Arxiv

ICCV HONOLULU HAWAII OCT 19-23, 2025

UAR-Scenes: Uncertainty-Aware Diffusion Guided Refinement of 3D Scenes

Sarosij Bose, Arindam Dutta, Sayak Nag, Junge Zhang, Jiachen Li, Konstantinos Karydis, Amit K. Roy Chowdhury

UCR Vision and Learning Group

Center for Robotics & Intelligent Systems

UC RIVERSIDE Marlan and Rosemary Bourns College of Engineering

UC RIVERSIDE | ARCS Laboratory Autonomous Robots and Control Systems Lab

Introduction/Motivation

- State-of-the-art feed forward algorithms (such as Szymanowicz et. al*, Flash3D) leverage feed-forward gaussian splatting to produce 3D scenes from sparse views or even a single image.

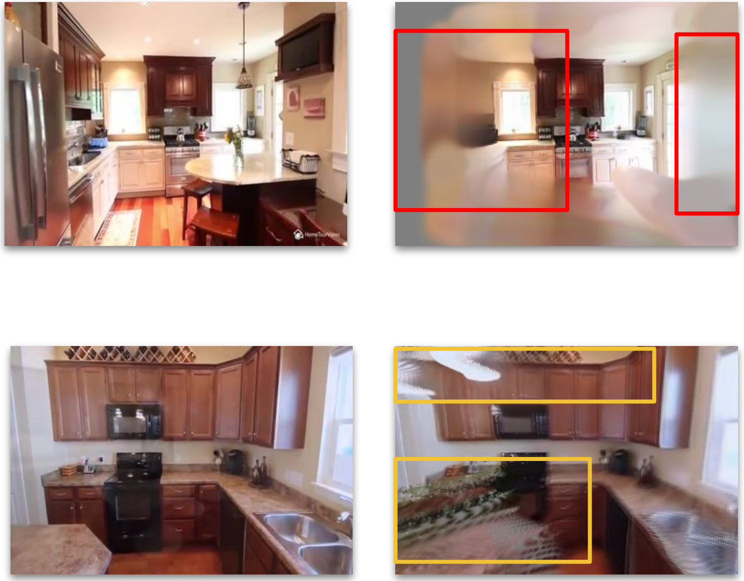
Source ***Novel View***

**Szymanowicz et. al, "Flash3D: Feed-Forward Generalizable 3D Scene Reconstruction from a Single Image", 3DV 2025*

Introduction/Motivation

- State-of-the-art feed forward algorithms (such as Szymanowicz et. al*, Flash3D) leverage feed-forward gaussian splatting to produce 3D scenes from sparse views or even a single image.

- This fast reconstruction comes at a cost: the reconstructed scenes contain artifacts and perform poorly in unseen and occluded regions far from the source view.



Source ***Novel View***

*Szymanowicz et. al, "Flash3D: Feed-Forward Generalizable 3D Scene Reconstruction from a Single Image", 3DV 2025

How to Synthesize Plausible Views?

Video Diffusion Models can synthesize new views!*

Source View

Synthesis

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*

How to Synthesize Plausible Views?

Video Diffusion Models can synthesize new views!*

But it can't be controlled in a particular trajectory

Source View

Uncontrolled Synthesis

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*

How to Synthesize Plausible Views?

Video Diffusion Models can synthesize new views!*

We inject Plücker Embeddings to condition along the trajectory

→

Source View

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*

How to Synthesize Plausible Views?

Video Diffusion Models can synthesize new views!*

We inject Plücker Embeddings to condition along the trajectory

$\mathbf{\Pi} = (\mathbf{R}, \mathbf{T})$

Source View

Controlled Synthesis

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*

How to Synthesize Plausible Views?

This can now be used for any trajectory

Video Diffusion Models can synthesize new views!*

We inject Plücker Embeddings to condition along the trajectory

$\Pi = (R, T)$

Source View

Controlled Synthesis

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*

Semantic Uncertainty Quantification

- Even if plausible, *the generative prior* is not actually aware of what it's output is. To improve it in a self-supervised manner, we provide additional guidance in the form of **uncertainty maps**.

