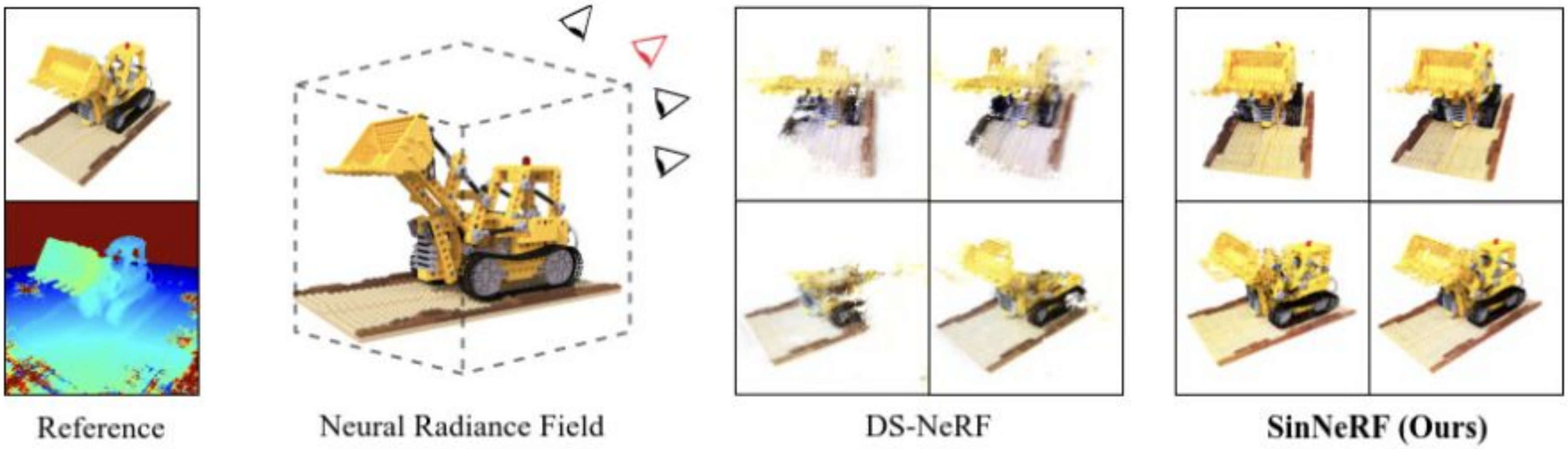
Team 2
Paper Presentation

Jinhyuk Jang, Prin Phunyaphibarn, Asiman Ziyaddinov

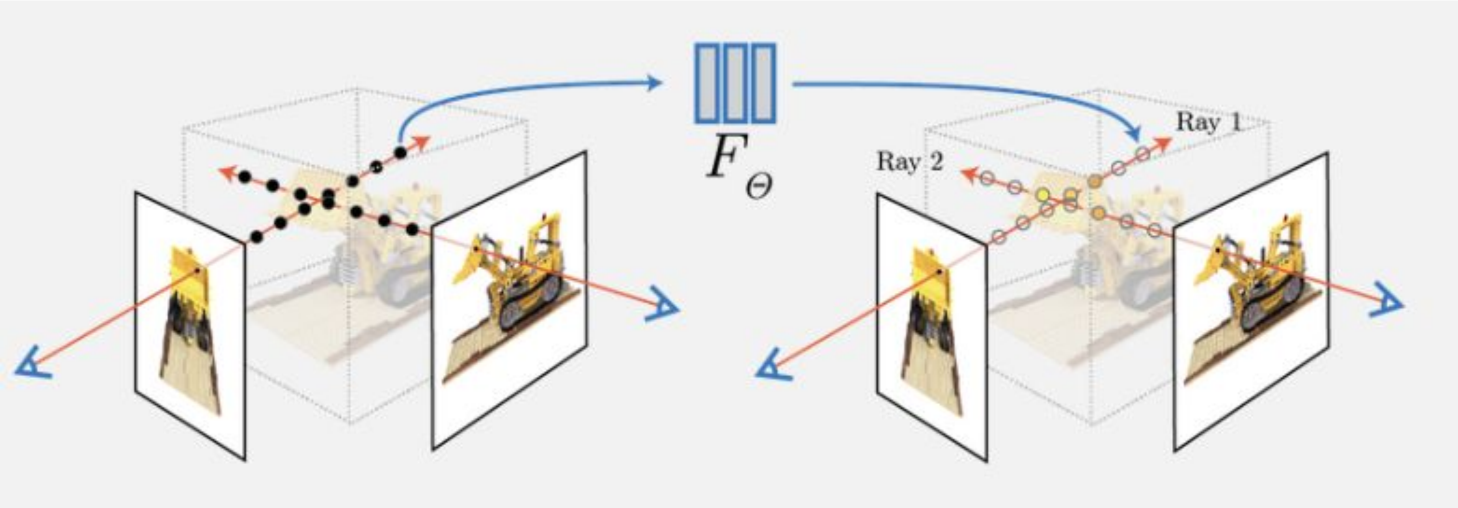
SinNeRF: Training Neural Radiance Fields from a Single Image (ECCV 2022)

