regions

Table 1 Quantitative comparison on LLFF under 3-, 6-, and 9-view settings. Higher PSNR and SSIM and lower LPIPS indicate better performance. Best and second-best results among the compared methods are shown in **bold** and underline, respectively.

	3-view			6-view			9-view		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF-based methods									
Mip-NeRF [11]	16.11	0.401	0.460	22.91	0.756	0.213	24.88	0.826	0.170
DietNeRF [21]	14.94	0.370	0.496	21.75	0.717	0.248	24.28	0.801	0.183
RegNeRF [22]	19.08	0.587	0.336	23.10	0.760	0.206	24.86	0.820	0.161
FreeNeRF [24]	19.63	0.612	0.308	23.73	0.779	0.195	25.13	0.827	0.160
SparseNeRF [26]	19.86	0.624	0.328	–	–	–	–	–	–
3DGS-based methods									
3DGS [1]	19.17	0.646	0.268	23.74	0.807	0.162	25.44	0.860	0.096
DNGaussian [2]	19.12	0.591	0.294	22.18	0.755	0.198	23.17	0.788	0.180
FSGS [3]	20.43	0.682	0.248	24.09	0.823	0.145	25.31	0.860	0.122
CoR-GS [7]	20.36	0.710	0.202	24.34	0.831	0.122	25.94	0.872	0.088
DropGaussian [5]	20.34	0.701	0.201	24.58	0.830	0.125	25.85	0.864	0.093
Ours	**20.86**	**0.730**	**0.179**	**24.71**	**0.840**	**0.107**	**26.31**	**0.881**	**0.083**

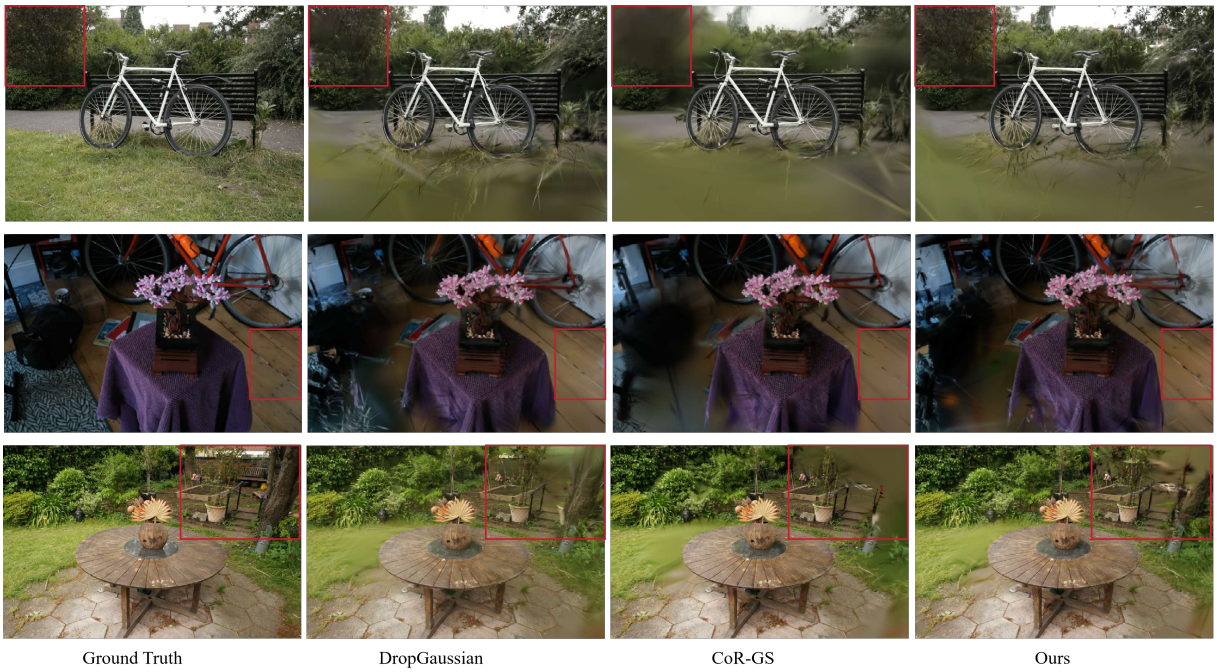

Fig. 2 Qualitative comparison on Mip-NeRF 360 with 12 training views. Each row shows one test scene, and the red boxes highlight challenging foreground-background boundaries, fine vegetation, and weakly observed regions.

with weak direct coverage, although some texture smoothing remains in heavily occluded foliage.

