Uncertainty-Aware Diffusion-Guided Refinement of 3D Scenes

Anonymous ICCV submission

Paper ID 18

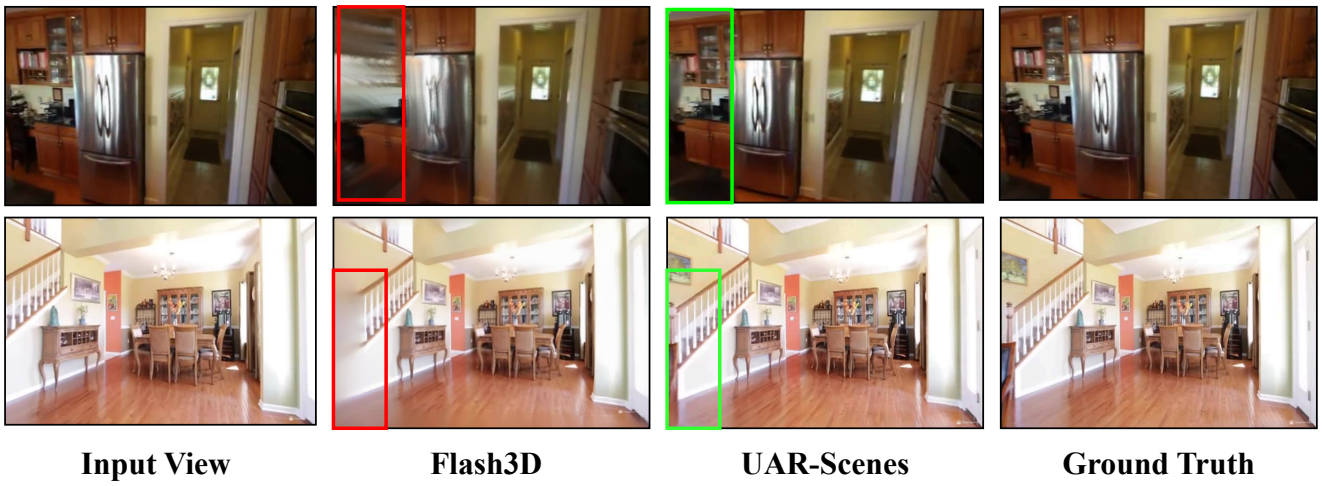


Figure 1. **Overview:** We introduce `UAR-Scenes`, a diffusion-guided refinement pipeline that enhances outputs from pre-trained single-image to 3D scene reconstruction models, such as Flash3D [1], which yield imperfect renderings (red box) under slight viewpoint variations (2nd column from left). By harnessing the generative power of a latent video diffusion model (LVDM) [2], our approach can sample plausible explanations for unseen regions (green box) to produce refined, high-quality novel views of 3D scenes (3rd column from left).

Abstract

Reconstructing 3D scenes from a single image is a fundamentally ill-posed task due to the severely under-constrained nature of the problem. Consequently, when the scene is rendered from novel camera views, existing single image to 3D reconstruction methods render incoherent and blurry views. This problem is exacerbated when the unseen regions are far away from the input camera. In this work, we address these inherent limitations in existing single image-to-3D scene feedforward networks. To alleviate the poor performance due to insufficient information beyond the input image's view, we leverage a strong generative prior, in the form of a pre-trained latent video diffusion model, for iterative refinement of a coarse scene represented by optimizable Gaussian parameters. To ensure that the style and texture of the generated images align with that of the input image, we incorporate on-the-fly Fourier-style transfer between the generated images and the input image. Additionally, we design a semantic uncertainty quantification module that calculates the per-pixel entropy and yields uncertainty maps used to guide the refinement process from the most confident pixels while discarding the remaining highly uncertain ones. We conduct extensive experiments on real-world scene datasets, including in-domain RealEstate-10K and out-of-domain KITTI-v2, showing that our approach can provide more realistic and high-fidelity novel view synthesis results compared to existing state-of-the-art methods.

1. Introduction

3D Scene Reconstruction is a long standing challenging task, playing a crucial role in holistic scene understanding [3, 4], generative content creation, mixed reality [5], robot navigation and path planning [6–9]. Reconstructing 3D scenes involves Novel View Synthesis (NVS), wherein the scene is rendered from unseen camera angles to obtain novel views of the scene. Although notable progress has been made towards improving NVS, most existing methods rely on multiple images from numerous viewpoints [10–12] to increase the

available information, which may not always be available. This leads us to the relatively underexplored task of 3D reconstruction from a single RGB image [1]. Although the existing single image to 3D scene method can generate reasonable results, it often fails when the scene is rendered from an unseen camera angle, due to scarcity of information limiting its capability to reconstruct scenes in views it has never seen before. This bottleneck leads us to ask the question: "*Can we design a pipeline that refines scenes so that it retains high fidelity in regions it has seen and generates consistent plausible completions for the unseen regions?*"

One way to compensate for this lack of multi-view information is to use generative priors so that they can provide valuable complementary information. The principle is to synthesize plausible 3D details even for regions not visible in the image, where *plausibility* refers to the fact that visible portions should remain multi-view consistent, while occluded or out-of-view areas should integrate seamlessly with the context of the scene (refer to supplementary for qualitative results). Some existing approaches toward solving this problem include text-guided inpainting [13] and distillation of 2D diffusion priors [14]. However, these methods either lack 3D correspondence awareness [15] or do not have any temporal continuity as they denoise each image separately. This leads to issues with multi-view consistency for large viewpoint changes as the camera moves further away from the source. One way to address this challenge is to reformulate the problem in a 2.5D reconstruction [16, 17] setting, which can be limited in terms of fully capturing unseen regions. Moreover, extending such approaches to fully unbounded real world scenes, where the scene may not loop back to the starting point, is non-trivial. These methods often rely heavily on just generated features, leading to mode collapse [18], content distortion [18, 19] and visible artifacts in rendered views [20, 21], especially when the camera motion spans a large range. Although diffusion priors are well-suited for object-level enhancement, these limitations indicate that priors primarily designed for object reconstruction, cannot be directly applied for scene refinement.

