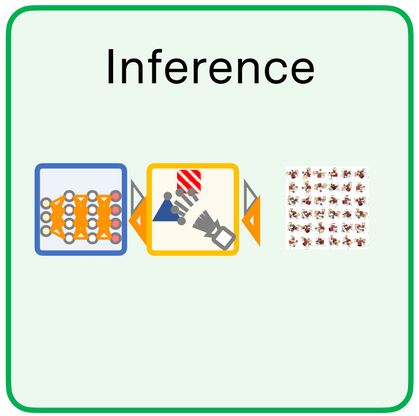
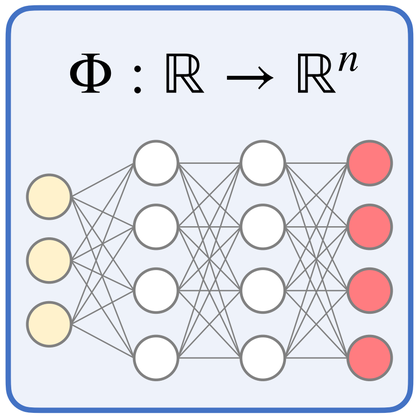
Conditional Generative Modeling

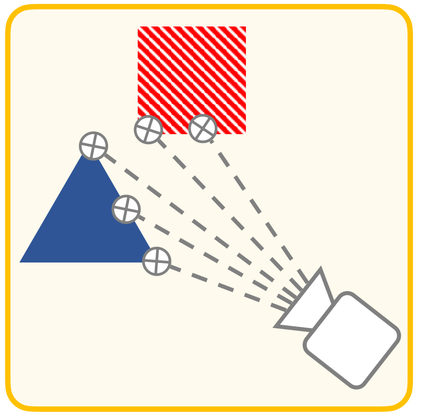
6.S980

Amortized 3D reconstruction

Observations







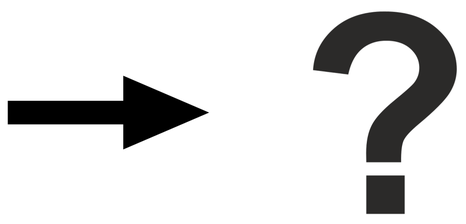
Renderings

Reconstructing 3D Scenes from a Single Image

→ ?

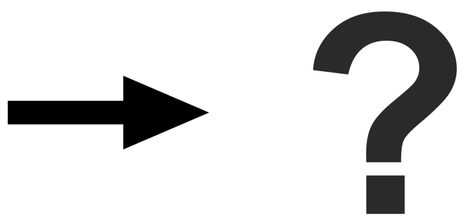
Diffusion with Forward Models: Solving Stochastic Inverse Problems Without Direct Supervision, Tewari et al. 2023

Reconstructing 3D Scenes from a Single Image



Diffusion with Forward Models: Solving Stochastic Inverse Problems Without Direct Supervision, Tewari et al. 2023

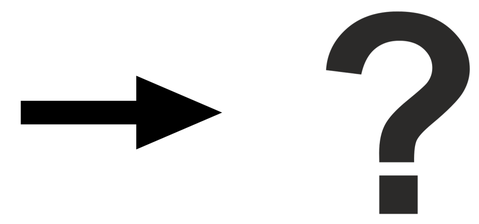
Reconstructing 3D Scenes from a Single Image



?

Diffusion with Forward Models: Solving Stochastic Inverse Problems Without Direct Supervision, Tewari et al. 2023

Reconstructing 3D Scenes from a Single Image

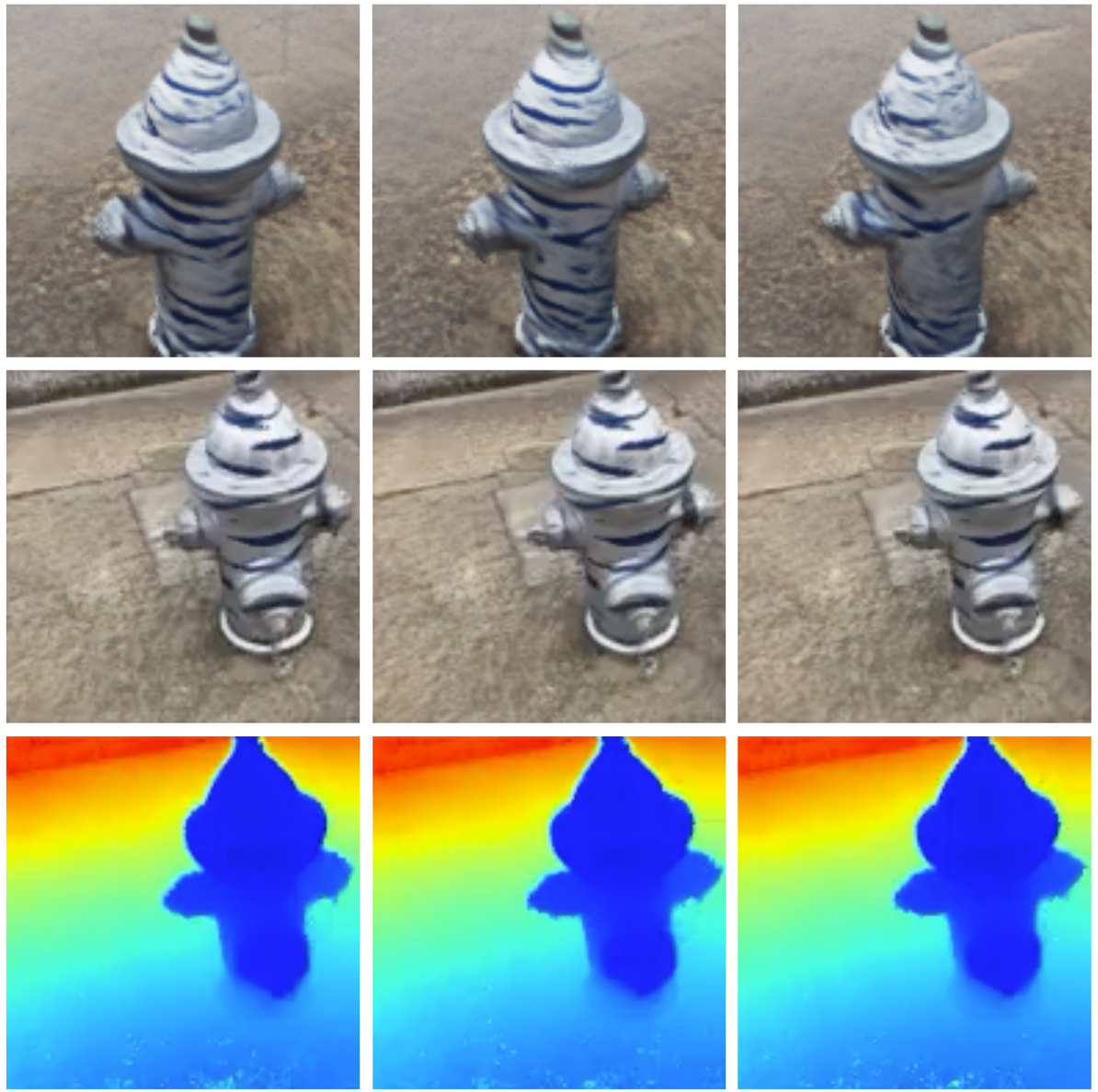


?

Diffusion with Forward Models: Solving Stochastic Inverse Problems Without Direct Supervision, Tewari et al. 2023

Reconstructing 3D Scenes from a Single Image

→ ?



Diffusion with Forward Models: Solving Stochastic Inverse Problems Without Direct Supervision, Tewari et al. 2023

The problem with deterministic novel view synthesis

Input: Single Image

Deterministic Reconstruction

Mean estimate: Averages over all possible reconstructions.
In practice, this means: Useless gradients…

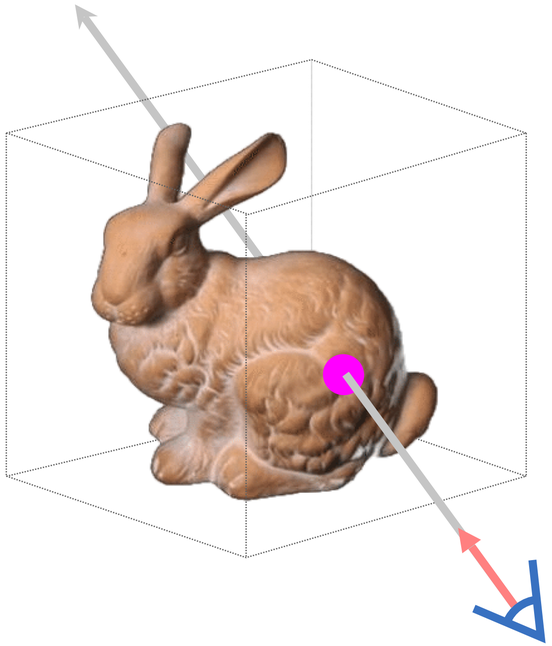
The problem with deterministic novel view synthesis

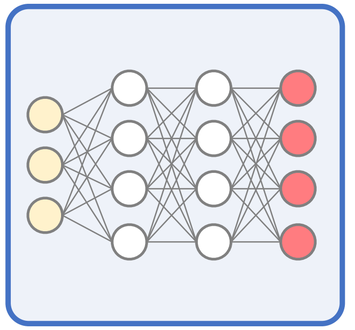
Input: Single Image

Deterministic Reconstruction

Mean estimate: Averages over all possible reconstructions.
In practice, this means: Useless gradients…

