One-Shot Refiner: Boosting Feed-forward Novel View Synthesis via One-Step Diffusion

Yitong Dong[1,2], Qi Zhang[2], Minchao Jiang[2,3], Zhiqiang Wu[2,4], Qingnan Fan[2], Ying Feng[2], Huaqi Zhang[2], Hujun Bao[1], Guofeng Zhang[1]*

[1]Zhejiang University
[2]Hangzhou VIVO Information Technology Co., Ltd
[3]Xidian University
[4]East China Normal University
dongyitong@zju.edu.cn, nwpuqzhang@gmail.com, zhangguofeng@zju.edu.cn

Abstract

We present a novel framework for high-fidelity novel view synthesis (NVS) from sparse images, addressing key limitations in recent feed-forward 3D Gaussian Splatting (3DGS) methods built on Vision Transformer (ViT) backbones. While ViT-based pipelines offer strong geometric priors, they are often constrained by low-resolution inputs due to computational costs. Moreover, existing generative enhancement methods tend to be 3D-agnostic, resulting in inconsistent structures across views, especially in unseen regions. To overcome these challenges, we design a Dual-Domain Detail Perception Module, which enables handling high-resolution images without being limited by the ViT backbone, and endows Gaussians with additional features to store high-frequency details. We develop a feature-guided diffusion network, which can preserve high-frequency details during the restoration process. We introduce a unified training strategy that enables joint optimization of the ViT-based geometric backbone and the diffusion-based refinement module. Experiments demonstrate that our method can maintain superior generation quality across multiple datasets.

1 Introduction

Scene understanding and novel view synthesis have seen rapid progress in recent years (Dong et al. 2022; Zhai et al. 2025b,a), largely driven by the transformative impact of differentiable rendering (Mildenhall et al. 2021) on digital content creation. Specifically, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has achieved an unprecedented breakthrough in this domain, enabling high-fidelity rendering with remarkable performance. However, this approach faces a critical limitation: 3DGS requires extensive, time-consuming optimization for each individual scene, which serves as a significant bottleneck, prohibiting efficient applications and limiting its generalizability.

To overcome this limitation, a new paradigm has emerged built upon 3D foundation models (Wang et al. 2024; Leroy, Cabon, and Revaud 2024; Wang et al. 2025), which typically leverage a pre-trained Vision Transformer (ViT) backbone. By integrating these powerful 3D geometric priors with 3DGS, such methods (Ye et al. 2024; Jiang et al. 2025a) can synthesize novel views from unposed sparse views via rapid single-pass inference. Although this advancement significantly boosts efficiency and generalization, a notable quality gap persists. These feed-forward models often struggle to generate high-fidelity novel views, particularly in rendering high-frequency details. A straightforward strategy might be to feed high-resolution images directly into a larger ViT backbone. However, this approach results in substantial memory overhead and restricts the network's applicability in real-world scenarios. It is essential to propose an efficient module to enhance the capability of ViT-based backbones to infer high-grained features more effectively.

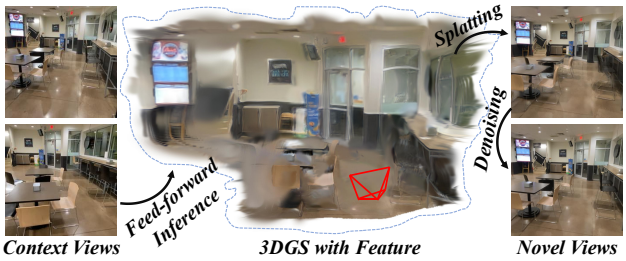

Figure 1: Starting from unposed input images, our method reconstructs 3D Gaussians within a canonical space, and leverages a one-step Stable Diffusion (SD) module to synthesize high-fidelity target views.

Moreover, 2D generative models, such as diffusion models, excel in capturing intricate details and textures, enabling them to produce high-quality 2D images. However, their direct application as a post-processing step for NVS presents a fundamental conflict: these models are 3D-agnostic, resulting in inconsistent structures across views. Compounding this issue is their notoriously slow, iterative sampling process. Our work overcomes these hurdles with a carefully designed Feature-Guided One-Step Diffusion architecture that is both fast and geometrically aware. The core of our innovation, however, lies in how we guide the generation to respect the scene structure. As illustrated in Fig. 3, we employ a two-pronged conditioning strategy. A dedicated guidance branch relays explicit geometric priors from the 3D back-

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

bone, anchoring the generative process to the scene's true structure. Simultaneously, the input view serves as a reference condition, guaranteeing that the final output preserves fine-grained information.

Consequently, we achieve an end-to-end framework to enhance the high-quality novel view synthesis from unposed sparse inputs, as shown in Fig. 1. Specifically, the main contributions of our work are as follows:

1. We design a *Dual-Domain Detail Perception Module* (DD-DPM), which enables handling high-resolution images without being limited by the ViT backbone, and endows Gaussians with additional features to store high-frequency details.
2. We develop a *Feature-Guided One-Step Diffusion* architecture, which can preserve high-frequency details during the restoration process.
3. We propose an integrated training framework that allows end-to-end training of the ViT reconstruction backbone and the diffusion-based image enhancement module.

2 Related Work

Radiance fields for novel view synthesis. Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) model continuous volumetric density and radiance, enabling high-quality novel view synthesis but relying on positional encoding and importance sampling, which limits real-time performance. 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) instead represents a scene with millions of anisotropic Gaussians and uses differentiable rasterization to achieve photorealistic rendering at over 30 FPS (1080p). Its efficiency has spurred extensions in rendering (Lu et al. 2024; Dong et al. 2024), surface reconstruction (Huang et al. 2024a; Chen et al. 2024a; Yu, Sattler, and Geiger 2024), generation (Yi et al. 2023), and scene understanding (Qin et al. 2024; Jiang et al. 2025b). Despite these advances, both NeRF and 3DGS rely on dense posed images and require per-scene optimization, limiting their practicality in sparse-view settings.