To this end, we propose *Uncertainty-Aware Diffusion-Guided Refinement of 3D Scenes:* `UAR-Scenes`, an uncertainty-aware Gaussian splatting based refinement framework that combines the strengths of regressive and generative approaches. We perform the refinement using *Adaptive Densification and Pruning (ADP)* [6], a mechanism that controls the density of 3D points by introducing or deleting Gaussian primitives (check supplementary Section 6.1). ADP plays a crucial role in expanding or shrinking Gaussians depending on whether the region is over-constructed or not. Therefore, in observed regions where the confidence per pixel is high, our method mostly preserves Gaussian primitives with high-fidelity details obtained from the reconstruction backbone, while in unobserved or occluded areas, it learns to introduce new Gaussians by using uncertainty-driven sampling from a Latent Video Diffusion Model (LVDM) [22] to produce plausible scene completions. We leverage an MLLM [23], along with an open-vocabulary segmentation model [24], to compute the per-pixel entropy present in LVDM generated views which gives us the per-pixel uncertainty associated with the generated images. In this way, the inherent ambiguity present in single-view reconstruction methods can be addressed, expanding the applicability of our pipeline from simple object-centric datasets or scenes with limited backgrounds [25] to arbitrarily large real-world unbounded scenes [26]. Furthermore, it does not rely on any form of ground truth supervision (2D or 3D), making the entire process self-supervised and thus, post-hoc compatible with any existing 3D reconstruction algorithm. This design effectively addresses ambiguities, such as missing artifacts and inconsistent 3D reconstructions inherent in single-view settings. Thus, in a nutshell, our proposed model *integrates a differentiable feed-forward scene reconstruction pipeline that operates from a single image with an uncertainty-aware LVDM for self-supervised refinement of noisy scenes*.

The **main contributions** of our work are as follows:

- We propose `UAR-Scenes`, a novel uncertainty-aware refinement pipeline which can take an existing single image to 3D scene reconstruction model and generate reliable and plausible explanations for unseen and naturally occluded regions present in unbounded real-world scenes. To the best of our knowledge this is the first optimization-based work for the refinement of single image to 3D reconstruction methods.
- `UAR-Scenes` utilizes a camera pose controlled Latent Video Diffusion Model (LVDM) which generates consistent posed images to serve as pseudo 2D supervision for refining a noisy scene. To address noisy pseudo-views, we design an innovative uncertainty quantification module which leverages an MLLM guided open vocabulary segmentation model for highlighting the most confident pixels for `UAR-Scenes` to learn from.
- To address any texture and style inconsistencies between the input image and the generated pseudo views from the LVDM, we propose a Fourier Style Transfer (FST)-based texture alignment methodology.
- We conduct extensive experiments on both real-world in-domain dataset RealEstate-10K [26] (on which the baseline method was trained) and an out-domain dataset KITTI-v2 [27] which is unseen for both our method and the baseline, demonstrating that `UAR-Scenes` enhances performance significantly.

2. Related Works

Regressive 3D Reconstruction from Sparse Views. Existing generalizable reconstruction methods [1, 12, 28–32] aim

to train a single feed-forward model which can predict a 3D scene directly. Such methods include PixelSplat [11] which utilizes epipolar transformers to reduce scale ambiguity and obtain better scene features. Similarly, MVSplat [12] builds cost volumes to obtain Gaussians in a fast and efficient manner. LatentSplat [30] utilizes a GAN decoder as a generative component but GANs are known to be highly unstable [33] and hence unsuitable for modeling real-world unbounded scenes. Recently, Flash3D [1] introduced a single image to 3D scene pipeline leveraging a pre-trained depth estimator. In this paper, we build upon this regressive framework and show how most works falling in this paradigm fail to generate realistic completions for unseen regions.

Generative Scene Reconstruction and Synthesis. Our method falls within an under-explored area of utilizing only 2D (pseudo) labels to supervise 3D tasks using latent diffusion models. Existing approaches like [34–37] denoise multi-view images using a 3D-aware denoiser. However, such methods do not produce a coherent 3D structure since the denoising process is done separately for each image leading to sub-optimal multi-view consistency. Works such as 3DGS-Enhancer [38] address a similar problem in a sparse view setting but they employ a mix of refined views along with un-refined views when training the video-diffusion model. Other works such as CAT3D [39], GeNVS [40], ReconFusion [14], LM-Gaussian [41], and ReconX [42][1] all tackle similar problems but they deal with the problem by employing expensive conditioning techniques [43] which may not always be multi-view consistent. *Further, none of these methods optimize the Gaussians directly but instead fine-tune an existing generative diffusion model on datasets to make it generalizable, which ties the models down to the particular distribution on which it is trained.*