Plausible Synthesis

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*
**Li et. al, "Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models" (BLIP2), ICML 2023*

Semantic Uncertainty Quantification

- Even if plausible, *the generative prior* is not actually aware of what it's output is. To improve it in a self-supervised manner, we provide additional guidance in the form of **uncertainty maps**.

- We distill semantics from an **open-set segmentation model** to gauge uncertainty by extracting classes with an MLLM.

Plausible Synthesis

**Blatmann et. al, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets" (SVD), ArXIV, 2024*
**Li et. al, "Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models" (BLIP2), ICML 2023*

Semantic Uncertainty Quantification

- Even if plausible, *the generative prior* is not actually aware of what it’s output is. To improve it in a self-supervised manner, we provide additional guidance in the form of **uncertainty maps**.

- We distill semantics from an **open-set segmentation model** to gauge uncertainty by extracting classes with an MLLM.

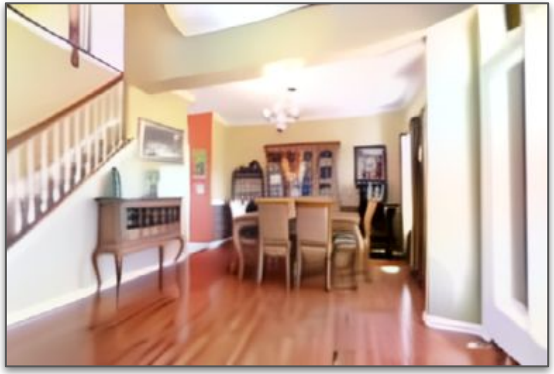
- The MLLM* is shown some in-context examples to act as a open-set object classifier as shown alongside.

Plausible Synthesis

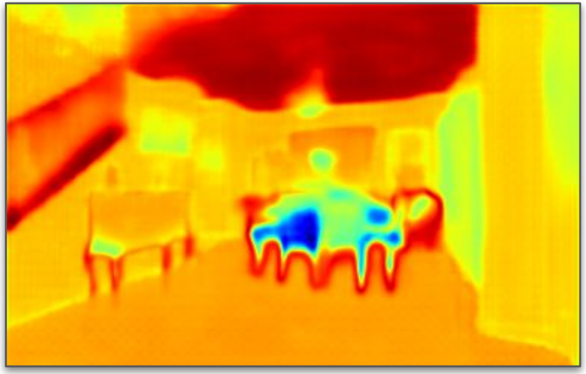
[“countertop, “table”, “clock”, fridge.”]

**Blatmann et. al, “Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets” (SVD), ArXIV, 2024*
**Li et. al, “Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models” (BLIP2), ICML 2023*