Generalizable 3D Reconstruction for Sparse View. With the rise of pre-trained foundation models (Wang et al. 2024; Leroy, Cabon, and Revaud 2024; Wang et al. 2025) that encode strong geometric priors, feed-forward novel view synthesis (NVS) from sparse inputs has received increasing attention. Unlike other modality-based 3D reconstruction methods (Zhang et al. 2024; Zhang et al.), these approaches (Zhang et al. 2025; Ye et al. 2024; Jiang et al. 2025a) generally leverage foundation models as backbones to extract geometric cues and improve reconstruction quality. MuRF (Xu et al. 2024) aggregates multi-view information through cost-volume construction, while pixelSplat (Charatan et al. 2024) exploits epipolar geometry to achieve more accurate depth estimation.

Diffusion priors for novel view synthesis. Incorporating diffusion models into reconstruction tasks has been shown to enhance the quality of novel view generation. ReconFusion (Wu et al. 2024) leverages 2D diffusion priors to recover high-fidelity NeRF from sparsely sampled input views. DiffusionNeRF (Wynn and Turmukhambetov 2023) employs a diffusion model to learn gradients of logarithmic RGBD patch priors, which serve as regularized geometry and color for a scene. Nerfbusters (Warburg et al. 2023) utilizes diffusion priors to remove artifacts. Methods such as ReconX (Liu et al. 2024), 3DGS-Enhancer (Liu, Zhou, and Huang 2024), and Difix3D+ (Wu et al. 2025a) require per-scene optimization (e.g., Difix3D+ takes 15–30 minutes per scene) and perform poorly under sparse inputs. MVSplat360 (Chen et al. 2024b) directly denoises rendered features via video diffusion, but is time-consuming, whereas LatentSplat (Wewer et al. 2024) relies on a lightweight VAE-GAN decoder and does not exploit the generative capabilities of a UNet for denoising. Our goal is to achieve feed-forward novel view synthesis under sparse viewpoints by using one-step Stable Diffusion for refinement.

3 Method

Given unposed sparse-view images $\{I_i\}_{i=1}^{N} \in \mathbb{R}^{H\times W\times 3}$ and their intrinsics $\{k_i\}_{i=1}^{N} \in \mathbb{R}^{3\times 3}$, Our method learns a feed-forward network to generate 3D Gaussians for novel view synthesis (NVS), with an additional refinement module applied to further improve rendering quality. The scene can be represented by 3D Gaussian Splatting (3DGS): $g_j := \{\mu_j,\ \mathbf{s_j},\ \mathbf{q_j},\ o_j,\ \mathbf{c_j}\}$. Here, $\mu_j \in \mathbb{R}^3$ and $o_j \in \mathbb{R}$ denote the Gaussian center and opacity, $\mathbf{s}_j \in \mathbb{R}^3$ and $\mathbf{q}_j \in \mathbb{R}^4$ define the 3D covariance, and $\mathbf{c}_j \in [0,1]^3$ represents RGB color via spherical harmonics coefficients. These primitives allow efficient modeling of 3D geometry and appearance for high-quality novel view synthesis. The overall pipeline of our method is illustrated in Fig. 2. Our network architecture consists of an encoder, a decoder, Gaussian parameter prediction heads, and a final enhancement module.

3.1 Geometry Transformer Backbone

Given unposed images I_i, we use a pretrained Vision Transformer (ViT) module (Leroy, Cabon, and Revaud 2024) to acquire the geometric information of the scene. Following the design in (Leroy, Cabon, and Revaud 2024), the geometry transformer module comprises an encoder and a decoder.

Encoder. We initially patchify each RGB image I_i into sequences of image tokens t_i^I. To enhance the network's capability of perceiving geometric information and thereby further optimize reconstruction quality, we inject the intrinsic parameter information of each image into the model. Specifically, we inject camera intrinsic parameters $[f_x.f_y, c_x, c_y]$ of each image into a linear layer to obtain global feature tokens t_i^C, which are then expanded to corresponding image tokens. Next, the concatenated tokens from each view are individually input to a ViT encoder, with the encoder employing shared weights across all views.

Decoder. The combined tokens from the encoder are then input into the ViT decoder, where cross-view information interaction is achieved through attention modules, resulting in features that contain global geometric information. This global feature is then used to estimate 3D scene parameters, such as point clouds and Gaussian parameters.