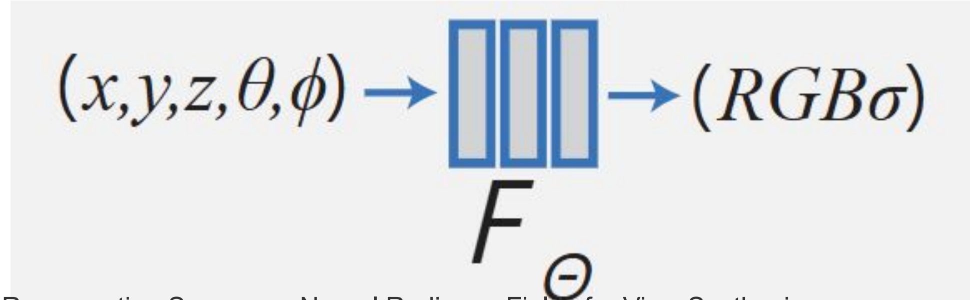


TL;DR: Given only a single reference view as input, our novel semi-supervised framework trains a neural radiance field effectively. In contrast, previous method shows inconsistent geometry when synthesizing novel views.

Dejia Xu, et al. 2022, SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image

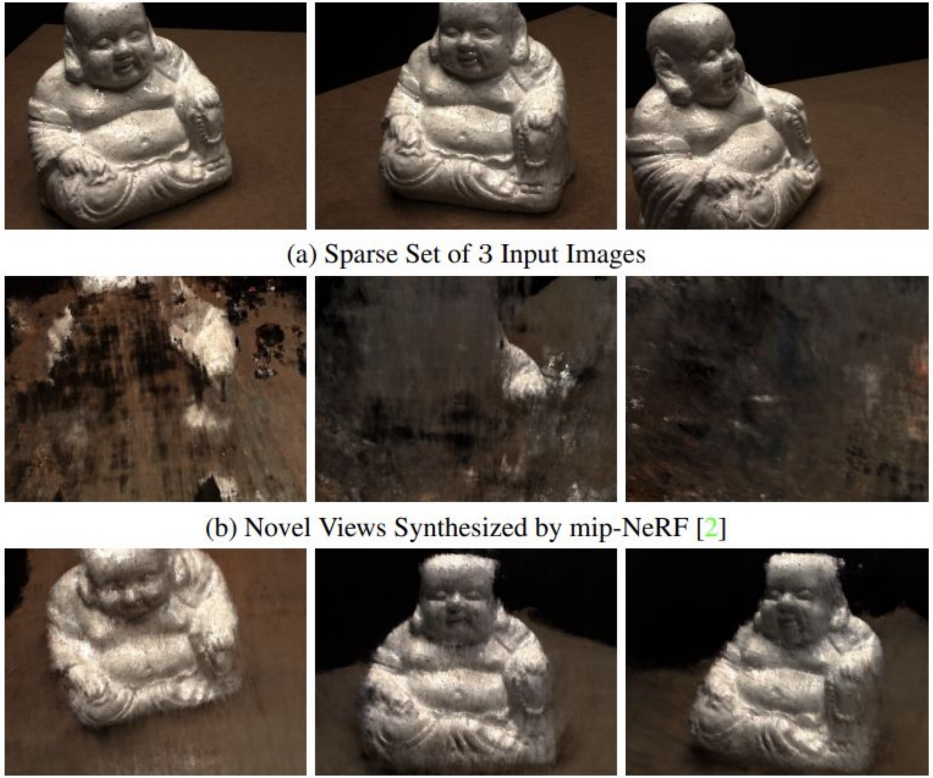
NeRF (Neural Radiance Fields)





Ben Mildenhall. et al, 2020, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

Sparse inputs cause problems!!



