

RESPLAT: LEARNING RECURRENT GAUSSIAN SPLATS

Haofei Xu^{1,2} **Daniel Barath¹** **Andreas Geiger²** **Marc Pollefeys^{1,3}**

¹ETH Zurich ²University of Tübingen, Tübingen AI Center ³Microsoft



Figure 1: **Learning recurrent Gaussian splats in a feed-forward manner.** We propose ReSplat, a feed-forward recurrent network that iteratively refines 3D Gaussian splats to improve sparse view settings where optimization-based 3DGS (Kerbl et al., 2023) struggles. As initialization (iteration 0), we introduce a compact reconstruction model that predicts Gaussians in a $16 \times$ subsampled space, producing $16 \times$ fewer Gaussians and $4 \times$ faster rendering than per-pixel MVSplat (Chen et al., 2024b) and DepthSplat (Xu et al., 2025). The reduced number of Gaussians makes subsequent refinement efficient. Compared to the optimization-based 3DGS, ReSplat is $100 \times$ faster thanks to its feed-forward design, while still benefiting from iterative updates. Here we show results for 8 input views (512×960 resolution) on DL3DV (Ling et al., 2023) dataset; see Table 1 for detailed metrics.

ABSTRACT

While feed-forward Gaussian splatting models provide computational efficiency and effectively handle sparse input settings, their performance is fundamentally limited by the reliance on a single forward pass during inference. We propose ReSplat, a feed-forward recurrent Gaussian splatting model that iteratively refines 3D Gaussians without explicitly computing gradients. Our key insight is that the Gaussian splatting rendering error serves as a rich feedback signal, guiding the recurrent network to learn effective Gaussian updates. This feedback signal naturally adapts to unseen data distributions at test time, enabling robust generalization. To initialize the recurrent process, we introduce a compact reconstruction model that operates in a $16 \times$ subsampled space, producing $16 \times$ fewer Gaussians than previous per-pixel Gaussian models. This substantially reduces computational overhead and allows for efficient Gaussian updates. Extensive experiments across varying of input views (2, 8, 16), resolutions (256×256 to 540×960), and datasets (DL3DV and RealEstate10K) demonstrate that our method achieves state-of-the-art performance while significantly reducing the number of Gaussians and improving the rendering speed. Our project page is at haofeixu.github.io/resplat.

1 INTRODUCTION

Feed-forward Gaussian splatting (Charatan et al., 2024; Szymanowicz et al., 2024) aims to directly predict 3D Gaussian parameters from input images, eliminating the need for expensive per-scene optimization (Kerbl et al., 2023) and enabling high-quality sparse-view reconstruction and view synthesis (Chen et al., 2024b; Liu et al., 2024; Zhang et al., 2024b; Wang et al., 2024). Very recently, significant progress has been made in this line of research: feed-forward models (Zhang et al., 2024a; Xu et al., 2025; Chen et al., 2025b; Ye et al., 2025a; Chen et al., 2024c) can now produce promising reconstruction and view synthesis results from sparse input views.

Despite these advances, the improved performance remains largely concentrated on standard in-domain benchmarks (Zhou et al., 2018; Ling et al., 2023), and existing feed-forward models often

struggle to generalize to new, unseen datasets and scenarios. A primary reason is that most current methods (Charatan et al., 2024; Chen et al., 2024b; Zhang et al., 2024a; Xu et al., 2025; Chen et al., 2025b) focus on learning a single-step mapping from images to 3D Gaussians. While conceptually simple, this approach is inherently limited by the capacity of the employed network, particularly when reconstructing complex and challenging scenes. In contrast, per-scene optimization methods (Kerbl et al., 2023) achieve high-quality results via many iterative updates but are expensive. This motivates our approach: using multiple learned recurrent steps to progressively improve reconstruction quality, balancing the efficiency of feed-forward methods with the adaptability of iterative optimization.

We first identify that the Gaussian splatting rendering error provides a valuable feedback signal informing the model about the quality of its prediction. This also allows the network to adapt to the test data, reducing the dependence on the training distribution and leading to robust generalization. Moreover, this process can be arbitrarily iterated: the model incrementally refines its prediction, improving quality. This recurrent mechanism not only reduces learning difficulty by decomposing the task into smaller, incremental steps, but also increases model expressiveness with each update, approaching the behavior of an infinitely deep network (Bai et al., 2019; Geiping et al., 2025).

Driven by this observation, we begin with a single-step feed-forward Gaussian reconstruction model to initialize the recurrent process, and then perform recurrent updates to improve the initial Gaussians. Since the recurrent updates occur in 3D space, where a large number of Gaussians would impose a significant computational burden, we design our initial model to predict Gaussians in a $16 \times$ subsampled space. This contrasts with most existing feed-forward models (Charatan et al., 2024; Szymanowicz et al., 2024; 2025; Zhang et al., 2024a) that predict one or multiple Gaussians per pixel, which scales poorly with increasing numbers of views and image resolutions. Our method achieves a $16 \times$ reduction in the number of Gaussians while maintaining performance.

Based on this compact initial reconstruction, we train a weight-sharing recurrent network that iteratively improves the reconstruction. Crucially, the recurrent network leverages the rendering error of input views to determine where and how to update the Gaussians. Specifically, we render the input views (available at test time) using the current prediction, compute the rendering error, and propagate it in feature space to 3D Gaussians. The recurrent network then predicts the parameter updates directly from this error and the current Gaussians, without requiring explicit gradient computation.

We validate our method through extensive experiments across diverse scenarios. On the challenging DL3DV (Ling et al., 2023) dataset, using 8 input views at 512×960 resolution, our learned recurrent model improves PSNR by +2.7 dB, while using only 1/16 of the Gaussians and achieving 4 \times faster rendering speed. We also demonstrate that our recurrent model leads to robust generalization to unseen datasets and image resolutions, where previous single-step feed-forward models usually struggle. With 16 input views at 540×960 resolution, we outperform Long-LRM (Chen et al., 2025b) while using 4 \times fewer Gaussians. On the commonly used two-view RealEstate10K benchmark, our ReSplat also achieves state-of-the-art results, demonstrating the strong performance of our method.