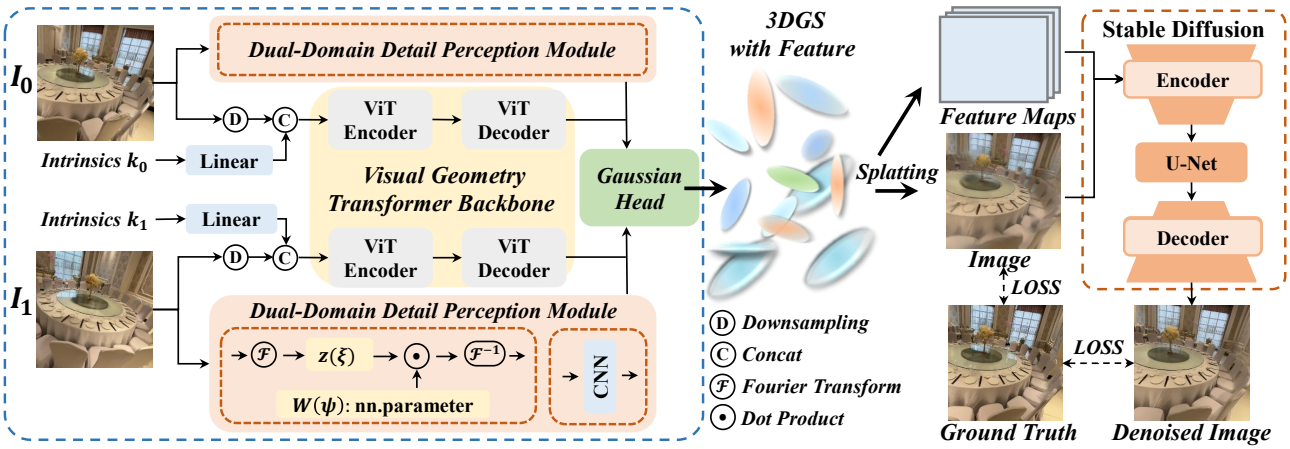

Figure 2: Overview of our pipeline. Starting from a set of unposed images, we first perform spatial downsampling and feed them into a Vision Transformer (ViT)-based backbone for global feature extraction. Simultaneously, we employ a Dual-Domain Detail Perception Module to enhance fine-grained detail perception from both spatial and frequency domains. The fused features are passed into Gaussian Parameter Prediction Heads to directly predict Gaussians with features in a canonical space. Finally, a Single-step Denoising module refines these outputs to produce higher-quality novel view synthesis (NVS) results.

3.2 Detail-Aware Scene Reconstruction

Due to memory constraints, existing feed-forward networks (Charatan et al. 2024; Ye et al. 2024) based on pretrained transformer foundation models (Wang et al. 2024; Leroy, Cabon, and Revaud 2024; Wang et al. 2025) commonly restrict input images to low resolutions such as (256×256). However, high-resolution images are readily available in real-world scenarios, and directly feeding their downsampled versions into the Geometry Transformer Backbone can lead to the loss of critical structural and appearance information. Feeding high-resolution images directly into the ViT backbone, or employing additional complex modules to handle them, results in substantial memory overhead. This increases the computational burden and restricts the network's applicability in real-world scenarios. To address this issue, we propose two methods to enable detail-aware reconstruction. First, we introduce a lightweight Dual-Domain Detail Perception Module (DD-DPM), which efficiently extracts informative cues from high-resolution inputs with low memory overhead and feeds them into the Gaussian Parameter Prediction Heads. Second, we extend the feature dimensionality of each Gaussian and implicitly inject information from neighboring pixels in the high-resolution image into the corresponding Gaussian features, thus enhancing the preservation of fine-grained details.

Dual-Domain Detail Perception Module. To better capture both local textures and global patterns, we propose a Dual-Domain Detail Perception Module (DD-DPM) that processes image data jointly in the spatial and frequency domains. Converting images into the frequency domain is a widely adopted strategy, particularly in image super-resolution tasks, as frequency components encode global structures without relying on a strict spatial correspondence. This property enables more flexible and effective modeling of image details. Motivated by this, we incorporate a frequency-domain adaptation module to enhance the expressiveness of the extracted features. The module begins by applying a 2D Fourier transform to obtain spectral representations of the input images.

$$z(\xi) = \mathcal{F}(I_i), \tag{1}$$

where $\mathcal{F}$ denotes the Fourier transforms and ξ is the frequency domain variables. To focus on the most informative spectral signals, we use an MLP to predict importance scores over normalized frequency coordinates. A top-k selection is applied to retain key components, which are then modulated by learned complex weights. The final output is transformed back through an inverse Fourier transform:

$$F_i' = \mathcal{F}^{-1}(z'(\xi)), \tag{2}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transforms. On the other hand, a lightweight CNN is employed to capture fine-grained texture information from the spatial domain, which is then combined with frequency-domain features and fed into the Gaussian Parameter Prediction Heads.

Gaussian Parameter Prediction Heads. We design heads based on the DPT decoder (Ranftl, Bochkovskiy, and Koltun 2021) to predict Gaussian parameters. Given that the features derived from the ViT encoder and decoder already contain strong geometric priors, we directly utilize these global features through a depth head to obtain depth values, which serve as the center of each Gaussian μ_j. This can largely ensure the spatial consistency of Gaussians.

We utilize a second head to predict the remaining Gaussian parameters. To preserve as much valid information contained in high-resolution images as possible, we have designed two strategies: First, in addition to the global features obtained from the ViT Transformer and the images themselves, we fuse the features derived from the Weighted Fourier Neural Operator, thereby further enhancing the representational capability of Gaussians. Second, we assign ad-

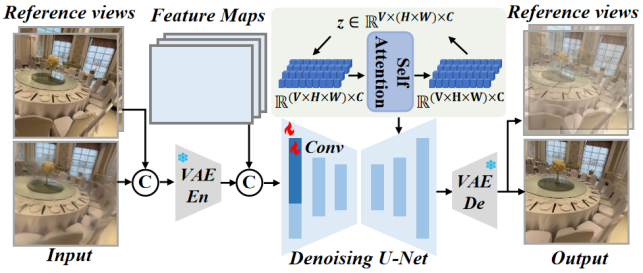

Figure 3: Structure of the diffusion model in NVS.

ditional features to each Gaussian to contain detailed information, which in turn assists our enhancement module.