(a) Sparse Set of 3 Input Images

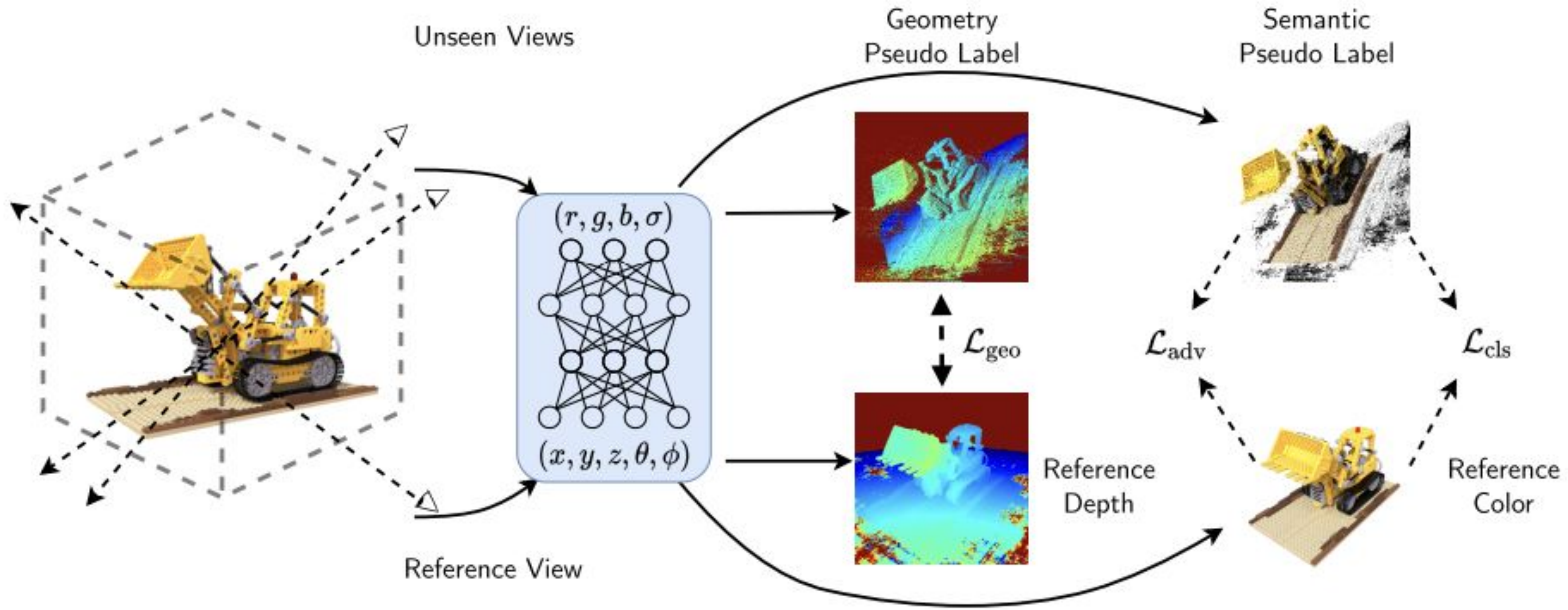
(b) Novel Views Synthesized by mip-NeRF [2]

(c) Same Novel Views Synthesized by Our Method

Michael Niemeyer. et al, 2022, RegNeRF:Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs

Solution:

Provide necessary constraints on unseen views

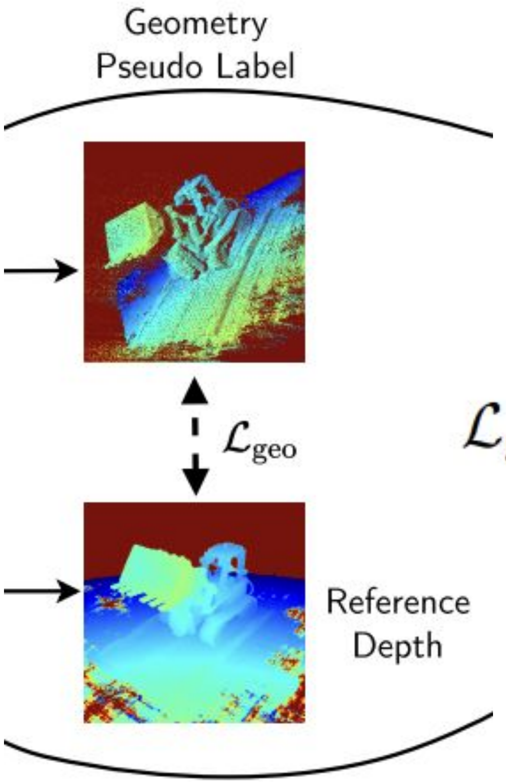


Geometry Pseudo Label (Pseudo Depth Label)