Uncertainty Quantification in 3D Reconstruction. Quantifying uncertainty has received relatively less attention in 3D scene reconstruction, especially considering it's importance in the context of single image to 3D reconstruction where there is inherently a lot of ambiguity. One of the first approaches in estimating epistemic uncertainty in NeRFs from sparse views using variational inference was studied in works [44–46] but owing to increasing parameters alongwith using different models, these approaches are inherently expensive. BayesRays [47] proposed a method to quantify uncertainty in any NeRF based reconstruction module in a post-hoc fashion. Further, uncertainty has also been explored in the context of active-view selection and path planning in the field of 3D mapping in robotics [7, 48] to estimate which next view maximizes the information gain over a given pre-trained model by approximating the hessian matrix of the log-likelihood function. However, this becomes un-scalable to compute for each Gaussian primitive. Inspired by these approaches, we opt for a semantic scene understanding based approach for quantifying the uncertainty in novel test-time views for refinement.

[1]Code and models are not publicly available for these methods. ReconFusion was trained on an internal dataset.

3. Methodology

Problem Formulation. Consider a single input image $\mathcal{I}$ with camera pose intrinsic $\mathcal{K}$, a sequence of test-camera trajectory extrinsics $\mathcal{P}_{1\ldots M}$, and any given pre-trained deterministic feed-forward reconstruction model $\mathcal{F}(\cdot)$, from which we obtain N noisy Gaussian splats each representing a scene. Our goal is to refine these N noisy 3D Gaussian splats, $\phi_n \in \mathbb{R}^{N\times d}$ to better representations $\phi_n^{'} \in \mathbb{R}^{N\times d}$ leveraging the strong generative prior $\mathcal{G}$. We further sample k supplementary views derived from the LVDM for providing pseudo 2D supervision, which refines the noisy Gaussians *without access to any 2D or 3D ground truth*.

Overview. We provide an overview of our uncertainty-sensitive refinement pipeline in Figure 2. To get an accurate target representation of the scene, the geometric features from the backbone deterministic reconstruction model are preserved and the textures from the LVDM are utilized. The discussion starts with the deterministic model, serving as the initial point for refining scene Gaussians, detailed in Section 3.1. Then, we briefly discuss the topic of camera controllable LVDMs in Section 3.2 since it is crucial to align the pseudo views and the noisy rendered views so that they are in the same camera space. The next part addresses the method of weighting the optimization objective based on the confidence associated with each pixel, as outlined in Section 3.3. The final section describes the application of the video diffusion model to generate pseudo-views and facilitate iterative refinement, further elaborated in Section 3.4.

3.1. Initial Coarse 3D Representation

Here, we briefly discuss the working formulation behind the 3D reconstruction model $\mathcal{F}(\cdot)$ which provides us with an initial set of gaussians representing the scene ϕ. Flash3D [1] uses an off-the-shelf monocular depth estimation network [50] to estimate the metric depth $\mathbf{D}$ of the input image $\mathcal{I}$. It utilizes a U-Net [51] with ResNet blocks [52] to produce multiple sets of depth-ordered layered Gaussians per pixel $\mathbf{p}$ estimating to a certain degree the occluded regions of the scene. The model is then trained over a collection of such scenes and, for each scene, it renders images along a known set of camera extrinsics to get the corresponding rendered images $\tilde{\mathcal{I}}$. These paired rendered and ground truth views $(\tilde{\mathcal{I}}, \mathcal{I}_{\text{gt}})$ can then be used to supervise the initialization model with the reconstruction objective:

$$\mathcal{L}_{\text{rec}} = \|\tilde{\mathcal{I}} - \mathcal{I}_{\text{gt}}\|_2 + \alpha\, \text{SSIM}\big(\tilde{\mathcal{I}}, \mathcal{I}_{\text{gt}}\big), \tag{1}$$

where $\tilde{\mathcal{I}}$ is the image rendered by the initialized model $\mathcal{F}(\cdot)$ rasterizing its layered Gaussians and $\mathcal{I}_{\text{gt}}$ is the corresponding

