

OracleGS: Grounding Generative Priors for Sparse-View Gaussian Splatting

Atakan Topaloğlu^{1,2,3 *} Kunyi Li^{4,6}
A. Murat Tekalp^{2,3}

Michael Niemeyer⁵ Nassir Navab^{4,6}
Federico Tombari^{4,5,6}

¹ETH Zürich ²Koç University ³KUIS AI Center ⁴Technical University of Munich
⁵Google ⁶Munich Center for Machine Learning

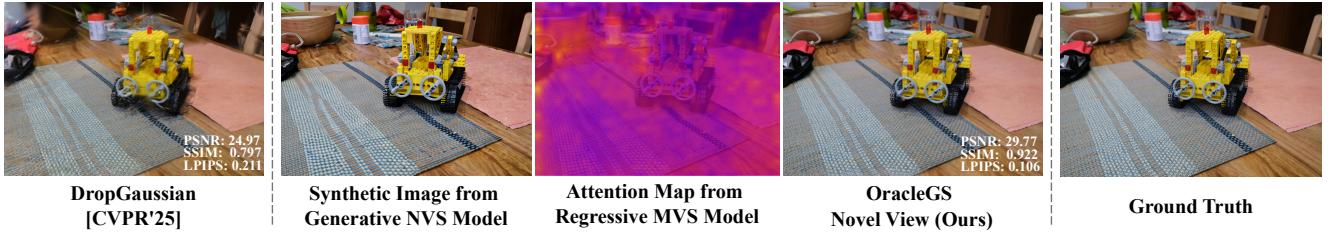


Figure 1. **OracleGS Reconciles Generative Completeness with Regressive Fidelity.** First, a generative model proposes a potentially flawed, synthetic view (middle left). Our MVS-based oracle then grounds this proposal by quantifying a 3D uncertainty map (middle), effectively identifying various sources of generative errors; including faulty textures on the mat, inconsistent structures on the lego, and under-observed backgrounds, as regions of high uncertainty (highlighted in blue). Using this signal to guide a confidence-weighted optimization, OracleGS filters these artifacts, producing novel views (middle right) with superior fidelity compared to both the synthetic input and prior state-of-the-art DropGaussian [31] (left).

Abstract

Sparse-view novel view synthesis is fundamentally ill-posed due to severe geometric ambiguity. Current methods are caught in a trade-off: regressive models are geometrically faithful but incomplete, whereas generative models can complete scenes but often introduce structural inconsistencies. We propose OracleGS, a novel framework that reconciles generative completeness with regressive fidelity for sparse view Gaussian Splatting. Instead of using generative models to patch incomplete reconstructions, our “propose-and-validate” framework first leverages a pre-trained 3D-aware diffusion model to synthesize novel views to propose a complete scene. We then repurpose a multi-view stereo (MVS) model as a 3D-aware oracle to validate the 3D uncertainties of generated views, using its attention maps to reveal regions where the generated views are well-supported by multi-view evidence versus where they fall into regions of high uncertainty due to occlusion, lack of texture, or direct inconsistency. This uncertainty signal directly guides the optimization of a 3D Gaussian Splatting model via an uncertainty-weighted loss. Our approach conditions the powerful generative prior on multi-view geo-

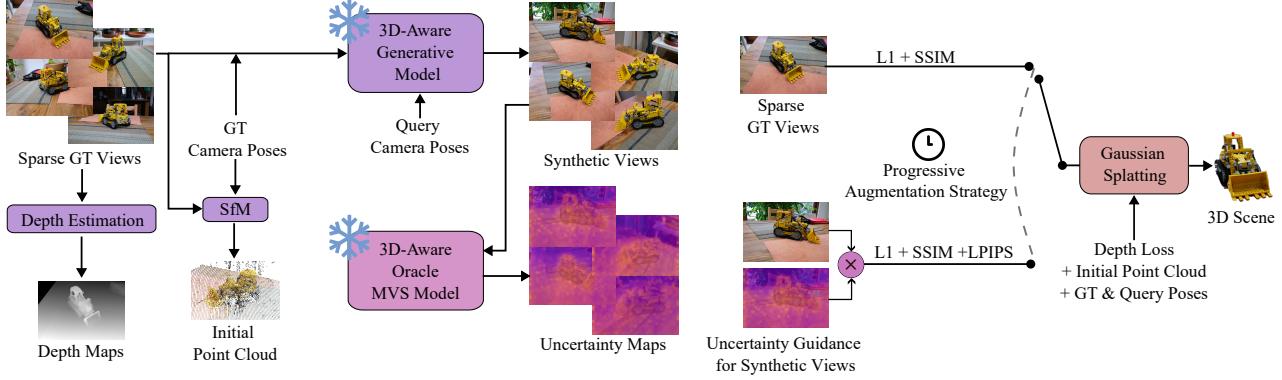
metric evidence, filtering hallucinatory artifacts while preserving plausible completions in under-constrained regions, outperforming state-of-the-art methods on datasets including Mip-NeRF 360 and NeRF Synthetic.

1. Introduction

Novel-view synthesis (NVS) [26] constitutes a fundamental problem in computer vision, which seeks to reconstruct the 3D scenes from given viewpoints and render photo-realistic observations from novel perspectives. This capability is central to a variety of downstream applications, including large-scale 3D content generation [13, 23, 25, 62], digital twin construction [9, 35], and immersive virtual reality [7, 14]. A particularly challenging yet practically significant setting is sparse-view NVS, where only a limited number of casually captured images are available acquired via handheld devices in uncontrolled environments. Addressing this setting is critical for enabling scalable and accessible 3D content creation in real-world scenarios.

However, when only a few input views are available, the underlying 3D scene structure becomes severely under-constrained, hindering high-quality synthesis. This transforms the task from a well-posed interpolation problem into a highly ill-posed extrapolation problem, where vast

*Work conducted as part of a research collaboration with Google and the Technical University of Munich (TUM).