General structure of Neural Renderers for 3D-structured Representations

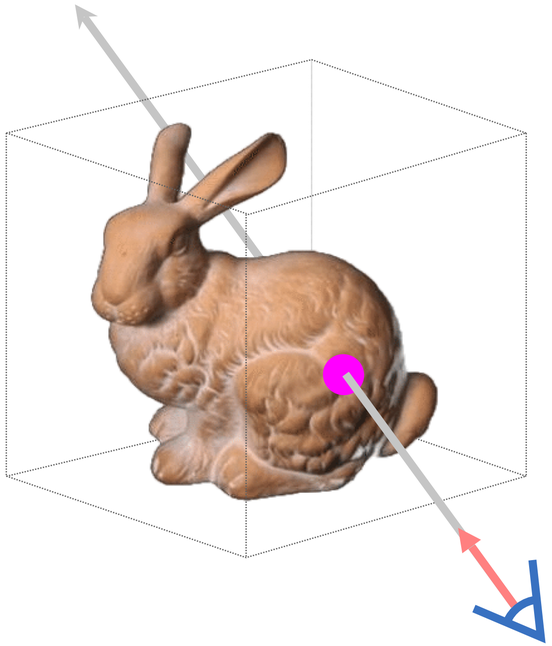


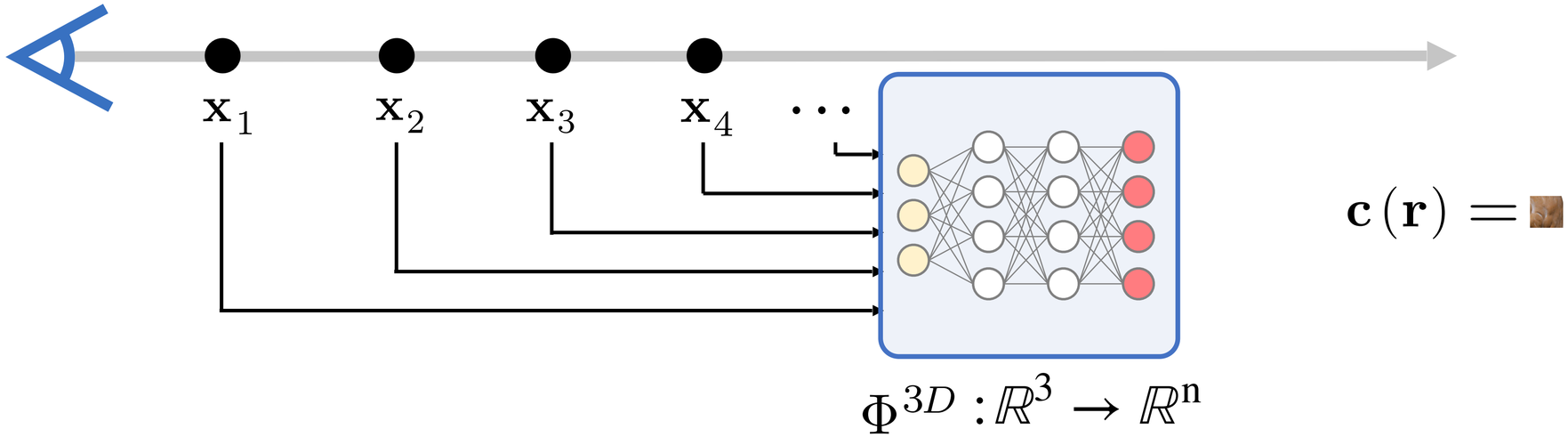


$$\Phi^{3D} : \mathbb{R}^3 \rightarrow \mathbb{R}^n$$

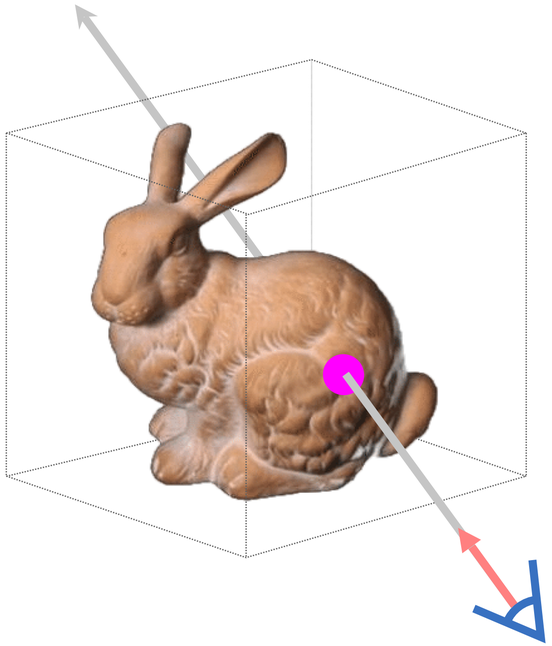
$$\mathbf{c}(\mathbf{r}) =$$ 

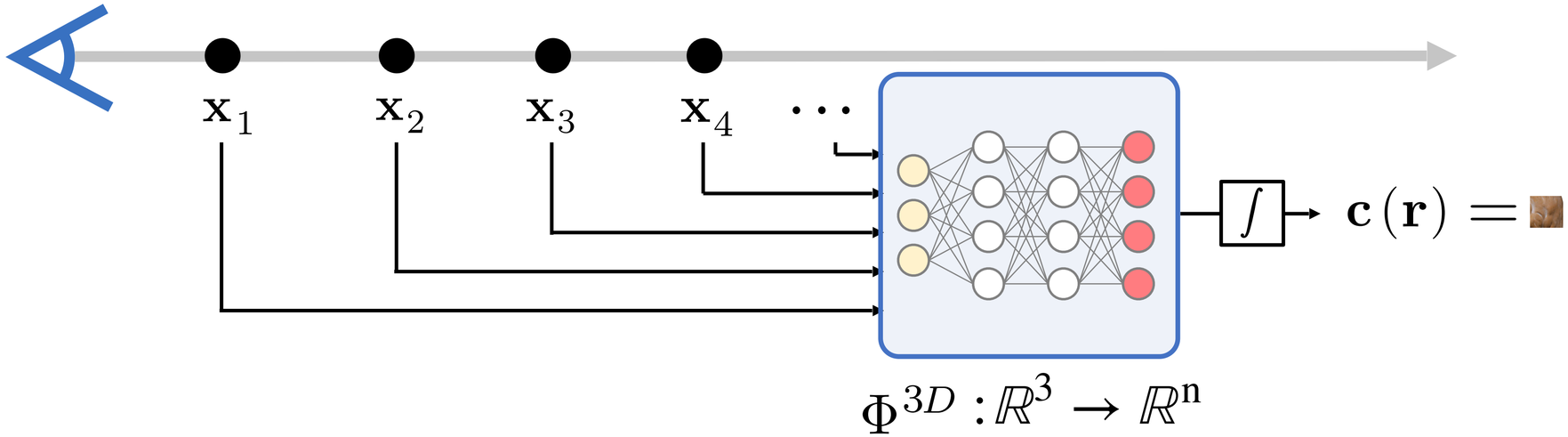
General structure of Neural Renderers for 3D-structured Representations