2 RELATED WORK

Feed-Forward Gaussian Splatting. Significant progress has recently been made in feed-forward Gaussian models (Zhang et al., 2024a; Chen et al., 2025b; Xu et al., 2025). However, two major limitations persist: First, most existing feed-forward models predict one or multiple Gaussians for each pixel (Charatan et al., 2024; Szymanowicz et al., 2025; Chen et al., 2025b; 2024b; Zhang et al., 2024a), which will produce millions of Gaussians when handling many input views and/or high-resolution images and thus limit the scalability. Second, most existing methods are developed with single-step feed-forward inference, while conceptually simple, the achievable quality is bounded by the network capability for challenging and complex scenes. In this paper, we overcome these two limitations by first reconstructing Gaussians in a $16 \times$ subsampled space, and then performing recurrent Gaussian update based on the rendering error, which significantly improves the efficiency and quality compared to previous methods. Different from the recent method SplatFormer (Chen et al., 2025a) that introduces a single-step refinement network for optimized 3DGS parameters, we propose a weight-sharing recurrent network to iteratively improve the results by using a feed-forward reconstruction as initialization. In addition, SplatFormer is evaluated only on object-centric datasets and it's not straightforward to make it work for complex scenes. In contrast, our ReSplat targets

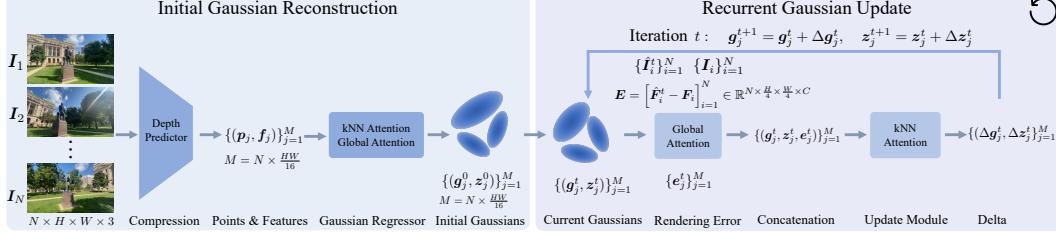


Figure 2: **Learning to recurrently update 3D Gaussians.** Given N posed input images, we first predict per-view depth maps at $1/4$ resolution and then unproject and transform them to a point cloud with image features $\{(p_j, f_j)\}_{j=1}^M$, where $M = N \times \frac{HW}{16}$ is the number of points. We then reconstruct an initial set of 3D Gaussians $\{(g_j^0, z_j^0)\}_{j=1}^M$ in a $16 \times$ subsampled 3D space with a kNN and global attention-based Gaussian regressor. Next, we learn to refine the initial Gaussians recurrently. At each recurrent step t , we use the current Gaussian prediction to render input views and then compute the feature-level rendering errors E between rendered and ground-truth input views. A global attention is next applied on the rendering error to propagate the rendering errors to the 3D Gaussians. A kNN attention-based update module next takes as input the concatenation of current Gaussian parameters g_j^t , the hidden state z_j^t , and the rendering error e_j^t , and predicts the incremental updates Δg_j^t and Δz_j^t . We iterate this process until a total number of T steps.

scene-level benchmarks and we demonstrate the effectiveness of the rendering error as an informative feedback signal, which we found crucial but is missing in SplatFormer.

Learning to Optimize. Many tasks in machine learning and computer vision can be formulated as a minimization problem with an optimization objective, where the solutions are found by iterative gradient decent (Andrychowicz et al., 2016; Lucas & Kanade, 1981; Sun et al., 2010). Modern approaches (Ma et al., 2020; Teed & Deng, 2020; Metz et al., 2022; Harrison et al., 2022) try to simulate the optimization process by iteratively updating an initial prediction with a weight-sharing network, which usually achieves superior results compared to single-step regression methods, especially for out-of-distribution generalization. In vision, such a framework has been successfully applied to optical flow (Teed & Deng, 2020), stereo matching (Lipson et al., 2021; Wen et al., 2025a), scene flow (Teed & Deng, 2021b), SLAM (Teed & Deng, 2021a), Structure-from-Motion (Li et al., 2025), Multi-View Stereo (Wang et al., 2022), etc. Unlike prior work that often relies on feature correlations (Teed & Deng, 2020) for the recurrent process, we identify the Gaussian rendering error as an informative feedback signal to guide the update of 3D Gaussian parameters.

Learning to Optimize for View Synthesis. In the context of view synthesis, DeepView (Flynn et al., 2019) predicts multi-plane images with learned gradient decent, where explicit gradient computation is necessary. In addition, G3R (Chen et al., 2024d) proposes to learn to iteratively refine the 3D Gaussians with the guidance of the explicitly computed gradients. However, our method is gradient free. Moreover, G3R requires well-covered 3D points for initialization and struggles with sparse points, while we directly predict initial Gaussians from posed images, without requiring any initial 3D points. Similar to G3R, QuickSplat (Liu et al., 2025) also relies on gradient computation but focuses on surface reconstruction. Another related work LIFe-GOM (Wen et al., 2025b) tries to iteratively update the 3D representation in a gradient-free manner, but it focuses on the task of human avatars with a hybrid Gaussian-mesh 3D representation. In contrast, our method aims to improve the quality and generalization of feed-forward Gaussian splatting models for general scenes.

3 APPROACH

Given N input images $\{I^i\}_{i=1}^N$ ($I^i \in \mathbb{R}^{H \times W \times 3}$) with their intrinsic $\{K^i\}_{i=1}^N$ ($K^i \in \mathbb{R}^{3 \times 3}$) and extrinsic $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=1}^N$ ($\mathbf{R}_i \in \text{SO}(3)$, $\mathbf{t}_i \in \mathbb{R}^3$) matrices, our goal is to predict a set of 3D Gaussian primitives (Kerbl et al., 2023) $\mathcal{G} = \{(\mu_j, \alpha_j, \Sigma_j, \mathbf{sh}_j)\}_{j=1}^M$ to model the scene, where M is the total number of Gaussian primitives and μ_j , α_j , Σ_j and \mathbf{sh}_j are the 3D Gaussian’s position, opacity, covariance, and spherical harmonics, respectively. The reconstructed 3D Gaussians can be efficiently rasterized, enabling fast and high-quality novel view synthesis.