1. Preprocess

2. Propose and Validate

3. Uncertainty-Enforced 3DGS Optimization

Figure 2. Overview. Given sparse input views with poses, we first estimate initial point cloud and depth maps. Afterwards, a 3D-Aware generative model proposes novel synthetic views, while the 3D-Aware Oracle’s attention maps are used as a proxy for 3D uncertainty. Finally, we train the 3DGS model using a standard loss on the GT views and our novel uncertainty-guided loss on the synthetic views. We employ a progressive augmentation strategy over the course of the optimization to control the ratio of GT and synthetic images at each iteration, which helps to stabilize training and guide scene structure.

unobserved regions lead to significant structural ambiguity. This is particularly evident for explicit representations like 3D Gaussian Splatting (3DGS) [17], which, despite offering high quality real-time rendering, are prone to severe artifacts when optimized with sparse data [52, 60]. Consequently, a significant body of recent work has focused on regularizing the 3DGS optimization process using techniques like generating pseudo-views [64], applying dropout-based strategies [31], or incorporating external priors from pre-trained models [52, 60] in the sparse view regime.

To address the challenge, recent research has diverged into two main paradigms, each with fundamental limitations. On the one hand, feedforward regression-based models [4, 6, 15, 57, 65] are effective in extracting information from visible regions, but their performance degrades in regions of high 3D uncertainty, where robust correspondences cannot be established. This leads to poorly estimated surfaces and textures in occluded or under-observed parts of the scene. On the other hand, generative models, particularly diffusion-based approaches [8, 24, 33, 39, 48, 49, 51, 58, 63] excel in hallucinating plausible content for these unobserved regions. However, they often lack strict 3D consistency with the input views, producing results that are visually compelling but structurally incorrect. This leaves a critical gap: regressive methods are geometrically faithful but incomplete, while generative methods are complete but often unfaithful.

In this paper, we resolve this conflict by inverting the conventional paradigm. Instead of using a regressive model to form an incomplete reconstruction and a generative prior to fill the gaps, our “propose-and-validate” framework first

leverages a powerful 3D-aware generative model to *propose* a complete and plausible scene by synthesizing a dense set of novel views. We then repurpose a classical multi-view stereo (MVS) model not as a reconstructor, but as a 3D-Aware oracle that assesses where the generated views are well-supported by multi-view evidence, thereby *validating* these proposals. It distills this assessment into per-pixel uncertainty maps, which directly guide the optimization of a 3DGS [17] representation via a uncertainty-weighted loss. This approach grounds the powerful generative prior on multi-view evidence, filtering out inconsistent artifacts while preserving plausible completions, leading to novel views that are both complete and structurally sound. In summary, our main contributions are:

- **A “propose-and-validate” framework for sparse-view NVS.** Our method, OracleGS, synergizes generative priors with geometric validation by using a 3D-aware diffusion model to propose novel views and a multi-view stereo model to validate their consistency, achieving state-of-the-art results in sparse-view-3DGS.
- **A novel uncertainty-weighted loss derived from MVS attention.** We introduce a novel framework that repurposes a multi-view stereo (MVS) model as a 3D uncertainty oracle, using its attention maps to formulate a new uncertainty-weighted loss for guiding generative novel view synthesis.
- **A progressive augmentation schedule to stabilize optimization.** We introduce a training curriculum that treats synthetic views as a temporary structural scaffold. By modulating their influence over the course of the training, we first build a coherent global structure before yielding to ground-truth data for high-frequency refinement.

2. Related Work

2.1. 3D Scene Reconstruction

Neural Radiance Fields. Neural Radiance Fields (NeRF) [26], represents a 3D scene as a continuous, implicit function learned by a neural network that maps 3D coordinates and viewing directions to color and density. This approach led to a series of extensions that improved rendering quality with anti-aliasing [27], handling complex unbounded scenes [1], and accelerating training [2] or improving performance [28, 32]. Despite their high fidelity, NeRF-based methods typically suffer from slow training and rendering times, limiting their practical applicability.

3D Gaussian Splatting. To address these efficiency bottlenecks, 3D Gaussian Splatting (3DGS) [17] was introduced. Instead of a neural network, 3DGS models a scene explicitly with a set of 3D Gaussian primitives that are optimized via a differentiable rasterizer, enabling real-time rendering and significantly faster training while achieving state-of-the-art quality, which has led to their rapid adoption for novel view synthesis, including improved rendering quality [19, 59], implicit surface reconstruction [22] dynamic scene modeling [50], text-to-3D generation [38, 56] while hybrid methods [30] use radiance fields as a prior to improve the robustness of the Gaussian optimization process.

While both NeRF and 3DGS have demonstrated success in the dense-view regime, their performance degrades significantly when the number of training images are reduced, under sparse-view conditions. This vulnerability highlights the need for strong priors to regularize the reconstruction process, a challenge our work directly addresses.

2.2. Sparse-view Novel View Synthesis

To combat the ill-posedness of sparse-view reconstruction, a significant body of work has focused on introducing priors through internal regularization. These methods attempt to constrain the solution space using only the information present in the input views or by applying general-purpose heuristics. Early approaches for NeRF introduced semantic losses from CLIP such as DietNeRF [10], Reg-NeRF [29] uses appearance regularization via normalizing flows, while geometric priors like local depth ranking and spatial continuity are used in Sparse-NeRF [40]. Others focused on training-based strategies, such as FreeN-eRF [53] which uses frequency annealing to reduce artifacts, whereas SimpleNeRF [36] and DNGaussian [21] utilize depth supervision. More recent methods have adapted these ideas for 3D Gaussian Splatting, using diffusion-based losses for guidance such as Sparse-GS [52], or generating pseudo-views to densify the input such as FSGS [64]. CoR-GS [60] co-regularizes the Gaussians to suppress er-

rors, CoMapGS [12] uses uncertainty-aware supervision derived from covisibility maps, Intern-GS [37] uses dense DUS3R [46] initialization and foundation-model-guided optimization, while DropGaussian [31] applies dropout to Gaussians to alleviate overfitting. While these approaches improve stability, they are fundamentally data-starved and constrained by regularization. They can refine what is visible but cannot reliably hallucinate the complex, unobserved geometry present in the scene’s information voids.