The three-view LLFF results in Fig. 3 constitute a more extreme test because large portions of each forward-facing scene are unobserved. Our method better maintains repeated and thin structures (e.g., leaves, dinosaur ribs, and railings) and object boundaries, with less cross-surface blending and fewer patchy artifacts. Overall, the visual results agree with the SSIM and LPIPS improvements in Tables 1 and 2: the main benefit is



Fig. 3 Qualitative comparison on LLFF with only three training views. Red boxes emphasize fine repeated structures, object boundaries, reflective regions, and small ceiling details.

Table 2 Quantitative comparison on Mip-NeRF 360 under 12- and 24-view settings. Best and second-best results among the compared methods are shown in **bold** and underline, respectively.

	12-view			24-view		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS [1]	18.52	0.523	0.415	22.80	0.708	0.276
FSGS [3]	18.80	0.531	0.418	23.28	0.715	0.274
CoR-GS [7]	19.52	0.558	0.418	23.39	0.727	0.271
DropGaussian [5]	19.57	0.562	0.391	23.71	0.732	0.274
Ours	**20.08**	**0.594**	**0.375**	**24.01**	**0.740**	**0.265**

improved structural coherence at novel viewpoints rather than a global color change.

4.4 Ablation Study

We evaluate the contribution of depth-based geometry initialization and pseudo-view supervision (Sec. 3.3) in the sparsest setting of the two datasets. Specifically, *w/o P* removes the entire pseudo-view supervision module: no pseudo views are rendered and the confidence supervision loss $\mathcal{L}_{\text{conf}}$ is disabled throughout training, while the EMA model is still maintained for parameter averaging and inference. The *w/o C* variant retains pseudo-view supervision but removes the cross-view confidence mask, so all EMA predictions are used as supervision targets without reliability filtering. The *w/o D* variant removes only the Depth-Prior-Guided Warm-up and keeps the second training stage unchanged.

Table 3 Ablation study on LLFF (3-view) and Mip-NeRF 360 (12-view). *Full* denotes the complete model. "w/o P", "w/o C", and "w/o D" remove pseudo-view supervision, the cross-view confidence mask, and the Depth-Prior-Guided Warm-up, respectively.

Dataset	Setting	PSNR↑	SSIM↑	LPIPS↓
LLFF	Full	20.86	0.730	0.179
	w/o P	20.68	0.722	0.188
	w/o C	20.72	0.723	0.185
	w/o D	20.57	0.717	0.198
Mip-NeRF 360	Full	20.08	0.594	0.375
	w/o P	19.75	0.561	0.393
	w/o C	19.79	0.570	0.393
	w/o D	19.90	0.585	0.373

As shown in Table 3, removing any component consistently degrades reconstruction quality. Comparing *w/o P* and *w/o C* isolates the role of pseudo-view supervision itself in the absence of the confidence mask. The two variants perform close to each other on both datasets, whereas the full model clearly outperforms *w/o C*. This indicates that pseudo-view supervision is effective only when it is restricted to reliable EMA regions, and that the cross-view confidence mask, rather than pseudo-view supervision itself, accounts for most of the gain of this module. Comparing *w/o D* with *w/o P* further reveals complementary behavior across datasets. On LLFF, removing depth warm-up causes the larger PSNR drop,

suggesting that reliable geometry initialization is particularly important under extremely sparse forward-facing observations. On Mip-NeRF 360, in contrast, removing the pseudo-view supervision module leads to a larger degradation, especially in SSIM and LPIPS, indicating that confidence-filtered self-supervision becomes more important for unbounded scenes. Overall, these results confirm that the two proposed stages provide complementary benefits.

5 Conclusion

We presented DCE-GS, a two-stage framework for sparse-view Gaussian splatting that aims to reduce geometric uncertainty and overfitting. The first stage employs depth-prior-guided warm-up to obtain a better initial geometry, after which the depth prior is removed to avoid propagating potentially inaccurate monocular predictions. The second stage maintains a complementary online–EMA scheme, where Gaussian dropout diversifies the online model and the EMA model supplies stable pseudo-view predictions. A patch-level confidence mechanism then selects reliable pseudo-view regions to supervise the online model. Together, these stages combine geometry-aware initialization with confidence-guided pseudo-view regularization.

Nonetheless, DCE-GS has several limitations. Its geometric initialization depends on a pre-trained monocular depth model, whose errors may propagate to challenging regions. Moreover, the generated pseudo views add no new observations and can only partially compensate for information missing from extremely sparse views, severe occlusions, or complex non-Lambertian surfaces.