Unlike previous feed-forward models (Charatan et al., 2024; Zhang et al., 2024a; Chen et al., 2025b) that perform a single-step feed-forward prediction, we propose to learn to estimate the Gaussian parameters recurrently. This not only reduces learning difficulty by decomposing the task into smaller, incremental steps but also enables higher reconstruction quality. In particular, we first predict an initial set of 3D Gaussians and then iteratively refine them in a gradient-free, feed-forward manner. Given that the Gaussian update occurs in the 3D space, a large number of 3D Gaussians will introduce significant computational overhead during the update process. Thus, in our initial reconstruction stage, we predict a compact set of 3D Gaussians in a $16 \times$ subsampled space. More specifically, we perform $4 \times$ spatial compression when predicting per-view depth maps which leads to $16 \times$ fewer number of Gaussians compared to previous per-pixel representation (Charatan et al., 2024; Szymonowicz et al., 2024). Thus, the number of Gaussians M in our model is $N \times \frac{HW}{16}$, which scales efficiently to many input views and high-resolution images. Fig. 2 provides an overview of our pipeline.

3.1 INITIAL GAUSSIAN RECONSTRUCTION

Subsampled 3D Space. Our initial Gaussian reconstruction model is based on the DepthSplat (Xu et al., 2025) architecture, which includes both multi-view and monocular branches for per-pixel Gaussian prediction. However, unlike DepthSplat, we predict Gaussians in a spatially $16 \times$ subsampled 3D space ($N \times \frac{HW}{16}$), and thus our number of Gaussians is $16 \times$ fewer than DepthSplat. To achieve $16 \times$ subsampling, we resize the full resolution depth predictions from the depth model in DepthSplat to $1/4$ resolution ($N \times \frac{H}{4} \times \frac{W}{4}$), and then unproject and transform them in 3D via camera parameters to obtain a point cloud with $M = N \times \frac{HW}{16}$ points. Each 3D point $\mathbf{p}_j \in \mathbb{R}^3$ is also associated with a feature vector $\mathbf{f}_j \in \mathbb{R}^{C_1}$ extracted from the input images. This process can be expressed as

$$\{\mathbf{I}_i, \mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}_{i=1}^N \rightarrow \{(\mathbf{p}_j, \mathbf{f}_j)\}_{j=1}^M. \quad (1)$$

Since we now have $16 \times$ fewer 3D points, naïvely predicting Gaussian parameters from the point features \mathbf{f}_j will lead to considerable performance loss. However, we found that using additional kNN attention (Zhao et al., 2021) and global attention (Vaswani et al., 2017) layers on the point cloud to encode the 3D context (Xu et al., 2024; Chen et al., 2024a) information can compensate for this loss.

Aggregating the 3D Context. We use six alternative blocks of kNN attention and global attention to model both local and global 3D contexts, which enables communication between different 3D points and produce 3D context-aggregated features $\mathbf{f}_j^* \in \mathbb{R}^{C_1}$ with increased expressiveness:

$$\{(\mathbf{p}_j, \mathbf{f}_j)\}_{j=1}^M \rightarrow \{(\mathbf{p}_j, \mathbf{f}_j^*)\}_{j=1}^M. \quad (2)$$

Decoding to Gaussians. We use the point cloud $\{\mathbf{p}_j\}_{j=1}^M$ as the Gaussian centers and other Gaussian parameters are decoded from the 3D context-aggregated features $\{\mathbf{f}_j^*\}_{j=1}^M$. Accordingly, we obtain an initial set of 3D Gaussians with parameters $\{(\mu_j, \alpha_j, \Sigma_j, \mathbf{sh}_j)\}_{j=1}^M$ and feature vectors $\{\mathbf{f}_j^*\}_{j=1}^M$. We use $\mathbf{g}_j^0 \in \mathbb{R}^{C_2}$ to denote the concatenation of all the Gaussian parameters $(\mu_j, \alpha_j, \Sigma_j, \mathbf{sh}_j)$ for the j -th Gaussian at initialization, where C_2 is the total number of parameters for each Gaussian. We use \mathbf{z}_j^0 to denote the initial hidden state of the j -th Gaussian for the subsequent recurrent process, and initialize it with feature \mathbf{f}_j^* : $\mathbf{z}_j^0 = \mathbf{f}_j^* \in \mathbb{R}^{C_1}$. Thus, the initial Gaussians can be represented as

$$\mathcal{G}^0 = \{(\mathbf{g}_j^0, \mathbf{z}_j^0)\}_{j=1}^M. \quad (3)$$

3.2 RECURRENT GAUSSIAN UPDATE

Based on the initial Gaussian prediction in Section 3.1 (Eq. (3)), we train a recurrent network which iteratively refines the initial prediction. In particular, at iteration t , ($t = 0, 1, \dots, T - 1$, T is the total number of iterations), the recurrent network predicts incremental updates to all Gaussian parameters $\Delta \mathbf{g}_j^t \in \mathbb{R}^{C_2}$ and their hidden state $\Delta \mathbf{z}_j^t \in \mathbb{R}^{C_1}$ as:

$$\mathbf{g}_j^{t+1} = \mathbf{g}_j^t + \Delta \mathbf{g}_j^t, \quad \mathbf{z}_j^{t+1} = \mathbf{z}_j^t + \Delta \mathbf{z}_j^t. \quad (4)$$

To predict the incremental updates $\Delta \mathbf{g}_j^t$ and $\Delta \mathbf{z}_j^t$, we propose to learn the update in a gradient-free, feed-forward manner from the rendering error of input views.

Computing the Rendering Error. Given that we have access to the input views at test time, we are able to create a feedback loop to guide the recurrent network to learn the incremental updates. More specifically, we first render the input views $\{\hat{I}_i^t\}_{i=1}^N$ based on the current Gaussian parameters at iteration t , and then measure the difference between the rendered and ground-truth input views. We observe that measuring the rendering error in the feature space performs better than in the raw RGB space. In particular, we extract the first-three-stage features (at 1/2, 1/4 and 1/8 resolutions) of the ImageNet (Deng et al., 2009) pre-trained ResNet-18 (He et al., 2016) for the rendered input views and ground-truth input views, and bilinearly resize the three-scale features to the same 1/4 resolution, followed by concatenation. We denote the extracted features as $\{\hat{F}_i^t\}_{i=1}^N$ and $\{F_i\}_{i=1}^N$ ($\hat{F}_i^t, F_i \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_3}$) for rendered and ground-truth input views, respectively. We then compute the difference between features with subtraction $\{\hat{F}_i^t - F_i\}_{i=1}^N$. We denote all rendering errors as $\hat{E}^t = \{\hat{e}_j^t\}_{j=1}^{N \times \frac{H}{4} \times \frac{W}{4}}$, where $\hat{e}_j^t \in \mathbb{R}^{C_3}$ is the j -th feature difference of dimension C_3 at iteration t .