2.3. External Priors from Auxiliary Vision Models

Instead of relying solely on internal regularization, our work pioneers an approach that leverages powerful, pre-trained models from two distinct vision tasks to provide strong external priors.

Multi-View Stereo (MVS). The primary goal of MVS is to estimate dense geometry by establishing robust correspondences across multiple views. Modern MVS architectures, often transformer-based with the advent of Dust3R [46] and 3R methods [3, 11, 20, 41, 54], and line of work [16, 42–44] culminating most recently in VGGT [45] provide a powerful source of geometric evidence for well-observed regions. Their strength is in producing geometrically accurate outputs where there is sufficient visual information. However, their performance degrades in textureless areas and fails completely in occluded regions where no correspondences can be found, resulting in incomplete geometric guidance.

Novel View Synthesis (NVS). In a complementary direction, regressive feedforward NVS Models [4, 6, 15, 57, 65] and generative feedforward NVS Models [8, 24, 33, 39, 48, 49, 51, 58] and most recently Stable-Virtual-Camera (SEVA) [63] can synthesize photorealistic and semantically coherent images from arbitrary new camera poses. Their strength lies in their ability to plausibly complete a scene, and for diffusion-based generative models, this means hallucinating content for unobserved areas based on learned priors of the visual world. This generative capability, however, comes at the cost of fidelity; they are not constrained by multi-view consistency and can produce plausible-looking “fictions” that contradict the true scene structure.

Thus, these two classes of models are complementary. MVS provides geometrically accurate but incomplete data, while generative NVS provides complete but potentially inaccurate data. Our work introduces a framework to resolve this tension. While previous methods use MVS for direct supervision or generative models for naive inpainting, we repurpose the MVS model as a 3D-aware oracle to validate proposals from the generative model, achieving a synthesis that is both structurally sound and visually complete.

| Methods | 12-view | | | 24-view | | |
|-------------------------------------|--------------------|--------------------|------------------------|--------------------|--------------------|------------------------|
| | PSNR(\uparrow) | SSIM(\uparrow) | LPIPS (\downarrow) | PSNR(\uparrow) | SSIM(\uparrow) | LPIPS (\downarrow) |
| Mip-NeRF 360 [1] | 17.73 | 0.432 | 0.520 | 19.78 | 0.530 | 0.431 |
| RegNeRF [29] | 18.84 | 0.437 | 0.544 | 20.55 | 0.546 | 0.398 |
| SparseNeRF [40] | 17.44 | 0.395 | 0.609 | 21.13 | 0.600 | 0.389 |
| 3DGS [17] | 18.52 | 0.523 | 0.415 | 22.80 | 0.708 | 0.276 |
| DNGaussian [21] | 16.28 | 0.432 | 0.549 | 19.26 | 0.550 | 0.440 |
| FSGS [64] | 18.80 | 0.531 | 0.418 | 22.82 | 0.693 | 0.293 |
| SparseGS [52] | 19.37 | 0.577 | 0.398 | 23.02 | 0.713 | 0.290 |
| CoR-GS [60] | 19.52 | 0.558 | 0.418 | 23.39 | 0.727 | 0.271 |
| CoMapGS [12] | 19.68 | 0.591 | 0.394 | 23.46 | 0.734 | 0.264 |
| DropGaussian (Our Replication) [31] | 19.38 | 0.583 | 0.402 | 23.44 | 0.736 | 0.273 |
| Ours | 20.32 | 0.596 | 0.350 | 23.72 | 0.723 | 0.244 |

Table 1. Quantitative comparison on Mip-NeRF360 dataset [1] for 12 and 24 input views. The best, second-best, and third-best entries are marked in red, orange, and yellow respectively.

3. Method

3.1. Preliminaries

3D Gaussian Splatting (3DGS) [17] is an explicit scene representation that uses a collection of 3D Gaussians for high-fidelity, real-time novel view synthesis. Each Gaussian is defined by a set of optimizable attributes: a position (mean) $\mu \in \mathbb{R}^3$, a covariance matrix $\Sigma \in \mathbb{R}^{3 \times 3}$, an opacity $\alpha \in [0, 1]$, and a color $\mathbf{c} \in \mathbb{R}^3$. To model view-dependent effects, the color is typically represented by Spherical Harmonics (SH) coefficients. For efficient optimization, the covariance matrix Σ is parameterized by a 3D scaling vector s and a rotation quaternion \mathbf{q} . To render a novel view, the 3D Gaussians are projected onto the 2D image plane and then blended in depth-sorted order. The final color \mathbf{C} for a pixel is computed via alpha-blending:

$$\mathbf{C} = \sum_{i=1}^N T_i \alpha_i \mathbf{c}_i, \quad (1)$$

where N is the number of Gaussians overlapping the pixel, sorted by depth. The color of the i -th Gaussian is \mathbf{c}_i , and its opacity is α_i . The accumulated transmittance T_i ensures that Gaussians closer to the camera have a greater contribution.

While 3DGS achieves state-of-the-art results, its performance is contingent upon a dense set of input views that provide robust geometric and photometric constraints. In the sparse-view setting, the optimization becomes severely under-constrained, often leading to undesirable artifacts such as floater artifacts and incomplete, blurry renderings.

We introduce **OracleGS**, with a “propose-and-validate” framework that synergizes generative and regressive frameworks. First, a generative model *proposes* a complete scene by synthesizing novel views. An MVS model, repurposed as a 3D-aware oracle, then *validates* these pro-

posals by quantifying their 3D uncertainty. This signal guides the optimization of a 3DGS [17] representation, optimized using 3DGS-MCMC [19], which is initialized from a COLMAP [34] point cloud generated using ground-truth poses. Our method is illustrated in Figure 2.

3.2. View Augmentation & Uncertainty Estimation

The foundation of our method is a new architecture that assigns synergistic roles to its generative and regressive components via the “Propose-and-validate” framework.