Author contributions. M. R. and W. L. contributed to conceptualization and writing. Y. L. and Y. Y. contributed to investigation and data curation. All authors reviewed the manuscript.

Funding. The authors did not receive support from any organization for the submitted work.

Data availability. The datasets used in this study are publicly available. The LLFF dataset is available at https://github.com/Fyusion/LLFF, and the Mip-NeRF 360 dataset is available at https://jonbarron.info/mipnerf360/.

Code availability. The code used in this study is not publicly available at this time.

Declarations. Conflict of interest The authors declare no conflict of interest.

References

[1] 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) https://doi.org/10.1145/3592433

[2] Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: DNGaussian: Optimizing sparse-view 3D gaussian radiance fields with global-local depth normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20775–20785. IEEE Computer Society, Los Alamitos, CA, USA (2024)

[3] Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: FSGS: Real-time few-shot view synthesis using gaussian splatting. In: Proceedings of the European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 15097, pp. 145–163. Springer, Cham, Switzerland (2024). https://doi.org/10.1007/978-3-031-72933-1_9

[4] Xu, Y., Wang, L., Chen, M., Ao, S., Li, L., Guo, Y.: DropoutGS: Dropping out gaussians for better sparse-view rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 701–710. IEEE Computer Society, Los Alamitos, CA, USA (2025)

[5] Park, H., Ryu, G., Kim, W.: DropGaussian: Structural regularization for sparse-view gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21600–21609. IEEE Computer Society, Los Alamitos, CA, USA (2025)

[6] Li, H., Zhu, Q., Meng, X., Wang, X., Zhao, D., Fan, X.: PairDropGS: Paired Dropout-Induced Consistency Regularization for Sparse-View Gaussian Splatting (2026). https://doi.org/10.48550/arXiv.2605.12072 . https://arxiv.org/abs/2605.12072

[7] Zhang, J., Li, J., Yu, X., Huang, L., Gu, L., Zheng, J., Bai, X.: CoR-GS: Sparse-view 3D

gaussian splatting via co-regularization. In: Proceedings of the European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 15059, pp. 335–352. Springer, Cham, Switzerland (2024). https://doi.org/10.1007/978-3-031-73232-4_19

[8] 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 (ICCV), pp. 4940–4950. IEEE Computer Society, Los Alamitos, CA, USA (2025)

[9] Li, C., Shi, Z., Lu, Y., He, W., Xu, X.: Robust neural rendering in the wild with asymmetric dual 3D gaussian splatting. In: Advances in Neural Information Processing Systems, vol. 38. Curran Associates, Inc., Red Hook, NY, USA (2025). https://doi.org/10.52202/085713-1672

[10] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 12346, pp. 405–421. Springer, Cham, Switzerland (2020). https://doi.org/10.1007/978-3-030-58452-8_24

[11] Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5470–5479. IEEE Computer Society, Los Alamitos, CA, USA (2022)

[12] 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 (CVPR), pp. 4578–4587. IEEE Computer Society, Los Alamitos, CA, USA (2021)

[13] 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 (CVPR), pp. 4690–4699. IEEE Computer Society, Los Alamitos, CA, USA (2021)

[14] 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 (ICCV), pp. 14124–14133. IEEE Computer Society, Los Alamitos, CA, USA (2021)

[15] Johari, M.M., Lepoittevin, Y., Fleuret, F.: GeoNeRF: Generalizing NeRF with geometry priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18365–18375. IEEE Computer Society, Los Alamitos, CA, USA (2022)

[16] Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3D gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19447–19456. IEEE Computer Society, Los Alamitos, CA, USA (2024)

[17] Cao, Q., Ding, J., Liu, Z., Kuang, Z., Shao, Y., Shen, Y.: VC-GS: View-consistent deblurring gaussian splatting via alternating branch optimization. The Visual Computer **41**(10), 7303–7317 (2025) https://doi.org/10.1007/s00371-025-03806-y

[18] Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-GS: Structured 3D gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20654–20664. IEEE Computer Society, Los Alamitos, CA, USA (2024)

[19] Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2D gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 Conference Papers, pp. 1–11. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3641519.3657428

[20] Guédon, A., Lepetit, V.: SuGaR: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5354–5363. IEEE Computer Society, Los Alamitos, CA, USA