Propagating the Rendering Error to Gaussians. To propagate the rendering error to 3D Gaussians such that they can guide the network to update the Gaussians, a straightforward approach is to concatenate the rendering error \hat{e}_j^t with the Gaussians (g_j^t, z_j^t) in a spatially aligned manner since they have the same number of points ($N \times \frac{HW}{16}$). However, in this way, the j -th Gaussian can only receive information about the j -th rendered pixel, while it can also contribute to other rendered pixels during the rendering process. To propagate the rendering error more effectively, we propose to apply global attention on all the $N \times \frac{H}{4} \times \frac{W}{4}$ rendering errors \hat{E}^t , which enables each Gaussian to receive information from all rendering errors. This process can be formulated as

$$\mathbf{E} = \text{global_attention}(\hat{\mathbf{E}}) = \{\mathbf{e}_j^t\}_{j=1}^{N \times \frac{HW}{16}}, \quad (5)$$

where \mathbf{e}_j^t is the j -th rendering error which has aggregated the original point-wise rendering error \hat{e}_j^t globally. We then concatenate the Gaussians with the globally aggregated rendering errors as $\{(g_j^t, z_j^t, \mathbf{e}_j^t)\}_{j=1}^M$, which are next used to predict the incremental update (illustrated in Fig. 2).

Recurrent Gaussian Update. Let the Gaussians at iteration t be $\mathcal{G}^t = \{(g_j^t, z_j^t)\}_{j=1}^M$, our update module predicts the incremental updates of Gaussian parameters and hidden state as

$$\{(g_j^t, z_j^t, \mathbf{e}_j^t)\}_{j=1}^M \rightarrow \{(\Delta g_j^t, \Delta z_j^t)\}_{j=1}^M, \quad (6)$$

and then they are added to the current prediction (Eq. (4)). This process is iterated for T times. We observe that our model converges after 3 iterations. During training, we randomly sample the number of iterations T between 1 and 3, and our model supports different number of iterations at inference time, allowing a flexible trade-off between accuracy and speed with a single model. Since the recurrent process occurs in the 3D space, we choose to use four kNN attention (Zhao et al., 2021) blocks as the recurrent architecture to model the local structural details.

3.3 TRAINING LOSS

Our model is trained in two stages. In the first stage, we train an initial Gaussian reconstruction model to provide compact initializations to our subsequent updates. The training loss is a combination of a rendering loss ℓ_{render} and an edge-aware depth smoothness loss $\ell_{\text{depth_smooth}}$ on the predicted depth maps of the input views:

$$L_{1\text{st}} = \sum_{v=1}^V \ell_{\text{render}}(\hat{\mathbf{I}}_v, \mathbf{I}_v) + \alpha \cdot \sum_{i=1}^N \ell_{\text{depth_smooth}}(\mathbf{I}_i, \hat{\mathbf{D}}_i), \quad (7)$$

$$\ell_{\text{render}}(\hat{\mathbf{I}}, \mathbf{I}) = \ell_1(\hat{\mathbf{I}}, \mathbf{I}) + \lambda \cdot \ell_{\text{perceptual}}(\hat{\mathbf{I}}, \mathbf{I}), \quad (8)$$

$$\ell_{\text{depth_smooth}}(\mathbf{I}, \hat{\mathbf{D}}) = |\partial_x \hat{\mathbf{D}}| e^{-|\partial_x \mathbf{I}|} + |\partial_y \hat{\mathbf{D}}| e^{-|\partial_y \mathbf{I}|}. \quad (9)$$

where V is the number of target views to render in each training step. N is the number of input views. The perceptual loss $\ell_{\text{perceptual}}$ (Johnson et al., 2016) measures the distance in the VGG (Simonyan & Zisserman, 2014)'s feature space, which is also used in previous methods (Zhang et al., 2024a; Jin et al., 2025). The depth smoothness loss $\ell_{\text{depth_smooth}}$ is a regularization term on the estimated depth maps of the input views to encourage the depth gradient to be similar to the image gradient (Godard et al., 2017; 2019). We use $\alpha = 0.01$ and $\lambda = 0.5$ for all the experiments.

In the second stage, we freeze our initial reconstruction model and train only the recurrent model end-to-end. We use the rendering loss ℓ_{render} of rendered and ground truth target views to supervise the network. All the Gaussian predictions during the recurrent process are supervised with the rendering loss with exponentially ($\gamma = 0.9$) increasing weights:

$$L_{\text{2nd}} = \sum_{t=0}^{T-1} \gamma^{T-1-t} \sum_{v=1}^V \ell_{\text{render}}(\hat{\mathbf{I}}_v^t, \mathbf{I}_v). \quad (10)$$

4 EXPERIMENTS

Implementation Details. We implement our method in PyTorch and use Flash Attention 3 (Shah et al., 2024) for efficient attention computations. We choose $k = 16$ for kNN attentions following Point Transformer (Zhao et al., 2021), and our Gaussian splatting renderer is based on gsplat (Ye et al., 2025b)’s Mip-Splatting (Yu et al., 2024) implementation. We optimize our model with AdamW (Loshchilov & Hutter, 2017) optimizer with cosine learning rate schedule.

Our model contains several global attention layers. Considering that performing global attention on $N \times \frac{H}{4} \times \frac{W}{4}$ features would be expensive for high-resolution images, we first perform $4 \times$ spatial downsampling with pixel unshuffle (reshaping from the spatial dimension to the channel dimension) and then compute global attention on the $N \times \frac{H}{16} \times \frac{W}{16}$ features. Finally, we upsample the features back to the $1/4$ resolution with pixel shuffle (reshaping from the channel dimension to the spatial dimension). This implementation enables our model to scale efficiently to high-resolution images.

Our default model uses a ViT-B (Dosovitskiy et al., 2020; Yang et al., 2024) backbone as the monocular branch of our depth prediction model (Xu et al., 2025), which has 223M parameters in total (209M for the initialization model and 14M for the recurrent model). For ablation experiments, we use a ViT-S backbone to save compute, which has 76M parameters in total (62M for the initialization model and 14M for the recurrent model).

We will release our code, pre-trained models, training and evaluation scripts to ease reproducibility.