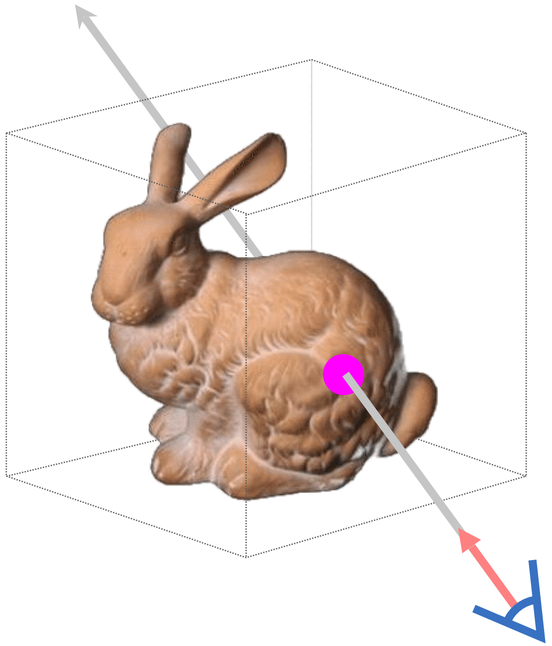




General structure of Neural Renderers for 3D-structured Representations

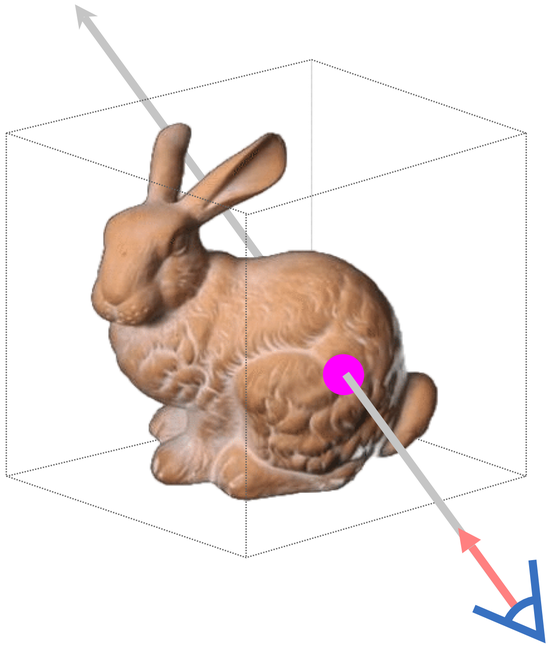






Light Field

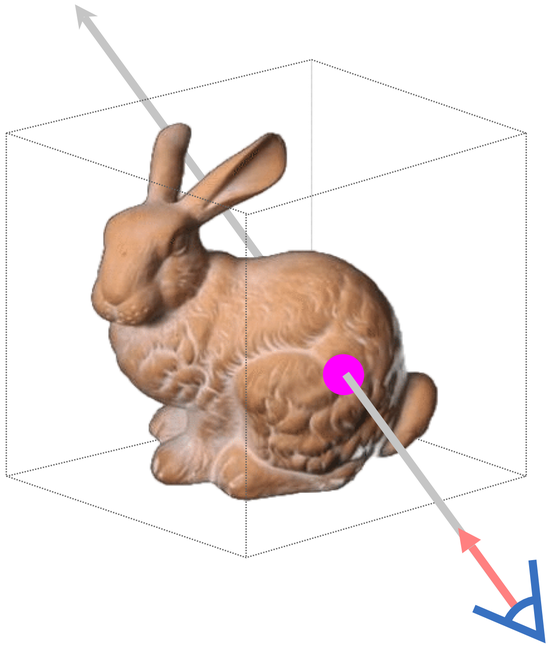
[Adelson et al. 1991, Levoy et al. 1996, Gortler et al. 1996]

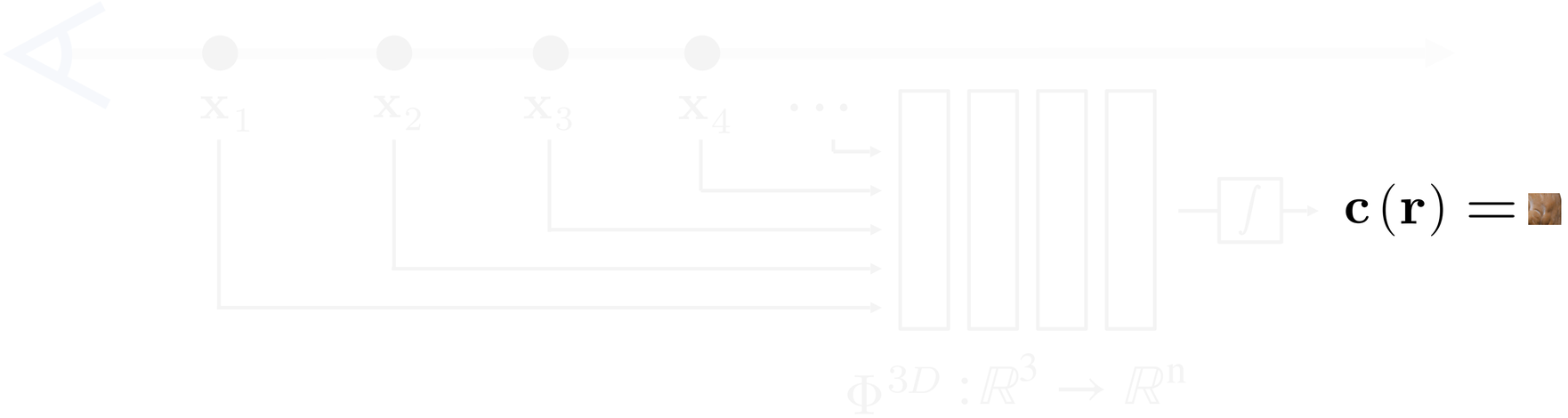


Light Field Networks

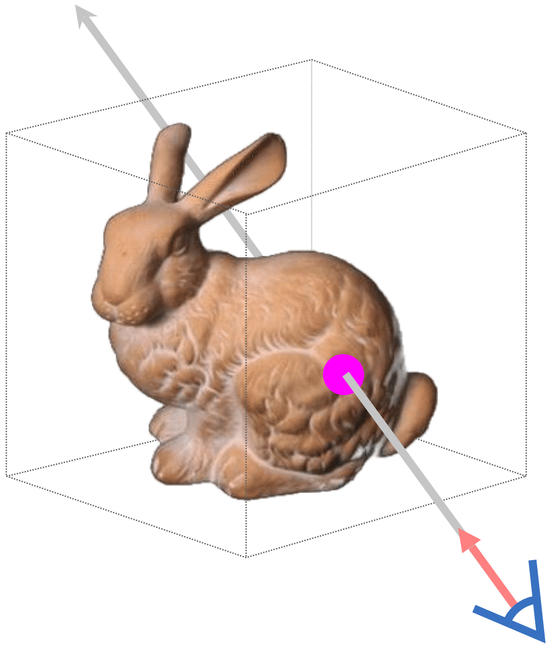
$\mathbf{c}(\mathbf{r}) =$

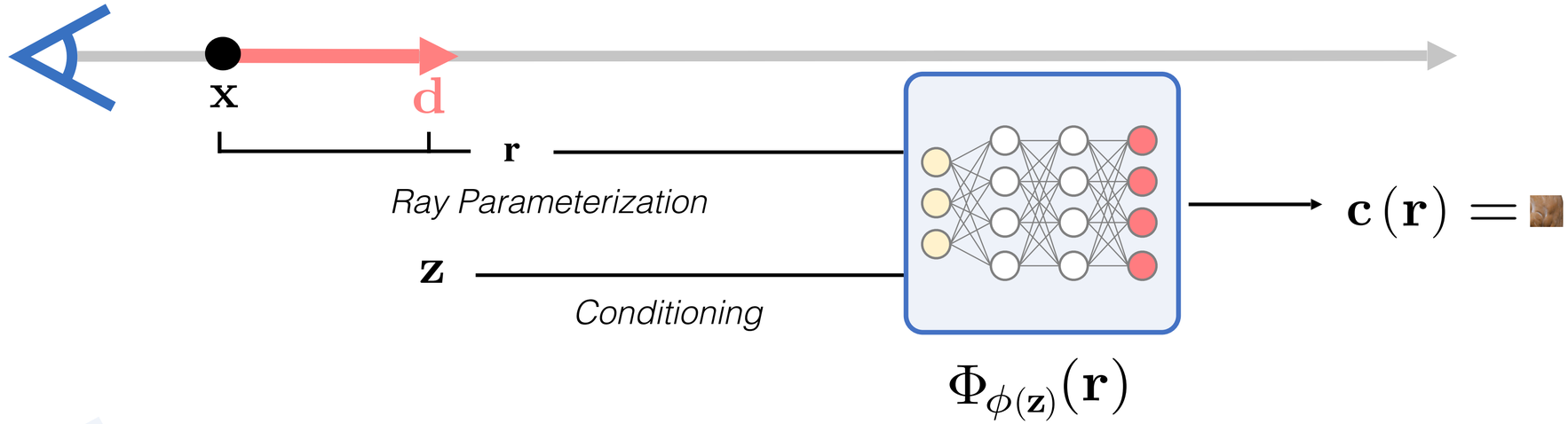
Light Field Networks





Now: *Learn Prior over Light Fields!*





Light Field Networks
500 FPS
1 evaluation per ray

Volumetric Rendering (pixelNeRF)
0.033 FPS
196 evaluations per ray

100x speed

Real-time.
>100x reduction in memory: Can be trained on small GPUs!

Light Field Networks
500 FPS
1 evaluation per ray

Volumetric Rendering (pixelNeRF)
0.033 FPS
196 evaluations per ray

100x speed

Real-time.
>100x reduction in memory: Can be trained on small GPUs!

Background: Diffusion Models

Forward SDE (data → noise)

$\mathbf{x}(0)$ ——— $\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$ ———→ $\mathbf{x}(T)$