Image warping

<u>Pseudo depth label</u> is acquired during this process

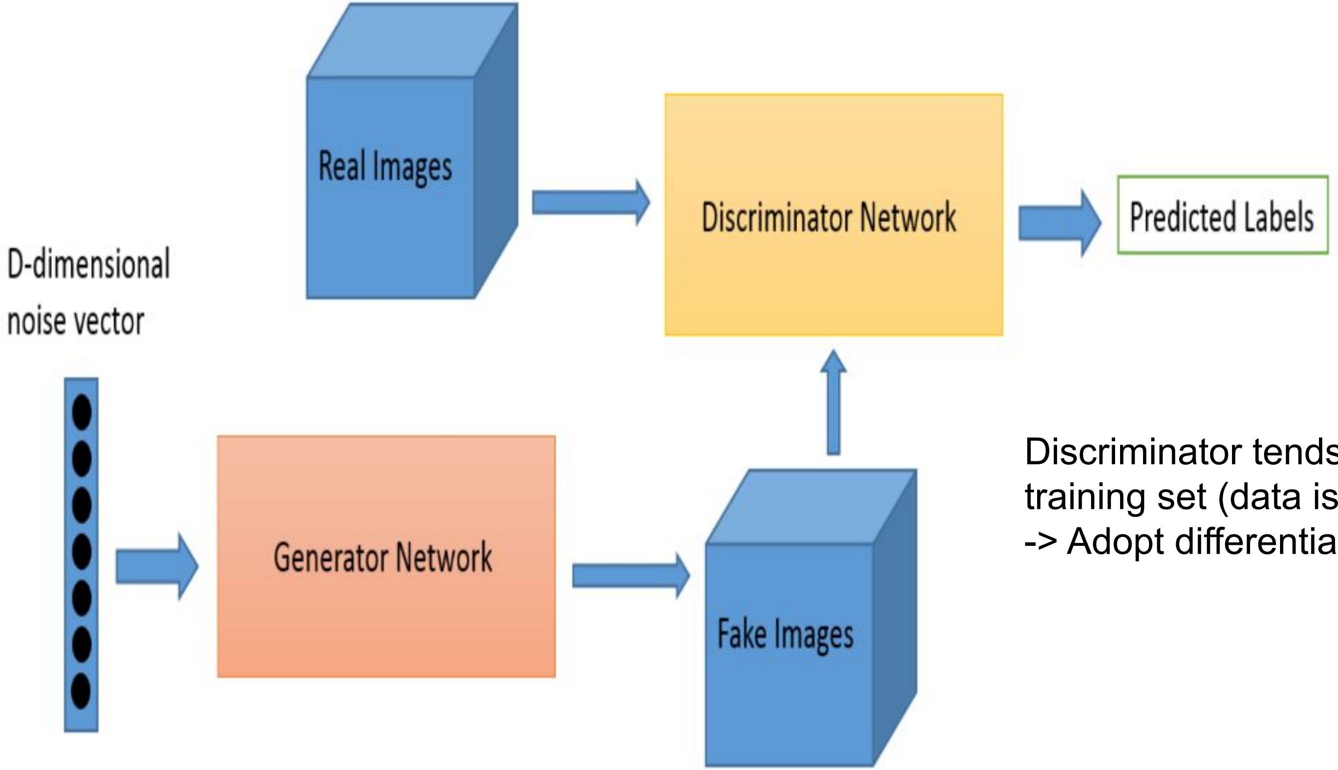
Geometry Pseudo Label



Geometric consistency with reference view and unseen view

$$\mathcal{L}_{\text{geo}} = \underbrace{\mathcal{L}_1(d_1, f(d_2))}_{\substack{\text{Depth of reference view} \\ \text{Vs} \\ \text{Depth of Image warping} \\ \text{on unseen view}}} + \underbrace{\mathcal{L}_1(f(d_1), d_2)}_{\substack{\text{Depth of unseen view} \\ \text{Vs} \\ \text{Depth of Image warping} \\ \text{on reference view}}} + \underbrace{\lambda_4 \mathcal{L}_{\text{smooth}}}_{\substack{\text{Regularize} \\ \text{uncertain regions} \\ \text{in warped results}}}$$

Semantic Pseudo Label - Local Texture Guidance



Discriminator tends to memorize entire training set (data is too limited)
-> Adopt differentiable augmentation!