View Augmentation via Generative NVS. The first stage addresses the information gap inherent in sparse inputs. We use a pre-trained, 3D-aware diffusion model (SEVA) [63] to *propose* a complete version of the scene. Given the sparse set of N ground-truth (GT) images $\{I_i\}_{i=1}^N$ and their corresponding camera poses $\{P_i\}_{i=1}^N$, the model generates a dense set of M synthetic images $\{I'_j\}_{j=1}^M$ from new query camera poses $\{P'_j\}_{j=1}^M$. Though the generative model creates novel views, they are not geometrically perfect. The purpose of this stage is to provide a coarse proposal for the scene’s appearance and structure, effectively filling unobserved regions. However, these synthetic views $\{I'_j\}$ are a rich but potentially inconsistent source of data; directly using them without refinement would propagate geometric inaccuracies and visual artifacts, leading to a sub-optimal final reconstruction. The subsequent stages are designed to mitigate these issues.

Uncertainty Estimation via a Repurposed MVS Oracle. Proposed synthetic views contain inconsistencies, and using them naively destabilizes the training process. To validate the structural integrity of the proposed synthetic views, we introduce a *3D-Aware Oracle* whose task is to assess their 3D uncertainty. While traditional methods leverage

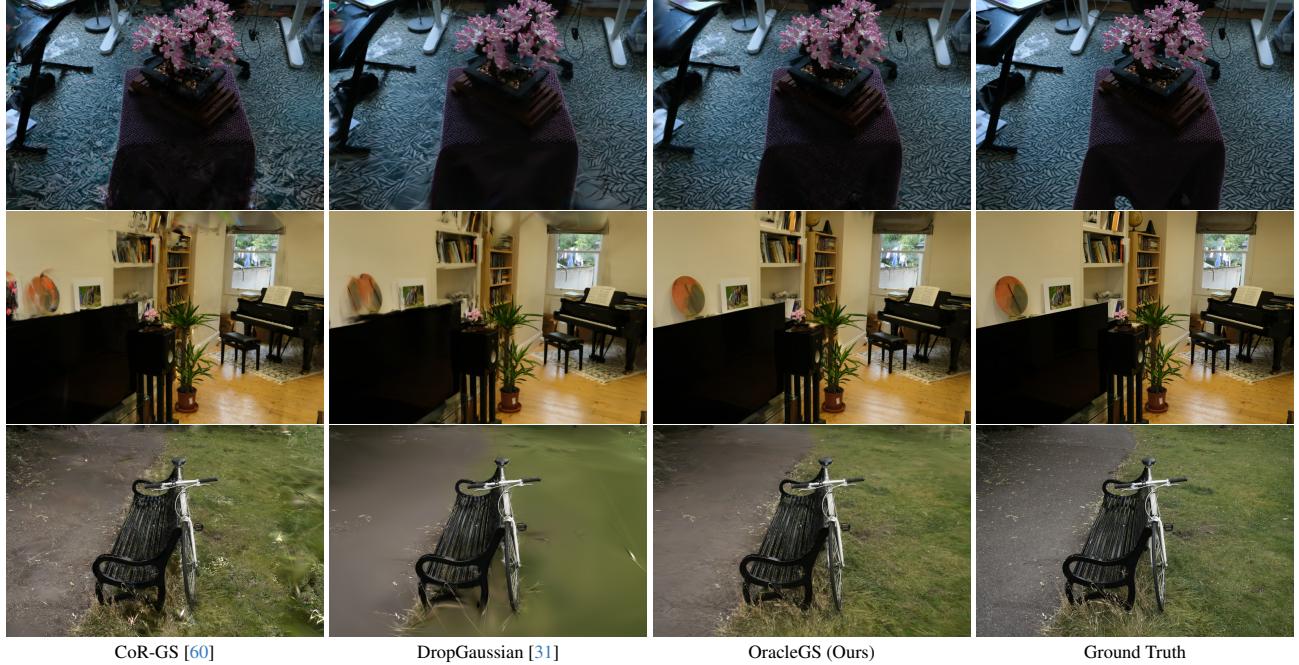


Figure 3. Visual comparison with state-of-the-art methods on the Mip-NeRF360 [1] dataset. Our method, OracleGS, demonstrates superior handling of common failure modes. **Top row (Bonsai, 12 views):** OracleGS accurately reconstructs the challenging carpet texture and background regions while competing methods produce noticeable artifacts. **Middle row (Room, 12 views):** Our method avoids the distortions present in other reconstructions. **Bottom row (Bicycle, 24 views):** Our approach strikes a balance between detail and smoothness, preventing the noisy overfitting seen in CoR-GS [60] and the oversmoothing that erases fine details in DropGaussian [31].

MVS models for direct supervision via their final depth or point cloud outputs, these are often noisy and incomplete in sparse-view settings, providing unreliable guidance. We circumvent this limitation by extracting a more reliable signal directly from the MVS model’s intrinsics. Inspired by recent findings that attention maps in vision transformers encode rich structural cues [5], we repurpose the **global-attention maps** from a transformer-based MVS model, VGGT [45] as a powerful proxy for 3D uncertainty. Specifically, a high global-attention score between views signifies strong multi-view consistency and thus high 3D confidence. Conversely, a low score suggests high 3D uncertainty, effectively identifying generative hallucinations that lack robust multi-view support. This fine-grained uncertainty signal is then used to guide the final optimization.