(2024)
[21] Jain, A., Tancik, M., Abbeel, P.: Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5885–5894. IEEE Computer Society, Los Alamitos, CA, USA (2021)
[22] Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., Radwan, N.: RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5480–5490. IEEE Computer Society, Los Alamitos, CA, USA (2022)
[23] Kim, M., Seo, S., Han, B.: InfoNeRF: Ray entropy minimization for few-shot neural volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12912–12921. IEEE Computer Society, Los Alamitos, CA, USA (2022)
[24] Yang, J., Pavone, M., Wang, Y.: FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8254–8263. IEEE Computer Society, Los Alamitos, CA, USA (2023)
[25] Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D.: Depth-supervised NeRF: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12882–12891. IEEE Computer Society, Los Alamitos, CA, USA (2022)
[26] Wang, G., Chen, Z., Loy, C.C., Liu, Z.: SparseNeRF: Distilling depth ranking for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9065–9076. IEEE Computer Society, Los Alamitos, CA, USA (2023)
[27] Paliwal, A., Ye, W., Xiong, J., Kotovenko, D., Ranjan, R., Chandra, V., Kalantari, N.K.: CoherentGS: Sparse novel view synthesis with coherent 3D gaussians. In: Proceedings of the European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 15088, pp. 19–37. Springer, Cham, Switzerland (2024). https://doi.org/10.1007/978-3-031-73404-5_2
[28] Xiong, H., Muttukuru, S., Xiao, H., Upadhyay, R., Chari, P., Zhao, Y., Kadambi, A.: SparseGS: Sparse view synthesis using 3D gaussian splatting. In: Proceedings of the International Conference on 3D Vision (3DV), pp. 1032–1041. IEEE, Piscataway, NJ, USA (2025). https://doi.org/10.1109/3DV66043.2025.00100
[29] Bao, Z., Liao, G., Zhou, K., Liu, K., Li, Q., Qiu, G.: LoopSparseGS: Loop-based sparse-view friendly gaussian splatting. IEEE Transactions on Image Processing **34**, 3889–3902 (2025) https://doi.org/10.1109/TIP.2025.3574929
[30] Zhang, Q., Wysocki, O., Urban, S., Jutzi, B.: CDGS: Confidence-aware depth regularization for 3D gaussian splatting. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLVIII-2/W7-2024**, 189–196 (2024) https://doi.org/10.5194/isprs-archives-XLVIII-2-W7-2024-189-2024
[31] Zhan, C., Zhang, Y., Lin, Y., Wang, G., Wang, H.: RDG-GS: Relative depth guidance with gaussian splatting for real-time sparse-view 3D rendering. International Journal of Computer Vision **134** (2026) https://doi.org/10.1007/s11263-025-02594-1
[32] Zheng, Y., Jiang, Z., He, S., Sun, Y., Dong, J., Zhang, H., Du, Y.: NexusGS: Sparse view synthesis with epipolar depth priors in 3D gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 26800–26809. IEEE Computer Society, Los Alamitos, CA, USA (2025)
[33] Zhou, H., Zhu, H., Qin, J., Wang, Y., Liao, J., Shen, X., Xiao, C.: SARG-GS: Semantic-augmented and residual-guided 3D gaussian splatting for sparse view synthesis. The Visual Computer **42**(10) (2026) https://doi.org/10.1007/s00371-026-04646-0
[34] Li, B., Gao, T., Gu, Z., Liu, H.: Semantic-guided 3D gaussian splatting for sparse-view reconstruction in industrial digital twins. Visual Computing for Industry, Biomedicine, and Art **9**(1) (2026) https://doi.org/10.1186/s42492-026-00229-x

[35] Shi, C., Yang, C., Hu, X., Yang, Y., Ding, J., Tan, M.: MMGS: Multi-model synergistic gaussian splatting for sparse view synthesis. Image and Vision Computing **158**, 105512 (2025) https://doi.org/10.1016/j.imavis.2025.105512

[36] Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). OpenReview.net, Toulon, France (2017). https://openreview.net/forum?id=BJ6oOfqge

[37] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30, pp. 1195–1204. Curran Associates, Inc., Red Hook, NY, USA (2017)

[38] Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9492–9502. IEEE Computer Society, Los Alamitos, CA, USA (2024)

[39] Han, L., Zhou, J., Liu, Y.-S., Han, Z.: Binocular-guided 3D gaussian splatting with view consistency for sparse view synthesis. In: Advances in Neural Information Processing Systems, vol. 37. Curran Associates, Inc., Red Hook, NY, USA (2024). https://doi.org/10.52202/079017-2191

[40] Gu, M., Zhang, J., Li, J., Yu, X., Luo, H., Zheng, J., Bai, X.: SparseSurf: Sparse-view 3D gaussian splatting for surface reconstruction. Proceedings of the AAAI Conference on Artificial Intelligence **40**(6), 4311–4319 (2026) https://doi.org/10.1609/aaai.v40i6.42428