Hamed Alqahtani. 2019. An Analysis Of Evaluation Metric Of GANs

Semantic Pseudo Label - Local Texture Guidance

Loss of Discriminator

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{data}}(\boldsymbol{x})}\left[f_D(-D(T(\boldsymbol{x})))\right] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_D(D(T(G(\boldsymbol{z}))))\right],$$

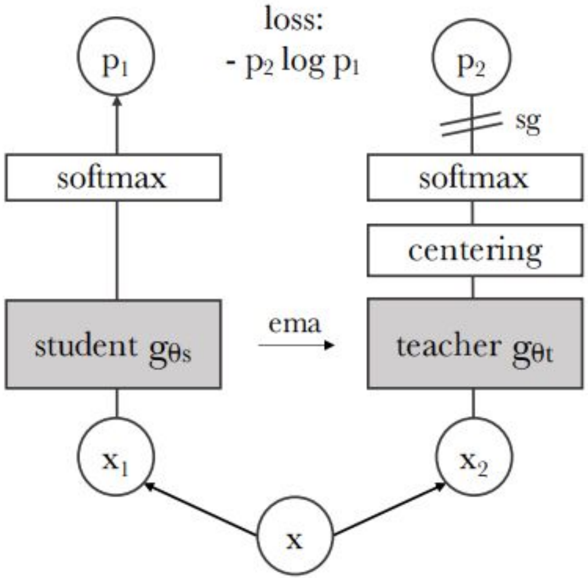
$$\mathcal{L}_{\mathrm{G}} = \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_G(-D(T(G(\boldsymbol{z}))))\right],$$

$$\mathcal{L}_{\mathrm{adv}} = \mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{G}},$$

Loss of Generator

Local textures are now similar between reference and unseen views

Semantic Pseudo Label - Global Texture Guidance



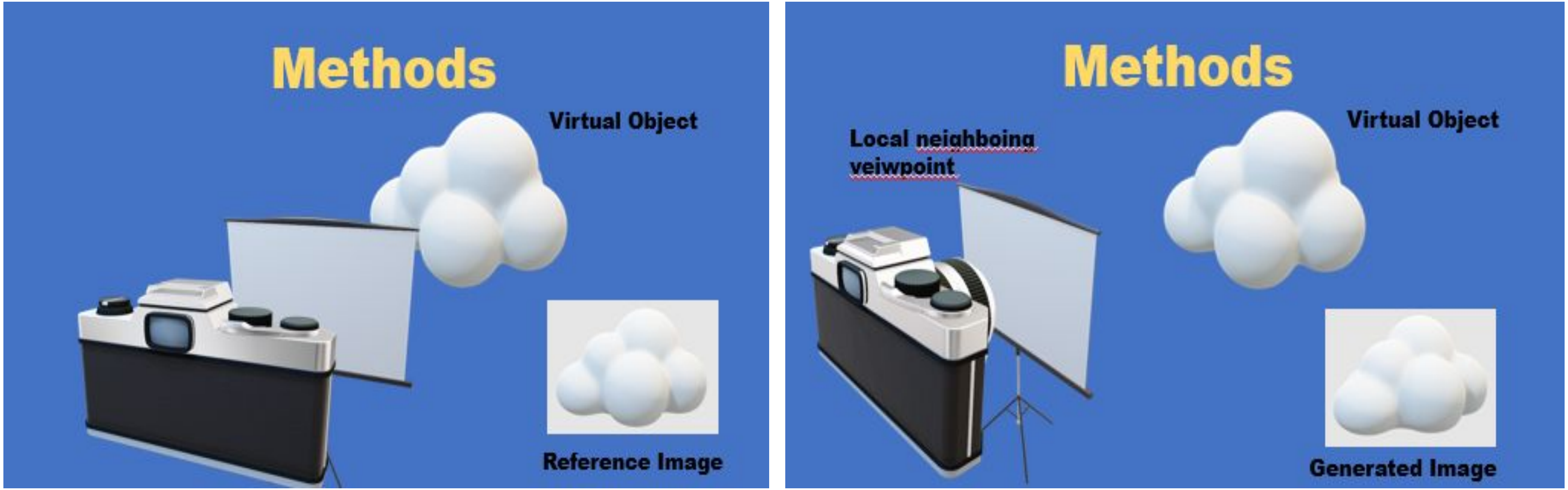
$$\mathcal{L}_{\text{cls}} = ||f_{\text{vit}}(A) - f_{\text{vit}}(B)||^2$$

Global Texture is now similar between reference and unseen view

DINO-ViT: self supervised vision transformer
CLS tokens from DINO-ViT's output = representation of entire image

Mathilde Caron, 2021, Emerging Properties in Self-Supervised Vision Transformers

Progressive Gaussion Pose Sampling



Progressive Sampling allows network to focus on dealing with confident regions

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \lambda_1 \mathcal{L}_{\text{geo}} + \lambda_2 \mathcal{L}_{\text{adv}} + \lambda_3 \mathcal{L}_{\text{cls}},$$