3D-Aware Oracle Implementation. To implement the geometric oracle Φ , we first partition the set of synthetic views $\{I'_j\}_{j=1}^M$ into disjoint chunks $\{C_k\}$. This is necessary because the GPU memory requirements of the MVS model [45] scales with the number of input views. Each chunk C_k is independently processed by the MVS model to assess the 3D uncertainty of the generated scene proposal. Let $\Psi^{(l)}(I'_j|C_k)$ be an operator that extracts the global attention maps from view $I'_j \in C_k$ from layer l of the MVS model, computed with respect to the other views within its

own chunk C_k . The final uncertainty map $U_j \in [0, 1]^{H \times W}$ is a weighted average of these layer normalized attention maps from a set of layers \mathcal{L} :

$$U_j = \sum_{l \in \mathcal{L}} w_l \cdot \Psi^{(l)}(I'_j|C_k), \quad \text{where } I'_j \in C_k. \quad (2)$$

In our implementation, we use layers $\mathcal{L} = \{0, 22\}$, reference view $I_j = I_0$, with corresponding weights $w_l = \{\frac{1}{4}, \frac{3}{4}\}$. We provide an analysis for this layer selection in the supplementary material. This map serves as the oracle’s final judgment, with $U_j(u, v) \approx 1$ signifying high confidence, low uncertainty and $U_j(u, v) \approx 0$ signifying low confidence, high uncertainty.

3.3. Uncertainty-Enforced 3DGS Optimization

Having generated the synthetic proposals and the corresponding per-pixel uncertainty maps, we now detail how these components are integrated to supervise the 3DGS optimization. The final stage uses sparse GT training views, synthetic proposals from non-test viewpoints, and the oracle’s uncertainty maps to optimize a 3DGS scene.

Ground-Truth and Depth Loss (\mathcal{L}_{GT}). For any sampled GT view I_i , and predicted view \hat{I}_i the loss combines a standard photometric term with a depth term.

$$\mathcal{L}_{GT} = \mathcal{L}_{photo} + \lambda_{depth}(t)\mathcal{L}_{depth} \quad (3)$$

The photometric loss, $\mathcal{L}_{\text{photo}}$, is a combination of L1 and SSIM:

$$\mathcal{L}_{\text{photo}} = (1 - \lambda_{\text{ssim}})\mathcal{L}_1(\hat{I}_i, I_i) + \lambda_{\text{ssim}}\mathcal{L}_{\text{SSIM}}(\hat{I}_i, I_i) \quad (4)$$

The depth loss for 3DGS $\mathcal{L}_{\text{depth}}$, as established in Hierarchical 3DGS [18], is applied when reliable monocular depth priors are available for a GT view. It is a masked L1 loss between the rendered inverse depth \hat{D}_i and the provided inverse depth map D_i :

$$\mathcal{L}_{\text{depth}} = \|(\hat{D}_i - D_i) \odot M_i\|_1 \quad (5)$$

where M_i is a mask indicating reliable depth regions and \odot indicates element-wise multiplication. The weight $\lambda_{\text{depth}}(t)$ is annealed with an exponential schedule with initial and final weights $\lambda_{\text{depth}_0} = 1$, $\lambda_{\text{depth}_1} = 0.01$ respectively.

Uncertainty-Weighted Synthetic Loss ($\mathcal{L}_{\text{synth}}$). For a sampled synthetic view I'_j , and predicted synthetic novel view \hat{I}'_j , the loss is modulated by the oracle’s uncertainty map U_j . Standard photometric losses are highly sensitive to pixel-level artifacts in generated images. We therefore incorporate a perceptual LPIPS [61] loss, which is more robust to minor texture variations. The total synthetic loss is:

$$\begin{aligned} \mathcal{L}_{\text{synth}} = & (1 - \lambda'_{\text{SSIM}} - \lambda_{\text{LPIPS}}) \cdot \|(\hat{I}'_j - I'_j) \odot U_j\|_1 \\ & + \lambda'_{\text{SSIM}} \cdot \|(1 - \text{SSIM}(\hat{I}'_j, I'_j)) \odot U_j\|_1 \\ & + \lambda_{\text{LPIPS}} \cdot \bar{U}_j \cdot \mathcal{L}_{\text{LPIPS}}(\hat{I}'_j, I'_j) \end{aligned} \quad (6)$$

Here, the L1 loss is weighted on a per-pixel basis by the uncertainty map U_j . We perform an element-wise multiplication of the map SSIM [47] creates within each patch with our uncertainty map U_j before averaging to a final loss value. This mechanism forces the optimization to prioritize structural consistency in regions deemed reliable by the oracle, while effectively down-weighting the contribution from patches in uncertain or artifact-prone areas. The LPIPS [61] loss, which operates on image patches, is modulated by a single scalar value $\bar{U}_j = \text{mean}(U_j)$, representing the overall geometric integrity of the entire synthetic view. This prevents low-uncertainty patches from subsidizing high-uncertainty ones in the perceptual loss, ensuring that low uncertainty images contribute significantly to the final appearance. We inherit the regularization terms on opacity and scale (\mathcal{L}_{reg}) from 3DGS-MCMC [19] to encourage a compact representation. The total loss becomes:

$$\mathcal{L} = \mathcal{L}_{\text{GT}} + \mathcal{L}_{\text{synth}} + \mathcal{L}_{\text{reg}} \quad (7)$$

3.4. Progressive Augmentation Schedule

To further stabilize training, we employ a dynamic training curriculum, governed by a scaled and shifted Beta distribution, to schedule the sampling probability of synthetic views

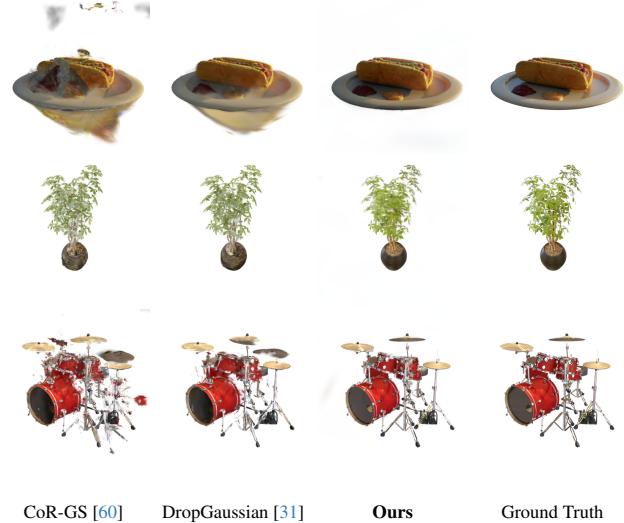


Figure 4. **Visual comparison with state-of-the-art methods on the NeRF Synthetic [26] dataset.** Our method consistently produces higher-fidelity reconstructions across diverse and challenging scenes compared to prior work. **Top row (Hotdog):** OracleGS eliminates the “floater” artifacts present in competing methods, preserving details on the condiments and plate. **Middle row (Ficus):** Our method reconstructs the intricate vase structures where competing methods fail. **Bottom row (Drums):** Our method reconstructs the thin structures of the drum kit’s stands and cymbals, which are fragmented or missing in other reconstructions.