Rendering novel view images with feature. We design feature-augmented 3D Gaussians, where each Gaussian is equipped with additional features to preserve high-frequency details. With this design, once the ViT backbone reconstructs the scene and generates the 3D Gaussians, novel-view rendering produces not only RGB color values but also the corresponding high-dimensional Gaussian features. These features can subsequently be used to support 3D-consistent image enhancement in downstream modules.

3.3 Boosting NVS with Diffusion module

While existing feed-forward methods benefit from the strong representation capability of ViT models to recover accurate geometric structures, the subsequent novel view rendering process still inevitably encounters 3D artifacts and blurred details. This problem largely arises from limited supervision provided by the input views, ultimately reducing the quality and perceptual fidelity of the generated novel views. Therefore, a super-resolution algorithm is needed to improve the fidelity and realism of the rendered images. To overcome this limitation, we utilize a one-step Stable Diffusion (SD) architecture to enhance the synthesized novel views, yielding higher-quality target images. Building on this foundation, we further optimize the SD module by incorporating Gaussian features as auxiliary inputs within UNet structure and integrating reference images into UNet's self-attention module to facilitate inter-image information exchange.

Stable Diffusion. Diffusion Models (DMs) learn data distributions through iterative denoising, and this process becomes significantly more efficient when performed in the latent space using a pre-trained autoencoder (Wu et al. 2025b). We employ the pre-trained latent diffusion model to mitigate 3D artifacts in novel view synthesis, typically caused by sparse supervision or geometric inconsistencies. Using the VAE encoder E_ϕ, latent diffusion network ϵ_ϕ, and the VAE decoder D_ϕ—where ϕ denotes the model parameters—the images can be denoised to achieve higher quality. Building on this, we further incorporate geometric alignment with 3D Gaussian features, enabling the SD module to generate a high-fidelity and 3D-consistent denoised target view I_d via latent-space diffusion decoding.

Gaussian Feature Integration in UNet. After encoding the low-quality rendering image I_r through the VAE encoder E_ϕ to obtain the latent feature $F_r \in \mathbb{R}^{C\times\frac{H}{8}\times\frac{W}{8}}$, we render Gaussian feature $F_g \in \mathbb{R}^{C'\times\frac{H}{8}\times\frac{W}{8}}$ along the corresponding dimension $(\frac{H}{8}, \frac{W}{8})$ and concatenate it with F_r along the channel dimension.

$$F = \mathcal{CAT}(F_r, F_g), \tag{3}$$

The fused features are then fed into the UNet module for cross-modal interaction. Notably, we extend the input channels of the UNet's initial convolution to accommodate the additional features, initializing the extra dimensions to zero to facilitate stable convergence.

Reference Image Interaction via Self-Attention. We utilize reference images as guidance to assist the diffusion model in generating high-quality images (Wu et al. 2025a; Agarwal et al. 2025). Specifically, we first concatenate the reference images with the Gaussian-rendered target image and feed the combined result into the VAE encoder to obtain latent features. We then simultaneously extract Gaussian features under both the reference and target views, and integrate them to produce fused features $z \in \mathbb{R}^{V\times C\times H\times W}$. Within the latent diffusion model, we modify the self-attention layers to transform the image interactions within the low-quality image $z \in \mathbb{R}^{V\times C\times(H\times W)}$ into a mixed attention mechanism between the low-quality image and the reference image $z \in \mathbb{R}^{C\times(V\times H\times W)}$. In this way, we can further capture fine-grained details from reference image.

Two-Stage Pipeline. In the first stage, we employ LoRA to fine-tune the latent diffusion network ϵ_ϕ of the Stable Diffusion (SD) module, encouraging it to restore images degraded by 3D rendering artifacts. To train the model effectively, we construct a paired dataset using DL3DV, comprising low-quality rendered images and their corresponding high-quality ground-truth (GT) images. Specifically, following the sampling strategy of (Ye et al. 2024), we encode sparse input views into 3D Gaussian scene representations using a geometry Transformer backbone and scene reconstruction module, and render novel views based on sampled camera extrinsics. These rendered images and their GT counterparts form the supervision pairs for training.

In the second stage, directly optimizing rendered images with the Stable Diffusion (SD) architecture can mitigate artifacts to some extent, yet it struggles to preserve fine-grained textures in high-resolution outputs and maintain geometric consistency. Gaussian features, however, contain both rich textural details and geometric cues. To address this limitation, we design a pipeline that jointly performs scene reconstruction and integrates a feature-guided SD module to achieve higher-fidelity image synthesis.

3.4 Training

The architecture comprises a front-end encoder, a decoder, and a back-end diffusion-based enhancement module. The network is fully trainable in an end-to-end manner. Given the pose, the 3D Gaussian constructed by the front-end is capable of rendering images I_r and their corresponding features. These images and features are fed into the enhancement module, which then leverages the priors of diffusion to enhance image quality, resulting in the final images I_d.