Quantitative evaluations

Scene	Pose Error					
	Translation($\times 10^{-2}$) $\downarrow$			Rotation($^\circ$) $\downarrow$		
	NeRFmm128	*NeRFmm256*	**SiNeRF**	*NeRFmm128*	*NeRFmm256*	**SiNeRF**
Fern	0.514	0.765	**0.438**	0.957	1.566	**0.743**
Flower	1.039	1.200	**0.796**	3.657	3.211	**0.506**
Fortress	6.463	6.046	**4.068**	2.590	2.410	**1.772**
Horns	1.607	**1.476**	2.153	3.806	3.044	**2.662**
Leaves	0.676	**0.608**	0.831	8.248	**6.782**	8.762
Orchids	1.627	2.243	**1.257**	4.140	5.459	**3.244**
Room	**1.315**	2.148	2.145	3.357	3.745	**2.075**
Trex	1.213	1.467	**0.462**	4.953	6.339	**0.856**
Mean	1.807	1.994	**1.519**	3.964	4.070	**2.578**

Scene	Image Quality								
	PSNR $\uparrow$			SSIM $\uparrow$			LPIPS $\downarrow$		
	NeRFmm128	*NeRFmm256*	**SiNeRF**	*NeRFmm128*	*NeRFmm256*	**SiNeRF**	*NeRFmm128*	*NeRFmm256*	**SiNeRF**
Fern	21.811	22.154	**22.482**	0.631	0.648	**0.665**	0.479	0.459	**0.437**
Flower	25.430	26.606	**27.229**	0.714	0.772	**0.798**	0.366	0.296	**0.295**
Fortress	26.173	25.596	**27.465**	0.653	0.602	**0.722**	0.438	0.538	**0.393**
Horns	22.949	23.174	**24.142**	0.626	0.635	**0.684**	0.492	0.506	**0.431**
Leaves	18.647	**19.741**	19.152	0.512	**0.609**	0.571	0.476	**0.385**	0.392
Orchids	16.695	15.858	**16.922**	0.391	0.350	**0.408**	0.540	0.550	**0.529**
Room	25.623	25.675	**26.101**	0.831	0.836	**0.844**	0.450	**0.411**	0.426
Trex	22.551	23.376	**24.939**	0.719	0.759	**0.816**	0.438	0.390	**0.356**
Mean	22.485	22.773	**23.554**	0.635	0.651	**0.689**	0.460	0.442	**0.407**

Qualitative results

Limitations

- Computational Intensity
- Generalization Constraints
- High Memory Usage
- Limited Performance in Large-Scale Scenes
- Potential Overfitting

Zero-1-to-3: Zero-shot One Image to 3D Object (ICCV 2023)

Ruoshi Liu
Columbia University

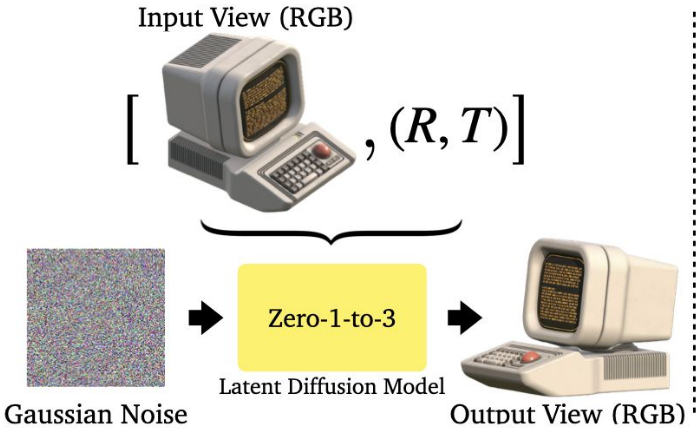
Rundi Wu
Columbia University

Basile Van Hoorick
Columbia University

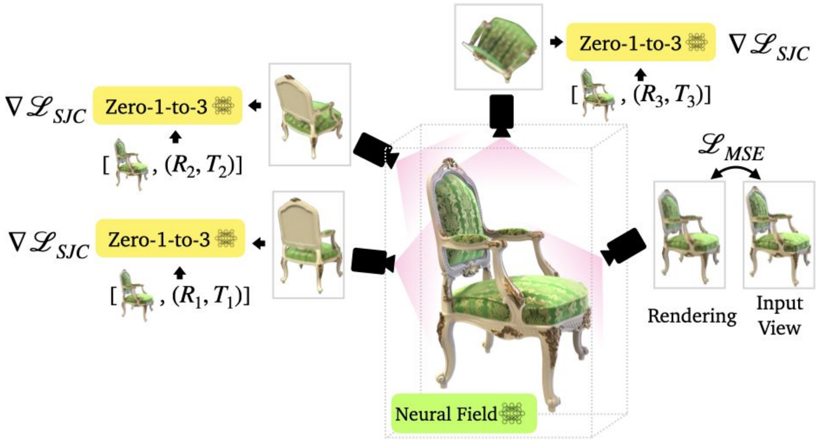
Pavel Tokmakov
Toyota Research Institute