Evaluation Setup. To compare with previous methods, we mainly consider three evaluation setups. First, view synthesis from 8 input views at 512×960 resolution on DL3DV, where we re-train 3DGS (Kerbl et al., 2023), MVSplat (Chen et al., 2024b) and DepthSplat (Xu et al., 2025) with their public code for fair comparisons. Second, view synthesis from 16 input views at 540×960 resolution on DL3DV following Long-LRM (Chen et al., 2025b), with which we could perform comparisons with Long-LRM since its model weights are not released and re-training would be expensive. Third, we also evaluate on the commonly used 2-view (256×256) setup on RealEstate10K, where we compare with related methods like GS-LRM (Zhang et al., 2024a) and LVSM (Jin et al., 2025).

Training Details. For experiments on DL3DV, we adopt a progressive training strategy with gradually increased image resolutions and number of input views for better efficiency. More specifically, we first train our model with 8 input views at 256×448 resolution, and then we fine-tune the model with 8 input views at 512×960 resolution, and finally we fine-tune the model with 16 input views at 512×960 resolution. For each stage, we train with 16 GH200 GPUs for 80K steps, with 50K steps for the initial reconstruction model and 30K steps for the recurrent model. For experiments on RealEstate10K at 256×256 resolution, we first train the initial model with 16 GH200 GPUs for 200K steps and then train the recurrent model for 100K steps.

4.1 MAIN RESULTS

8 Views at 512×960 Resolution on DL3DV. We report the results on the DL3DV benchmark split (140 scenes) in Table 1. For experiments with 3DGS (Kerbl et al., 2023), we perform per-scene optimization on the 8 input views for all the 140 scenes, while for feed-forward models, we perform zero-shot feed-forward inference on the unseen scenes after training on the DL3DV training set. We observe that 3DGS optimization typically converges with 4K optimization steps, and optimizing longer could lead to overfitting due to the sparse input views, thus we report the best results at 4K iterations. As shown in Table 1, 3DGS optimization is considerably slow due to the large number of iterations required, while our feed-forward ReSplat is 100 \times faster and is able to benefit from additional iterations. We observe that our recurrent model converges after 3 iterations, which also

Table 1: **Evaluation of 8 input views (512×960) on DL3DV.** The standard optimization-based approach 3DGS (Kerbl et al., 2023) requires several thousands of iterations to reach convergence, while our feed-forward ReSplat is significantly faster and is able to benefit from additional iterations. Previous per-pixel feed-forward models MVsplat (Chen et al., 2024b) and DepthSplat (Xu et al., 2025) produces millions of Gaussians, while our ReSplat compresses the number of Gaussians by $16\times$, also leading to $4\times$ faster rendering speed. Time is measured on a single GH200 GPU.

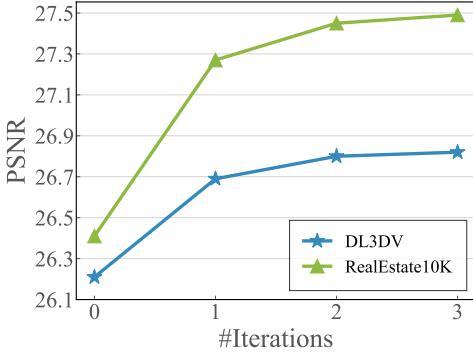
Method	Category	#Iterations	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	#Gaussians	Recon. Time (s)	Render Time (s)
3DGS	Optimization	1000	20.36	0.667	0.448	9K	15	0.0001
		2000	23.18	0.763	0.269	137K	31	0.0005
		3000	23.42	0.770	0.232	283K	50	0.0008
		4000	23.46	0.770	0.224	359K	70	0.0009
MVsplat	Feed-Forward	0	22.49	0.764	0.261	3932K	0.129	0.0030
DepthSplat	Feed-Forward	0	24.17	0.815	0.208	3932K	0.190	0.0030
ReSplat	Feed-Forward	0	26.21	0.842	0.185	246K	0.311	0.0007
		1	26.69	0.849	0.176	246K	0.440	0.0007
		2	26.80	0.850	0.174	246K	0.569	0.0007
		3	26.82	0.850	0.174	246K	0.693	0.0007



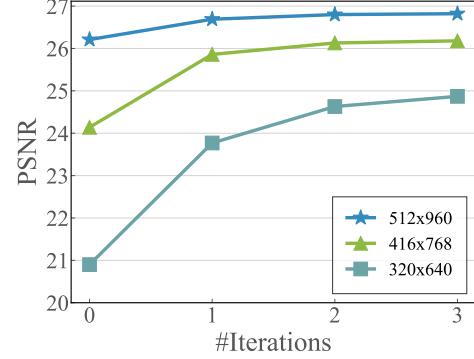
Figure 3: **Visual comparisons on DL3DV.** Our ReSplat renders higher quality images than both optimization and feed-forward methods.

makes it efficient thanks to the fast convergence. In addition, previous per-pixel feed-forward models MVsplat (Chen et al., 2024b) and DepthSplat (Xu et al., 2025) produces millions of Gaussians, while our ReSplat compresses the number of Gaussians by $16\times$, also leading to $4\times$ faster rendering speed. Overall, our ReSplat outperforms 3DGS by 3.3dB PSNR and DepthSplat by 2.6dB PSNR with superior efficiency on the number of Gaussians and the rendering speed. The visual comparisons are shown in Fig. 3, which demonstrates the higher rendering quality of our method. We provide additional visualizations of reconstructed Gaussians in appendix (Fig. 7 and Fig. 8).

Generalization Across Datasets and Image Resolutions. We further evaluate the generalization abilities of our feed-forward model trained on DL3DV at 512×960 resolution. First, when generalizing to the unseen RealEstate10K dataset (Fig. 4a), the improvement brought by our recurrent model is more significant since unlike single-step feed-forward models (iteration 0), our model is able to adapt to the test data using the rendering error and thus reduces the domain gap. Second, existing single-step feed-forward models usually suffer from performance drop when the testing image resolution is different from training. However, with our recurrent model, we can significantly improve the robustness to different testing resolutions (Fig. 4b). For example, our recurrent model improves 4dB PSNR when generalizing from 512×960 to 320×640 . Both experiments demonstrate the benefits of our recurrent model, which is able to adapt to the unseen test scenario using the rendering error as a feedback signal and thus greatly improves the robustness.



(a) **Cross-dataset generalization.** Our DL3DV-trained recurrent model improves more when generalizing to the unseen RealEstate10K dataset.



(b) **Cross-resolution generalization.** Our recurrent model is more robust to unseen image resolutions that are different from training (512×960).

Figure 4: **Generalization to unseen datasets and image resolutions.** The improvement brought by our recurrent model is more pronounced when generalizing to unseen scenarios. The model is trained on DL3DV at 512×960 resolution. We note that the RealEstate10K dataset is easier than DL3DV.