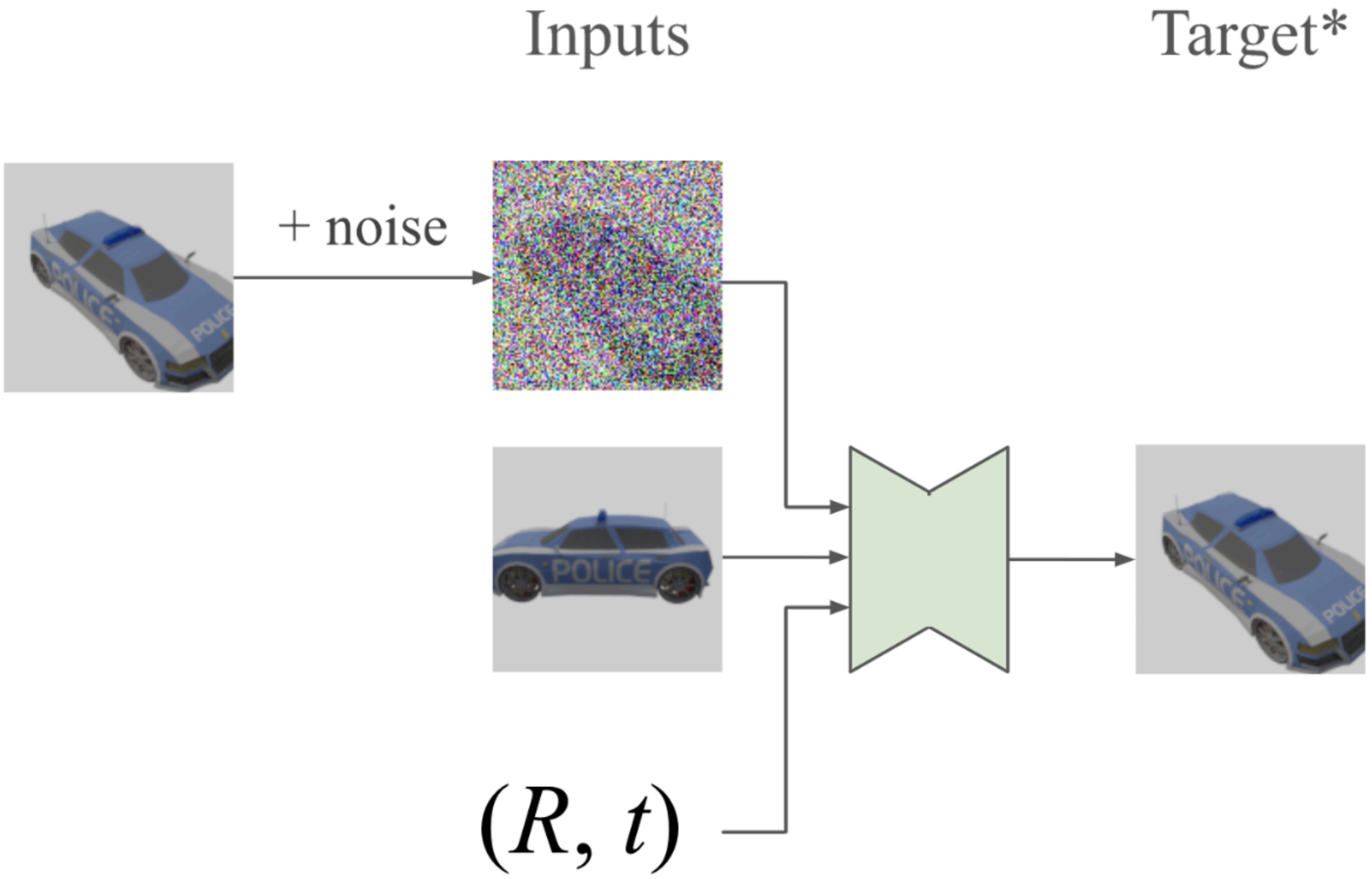
score function

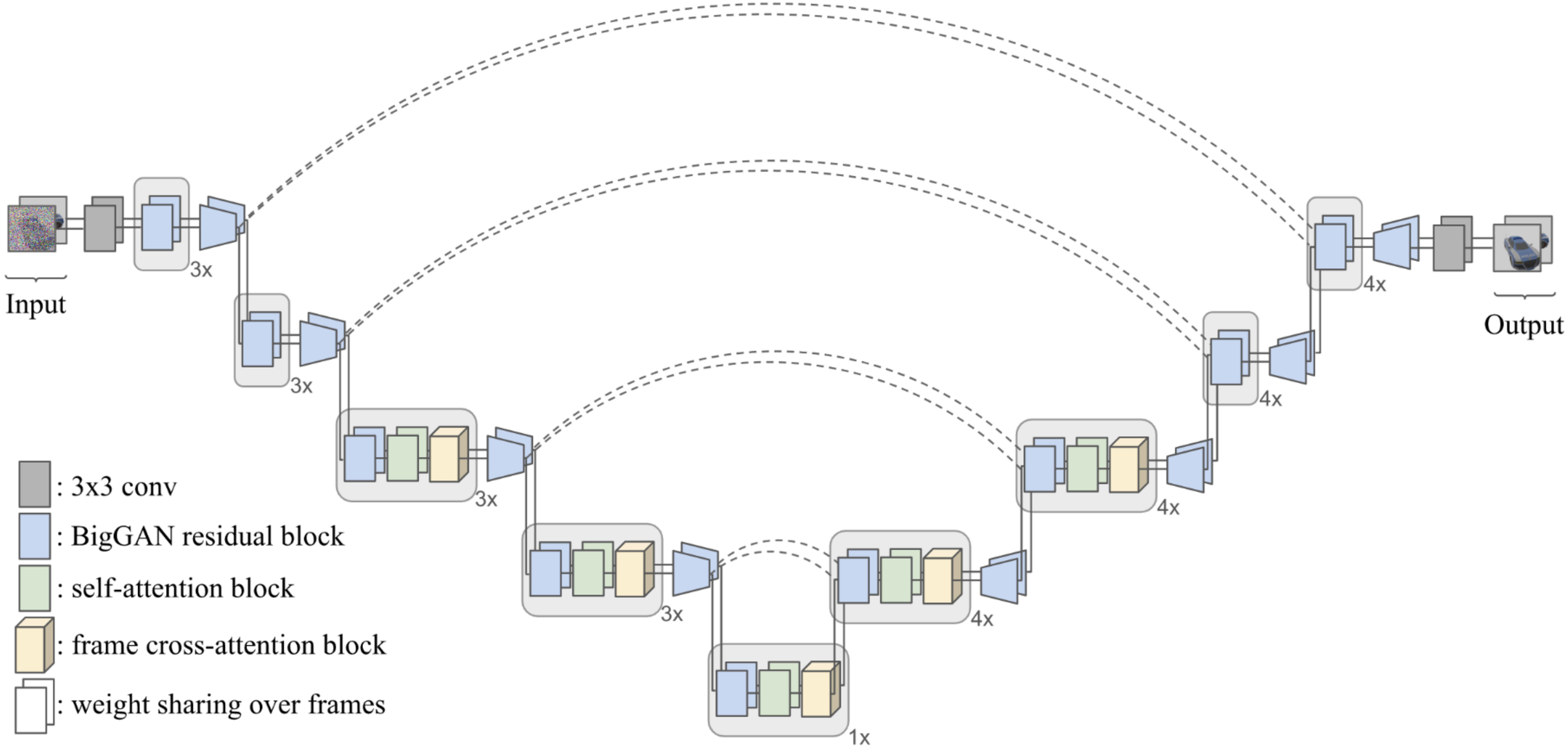
$\mathbf{x}(0)$ ←——— $\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$ ——— $\mathbf{x}(T)$

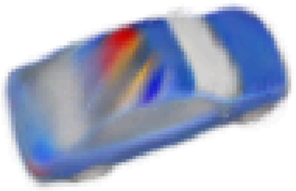
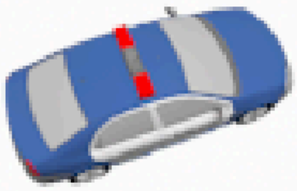
Reverse SDE (noise → data)

NOVEL VIEW SYNTHESIS WITH DIFFUSION MODELS

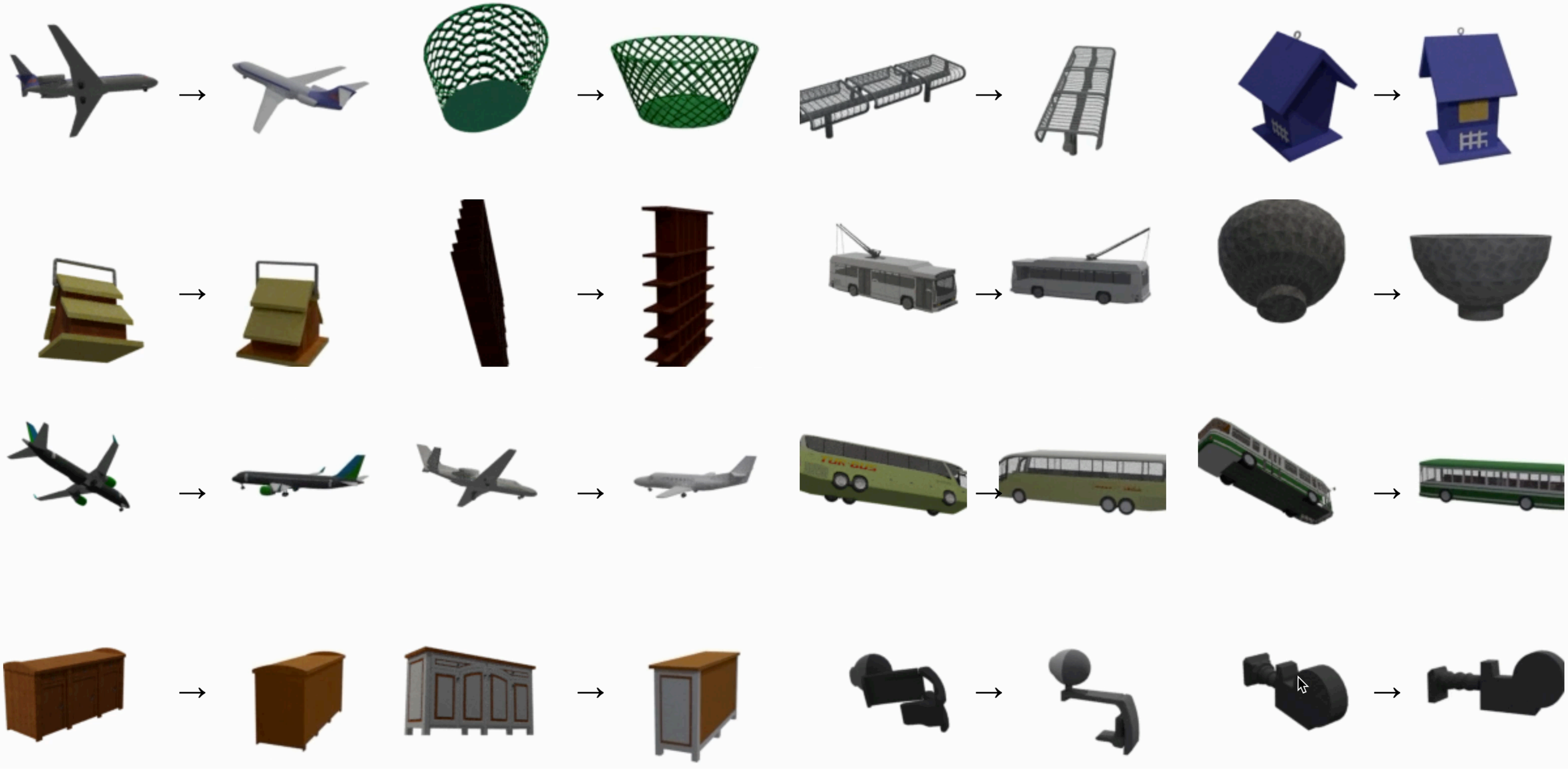
Daniel Watson **William Chan** **Ricardo Martin-Brualla**
Jonathan Ho **Andrea Tagliasacchi** **Mohammad Norouzi**