Sergey Zakharov
Toyota Research Institute

Carl Vondrick
Columbia University



Novel View Synthesis

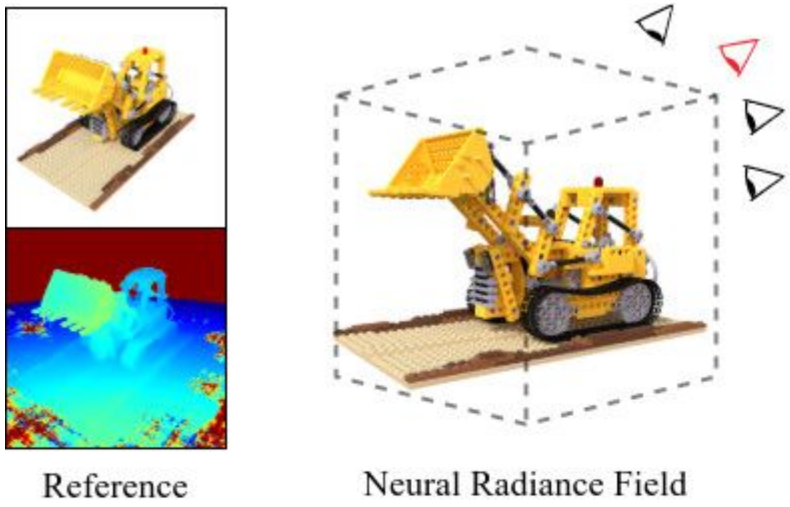


3D Reconstruction

Liu, Ruoshi, et al. "Zero-1-to-3: Zero-shot one image to 3d object." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Leveraging Stronger Priors

Pure NeRF/3DGS approaches are **ill-posed**



Not Enough Information!
(Ambiguities about Novel Views)

Solution:

Leverage Stronger Priors from Training Data

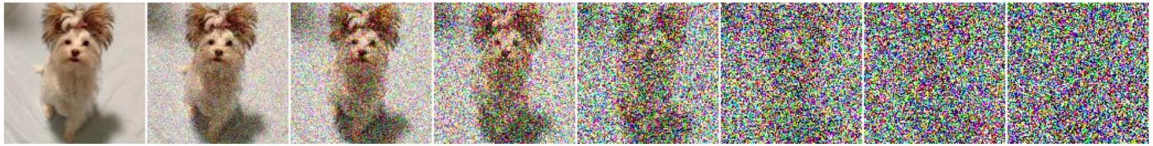
Objaverse-XL
> 10 million 3D objects

Deitke, Matt, et al. "Objaverse-xl: A universe of 10m+ 3d objects." Advances in Neural Information Processing Systems 36 (2024).

Diffusion Models

Forward SDE (data → noise)

$$\mathbf{x}(0) \xrightarrow{\quad d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \quad} \mathbf{x}(T)$$



score function

$$\mathbf{x}(0) \xleftarrow{\quad d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right] dt + g(t)d\bar{\mathbf{w}} \quad} \mathbf{x}(T)$$

Reverse SDE (noise → data)

Key Idea:

$\underbrace{\textbf{Reverse Noising}}_{\textbf{Learnable}}$ Converts $\underbrace{\textbf{Gaussian Noise}}_{\textbf{Easy to Sample}}$ into $\underbrace{\textbf{Clean Images}}_{\textbf{Hard to Sample}}$

Credit: https://yang-song.net/blog/2021/score/

Classifier Free Guidance

Given **(image, condition)** pairs, train to generate an image to match the condition