Table 2: **Evaluation of 16 input views** (540×960) on **DL3DV**. ReSplat produces $4 \times$ fewer Gaussians than Long-LRM but still outperforms it. The results of other baselines are borrowed from Long-LRM.

Method	#Iterations	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Recon. Time	#Gaussians
3DGS (Kerbl et al., 2023)	30000	21.20	0.708	0.264	13min	-
Mip-Splatting (Yu et al., 2024)	30000	20.88	0.712	0.274	13min	-
Scaffold-GS (Lu et al., 2024)	30000	22.13	0.738	0.250	16min	-
Long-LRM (Chen et al., 2025b)	0	22.66	0.740	0.292	0.4sec	2073K
	0	22.69	0.742	0.307	0.7sec	518K
ReSplat (Ours)	1	23.12	0.752	0.296	1.2sec	518K
	2	23.23	0.755	0.293	1.7sec	518K

Method	w/ 3DGS	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
pixelSplat	✓	25.89	0.858	0.142
MVSplat	✓	26.39	0.869	0.128
DepthSplat	✓	27.47	0.889	0.114
GS-LRM	✓	28.10	0.892	0.114
Long-LRM	✓	28.54	0.895	0.109
LVSM (enc-dec)	✗	28.58	0.893	0.114
LVSM (dec-only)	✗	29.67	0.906	0.098
ReSplat (Ours)	✓	29.72	0.911	0.100

Table 3: **Evaluation of two input views** (256×256) on **RealEstate10K**. ReSplat outperforms all previous feed-forward 3DGS models. Compared to the 3DGS-free model LVSM, results are similar, but ReSplat offers $20 \times$ faster rendering speed thanks to the 3DGS representation.

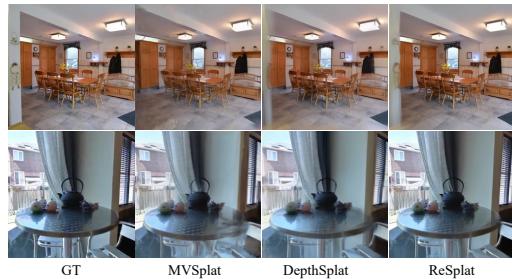


Figure 5: **Visual comparisons on RealEstate10K.** ReSplat produces sharper structures than MVSplat and DepthSplat.

16 Views at 540×960 Resolution on DL3DV. We follow Long-LRM (Chen et al., 2025b) for this evaluation setup such that a direct comparison is possible. The results of 3DGS (Kerbl et al., 2023), Mip-Splatting (Yu et al., 2024), and Scaffold-GS (Lu et al., 2024) are borrowed from Long-LRM paper. This experiment aims to reconstruct the full DL3DV scene from 16 input views, which is particularly challenging due to the large scale of the DL3DV dataset. However, our ReSplat still outperforms previous optimization and feed-forward methods, as shown in Table 2. Notably, Long-LRM uses Gaussian pruning based on the opacity values during training and evaluation, which leads to $4 \times$ reduction of number of Gaussians. In contrast, we compress the Gaussians by $16 \times$, thus our final reconstruction has $4 \times$ fewer Gaussians than Long-LRM while still outperforming it. Our reconstruction time is slower than Long-LRM mainly because of the kNN operation, which

Compression PSNR ↑ SSIM ↑ LPIPS ↓ Time (s)				
64×	24.77	0.797	0.226	0.096
16×	26.77	0.865	0.142	0.104
4×	28.36	0.900	0.103	0.206

Table 4: **Different compression factors.** 16× compression represents a good speed-accuracy trade-off.



Figure 6: **Different compression factors.**

Table 5: **Ablations.** We conduct several ablation experiments to validate our design choices.

(a) **Ablation of the initial reconstruction model.**

Method	PSNR ↑	SSIM ↑	LPIPS ↓	#Gaussians
DepthSplat	25.79	0.861	0.134	918K
Full	26.77	0.865	0.142	57K
w/o kNN attn	25.30	0.833	0.178	57K
w/o global attn	26.33	0.856	0.150	57K
w/o kNN, w/o global	24.50	0.814	0.200	57K

(b) **Ablation of the recurrent reconstruction model.**

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Initialization	26.77	0.865	0.142
Full	27.83	0.880	0.122
w/o render error	26.77	0.865	0.142
w/o ResNet feature	27.78	0.879	0.128
w/o kNN attn	27.64	0.877	0.124
w/o global attn	27.77	0.879	0.122

is expensive to compute with large number (*e.g.*, > 500K) of points. Further optimizing the implementation could potentially improve our reconstruction speed.

2 Views at 256 × 256 Resolution on RealEstate10K. Considering the redundancy is not a major issue with two input views at low resolution (256 × 256), we perform 4× subsampling in the 3D space and decode 4 Gaussians from each subsampled 3D point in our initial reconstruction model. Thus the number of Gaussians remains the same as previous per-pixel methods. The recurrent process remains the same as the many-view setups. Table 3 shows that our ReSplat outperforms previous feed-forward 3DGS models (*e.g.*, DepthSplat (Xu et al., 2025), GS-LRM (Zhang et al., 2024a) and Long-LRM (Chen et al., 2025b)) by clear margins. We present visual comparisons in Fig. 5, where our ReSplat produces better structures than MVsplat and DepthSplat. Compared to the 3DGS-free method LVSM (Jin et al., 2025), we outperform its encoder-decoder architecture by 1.1dB PSNR and our results are similar to its best-performing decoder-only model variant. However, our method offers the benefits of an explicit 3D Gaussian representation, enabling 20× faster rendering speed.

4.2 ANALYSIS AND ABLATION

We conduct several experiments to analyze the properties of our model and validate our design choices. To save compute, all experiments in this section are performed using 8 input views at 256 × 448 resolution on the DL3DV dataset, and we train all the models on 4 GH200 GPUs.

Compression Factor. Be default, we compress the number of Gaussians by 16× using depth maps at 1/4 resolution. In Table 4, we compare with 64× (with 1/8 depth maps) and 4× (with 1/2 depth maps) compression factors. We can observe that less compression leads to higher quality, as also visualized in Fig. 6. However, 4× compression is 2× slower than 16× for 8 input views at 256 × 448 resolution. It would be more expensive when handling higher resolution images (*e.g.*, 512 × 960). Thus, we choose 16× compression as a good speed-accuracy trade-off.