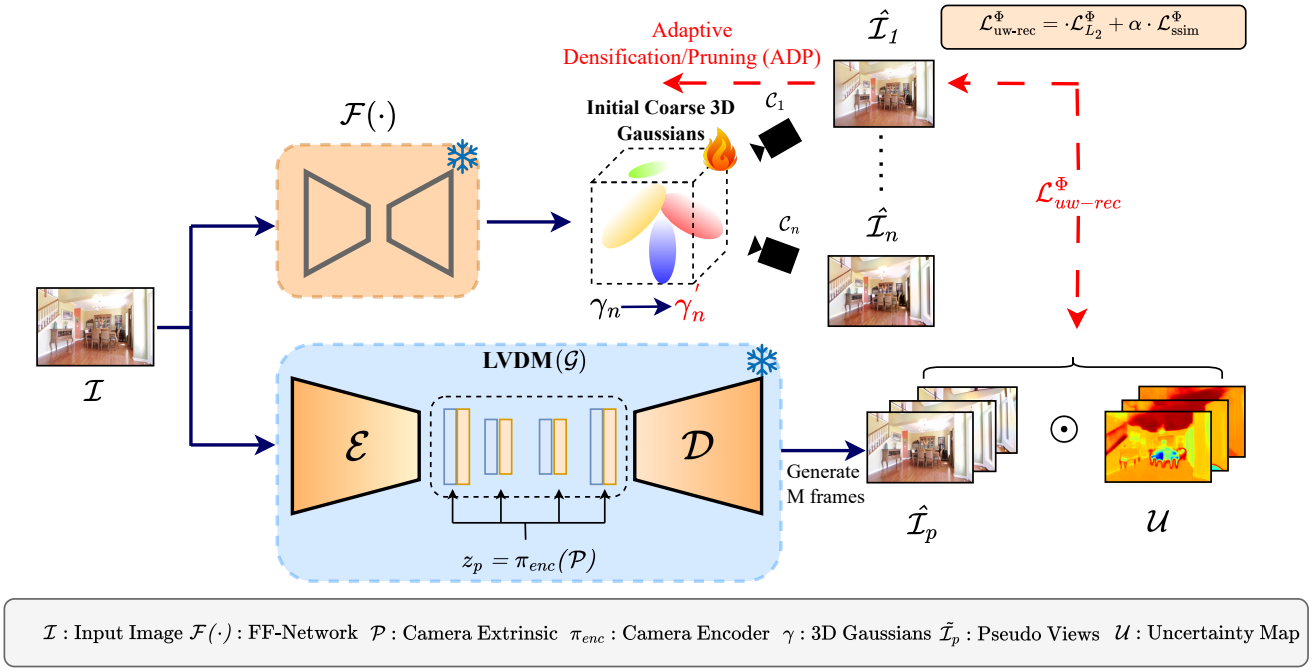


Figure 2. **Workflow of `UAR-Scenes`.** For a conditioning image $\mathcal{I}$, a pre-trained 3D reconstruction model $\mathcal{F}(\cdot)$ produces coarse gaussians (γ_n) representing the scene ϕ, represented by optimizable gaussian parameters. Using temporally consistent pseudo 2D supervisory images ($\widetilde{\mathcal{I}}_p$) sampled from the pre-trained camera extrinsic embedded LVDM model [22], we iteratively refine the gaussians γ_n using *Adaptive Densification and Pruning (ADP)* [49] to obtain clean gaussians ($\gamma_n^{'}$). To gauge the uncertainty of each pixel $\mathbf{p}$ in the generated pseudo images ($\widetilde{\mathcal{I}}_p$), we propose a semantic uncertainty quantification method. We estimate the entropy present in each ($\widetilde{\mathcal{I}}_p$) obtained by utilizing an off-the-shelf open-vocabulary segmentation model $\mathcal{S}$ [24] using which we obtain uncertainty maps $\mathcal{U}$ (as shown in Figure 3). We take the hadamard product between $\widetilde{\mathcal{I}}_p$ and $\mathcal{U}$ forming the target objective for the refinement loss in Equation 6, which guides the refinement process.

ground-truth image at train time. α is a hyper-parameter weighting the SSIM [53] loss.

Challenges. As illustrated in Figure 1, it is clear that relying solely on depth-based modeling to capture the structure and appearance of scenes is often insufficient in many cases. This issue becomes more pronounced in scenarios involving substantial camera motion, where a significant portion of the rendered image extends beyond the field of view of the input image. Moreover, merely utilizing the MSE loss as described in Equation 1 drives the model to average out pixels in unseen regions, which often results in suboptimal, blurry renderings.

3.2. LVDM with Camera Control

Existing LVDMs (refer to the supplementary for the general formulation of LVDMs) cannot generate denoised images with user-specified camera pose trajectories. MotionCtrl [22] addresses this issue which we have used in our method. It injects the Plücker embedding information of the corresponding relative camera extrinsics: $\{x^T,\ x_{(R,T)},\ \mathbf{R},\ \mathbf{T}\}$ and embeds them into the cross-attention layers of the U-Net which results in a modified revised objective (refer to Equation 2 of the supplementary for original objective of LVDMs):

$$\mathbb{E}_{x_0^{1..M},\, y_t^{1..M},\, \epsilon_t^{1..M} \sim \mathcal{N}(\mathbf{0},I)} \left\| \epsilon_t^{1..M} - \epsilon_\theta\left(z_t^{1..M}, t, y, \mathcal{C}_{(R,T)}\right) \right\|_2, \quad (2)$$

where $\mathcal{C}_{(R,T)}$ is the embedding of the camera extrinsics. Upon completion of the training process, the conditioning image y can be provided along with the camera's extrinsics $(\mathbf{R}, \mathbf{T})$, to generate the target views corresponding to the specified poses. For the sake of brevity, we refer to this version of camera controllable LVDM as LVDM ($\mathcal{G}$) unless mentioned otherwise.

3.3. Semantic Uncertainty Quantification

Even though the pseudo-images generated by the video diffusion model are plausible, they can generate over-saturated textures which adversely affects novel view synthesis (NVS) tasks (refer to Figure 5). Since we are directly optimizing the Gaussian parameters representing the scene, the aforementioned issues will lead to the degradation of the composition of the scenes during refinement, resulting in unrealistic renderings. To this end, to make the LVDM self-aware of its predictions and pass only the confident pixel information during refinement, we utilize a semantic

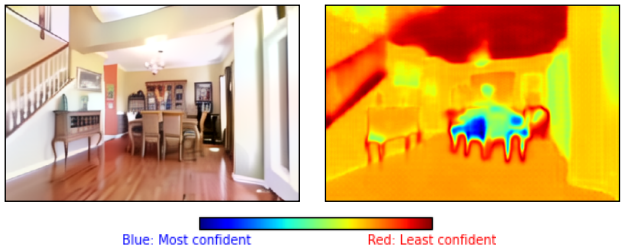


Figure 3. **Uncertainty Map Estimation.** On the right is the obtained uncertainty map from Pseudo Ground truth Image $\widetilde{\mathcal{I}}_{\text{p}}$ (on the left). It is generated by the LVDM after applying FST [54]. These maps are crucial for guiding the Gaussian refinement process to progressively focus more on the confident (blue and green regions) and not on blurry ceiling and stairs (red).

understanding-driven uncertainty quantification module as described below.