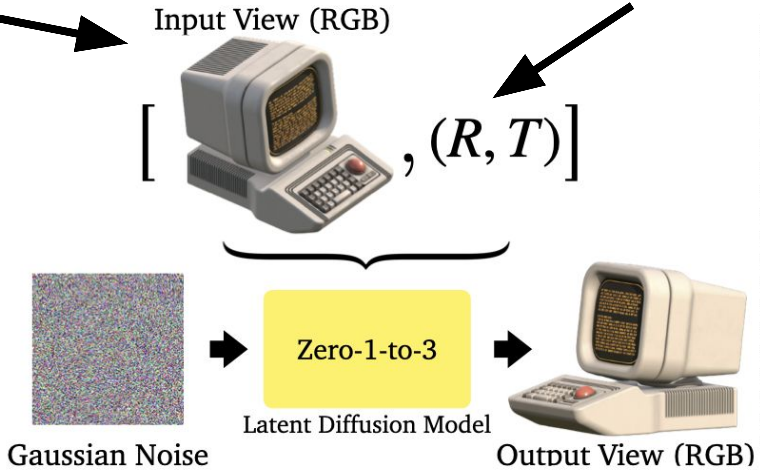
Text-to-Image

https://openai.com/index/dall-e-3/

Zero123: Conditional Diffusion Model

Reference Image

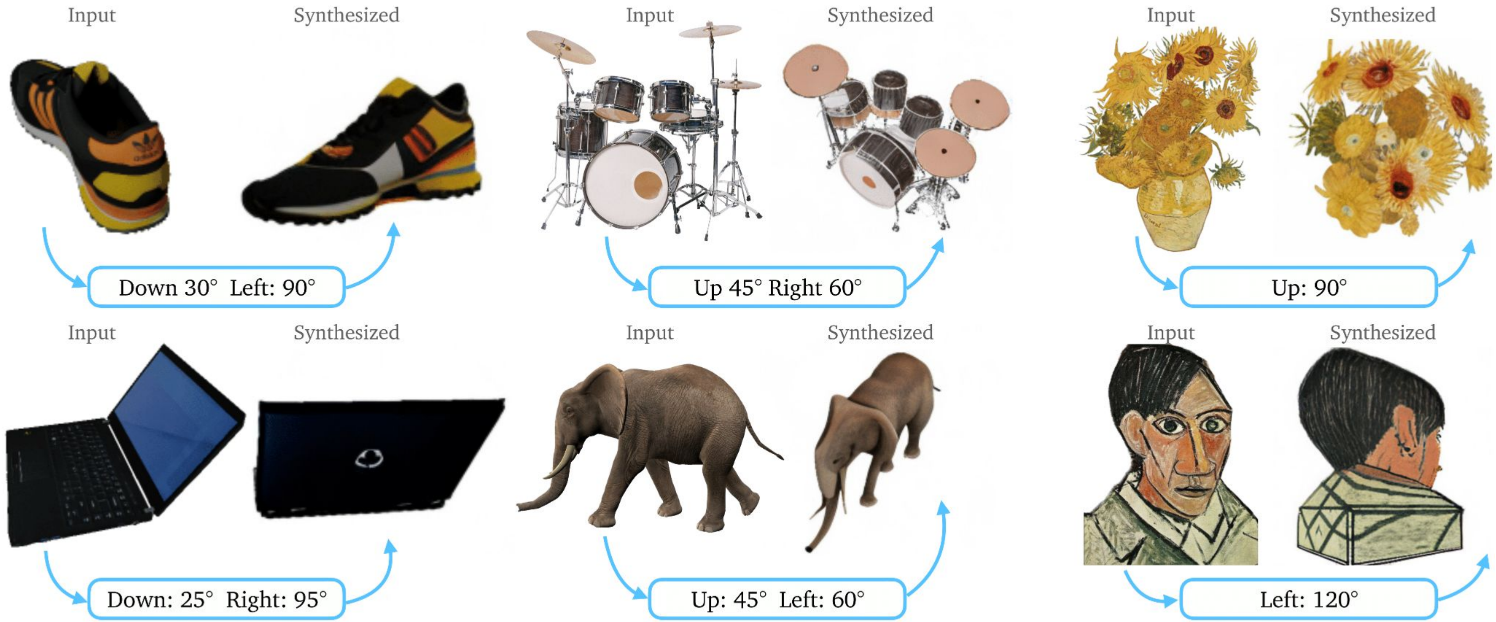
Relative Camera position and Orientation



Novel View Synthesis

Trained on Objaverse(-XL)

Single-Image Novel View Synthesis



Quantitative Results

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	Ours
PSNR ↑	<u>8.933</u>	5.914	6.573	**18.378**
SSIM ↑	<u>0.645</u>	0.540	0.552	**0.877**
LPIPS ↓	<u>0.412</u>	0.545	0.484	**0.088**
FID ↓	<u>12.919</u>	22.533	19.783	**0.027**

Google Scanned Objects Dataset (Single-Object)

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	Ours
PSNR ↑	7.130	6.561	<u>7.953</u>	**10.405**
SSIM ↑	0.406	0.442	<u>0.456</u>	**0.606**
LPIPS ↓	<u>0.507</u>	0.564	0.545	**0.323**
FID ↓	<u>5.143</u>	10.218	10.202	**0.319**

RTMV (Multi-Object)

Discussion

Pros

- By using training data, captures richer priors
- SoTA performance
- “Training-free” – No need to retrain for each object

Cons

- Only works on background-less images
- Generates random views depending on initial noise
- Slow generation speed