Ablation of the Initial Model. As shown in Table 5, we observe that the kNN attention is crucial to maintain our performance when compressing the number of Gaussians by 16×. The global attention also brings moderate performance gains, which indicates that both local and global 3D context aggregation is important for learning compact 3D representations.

Ablation of the Recurrent Model. The rendering error plays an indispensable role in our recurrent model. As shown in Table 5, we didn’t observe any obvious improvement by removing the rendering

error as input to our recurrent model. Computing the error in ResNet feature space is beneficial for the LPIPS metric, and both kNN attention and global attention contribute to the performance.

5 CONCLUSION

We present a recurrent Gaussian splatting model that achieves efficient and high-quality view synthesis in a feed-forward, gradient-free manner. By leveraging the rendering error as a feedback signal and operating in a compact subsampled 3D space, our method significantly reduces the number of Gaussians while improving performance and generalization across diverse scenarios.

Limitations. Our current model relies on the kNN-based point attention (Zhao et al., 2021), which incurs high computational cost when the number of Gaussians becomes large (*e.g.*, $> 500K$). We anticipate that more efficient point-based attentions (Wu et al., 2022; 2024) and sparse structures (Ren et al., 2024) could further improve the scalability and efficiency of our approach. In addition, our current model saturates after three iterations, and we speculate that keeping the number of Gaussians fixed during the recurrent process to be a potential reason and it would be interesting to design more adaptive update strategies and explore how to further scale up test-time compute. We view these directions as promising avenues for future research.

Acknowledgments. We thank Naama Pearl, Xudong Jiang, and Stefano Esposito for the insightful comments, and Yung-Hsu Yang and Kashyap Chitta for the fruitful discussions. Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645. This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a144 on Alps.

REFERENCES

- Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 2016.
- Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *NeurIPS*, 2019.
- David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.
- Yihang Chen, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Hac: Hash-grid assisted context for 3d gaussian splatting compression. In *ECCV*, 2024a.
- 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. In *ECCV*, 2024b.
- Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In *NeurIPS*, 2024c.
- Yun Chen, Jingkang Wang, Ze Yang, Sivabalan Manivasagam, and Raquel Urtasun. G3r: Gradient guided generalizable reconstruction. In *ECCV*, 2024d.
- Yutong Chen, Marko Mihajlovic, Xiyi Chen, Yiming Wang, Sergey Prokudin, and Siyu Tang. Splatformer: Point transformer for robust 3d gaussian splatting. In *ICLR*, 2025a.
- Ziwen Chen, 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 *ICCV*, 2025b.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.

- John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, 2019.
- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.
- Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.
- Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.
- James Harrison, Luke Metz, and Jascha Sohl-Dickstein. A closer look at learned optimization: Stability, robustness, and inductive biases. In *NeurIPS*, 2022.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- 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 *ICLR*, 2025.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.
- Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. In *CVPR*, 2025.
- Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. *arXiv*, 2023.
- Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, 2021.
- Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *ECCV*, 2024.
- Yueh-Cheng Liu, Lukas Höller, Matthias Nießner, and Angela Dai. Quicksplat: Fast 3d surface reconstruction via learned gaussian initialization. In *ICCV*, 2025.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv*, 2017.
- Tao Lu, Mulin Yu, Lining Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *CVPR*, 2024.
- Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.
- Wei-Chiu Ma, Shenlong Wang, Jiayuan Gu, Sivabalan Manivasagam, Antonio Torralba, and Raquel Urtasun. Deep feedback inverse problem solver. In *ECCV*, 2020.
- Luke Metz, James Harrison, C. Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, and Jascha Sohl-Dickstein. Velo: Training versatile learned optimizers by scaling up. In *NeurIPS*, 2022.
- Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *CVPR*, 2024.

- Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In *NeurIPS*, 2024.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010.
- Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, 2024.
- Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. In *3DV*, 2025.
- Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.
- Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgbd cameras. *NeurIPS*, 2021a.
- Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *CVPR*, 2021b.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.
- Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *CVPR*, 2022.
- Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *NeurIPS*, 2024.
- Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *CVPR*, 2025a.
- Jing Wen, Alexander G Schwing, and Shenlong Wang. Life-gom: Generalizable human rendering with learned iterative feedback over multi-resolution gaussians-on-mesh. In *ICLR*, 2025b.
- Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022.
- Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024.
- Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: Multi-baseline radiance fields. In *CVPR*, 2024.
- Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, 2025.
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024.
- Botao Ye, Sifei Liu, Haofei Xu, Xuetong Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025a.
- Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for gaussian splatting. *JMLR*, 26(34):1–17, 2025b.
- Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu.
Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, 2024a.

Shengjun Zhang, Xin Fei, Fangfu Liu, Haixu Song, and Yueqi Duan. Gaussian graph network:
Learning efficient and generalizable gaussian representations from multi-view images. *NeurIPS*,
2024b.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In
ICCV, 2021.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification:
learning view synthesis using multiplane images. *ACM TOG*, 2018.

A ADDITIONAL VISUALIZATIONS

We provide visualizations of reconstructed Gaussians in Fig. 7 and Fig. 8, where we can observe that our ReSplat reconstructs cleaner Gaussians and renders higher-quality images than DepthSplat.