Semantic Uncertainty Quantification

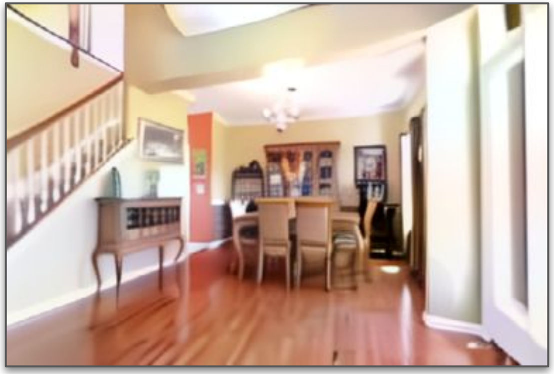


Novel View

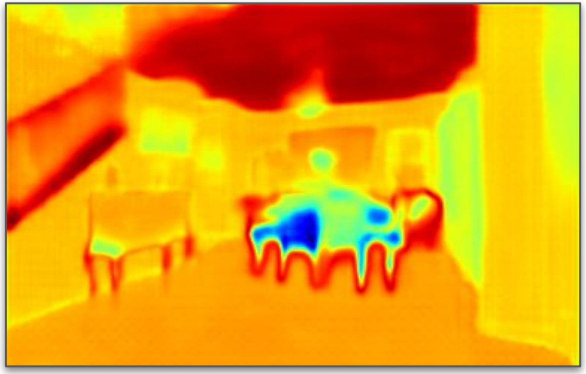


Uncertainty Map

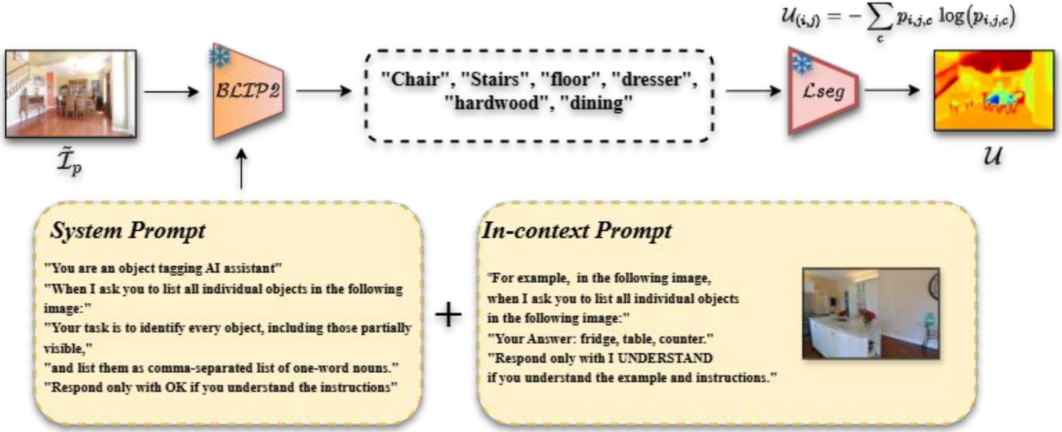
Semantic Uncertainty Quantification



Novel View

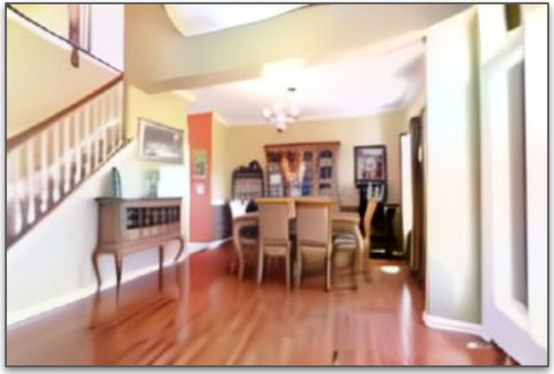


Uncertainty Map

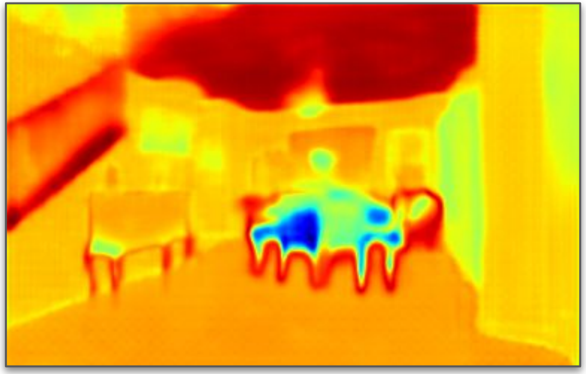


UQ Estimation Pipeline

Semantic Uncertainty Quantification

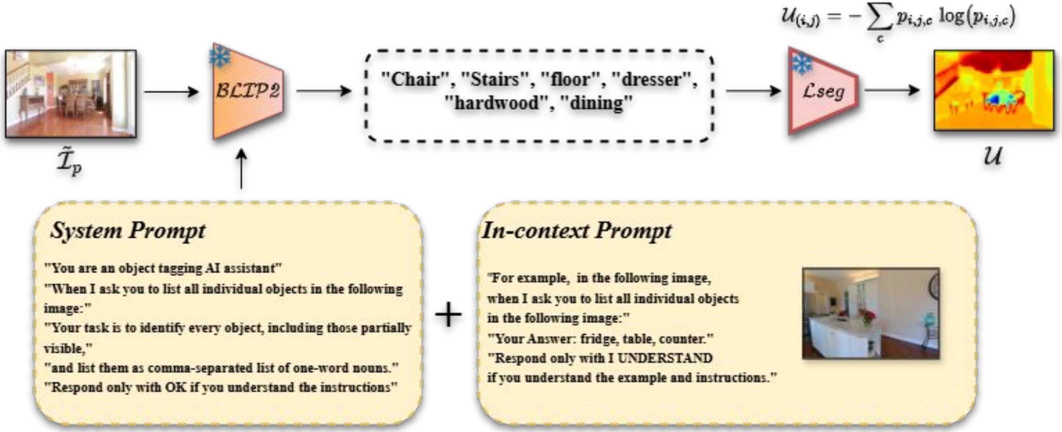


Novel View



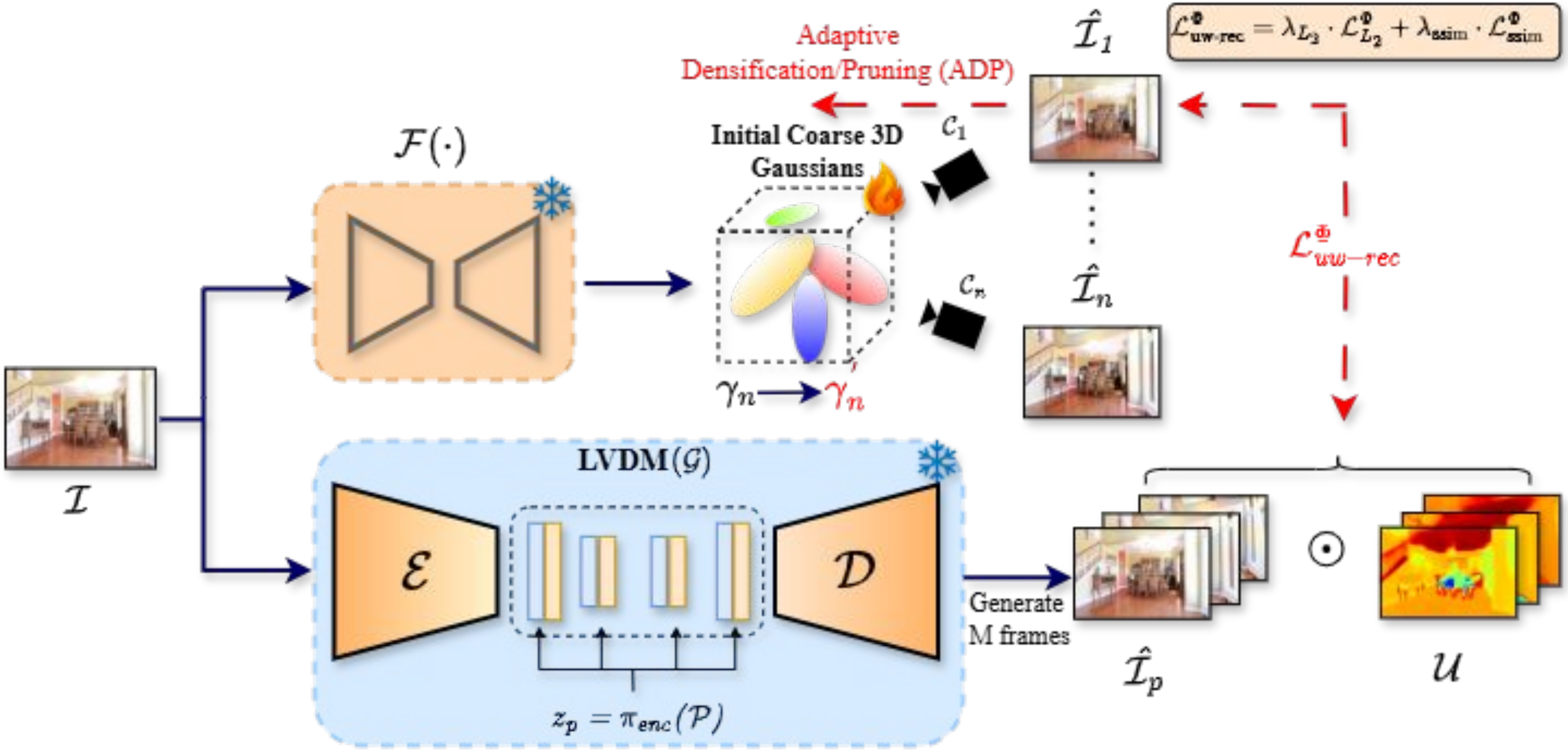
Uncertainty Map

Simple Distillation of semantics from an open-set segmentation model using a MLLM driven prior is enough to estimate uncertainty!



UQ Estimation Pipeline

Pipeline Overview