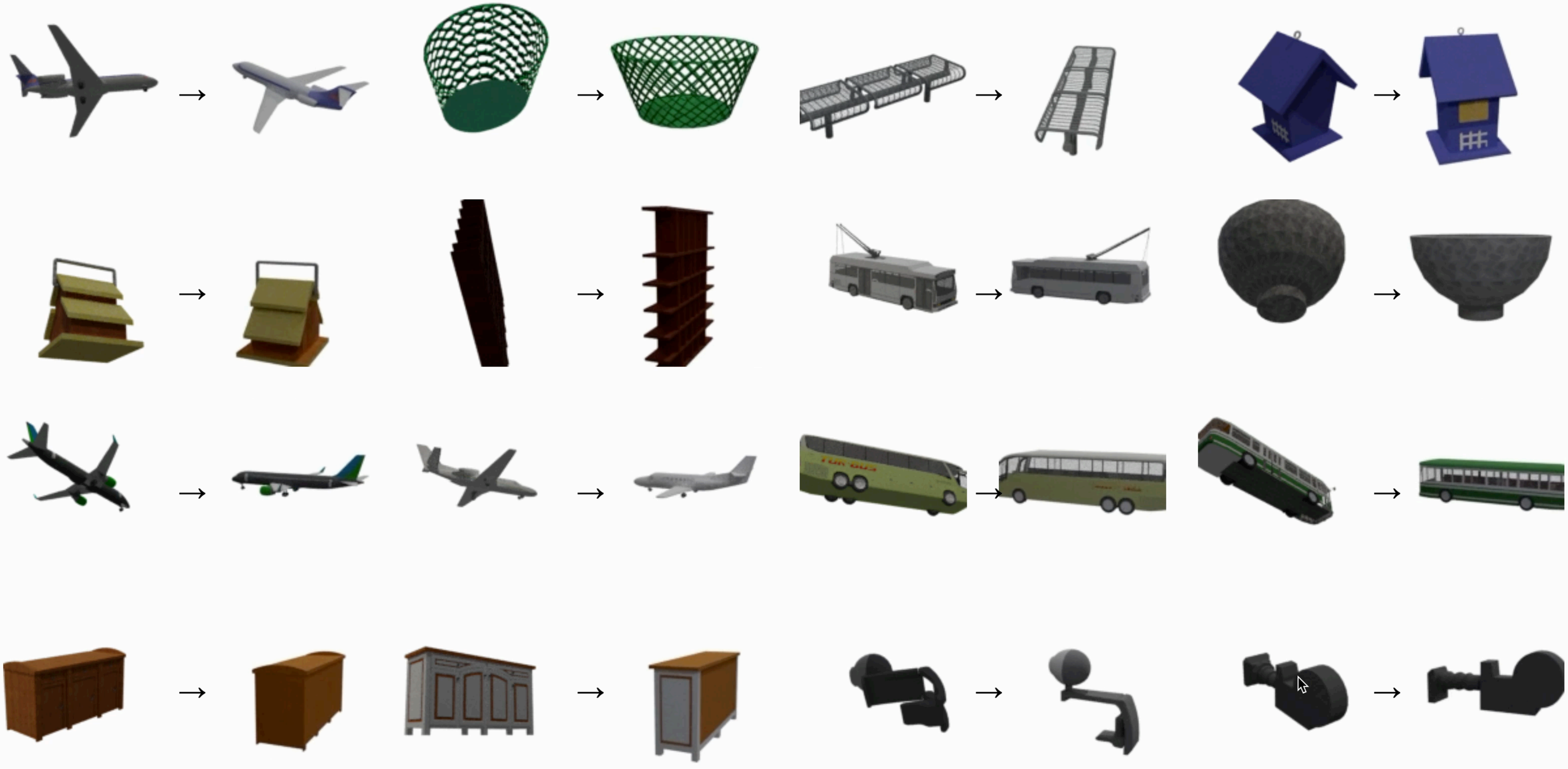



Input
3x
3x
3x
3x
4x
4x
4x
1x
Output
: 3x3 conv
: BigGAN residual block
: self-attention block
: frame cross-attention block
: weight sharing over frames

Input View	SRN	PixelNeRF	VisionNeRF	**3DiM (ours)**	Ground Truth
					
					

	SRN cars			SRN chairs		
	PSNR (↑)	SSIM (↑)	FID (↓)	PSNR (↑)	SSIM (↑)	FID (↓)
Geometry-aware						
SRN	22.25	0.88	41.21	22.89	0.89	26.51
PixelNeRF	23.17	0.89	59.24	23.72	0.90	38.49
VisionNeRF	22.88	0.90	21.31	**24.48**	**0.92**	10.05
CodeNeRF	**23.80**	***0.91**	–	23.66	*0.90	–
Geometry-free						
LFN	22.42	*0.89	–	22.26	*0.90	–
ENR	22.26	–	–	22.83	–	–
3DiM (ours)	21.01	0.57	**8.99**	17.05	0.53	**6.57**





We already have great text-conditional image generative models…

Zero-1-to-3: Zero-shot One Image to 3D Object

Ruoshi Liu
Columbia University

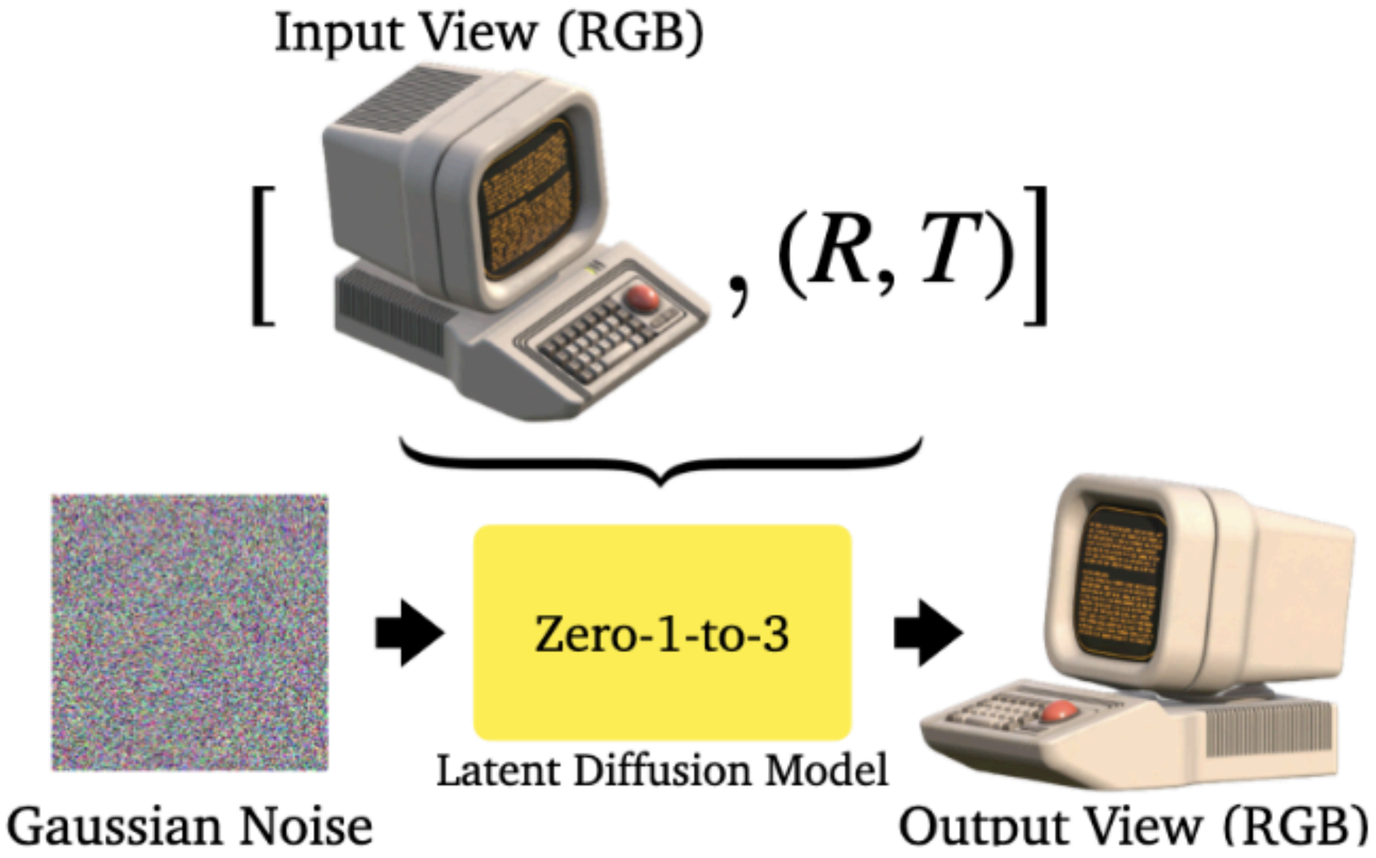
Rundi Wu
Columbia University

Basile Van Hoorick
Columbia University

Pavel Tokmakov
Toyota Research Institute

Sergey Zakharov
Toyota Research Institute

Carl Vondrick
Columbia University



- Core novelties: Pre-trained image generative model (StableDiffusion)
- Train on larger dataset, Objaverse (800k 3D models)