Training Loss. For reconstruction, we use the standard MSE loss together with the perceptual LPIPS loss between the

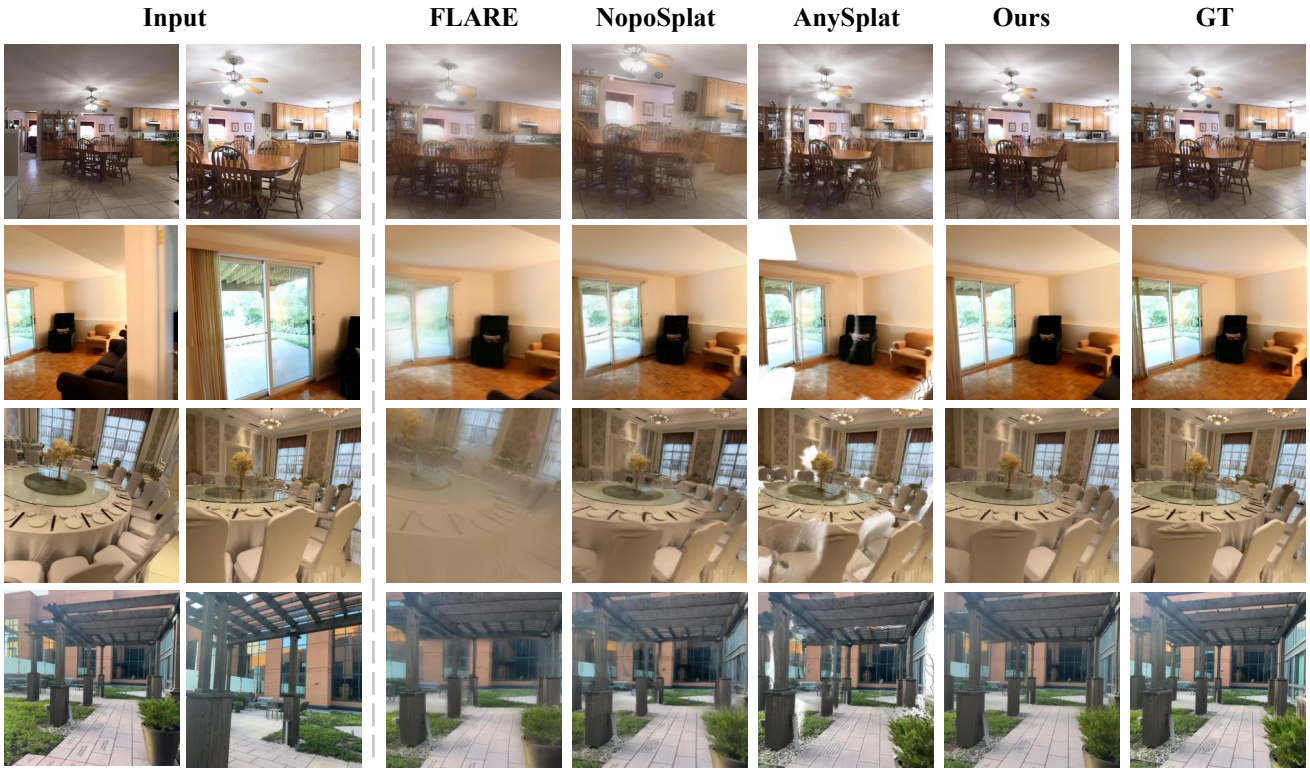

Figure 4: Qualitative comparison on DL3DV and RE10K datasets, all evaluated at a uniform resolution of $\mathbf{512} \times \mathbf{512}$. Compared with other methods, our approach is capable of recovering finer texture details.

rendered image I_r and the ground-truth image $\hat{I}$:

$$L_r = \lambda 1 \cdot MSE(I_r, \hat{I}) + \lambda 2 \cdot LPIPS(I_r, \hat{I}), \quad (4)$$

For the enhancement module, we first directly introduce the MSE loss and LPIPS loss between the denoising image I_d and the target image $\hat{I}$:

$$L_d = \lambda 3 \cdot MSE(I_d, \hat{I}) + \lambda 4 \cdot LPIPS(I_d, \hat{I}), \quad (5)$$

To enhance perceptual quality, we incorporate a DINOv2 pre-trained network as a feature extractor to guide the training of the GAN (Kumari et al. 2022). This encourages the generator to produce outputs that are not only visually realistic but also semantically consistent with the input. The GAN loss is formulated as:

$$L_g = \lambda_5 \mathbb{E}[\log D(\hat{I})] + \lambda_5 \mathbb{E}[\log(1 - D(G(I)))], \quad (6)$$

where G and D are the generator and discriminator using DINOv2 backbone for training, respectively. The total loss can be expressed as:

$$L_{total} = \lambda_r \cdot L_r + \lambda_d \cdot L_d + \lambda_g \cdot L_g, \quad (7)$$

Training pipeline. We first train the ViT-based 3D reconstruction module independently to obtain a stable 3D representation. Specifically, images resized to 256×256 are input to the ViT backbone to extract global semantic features. Simultaneously, the Dual-Domain Detail Perception Module processes higher-resolution images at 512×512 to capture fine-grained details. Both feature streams are subsequently fused in the prediction heads to reconstruct the 3D Gaussian representation. Following the rendering of images and features, the SD module is first trained independently. Subsequently, we perform joint training of the SD module guided by the Gaussian features, as detailed in Sec. 3.3.

4 Experiments

4.1 Training Details

Dataset. We trained our model on the 2K subset of the DL3DV dataset and evaluated it on the benchmark subset (Ling et al. 2024). DL3DV is a large-scale 3D scene dataset widely used for NVS, featuring a variety of reflection, transparency, and lighting conditions. To further assess generalization, we evaluated our approach on the RealEstate10k (RE10K) dataset (Zhou et al. 2018), which comprises large-scale indoor real estate videos with multi-view images and accurate camera poses. This setup enables testing the robustness of our method in NVS under complex real-world lighting, textures, and geometric variations.

Evaluation Metrics. The quality of novel view synthesis is first measured using three standard metrics: PSNR, SSIM, and LPIPS. In addition, we have incorporated parameter-free metrics, including DISTS, FID, NIQE, MUSIQ, M-IQA, and C-IQA. Detailed analyses are provided in Sec. 4.2.