Uncertainty Distillation. For each pseudo-view $\widetilde{\mathcal{I}}_{\text{p}}$ generated by the LVDM, we use a Multimodal Large Language Model (MLLM) [23] to detect objects (even if partially visible) and obtain their object tags $O_{1\ldots N}$. These $N{+}1$ classes (including background) are then passed to an open-vocabulary segmentation model $\mathcal{S}$ [24], producing pixel-level logits:

$$o_t = \mathcal{S}\Big(\widetilde{\mathcal{I}}_{\text{p}},\, O\Big) \in \mathbb{R}^{n\times(N+1)}, \tag{3}$$

where n is the number of pixels.

Let $p_{i,j,c}$ denote the predicted probability that the pixel at (i,j) belongs to class c. The per-pixel Shannon entropy is defined as:

$$\mathcal{U}_{(i,j)} = -\sum_{c} p_{i,j,c}\,\log\big(p_{i,j,c}\big), \tag{4}$$

which forms the uncertainty map $\mathcal{U}\in\mathbb{R}^{H\times W}$ which is then subsequently used to perform pixel-wise multiplication with the LVDM image during refinement. Intuitively, $\mathcal{U}_{(i,j)}$ captures the uncertainty present in the segmentation map, whereby a higher entropy indicates a more uniform (i.e. less confident) class distribution for a pixel, while a lower entropy implies higher confidence in the predicted semantic classes. This entropy-based uncertainty measure thus provides a principled way to assess the reliability of the pseudo-view segmentation for our uncertainty-aware refinement. Further implementation details on how the MLLM was used for object tagging alongwith an open-set segmentation model [24] is explained in the supplementary.

Uncertainty-Weighted Reconstruction. Instead of relying solely on the standard reconstruction loss introduced in Section 3.1, we leverage the derived per-pixel uncertainty map $\mathcal{U}_{(i,j)}$ to guide the refinement process. In particular, given the rendered image $\tilde{\mathcal{I}}$ from the pre-trained reconstruction model $\mathcal{F}(\cdot)$ and the pseudo image $\widetilde{\mathcal{I}}_{\text{p}}$, we modify Equation 1 to compute an uncertainty-weighted reconstruction loss by taking the Hadamard (pixel-wise) product between the reconstruction error and the complement of the uncertainty map to get the pixel-wise confidence:

$$\mathcal{L}_{\text{uw-rec}} = \Big\|\big(1-\mathcal{U}_{(i,j)}\big)\odot\big(\tilde{\mathcal{I}}-\widetilde{\mathcal{I}}_{\text{p}}\big)\Big\|_2 + \alpha\,\text{SSIM}\big(\tilde{\mathcal{I}},\widetilde{\mathcal{I}}_{\text{p}}\big), \tag{5}$$

where $\odot$ denotes element-wise multiplication. This loss formulation ensures that regions with high segmentation uncertainty $U(i,j)$ contribute less to the overall loss, thus emphasizing the confident areas in the image. The optimization is steered towards refining the scene Gaussians in regions where the pseudo-view segmentation is reliable, leading to more accurate and visually coherent reconstruction.

3.4. Iterative Refinement of Gaussians

After obtaining the coarse Gaussians γ_n from $\mathcal{F}(\cdot)$ in Section 3.1, utilizing camera controlled LVDM $\mathcal{G}$ for generating pseudo views for guidance (in Section 3.2) and defining our uncertainty weighted refinement objective in Section 3.3, we optimize the parameters of γ_n iteratively to obtain $\gamma_n^{'}$. However, it is observed that naive refinement of all the parameters leads to distorted renderings. We discuss the optimization strategy, style, texture alignment and the final refining objective for our pipeline below.

Selective Optimization of Scales. We observe that naive refinement of Gaussian scales obtained from the baseline method $\mathcal{F}(\cdot)$ can lead to an explosion in gradients and distorted renderings. This is primarily due to the tendency to fit multiple small Gaussians per pixel during the warm-up phase which rapidly increases the total number of Gaussians representing the scene. We limit this by optimizing the scales within a fixed range which we vary in an order of magnitude of 10^{-2} from the initial scales obtained from $\mathcal{F}(\cdot)$.

Texture alignment with FST (Φ). LVDMs tend to exhibit over-saturated colors and features which may become more pronounced in indoor scenes (as shown in Figure 5). While the generated images may be plausible, they might hurt the performance quantitatively in metrics like PSNR and SSIM which evaluate pixel-level accuracy. To address this, we perform *Fourier Style Transfer (FST)* adaptation [54] on the fly between the input image $\mathcal{I}$ and the output images generated by the LVDM model. This helps align the pseudo-images $\widetilde{\mathcal{I}}_{\text{p}}$ with the texture of the incoming source images. Aligning textures in this manner contributes to improving the overall performance, as shown in Table 4.