Zero-1-to-3: Zero-shot One Image to 3D Object

Ruoshi Liu
Columbia University

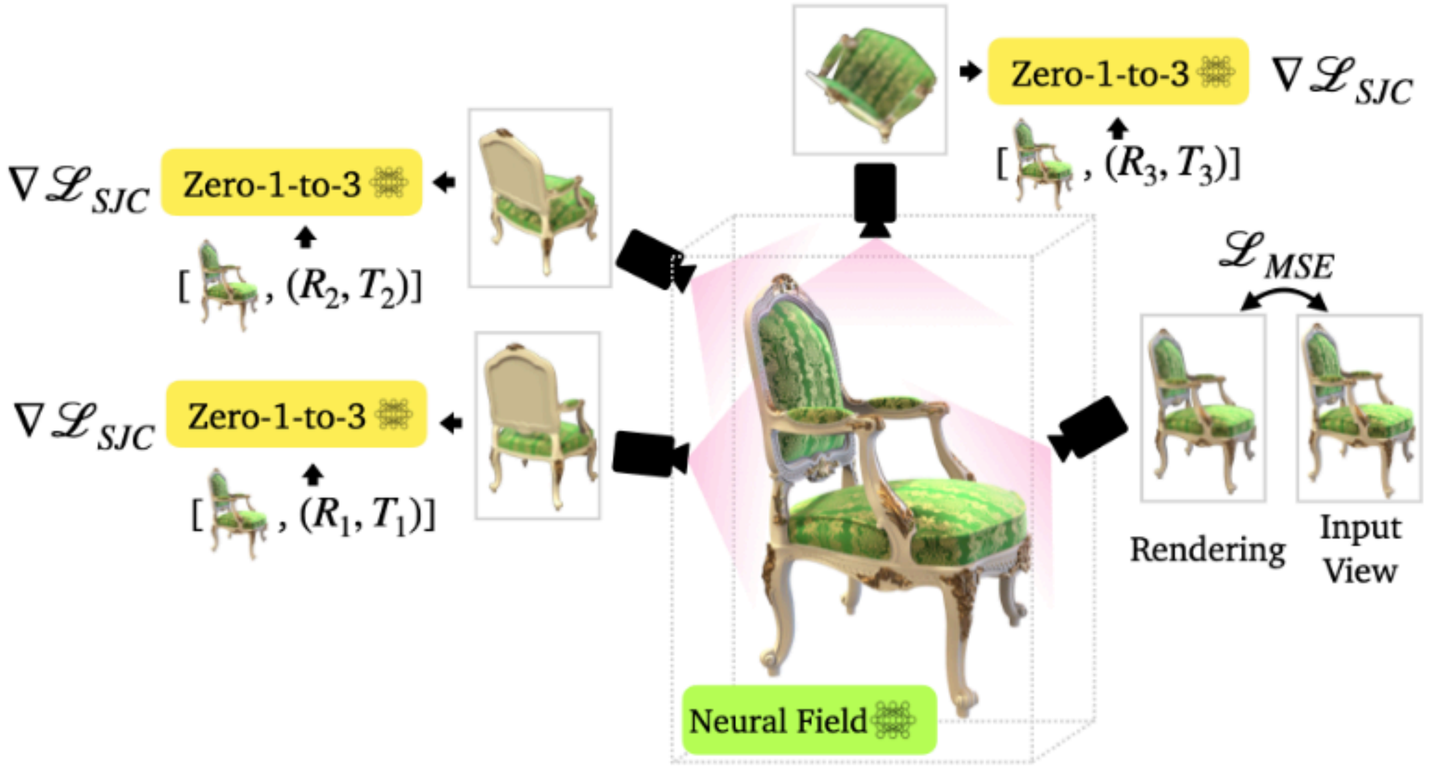
Rundi Wu
Columbia University

Basile Van Hoorick
Columbia University

Pavel Tokmakov
Toyota Research Institute

Sergey Zakharov
Toyota Research Institute

Carl Vondrick
Columbia University



- To get actual 3D representation: Score Distillation

Zero-1-to-3: Zero-shot One Image to 3D Object

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

Zero-1-to-3: Zero-shot One Image to 3D Object

Ruoshi Liu
Columbia University

Rundi Wu
Columbia University

Basile Van Hoorick
Columbia University

Pavel Tokmakov
Toyota Research Institute

Sergey Zakharov
Toyota Research Institute

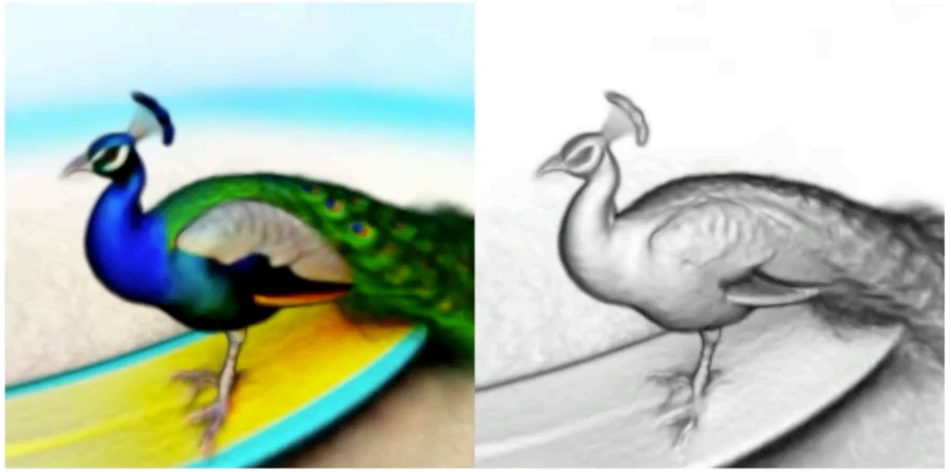
Carl Vondrick
Columbia University

Score Distillation Sampling

Poole et al. 2023

"a DSLR photo of a peacock on a surfboard"

DreamFusion
Automatic text-to-3D

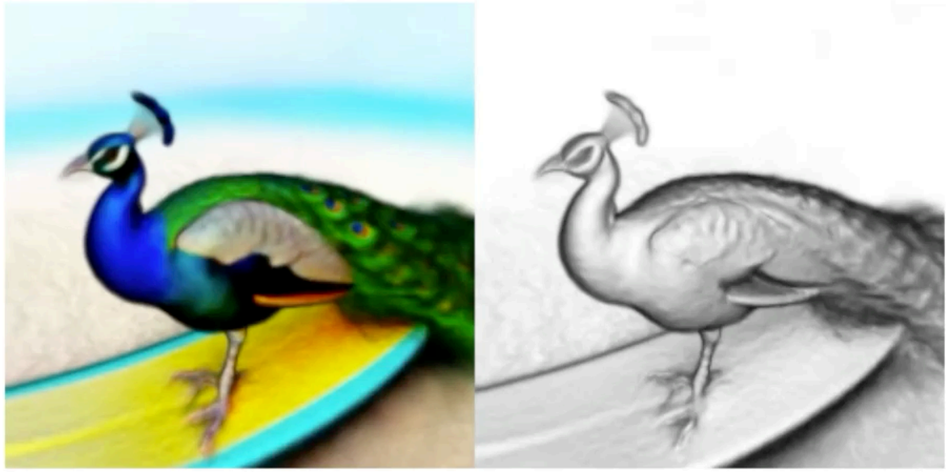


Score Distillation Sampling

Poole et al. 2023

"a DSLR photo of a peacock on a surfboard"