over time. This schedule is designed to use the synthetic data as an intermediary scaffold for reconstruction. Initially, the probability of sampling synthetic views is low, allowing the initial point cloud to guide the scene. Afterwards, the probability of selecting a synthetic view increases, allowing their dense geometric and appearance information to rapidly structure the scene, including the unobserved regions, providing a coherent foundation that the sparse initial point cloud cannot. As optimization progresses and the learning rate anneals, the sampling probability of a synthetic view decays, reducing the influence of the potentially imperfect synthetic data, allowing the model to dedicate its capacity to fitting the fine-grained, high-fidelity details present primarily in the ground-truth images. This curriculum ensures that the generative prior provides its main contribution of global structure when most needed, before gracefully yielding to the ground-truth data for final refinement. Please see the experimental settings and supplementary material for details.

4. Experiments

4.1. Experimental Settings

Datasets We conduct our experiments on two datasets, Mip-NeRF360 [1] dataset, which features seven challenging 360° scenes; and NeRF Synthetic [26] dataset which

| Methods | PSNR(\uparrow) | SSIM(\uparrow) | LPIPS (\downarrow) |
|-------------------|--------------------|--------------------|------------------------|
| RegNeRF [29] | 23.86 | 0.852 | 0.105 |
| FreeNeRF [53] | 24.26 | 0.883 | 0.098 |
| SparseNeRF [40] | 24.04 | 0.876 | 0.113 |
| 3DGS [17] | 21.56 | 0.847 | 0.130 |
| DNGaussian [21] | 24.31 | 0.886 | 0.088 |
| FSGS [64] | 24.64 | 0.895 | 0.095 |
| CoR-GS [60] | 24.43 | 0.896 | 0.084 |
| DropGaussian [31] | 25.42 | 0.888 | 0.089 |
| Ours | 24.75 | 0.905 | 0.067 |

Table 2. **Quantitative comparison on the Blender [26] dataset** for 8 input views. Best, second, and third results are marked in red, orange, and yellow respectively.

features eight scenes of 360° pathtraced objects with realistic non-Lambertian materials. For the Mip-NeRF 360 dataset, we use every 8th image as testing view and evenly sample 12 or 24 views from the remaining views as the training set, and all input images are downsampled to $\frac{1}{4}$ of the original height and width. For the NeRF Synthetic dataset, we follow the protocols established in Diet-NeRF [10] and FreeNeRF [53] by using 8 images for training and 25 for testing, all input images are downsampled to $\frac{1}{2}$ of the original height and width.

Implementation Details We inherit the standard parameters from [19] and additionally set $\lambda_{\text{SSIM}} = \lambda_{\text{LPIPS}} = 0.3$ and train up to 22,000 iterations. During training, the probability of sampling a synthetic image is modulated over time according to $p_{\text{synthetic}} \sim 0.1 + 20 \text{Beta}(t; \alpha = 2, \beta = 4)$, where t is the normalized training progress between 0 and 1. We use Depth-Anything V2 [55] to predict pseudo inverse depth and generate synthetic images I' using SEVA [63]. We report PSNR, SSIM [47], and LPIPS [61] to evaluate reconstruction performance quantitatively. We used a single A40 GPU for all experiments. For more implementation details, please refer to the supplementary material.

4.2. Comparison

Evaluation on Mip-NeRF 360 We conduct our evaluation on the challenging Mip-NeRF 360 dataset [1], which features 7 complex, unbounded indoor and outdoor 360° scenes with significant occlusions, making it an ideal benchmark for sparse-view reconstruction. We evaluate all scenes for 12-view setting at 15k and all scenes for 24-view at 22k iterations. As shown in Table 1, our method, **OracleGS**, establishes a new state-of-the-art in the extremely sparse **12-view** setting, outperforming all prior work across all three metrics. This strong performance is particularly notable given the framework’s robustness to generative failures. In the **24-view** setting, the reliance on generative pri-

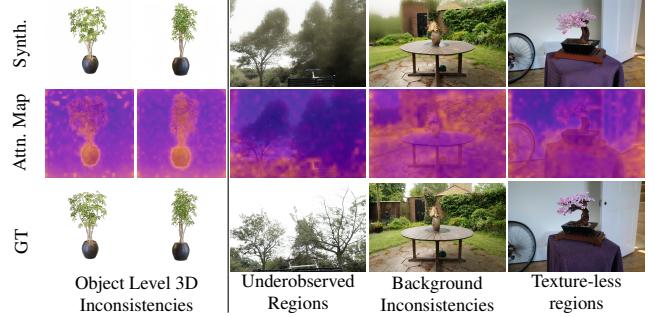


Figure 5. **Our 3D-aware Oracle quantifies diverse sources of 3D uncertainty.** We visualize extracted uncertainty maps (middle row) on synthetic images from the generative model using global-attention layers from the repurposed MVS model [45] normalized from one to zero where low uncertainty is shown in yellow and high uncertainty is shown in purple. Each column demonstrates the oracle’s ability to identify a specific failure mode in the synthetic proposals by comparing against the ground truth.

ors diminishes, but **OracleGS** remains highly competitive, achieving the best performance for PSNR and LPIPS by a margin. This demonstrates that while our approach provides the most significant advantage in the critically underdetermined low-data regime, it maintains the state-of-the-art performance as view density increases. For a fair comparison, we report results for DropGaussian [31] using the standard $\frac{1}{4}$ image downsampling, consistent with prior work, as the official paper evaluates using a non-standard $\frac{1}{8}$ resolution. Figure 3 shows our qualitative results on Mip-NeRF 360 [1] dataset. Please refer to the supplementary material for per scene results on Mip-NeRF 360.