$\mathcal{I}$: Input Image $\mathcal{F}(\cdot)$: FF-Network $\mathcal{P}$: Camera Extrinsic π_{enc} : Camera Encoder γ : 3D Gaussians $\hat{\mathcal{I}}_p$: Pseudo Views $\mathcal{U}$: Uncertainty Map

Quantitative Comparisons

NVS on RealEstate-10K (In-domain)

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 <u>underline</u>.

Model	5 frames			10 frames			$\mathcal{U}$[-30,30]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Syn-Sin [52]	–	–	–	–	–	–	22.30	0.740	–
SV-MPI [53]	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 [54]	28.45	0.897	0.111	25.89	0.850	0.150	24.75	0.820	0.179
Flash3D [1]	<u>28.46</u>	<u>0.899</u>	<u>0.100</u>	<u>25.94</u>	<u>0.857</u>	<u>0.133</u>	<u>24.93</u>	<u>0.833</u>	<u>0.160</u>
`UAR-Scenes`	**28.67**	**0.902**	**0.095**	**26.54**	**0.861**	**0.112**	**27.81**	**0.887**	**0.107**

Interpolation/Extrapolation on RealEstate-10K

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 <u>underline</u>.

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. [55]	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	<u>2.97</u>
MVSplat [12]	**26.39**	<u>0.869</u>	<u>0.128</u>	24.04	0.812	0.185	3.87
Flash3D [1]	23.87	0.811	0.185	<u>24.10</u>	<u>0.815</u>	<u>0.185</u>	4.02
`UAR-Scenes`	<u>26.37</u>	**0.871**	**0.125**	**24.37**	**0.819**	**0.144**	**2.55**

Quantitative Comparisons

NVS on RealEstate-10K (In-domain)

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 [52]	–	–	–	–	–	–	22.30	0.740	–
SV-MPI [53]	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 [54]	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**