DreamFusion
Automatic text-to-3D



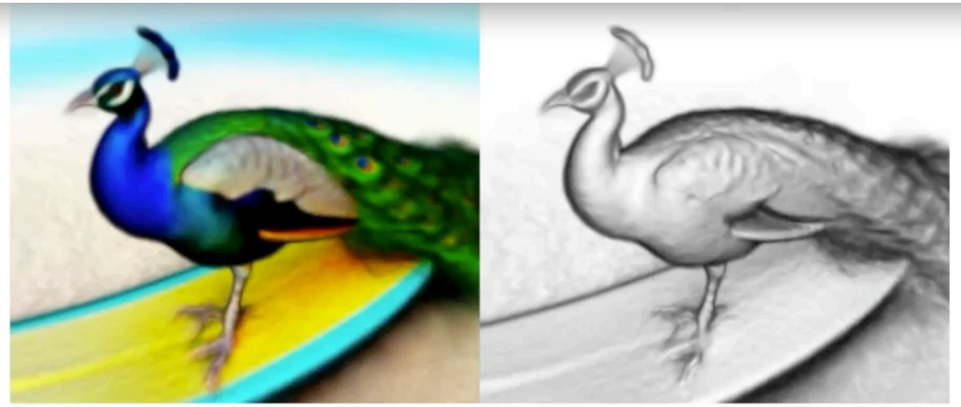
Score Distillation Sampling

Poole et al. 2023

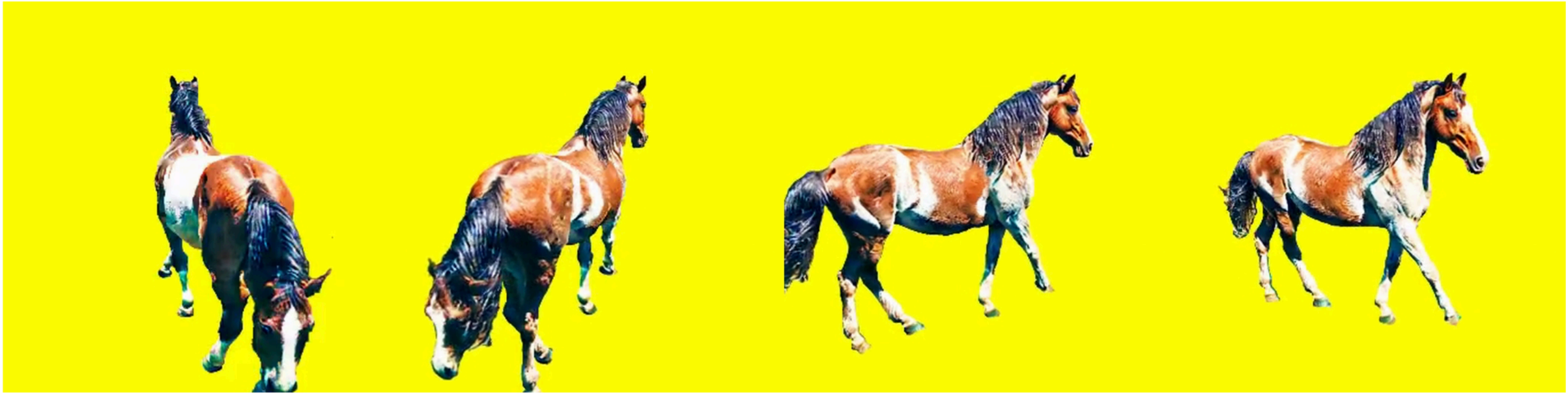
"a DSLR photo of a peacock on a surfboard"

DreamFusion

Diffusion is in 2D
Text-conditional
This CVPR: 3D Diffusion Models for 3D Scenes!

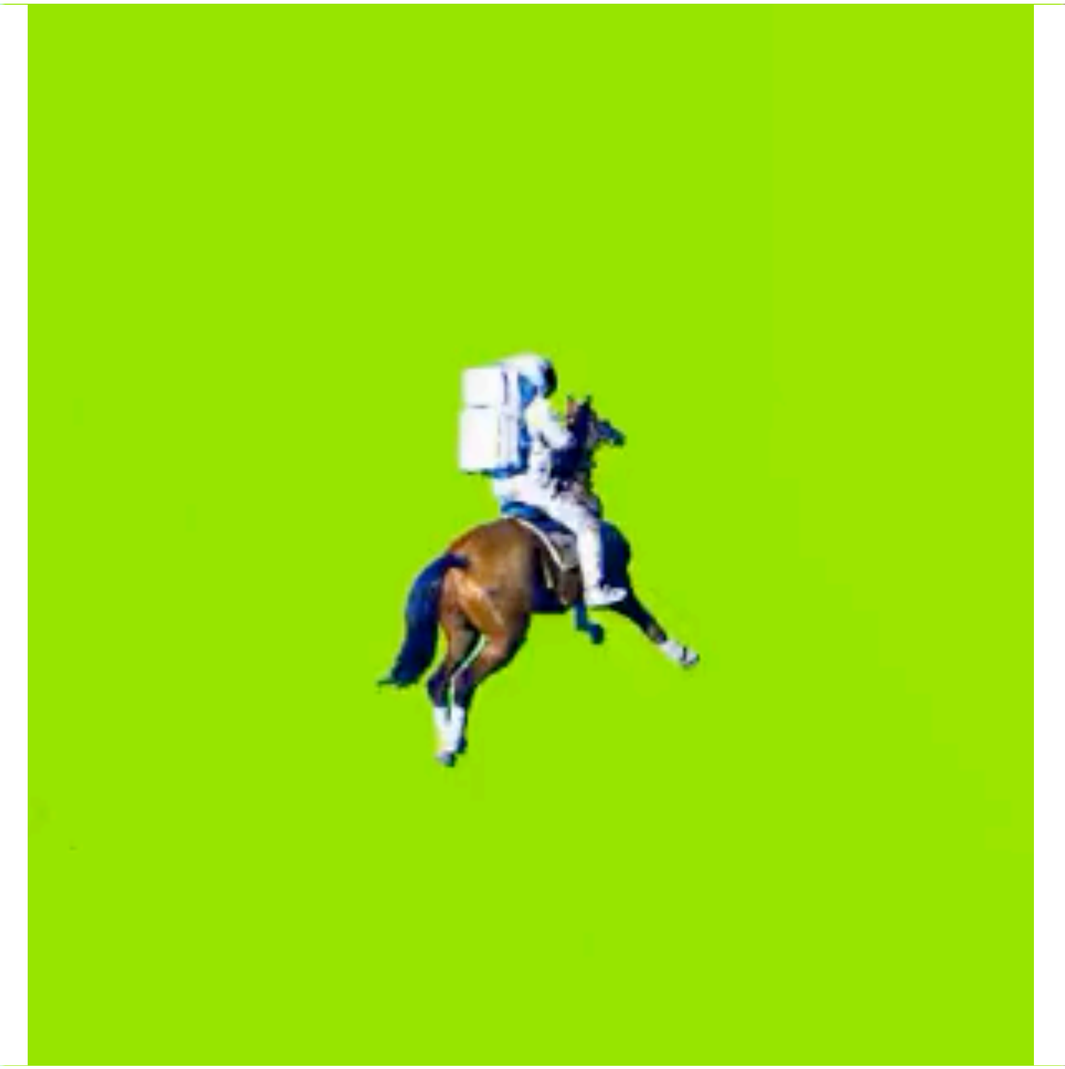


The “Janus Problem”

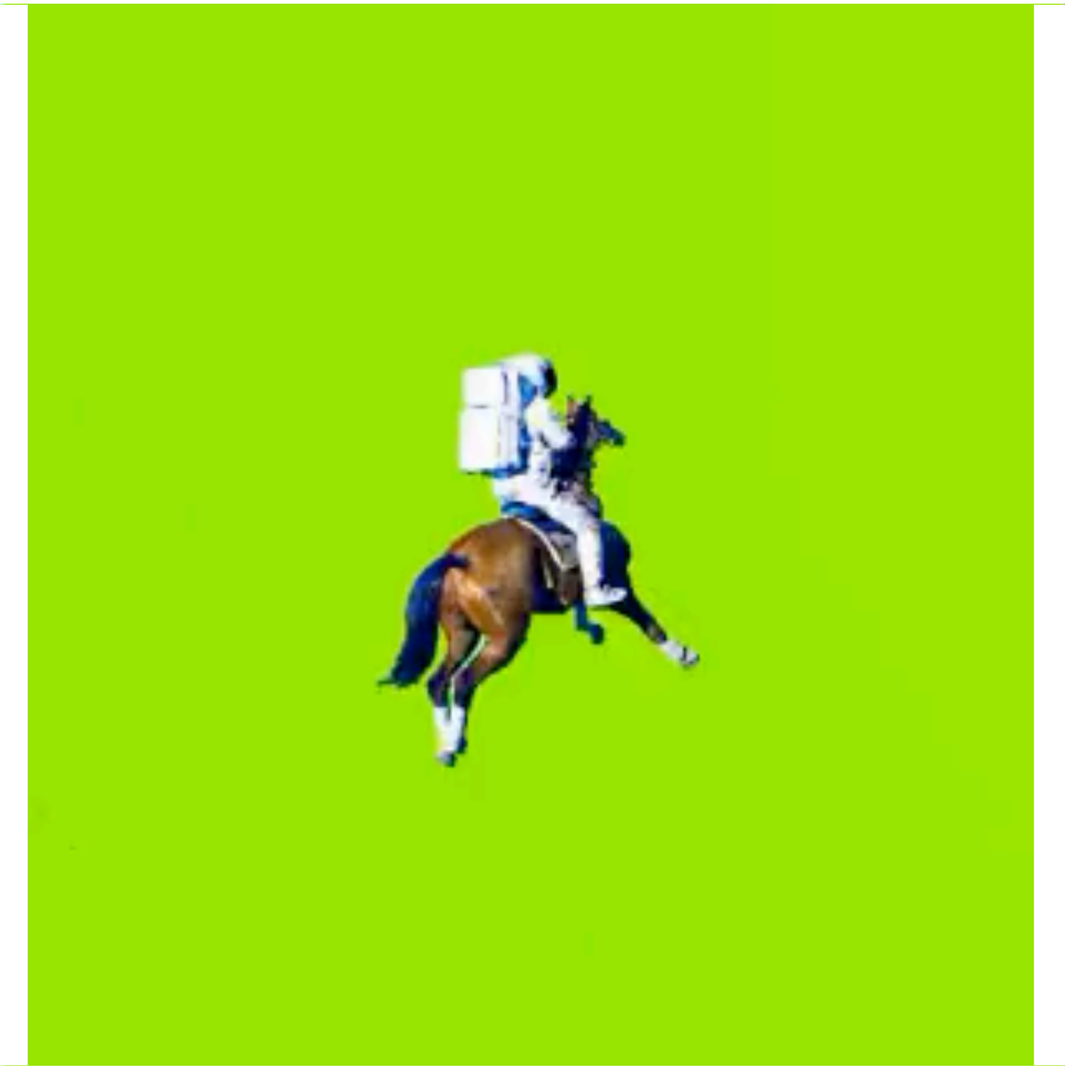


Multi-face Janus Problem

The “Janus Problem”