Evaluation on NeRF Synthetic We conduct our secondary evaluation on NeRF Synthetic dataset [26]. Quantitative results on the NeRF Synthetic dataset with 8 training views are reported in Table 2. Our method achieves the best scores in SSIM and LPIPS with second best results in PSNR. This indicates that our method preserves finer, perceptually important details at the expense of minute pixel-level deviations that PSNR is sensitive to as demonstrated in Figure 4 in our qualitative results on NeRF Synthetic [26] dataset.

4.3. Analysis of the 3D-Aware Oracle

To provide a deeper insight into the effectiveness of our proposed 3D-aware oracle, we visualize its output on several challenging scenes. As shown in Figure 5, the oracle is remarkably effective in identifying diverse sources of 3D uncertainty. For instance, the **Ficus** scene (left) demonstrates the oracle’s discriminative power: it assigns low uncertainty to a 3D *consistent* proposal on the right while correctly identifying a hallucinated, *inconsistent* leaf structure in another proposal on the left for the same scene.

| Method | 12-view | | | 24-view | | |
|---------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|
| | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow |
| Baseline [19] | 14.10 | 0.285 | 0.560 | 17.95 | 0.479 | 0.409 |
| + Syn. Views | 16.72 | 0.382 | 0.596 | 16.88 | 0.428 | 0.565 |
| + Schedule | 17.32 | 0.433 | 0.515 | 29.85 | 0.585 | 0.362 |
| + LPIPS Loss | 17.60 | 0.446 | 0.459 | 21.34 | 0.598 | 0.328 |
| + SfM Init. | 19.96 | 0.583 | 0.362 | 22.88 | 0.691 | 0.284 |
| + Depth Loss | 20.28 | 0.592 | 0.354 | 23.64 | 0.717 | 0.258 |
| + Unc. Guide | 20.32 | 0.596 | 0.350 | 23.72 | 0.723 | 0.244 |

Table 3. **Ablation study on Mip-NeRF 360 [1]**, showing the progressive impact of each component for 12-view and 24-view settings. Our full method is shown in bold.

In the **Bicycle** scene (middle left), the oracle identifies regions of high *epistemic uncertainty* that are simply *under-observed* in the sparse input views. By assigning high uncertainty to the background, our method prevents overfitting to a potentially inaccurate generative completion. The oracle is also effective at identifying inconsistencies at various distances, as shown in the **Garden** scene (middle right), where it successfully flags inconsistencies in the distant background foliage on the top left part of the image. Finally, the **Bonsai** scene (right) shows the oracle correctly identifying *textureless* surfaces as inherently uncertain. This is advantageous during optimization because it prevents the model from wasting capacity trying to perfectly replicate potentially noisy generative details on these low-information surfaces, allowing it to focus on reconstructing the high-confidence foreground object. This uncertainty signal is what allows our uncertainty-weighted loss to ground the generative prior.

4.4. Ablation Study

We conduct a comprehensive cumulative ablation study on the Mip-NeRF 360 benchmark [1]. We first demonstrate that naively incorporating **Synthetic Views (+Syn. Views)** into the training doesn’t improve the performance significantly in the sparse-view case as demonstrated by the 12-view ablations. As the number of training images increase, naive incorporation of synthetic images actually hurts performance. This confirms that raw generative artifacts can corrupt the optimization. To address this instability, we introduce a **Progressive Augmentation Schedule (+Schedule)**. This schedule provides a significant performance recovery, and its critical role is visually demonstrated in the supplementary material, where omitting it leads to catastrophic geometric collapse. We further stabilize training with **LPIPS Loss**, providing tolerance to minor generative artifacts that pixel-wise losses penalize. We then incorporate established techniques like **SfM Initialization (+SfM Init.)** and **Depth Regularization (+Depth Loss)**, which, as expected, provide improvements by grounding the op-

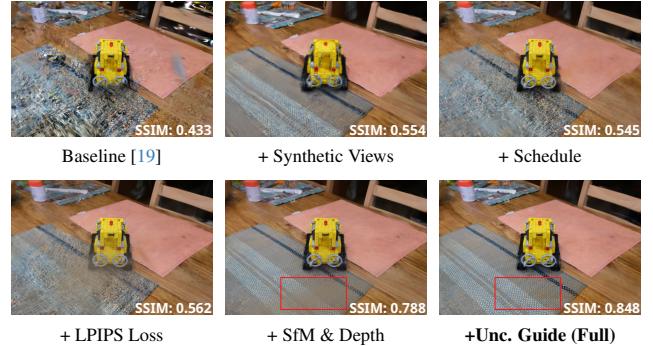


Figure 6. **Qualitative Ablation Study on the Mip-NeRF 360 [1]** ‘kitchen’ scene for 24 views setting.

timization in a plausible geometric state. Finally, we introduce **uncertainty-guided loss (+Unc. Guide)**. Despite being applied after several strong stabilization components which have already significantly improved the baseline, it improves the quantitative and qualitative results. The uncertainty guidance acts as an intelligent filter, selectively integrating the now-stabilized synthetic data based on its 3D uncertainty to achieve our final state-of-the-art performance.

Figure 6 provides a visual breakdown of our ablation study. The baseline 3DGS model struggles with high-frequency texture on the woven mat. Naively incorporating synthetic views leads to oversmoothing and significant geometric errors, such as the distorted, widened shape of the Lego model. Our training schedule corrects this geometric corruption and oversmoothing, but results in similar distortions with the baseline, reducing the SSIM score. The addition of a perceptual LPIPS loss refines texture of the woven mat. While grounding the optimization with SfM initialization and a depth prior correctly scales the scene, it also introduces brown ‘floater’ artifacts on the mat, highlighted in the red rectangle. Finally, our final contribution, uncertainty-guided loss down-weights inconsistent regions in the synthetic data, and removes the brown floaters on the mat, producing a clean, sharp final reconstruction.