Dataset	Methods	Full-Reference			No-Reference				Perceptual Quality
		PSNR↑	SSIM↑	LPIPS↓	NIQE↓	MUSIQ↑	M-IQA↑	C-IQA↑	FID↓
DL3DV	NopoSplat	15.53	0.47	0.57	4.94	59.77	0.51	0.29	118.42
	AnySplat	18.27	0.55	0.27	**3.53**	64.37	0.68	0.36	64.91
	FLARE	15.90	0.48	0.55	4.41	59.73	0.54	0.31	122.24
	Ours	**22.67**	**0.69**	**0.16**	3.87	**72.86**	**0.72**	**0.49**	**40.46**
RE10K	NopoSplat	15.81	0.60	0.54	5.85	53.97	0.49	0.31	79.81
	AnySplat	19.07	0.67	0.23	**4.31**	60.96	0.65	0.29	42.20
	FLARE	16.87	0.62	0.46	5.39	52.83	0.52	0.28	67.87
	Ours	**20.67**	**0.70**	**0.21**	4.71	**69.05**	**0.68**	**0.39**	**33.52**

Table 1: Novel view synthesis performance on DL3DV and RE10K datasets, all evaluated at a uniform resolution of $\mathbf{512} \times \mathbf{512}$.

Methods	LPIPS↓	MS↑	SC↑	BC↑
MVSplat360	0.35	**0.95**	0.90	0.92
LatentSplat	0.27	**0.95**	0.90	0.91
Ours	**0.19**	**0.95**	**0.92**	**0.93**

Table 2: Quantitative evaluation of the 3D consistency performance of generated videos on the DL3DV dataset.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓
MARINER	19.99	0.64	0.20	63.04
Difix3d+	21.67	0.67	0.18	50.30
Ours	**22.67**	**0.69**	**0.16**	**40.46**

Table 3: Quantitative comparison with enhancement methods (MARINER and the SD model used in Difix3D+).

4.2 Results

Novel View Synthesis. We selected state-of-the-art (SOTA) feed-forward methods to compare the performance of novel view synthesis (NVS) results with our method on the test set of DL3DV, where the resolution of the final output is 512×512. As shown in Table. 1, our method attains the superior performance among all compared approaches.

3D Consistent. To assess multi-view 3D consistency, consecutive frames are sampled as target views to generate videos. We compare our approach with NVS methods that also perform refinement on rendered images, including MVSplat360 and LatentSplat. The Motion Smoothness (MS), Subject Consistency (SC), and Background Consistency (BC) metrics from VBench (Huang et al. 2024b; Zheng et al. 2025; Huang et al. 2024c) are employed to jointly evaluate 3D consistency. As shown in Table 2, our method demonstrates substantially improved geometric consistency under sparse viewpoints.

Enhancement Module. We compare our feature-guided SD enhancement module with existing enhancement methods by substituting it with MARINER (Bösiger et al. 2024) and the SD model from Difix3D+ (Wu et al. 2025a). Table 3 shows that our approach surpasses these methods, effectively utilizing 2D reference cues while maintaining 3D consistency via Gaussian features.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓
NopoSplat	15.16	0.59	0.77	246.80
AnySplat	15.36	0.45	0.54	108.74
FLARE	15.83	0.59	0.73	230.59
Ours	**21.64**	**0.69**	**0.28**	**41.74**

Table 4: NVS performance comparison on the DL3DV dataset at an extended resolution of $\mathbf{1024} \times \mathbf{1024}$.

Cross-Dataset Generalization. We assess the generalization performance of our network on the RE10K dataset without fine-tuning. As shown in Table 1, the results indicate that our method exhibits strong generalization.

Resolution Expansion. We expand the output image into a higher-dimensional space without requiring additional training. Specifically, we first render low-resolution images and upsample them to high resolution (1024×1024), followed by generating high-quality outputs using Tile-VAE.

4.3 Ablation Studies

We perform ablation studies on a baseline trained with the DL3DV dataset to assess the impact of the proposed module.

Ablation on DD-DPM. To evaluate the effectiveness of the Detail-Preserving Module (DPM), we construct two ablation variants: (1) CNN-DPM, which relies solely on CNN to process high-resolution images and fuses its output with the low-resolution features from the ViT backbone; and (2) DD-DPM, which additionally introduces a frequency-domain processing branch to enable dual-domain detail perception. Table 5 indicates the clear advantage of DD-DPM, emphasizing the importance of frequency-domain features in maintaining fine details.

Ablation on Gaussian Feature. To verify the effectiveness of introducing features into Gaussian parameters, we directly connected a simple two-layer CNN structure at the backend, which is used to process the rendered features and images. This lightweight architecture serves as a baseline to assess whether the inclusion of Gaussian feature can bring tangible improvements. By comparing its performance with our full model, we can explicitly quantify the gains from embedding features into Gaussian representations during the rendering and post-processing pipeline.