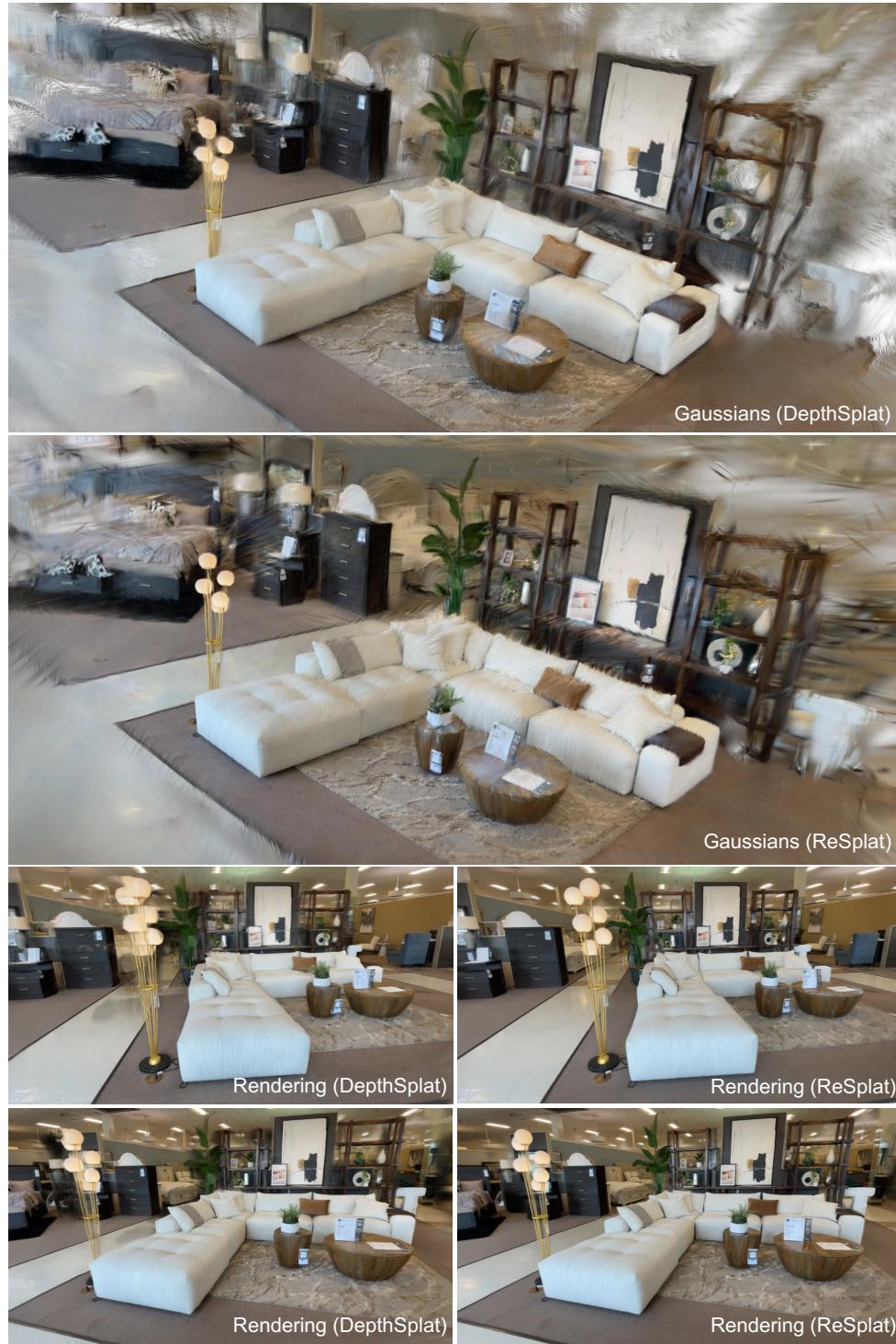


Figure 7: Comparison of reconstructed Gaussians and view synthesis results.

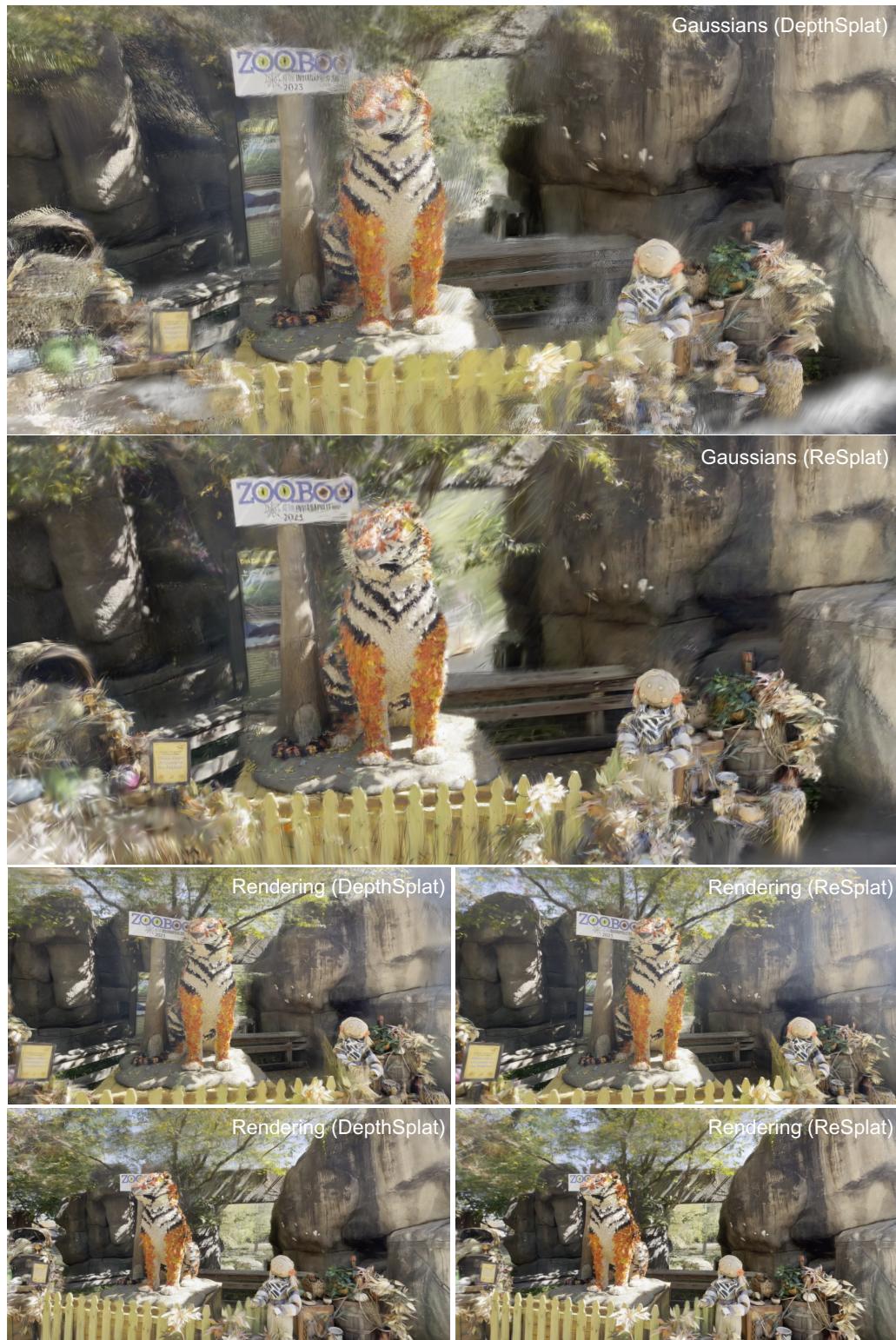


Figure 8: Comparison of reconstructed Gaussians and view synthesis results.