5. Conclusion

In this work, we introduced OracleGS, a “propose-and-validate” framework that resolves the trade-off between generative completion and regressive fidelity in sparse-view novel view synthesis. Our method uses a pretrained 3D-aware diffusion model to propose complete scenes and repurposes an MVS model as an oracle to validate them. By distilling 3D uncertainty from the oracle’s global-attention maps, we guide 3D Gaussian Splatting optimization to produce reconstructions that are both visually complete and geometrically sound, achieving state-of-the-art results on the Mip-NeRF 360 [1] and NeRF Synthetic [26] benchmarks.

Acknowledgments. A. Murat Tekalp acknowledges support from Turkish Academy of Sciences (TUBA).

References

- [1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. [3](#)
- [3] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *CVPR*, 2025. [3](#)
- [4] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024. [2](#), [3](#)
- [5] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv preprint arXiv:2503.24391*, 2025. [5](#)
- [6] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. [2](#), [3](#)
- [7] Linus Franke, Laura Fink, and Marc Stamminger. Vr-splatting: Foveated radiance field rendering via 3d gaussian splatting and neural points. *Proc. ACM Comput. Graph. Interact. Tech.*, 8(1), 2025. [1](#)
- [8] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024. [2](#), [3](#)
- [9] Junfu Guo, Yu Xin, Gaoyi Liu, Kai Xu, Ligang Liu, and Ruizhen Hu. Articulatedgs: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting, 2025. [1](#)
- [10] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. [3](#), [7](#)
- [11] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *CVPR*, 2025. [3](#)
- [12] Youngkyoon Jang and Eduardo Pérez-Pellitero. Comapgs: Covisibility map-based gaussian splatting for sparse novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26779–26788, 2025. [3](#), [4](#)
- [13] Lihan Jiang, Kerui Ren, Mulin Yu, Linning Xu, Junting Dong, Tao Lu, Feng Zhao, Dahua Lin, and Bo Dai. Horizon-gs: Unified 3d gaussian splatting for large-scale aerial-to-ground scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26789–26799, 2025. [1](#)
- [14] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, and Chenfanfu Jiang. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. *arXiv preprint arXiv:2401.16663*, 2024. [1](#)
- [15] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025. [2](#), [3](#)
- [16] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. In *Proc. ECCV*, 2024. [3](#)
- [17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [2](#), [3](#), [4](#), [7](#)
- [18] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics*, 43(4), 2024. [6](#)
- [19] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Spotlight Presentation. [3](#), [4](#), [6](#), [7](#), [8](#)
- [20] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. [3](#)
- [21] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *arXiv preprint arXiv:2403.06912*, 2024. [3](#), [4](#), [7](#)
- [22] Kunyi Li, Michael Niemeyer, Zeyu Chen, Nassir Navab, and Federico Tombari. Monogsdf: Exploring monocular geometric cues for gaussian splatting-guided implicit surface reconstruction, 2025. [3](#)
- [23] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *CVPR*, 2024. [1](#)
- [24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. [2](#), [3](#)
- [25] Yang Liu, He Guan, Chuanchen Luo, Lue Fan, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians, 2024. [1](#)
- [26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. [1](#), [3](#), [6](#), [7](#), [8](#)

- [27] Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, Peter Hedman, Ricardo Martin-Brualla, and Jonathan T. Barron. MultiNeRF: A Code Release for Mip-NeRF 360, Ref-NeRF, and RawNeRF, 2022. 3
- [28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. 3
- [29] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4, 7
- [30] Michael Niemeyer, Fabian Manhardt, Marie-Julie Rakotosaona, Michael Oechsle, Daniel Duckworth, Rama Gosula, Keisuke Tateno, John Bates, Dominik Kaeser, and Federico Tombari. Radsplat: Radiance field-informed gaussian splatting for robust real-time rendering with 900+ fps. In *International Conference on 3D Vision 2025*, 2025. 3
- [31] Hyunwoo Park, Gun Ryu, and Wonjun Kim. Dropgaussian: Structural regularization for sparse-view gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21600–21609, 2025. 1, 2, 3, 4, 5, 6, 7
- [32] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 3
- [33] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagnu, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 2, 3
- [34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4
- [35] Suguru Shimomura, Kazuki Yamanouchi, and Jun Tanida. Digital-twin imaging based on descattering gaussian splatting. *Optics Express*, 33(14):29351, 2025. 1
- [36] Nagabhushan Somraj, Adithyan Karanayil, and Rajiv Soundararajan. Simplenerf: Regularizing sparse input neural radiance fields with simpler solutions. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery. 3
- [37] Xiangyu Sun, Runnan Chen, Mingming Gong, Dong Xu, and Tongliang Liu. Intern-gs: Vision model guided sparse-view 3d gaussian splatting, 2025. 3
- [38] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3
- [39] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *34th European Conference on Computer Vision*, pages 439–457, 2024. 2, 3
- [40] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 4, 7
- [41] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3
- [42] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. *CVPR*, 2021. 3
- [43] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment, 2023.
- [44] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggfsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 3
- [45] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 5, 7
- [46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3
- [47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6, 7
- [48] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3
- [49] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J. Fleet. Controlling space and time with diffusion models, 2025. 2, 3
- [50] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 3
- [51] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. *arXiv*, 2023. 2, 3
- [52] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360° sparse view synthesis using gaussian splatting, 2023. 2, 3, 4
- [53] Jiawei Yang, Marco Pavone, and Yue Wang. 3, 7
- [54] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

- [55] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 7
- [56] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024. 3
- [57] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2, 3
- [58] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 3
- [59] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3
- [60] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. Cor-gs: Sparse-view 3d gaussian splatting via co-regularization. *arXiv preprint arXiv:2405.12110*, 2024. 2, 3, 4, 5, 6, 7
- [61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6, 7
- [62] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21336–21345, 2024. 1
- [63] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 2, 3, 4, 7
- [64] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting, 2023. 2, 3, 4, 7
- [65] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2, 3