Refinement Objective. During the iterative refinement stage, we adjust the parameters of each Gaussian primitive, $\gamma_n = (\mu_n, R_n, S_n, \alpha_n)$, to better align the synthesized pseudo target view with the reference. Specifically, we optimize using the uncertainty-weighted reconstruction loss

Table 1. **Novel View Synthesis.** Our model shows superior performance on RealEstate10k [26] for small, medium, and large baseline ranges. We highlight the best performance in **bold** and the second best performance in underline.

Model	5 frames			10 frames			$\mathcal{U}$[-30,30]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Syn-Sin [55]	–	–	–	–	–	–	22.30	0.740	–
SV-MPI [56]	27.10	0.870	–	24.40	0.812	–	23.52	0.785	–
BTS [16]	–	–	–	–	–	–	24.00	0.755	0.194
Splatter Image [29]	24.15	0.894	0.110	25.60	0.760	0.240	23.10	0.730	0.290
MINE [57]	28.45	0.897	0.111	25.89	0.850	0.150	24.75	0.820	0.179
Flash3D [1]	28.46	0.899	0.100	25.94	0.857	0.133	24.93	0.833	0.160
UAR-Scenes	**28.67**	**0.902**	**0.095**	**26.54**	**0.861**	**0.112**	**27.81**	**0.887**	**0.107**

(Eq.5), $\mathcal{L}_{\text{uw-rec}}$, after performing the color correction with FST, which leads to our final optimization objective.

$$\mathcal{L}^{\Phi}_{\text{uw-rec}} = \left\| \left(1 - U(i,j)\right) \odot \left(\tilde{\mathcal{I}} - \Phi(\widetilde{\mathcal{I}}_{\text{p}})\right) \right\|_2 + \alpha\, \text{SSIM}(\tilde{\mathcal{I}}, \Phi(\widetilde{\mathcal{I}}_{\text{p}})), \quad (6)$$

In practice, the Gaussian parameters are iteratively refined via a gradient-based scheme, ensuring that each update improves the fidelity of the rendered view while accounting for the uncertainty maps obtained using the segmentation model $\mathcal{S}$. The 2D supervisory updates rule progressively accounts for regions with higher confidence (lower uncertainty) that have a greater influence on the parameter updates, leading to refined, more accurate representations of the scene.

4. Experiments

We evaluate UAR-Scenes on the task of Novel View Synthesis (NVS) of scenes. Keeping in mind that multi-view consistency is a major focus to assess the quality of refined scenes, we present results on both narrow and wide baselines in Table 1. We further evaluate the interpolation and extrapolation performance following [1] and [30] after refinement in Table 2. To highlight the transferability of our method across datasets, we also show NVS results on KITTI-v2 in Table 3. This shows our refinement pipeline can be seamlessly integrated with any existing reconstruction method across a wide variety of tasks. Finally, we present ablation studies in showing the importance of each component in our framework and how it contributes towards overall performance.

4.1. Experimental Setup

Datasets. Our experiments involve in-domain results on the real-world scenes dataset RealEstate-10K [26] and out-domain results on the driving dataset KITTI-v2 [27] following the protocol of the base reconstruction model [1]. RealEstate-10K consists of 69893 real estate videos showing the indoor and outdoor views of houses collected from YouTube. KITTI-V2 consists of driving sequences recorded with a stereo camera per day. We follow the standard protocol of 1079 images for testing this dataset.

Evaluation Metrics. We adopt PSNR, SSIM [53] and Perceptual Similarity (LPIPS) [60] as our photo-metric evaluation standard as in [1, 11, 12]. Further, we report Frechet-Inception Distance (FID) [61] as well since we generate plausible explanations for unobserved regions which do not have a corresponding paired ground truth.

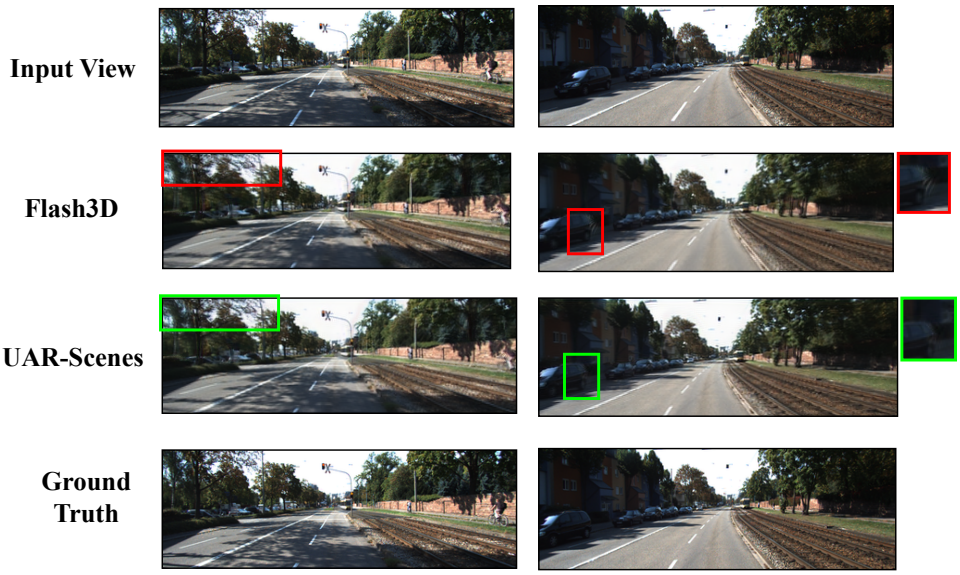
Pseudo-View Generation. We utilize the MotionCtrl [22] LVDM model which has fine-grained camera control for generating the pseudo views and provides supervision for refining the Gaussian primitives. This is essential to align the generated views with those obtained by rendering the coarse Gaussians of $\mathcal{F}(\cdot)$. Following this, we perform 50 inference sampling steps with a reduced batch size of 10 to address the increased computational load.