Interpolation/Extrapolation on RealEstate-10K

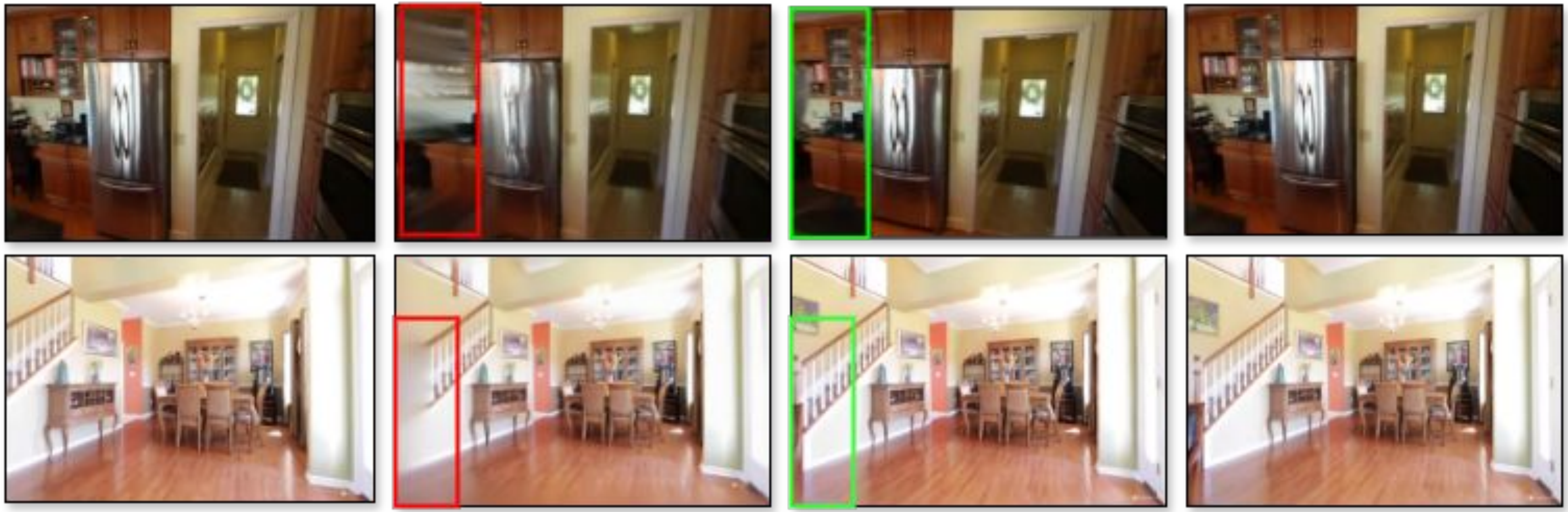
Table 2. **Interpolation vs. Extrapolation.** We compare our method (UAR-Scenes) on the RealEstate-10K dataset against baselines on PSNR, SSIM, LPIPS, and FID metrics. We highlight the best performance in **bold** and the second best performance in underline.

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. [55]	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**

NVS on KITTI-v2 (Out-domain)

Method	**KITTI**		
	PSNR↑	**SSIM↑**	**LPIPS↓**
LDI [56]	16.50	0.572	–
SV-MPI [53]	19.50	0.733	–
BTS [16]	20.10	0.761	0.144
MINE [54]	21.90	0.828	0.112
Flash3D [1]	21.96	0.826	0.132
UAR-Scenes	**22.31**	**0.844**	**0.128**

Qualitative Comparisons



Input View | Flash3D | UAR-Scenes | Ground Truth

Notice that there is significant camera motion between the input view and GT view

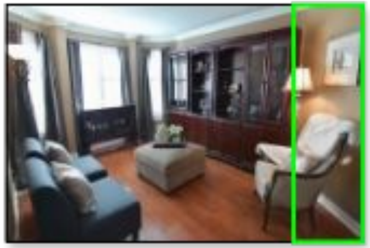
Novel View Synthesis using our method

Qualitative Comparisons

Plausible Prediction Capability

Input View

Flash3D



UAR-Scenes

Ground Truth

Input View

Flash3D

Robust out-domain performance!

UAR-Scenes

Ground Truth

Qualitative Comparisons

Flash3D LVDM LVDM-FST UAR-Scenes

Oversaturated Textures with vanilla LVDM

Texture alignment with FST

Authors

Sarosij Bose

Arindam Dutta

Sayak Nag

Junge Zhang

Jiachen Li

Konstantinos Karydis

Amit K. Roy Chowdhury

Thank You!

For more info, please read our paper!

UCR VLG | ARCS Laboratory