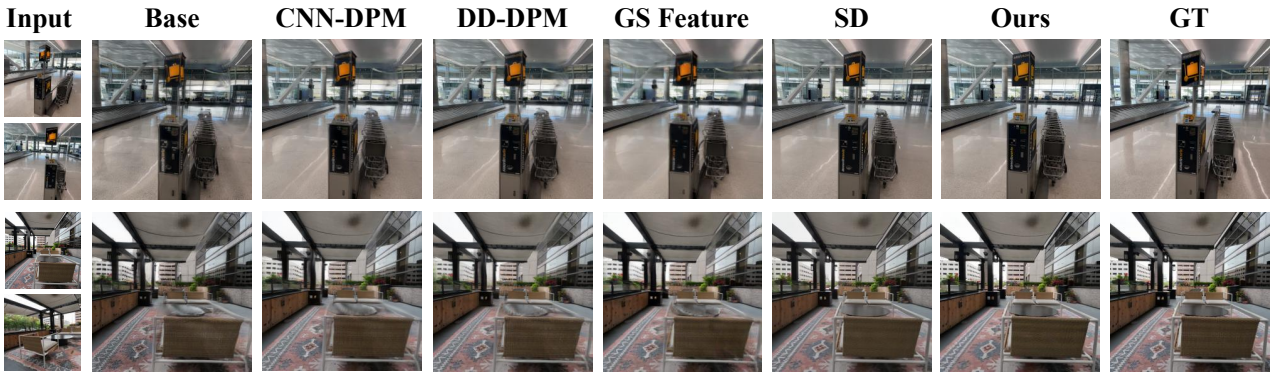

Figure 5: Qualitative ablation results validating the effectiveness of each component. Incorporating Dual-Domain Detail Perception Module (DD-DPM) and the feature-guided SD refine module yields substantially higher-quality novel view synthesis.

Methods	Full-Reference			No-Reference				Perceptual Quality
	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	MUSIQ↑	M-IQA↑	C-IQA↑	FID↓
Base	20.08	0.65	0.24	4.11	65.98	0.55	0.29	62.19
+CNN-DPM	22.10	0.69	0.23	4.24	64.95	0.59	0.30	62.60
+DD-DPM	22.14	0.69	0.22	4.15	63.47	0.60	0.30	55.32
+GS Feature(CNN)	<u>22.22</u>	<u>0.69</u>	0.27	4.93	56.15	0.55	0.30	70.51
+SD	22.17	0.68	<u>0.17</u>	<u>3.92</u>	<u>70.97</u>	<u>0.69</u>	<u>0.44</u>	<u>48.58</u>
+Feature guided SD	**22.67**	**0.69**	**0.16**	**3.87**	**72.86**	**0.72**	**0.49**	**40.46**

Table 5: Ablation study validating the effectiveness of each network component on the DL3DV dataset.

Ablation on SD Model. We constructed a dataset to train the SD network independently and connected it to the back-end of the reconstruction module to as an enhancement process for rendered images. This standalone training strategy allows the SD network to specialize in refining details and enhancing visual quality for novel view synthesis, while its integration with the reconstruction module ensures that the enhanced results remain consistent with the geometric structure predicted by the front-end, thereby achieving both accurate novel view synthesis and high-fidelity image quality.

Ablation on Feature-guided SD Refinement Module. We introduce a feature-guided SD network and perform joint training of both the ViT-based reconstruction module and the SD network. As illustrated in Table 5, incorporating the detailed features rendered by 3D Gaussians into the SD module enables efficient image refinement. This joint optimization strategy ensures that the SD network learns to leverage geometrically consistent detail cues from the rendering process, guiding its refinement efforts toward preserving structural integrity while enhancing visual fidelity, ultimately leading to more coherent and high-quality novel view synthesis compared to standalone enhancement approaches.

Analysis. We observe that although incorporating the SD module results in a slight decrease in simple metrics such as PSNR and SSIM, the perceptual quality of the outputs is significantly improved, producing clear and high-quality images. In contrast, omitting the SD module yields noticeably blurrier results. To quantitatively validate this improvement, we evaluate model-based and no-reference metrics including LPIPS, NIQE, MUSIQ, M-IQA, C-IQA, and FID, and find that the model with the SD module consistently outperforms the version without it across all metrics. Indeed, PSNR can no longer fully capture the true perceptual quality of images. LPIPS (Zhang et al. 2018) also points out that PSNR and SSIM often do not perfectly align with human perception. Consequently, when differences in simple metrics (e.g., PSNR) are marginal, model-based or no-reference metrics play a more important role in evaluation.

5 Conclusion

Summary. We propose a novel view synthesis method that integrates a feed-forward pipeline with a single-step Stable Diffusion (SD) model. This combination leverages the geometric efficiency of feed-forward methods and the generative strength of SD for detail refinement, producing results with accurate structure and realistic textures. To enhance consistency between geometry and appearance, we introduce a unified training framework that jointly optimizes geometric representation learning and image generation using features rendered from 3D Gaussians. This tight integration ensures more effective use of geometric features in guiding high-quality image synthesis. Our framework also paves the way for future work on simplifying parts of the SD architecture, aiming for tighter feature coupling and more efficient view synthesis. **Limitation.** The proposed method does not explicitly model dynamic objects, which limits its applicability in real-world scenarios involving dynamic scenes.

Acknowledgements

This work was partially supported by the National Key Research and Development Program of China (No. 2023YFF0905104) and NSF of China (No. 62441222).

References

Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; et al. 2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.

Bösiger, L.; Dusmanu, M.; Pollefeys, M.; and Bauer, Z. 2024. MaRINeR: Enhancing Novel Views by Matching Rendered Images with Nearby References. In *European Conference on Computer Vision*, 76–94. Springer.

Charatan, D.; Li, S. L.; Tagliasacchi, A.; and Sitzmann, V. 2024. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19457–19467.

Chen, D.; Li, H.; Ye, W.; Wang, Y.; Xie, W.; Zhai, S.; Wang, N.; Liu, H.; Bao, H.; and Zhang, G. 2024a. PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *arXiv: 2406.06521*.

Chen, Y.; Zheng, C.; Xu, H.; Zhuang, B.; Vedaldi, A.; Cham, T.-J.; and Cai, J. 2024b. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37: 107064–107086.

Dong, Y.; Li, Y.; Huang, Z.; Bian, W.; Liu, J.; Bao, H.; Cui, Z.; Li, H.; and Zhang, G. 2024. A Global Depth-Range-Free Multi-View Stereo Transformer Network with Pose Embedding. *arXiv: 2411.01893*.