Optimization Details. We optimize each scene for 1000 steps in which the learning rate of position information decays from 1×10^{-3} to 2×10^{-5}. We use a single NVIDIA L40 48 GB GPU for running all our experiments. Finally, our rendering resolution for all images is 256×384 following the protocol in [1]. We keep a batch size of 2 for all our experiments. Further implementation details are listed in the supplementary material.

4.2. Novel View Synthesis

RealEstate-10K. We benchmark our refinement model on the RealEstate dataset where we show improvements over stereo and sparse view methods while refining a single-image to 3D reconstruction pipeline. UAR-Scenes obtains around 1dB improvement on average across both the closer and wide baselines in PSNR showing the effectiveness of our pipeline in improving existing feed-forward models. Notice that while the performance of the baseline model Flash3D steadily decreases as the distance from the source increases, UAR-Scenes is able to still provide decent renderings. This is important as it suggests that these feed-forward methods fail to capture those areas of the scene which progressively start falling out of the range of the input conditioning image. We provide qualitative results in Figure 1 and Figure 4(a) where we clearly show how UAR-Scenes provides high quality rendering results in cases where Flash3D fails. We

(a) RealEstate-10K



(b) KITTI-v2

Figure 4. **Qualitative Results.** (a) Qualitative comparisons on the RealEstate-10K dataset shows that our method produces more *realistic results* which are more plausible and faithful to the original image (In 1st row, Flash3D's renderings are blurry outside the camera's seen frustum where `UAR-Scenes` is able to complete the window. Similarly `UAR-Scenes` can provide reasonable completions which may not always align with the GT as shown in 2nd row). (b) Qualitative comparison on the KITTI-v2 dataset which shows that our method can deliver *sharp results especially in edges* where there may be ambiguity (Back edge of car is distorted in Flash3D's prediction in 2nd column). Notice that despite the significant camera motion between the original input view and the target novel views, `UAR-Scenes` can render realistic and plausible renderings as highlighted above.

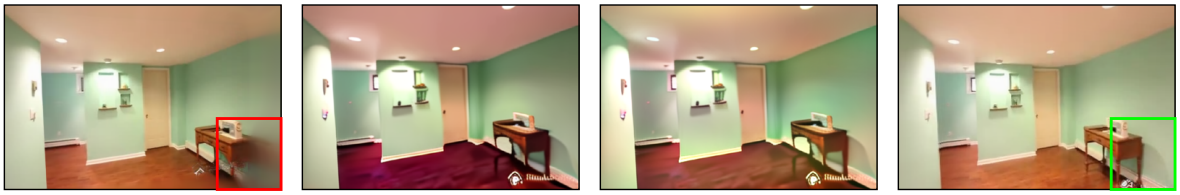


Figure 5. **Ablation Results.** The leftmost image is the rendered view from the baseline method Flash3D which fails in extrapolation. Next, we have the LVDM generated image which clearly has oversaturated textures which does not align with real world scenes. On the 3rd image from the left, FST alleviates this issue by performing style alignment which leads to better quality results in the final output on the right.

Table 2. **Interpolation vs. Extrapolation.** We compare our method (UAR-Scenes) on the RealEstate-10K dataset against baselines on PSNR, SSIM, LPIPS, and FID metrics. We highlight the best performance in **bold** and the second best performance in underline.

Method	Interpolation			Extrapolation			
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
PixelNeRF [10]	24.00	0.589	0.550	20.05	0.575	0.567	160.77
Du et al. [58]	24.78	0.820	0.410	21.23	0.760	0.480	14.34
pixelSplat [11]	25.49	0.794	0.291	22.62	0.777	0.216	5.78
latentSplat [30]	25.53	0.853	0.280	23.45	0.801	0.190	2.97
MVSplat [12]	**26.39**	0.869	0.128	24.04	0.812	0.185	3.87
Flash3D [1]	23.87	0.811	0.185	24.10	0.815	0.185	4.02
UAR-Scenes	26.37	**0.871**	**0.125**	**24.37**	**0.819**	**0.144**	**2.55**

Table 3. **Out-Domain Evaluation.** We evaluate on the KITTI-v2 [27] Dataset. We highlight the best performance in **bold** and the second best performance is underlined. We beat the baseline method comprehensively.

	KITTI		
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**
LDI [59]	16.50	0.572	–
SV-MPI [56]	19.50	0.733	–
BTS [16]	20.10	0.761	0.144
MINE [57]	21.90	0.828	0.112
Flash3D [1]	21.96	0.826	0.132
UAR-Scenes	**22.31**	**0.844**	**0.128**

show additional qualitative results in the supplementary.

KITTI-v2. We also perform an out-domain evaluation on the KITTI dataset showing that our refinement procedure is dataset agnostic and can be used in a post-hoc fashion to adapt to any setting. We obtain a PSNR improvement of around 0.35 dB over the baseline method. We provide further qualitative results in Figure 4(b).

4.3. Generative Results

We report interpolation and extrapolation results on the RealEstate-10K dataset using Flash3D as the baseline method. We are able to beat all the methods except MVSplat which uses stereo information to interpolate. In Extrapolation, over a wide baseline, our FID is significantly lower then all methods including LatentSplat [30] which uses a generative GAN denoiser but still fails to handle complex real-world scenes. Except Flash3D [1], we report FID numbers following existing feed forward methods [30, 62].

4.4. Ablation Studies

Architectural Choice. We conduct our ablation studies on the real world RealEstate-10K dataset to measure the effects of the various architectural designs on the overall performance. The results are listed in Table 4. It can be observed that adding the LVDM and performing uncertainity aware refinement contributes significantly to performance gains. The rendering effect behind each choice is shown in Figure 5. Note how the oversaturated texture information generated by the LVDM does not align with either the input image or the ground truth. We therefore perform the alignment operation using Φ to produce better quality supervision which is more practical for Novel View Synthesis tasks for scenes.

Table 4. **Ablation Studies.** We report PSNR (↑), SSIM (↑), and LPIPS (↓) on RealEstate-10K. The component columns indicate whether LVDM ($\mathcal{G}$), FST (Φ), and uncertainty ($\mathcal{U}$) are included.

Models	**Components**			**Metrics**		
	$\mathcal{G}$	Φ	$\mathcal{U}$	PSNR↑	SSIM↑	LPIPS↓
Baseline	✗	✗	✗	24.93	0.833	0.160
Baseline + LVDM	✓	✗	✗	27.24	0.867	0.126
Baseline + LVDM-FST	✓	✓	✗	27.33	0.869	0.119
UAR-Scenes	✓	✓	✓	**27.81**	**0.887**	**0.107**

5. Conclusion

We introduced UAR-Scenes, a novel 3D scene refinement pipeline that enhances scene Gaussians derived from a single image, thereby improving the quality of Gaussians produced by 3D reconstruction pipelines. Our results demonstrate the versatility of UAR-Scenes across both in-domain and out-domain datasets. Through Novel View Synthesis experiments, we outperform state-of-the-art feed-forward methods across small, medium, and large baseline settings. Additionally, UAR-Scenes excels in challenging interpolation and extrapolation tasks, yielding superior rendered views and high-quality generations at unseen camera angles, as evidenced by both qualitative and quantitative results. Furthermore, our ablation studies validate the effectiveness of individual components and design choices within the UAR-Scenes pipeline, highlighting the benefits of employing Fourier-style texture alignment with real-world scenes. Overall, our findings highlight UAR-Scenes's potential to advance 3D scene understanding and novel view synthesis.

References

[1] 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. *arXiv preprint arXiv:2406.04343*, 2024. 1, 2, 3, 6, 8

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[3] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 1

[4] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20507–20518, 2024. 1

[5] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 1

[6] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 1, 2

[7] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *arXiv*, 2023. 3

[8] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.

[9] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[10] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1, 8

[11] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 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*, pages 19457–19467, 2024. 3, 6, 8

[12] 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 *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 1, 2, 3, 6, 8

[13] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Reference-guided controllable inpainting of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17815–17825, 2023. 2

[14] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21551–21561, 2024. 2, 3

[15] Dongqing Wang, Tong Zhang, Alaa Abboud, and Sabine Süsstrunk. Innerf360: Text-guided 3d-consistent object inpainting on 360-degree neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12677–12686, 2024. 2

[16] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9076–9086, 2023. 2, 6, 8

[17] Rui Li, Tobias Fischer, Mattia Segu, Marc Pollefeys, Luc Van Gool, and Federico Tombari. Know your neighbors: Improving single-view reconstruction via spatial vision-language reasoning. In *CVPR*, 2024. 2

[18] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[19] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *ICLR*, 2024. 2

[20] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[21] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6796–6807, 2024. 2

[22] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 4, 6

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 5

[24] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic seg-

mentation. In *International Conference on Learning Representations*, 2022. 2, 4, 5

[25] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2

[26] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2, 6

[27] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 6, 8

[28] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20050, 2024. 2

[29] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10208–10217, 2024. 6

[30] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European Conference on Computer Vision*, pages 456–473. Springer, 2024. 3, 6, 8

[31] Shengji Tang, Weicai Ye, Peng Ye, Weihao Lin, Yang Zhou, Tao Chen, and Wanli Ouyang. Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025.

[32] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16453–16463, 2025. 2

[33] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans, 2017. 3

[34] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12608–12618, 2023. 3

[35] Titas Anciukevičius, Fabian Manhardt, Federico Tombari, and Paul Henderson. Denoising diffusion via image-based rendering. In *The Twelfth International Conference on Learning Representations*, 2024.

[36] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision, 2023.

[37] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8863–8873, 2023. 3

[38] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2025. 3

[39] Ruiqi Gao, Aleksander Hołyński, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: create anything in 3d with multi-view diffusion models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 75468–75494, 2024. 3

[40] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 3

[41] Hanyang Yu, Xiaoxiao Long, and Ping Tan. Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors. *arXiv preprint arXiv:2409.03456*, 2024. 3

[42] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 3

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[44] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations, 2021. 3

[45] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification, 2022.

[46] Luca Savant, Diego Valsesia, and Enrico Magli. Modeling uncertainty for gaussian splatting, 2024. 3

[47] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty quantification in neural radiance fields. *CVPR*, 2024. 3

[48] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 230–246. Springer, 2022. 3

[49] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4

[50] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 3

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 6

[54] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 5

[55] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 6

[56] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 6, 8

[57] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 6, 8

[58] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4970–4980, 2023. 8

[59] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 8

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[61] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[62] 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. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024. 8