Dong, Y.; Yan, D.; Li, T.; Xia, M.; and Shi, C. 2022. Pedestrian gait information aided visual inertial SLAM for indoor positioning using handheld smartphones. *IEEE Sensors Journal*, 22(20): 19845–19857.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024a. 2D gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, 1–11.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024b. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Huang, Z.; Zhang, F.; Xu, X.; He, Y.; Yu, J.; Dong, Z.; Ma, Q.; Chanpaisit, N.; Si, C.; Jiang, Y.; Wang, Y.; Chen, X.; Chen, Y.-C.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024c. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2411.13503*.

Jiang, L.; Mao, Y.; Xu, L.; Lu, T.; Ren, K.; Jin, Y.; Xu, X.; Yu, M.; Pang, J.; Zhao, F.; et al. 2025a. AnySplat: Feed-forward 3D Gaussian Splatting from Unconstrained Views. *arXiv preprint arXiv:2505.23716*.

Jiang, M.; Jia, S.; Gu, J.; Lu, X.; Zhu, G.; Dong, A.; and Zhang, L. 2025b. VoteSplat: Hough Voting Gaussian Splatting for 3D Scene Understanding. *arXiv preprint arXiv:2506.22799*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kumari, N.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2022. Ensembling Off-the-shelf Models for GAN Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Leroy, V.; Cabon, Y.; and Revaud, J. 2024. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, 71–91. Springer.

Ling, L.; Sheng, Y.; Tu, Z.; Zhao, W.; Xin, C.; Wan, K.; Yu, L.; Guo, Q.; Yu, Z.; Lu, Y.; et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22160–22169.

Liu, F.; Sun, W.; Wang, H.; Wang, Y.; Sun, H.; Ye, J.; Zhang, J.; and Duan, Y. 2024. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*.

Liu, X.; Zhou, C.; and Huang, S. 2024. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37: 133305–133327.

Lu, T.; Yu, M.; Xu, L.; Xiangli, Y.; Wang, L.; Lin, D.; and Dai, B. 2024. Scaffold-GS: Structured 3D gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20654–20664.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Qin, M.; Li, W.; Zhou, J.; Wang, H.; and Pfister, H. 2024. Langsplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20051–20060.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Warburg, F.; Weber, E.; Tancik, M.; Holynski, A.; and Kanazawa, A. 2023. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18120–18130.

Wewer, C.; Raj, K.; Ilg, E.; Schiele, B.; and Lenssen, J. E. 2024. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European conference on computer vision*, 456–473. Springer.

Wu, J. Z.; Zhang, Y.; Turki, H.; Ren, X.; Gao, J.; Shou, M. Z.; Fidler, S.; Gojcic, Z.; and Ling, H. 2025a. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26024–26035.

Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Srinivasan, P. P.; Verbin, D.; Barron, J. T.; Poole, B.; and Ho?y?ski, A. 2024. ReconFusion: 3D Reconstruction with Diffusion Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21551–21561.

Wu, Z.; Sun, Z.; Zhou, T.; Fu, B.; Cong, J.; Dong, Y.; Zhang, H.; Tang, X.; Chen, M.; and Wei, X. 2025b. OMGSR: You Only Need One Mid-timestep Guidance for Real-World Image Super-Resolution. *arXiv preprint arXiv:2508.08227*.

Wynn, J.; and Turmukhambetov, D. 2023. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4180–4189.

Xu, H.; Chen, A.; Chen, Y.; Sakaridis, C.; Zhang, Y.; Pollefeys, M.; Geiger, A.; and Yu, F. 2024. MuRF: multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20041–20050.

Ye, B.; Liu, S.; Xu, H.; Li, X.; Pollefeys, M.; Yang, M.-H.; and Peng, S. 2024. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*.

Yi, T.; Fang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2023. Gaussiandreamer: Fast generation from text to 3D gaussian splatting with point cloud priors. *arXiv: 2310.08529*.

Yu, Z.; Sattler, T.; and Geiger, A. 2024. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv: 2404.10772*.

Zhai, H.; Li, H.; Li, Z.; Pan, X.; He, Y.; and Zhang, G. 2025a. PanoGS: Gaussian-based Panoptic Segmentation for 3D Open Vocabulary Scene Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14114–14124.

Zhai, H.; Zhang, X.; Zhao, B.; Li, H.; He, Y.; Cui, Z.; Bao, H.; and Zhang, G. 2025b. SplatLoc: 3D Gaussian Splatting-based Visual Localization for Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 31(5): 3591–3601.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Wang, J.; Xu, Y.; Xue, N.; Rupprecht, C.; Zhou, X.; Shen, Y.; and Wetzstein, G. 2025. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21936–21947.

Zhang, Z.; Liu, J.; Wang, W.; Lin, B.; Xie, L.; Shen, C.; Cai, D.; et al. ???? GeoCAD: Local Geometry-Controllable CAD Generation with Large Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhang, Z.; Sun, S.; Wang, W.; Cai, D.; and Bian, J. 2024. Flexcad: Unified and versatile controllable cad generation with fine-tuned large language models. *arXiv preprint arXiv:2411.05823*.

Zheng, D.; Huang, Z.; Liu, H.; Zou, K.; He, Y.; Zhang, F.; Zhang, Y.; He, J.; Zheng, W.-S.; Qiao, Y.; and Liu, Z. 2025. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*.

Zhou, T.; Tucker, R.; Flynn, J.; Fyffe, G.; and Snavely, N. 2018. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*.