The “Janus Problem”

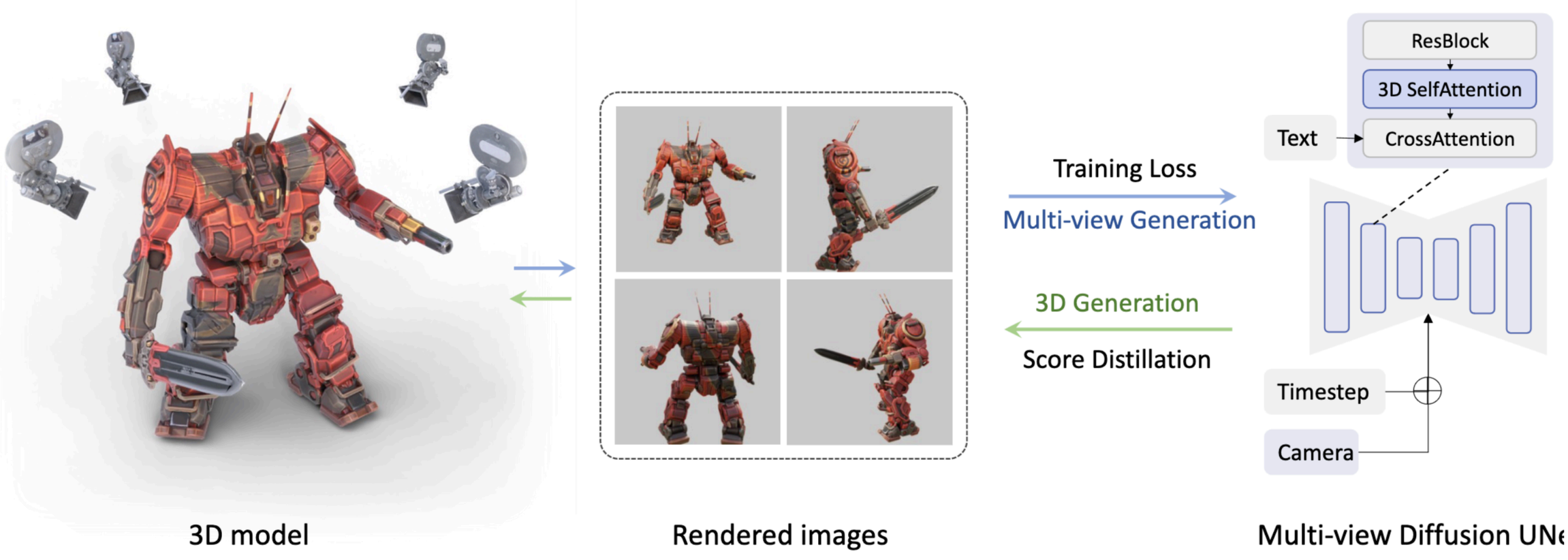


The “Janus Problem”

The “Janus Problem”

MVDream: Multi-View Diffusion for 3D Generation

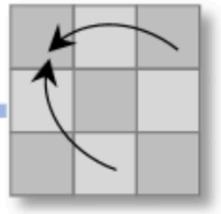
Shi et al. 2023



Multi-view Diffusion Training

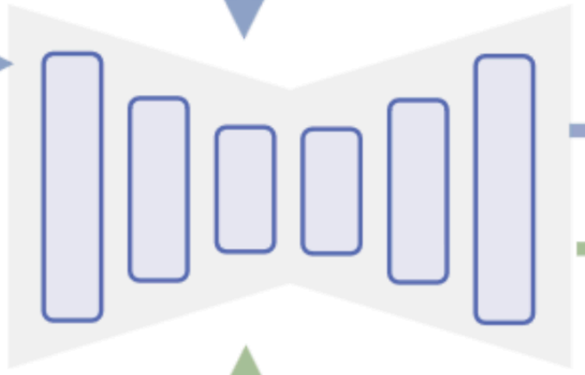
French cat

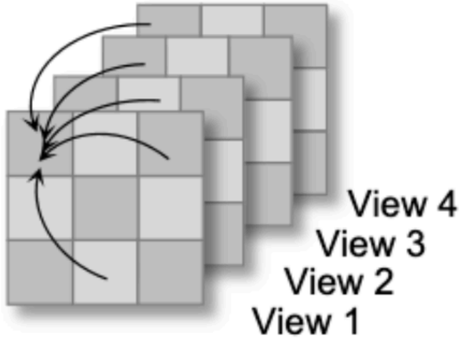
Robot



Image

Image
Diffusion Loss





View 4
View 3
View 2
View 1

camera

Multi-view
Diffusion Loss

MultiView Diffusion Largely Solves the Janus Problem

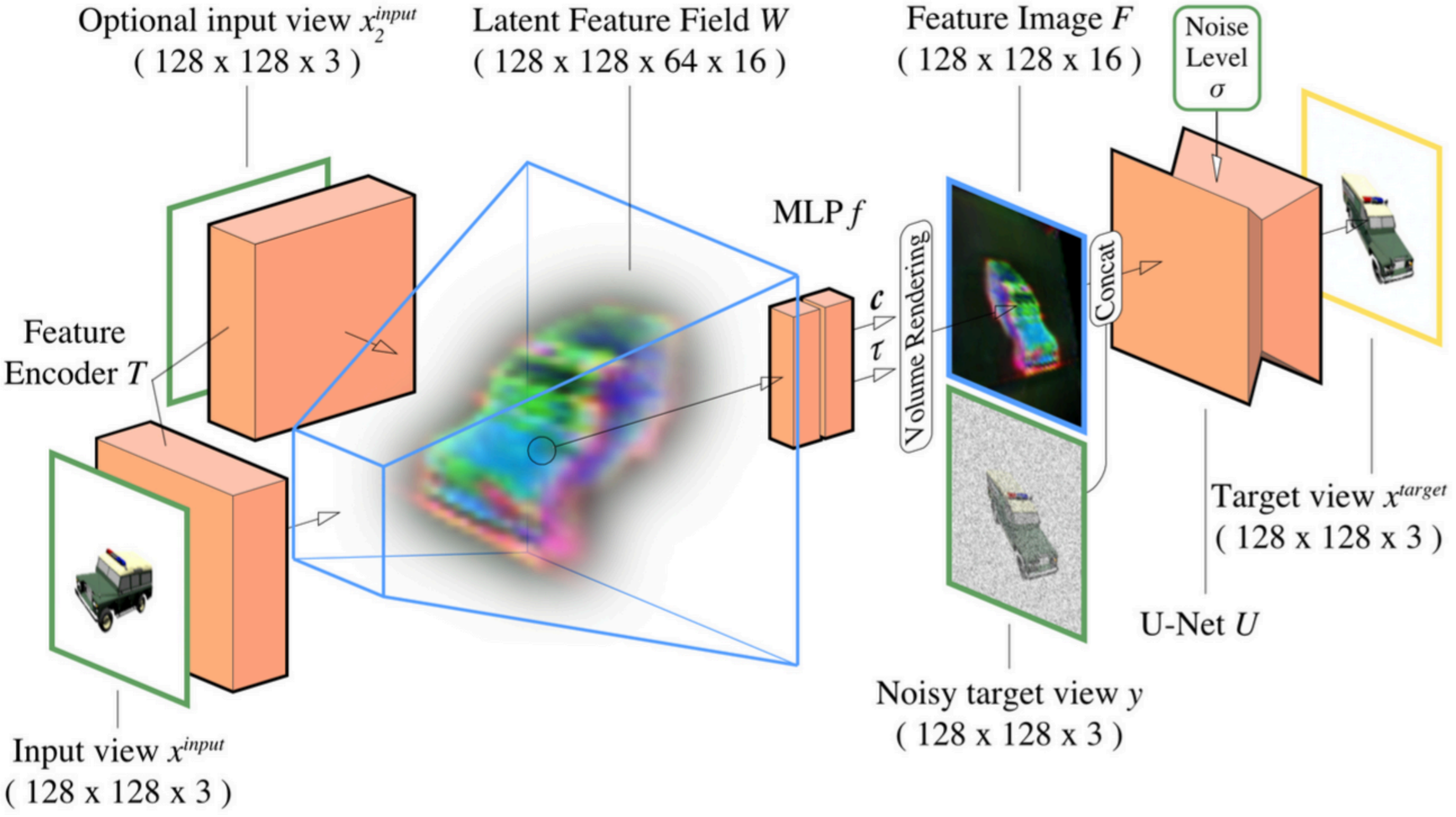
MultiView Diffusion Largely Solves the Janus Problem

`

`

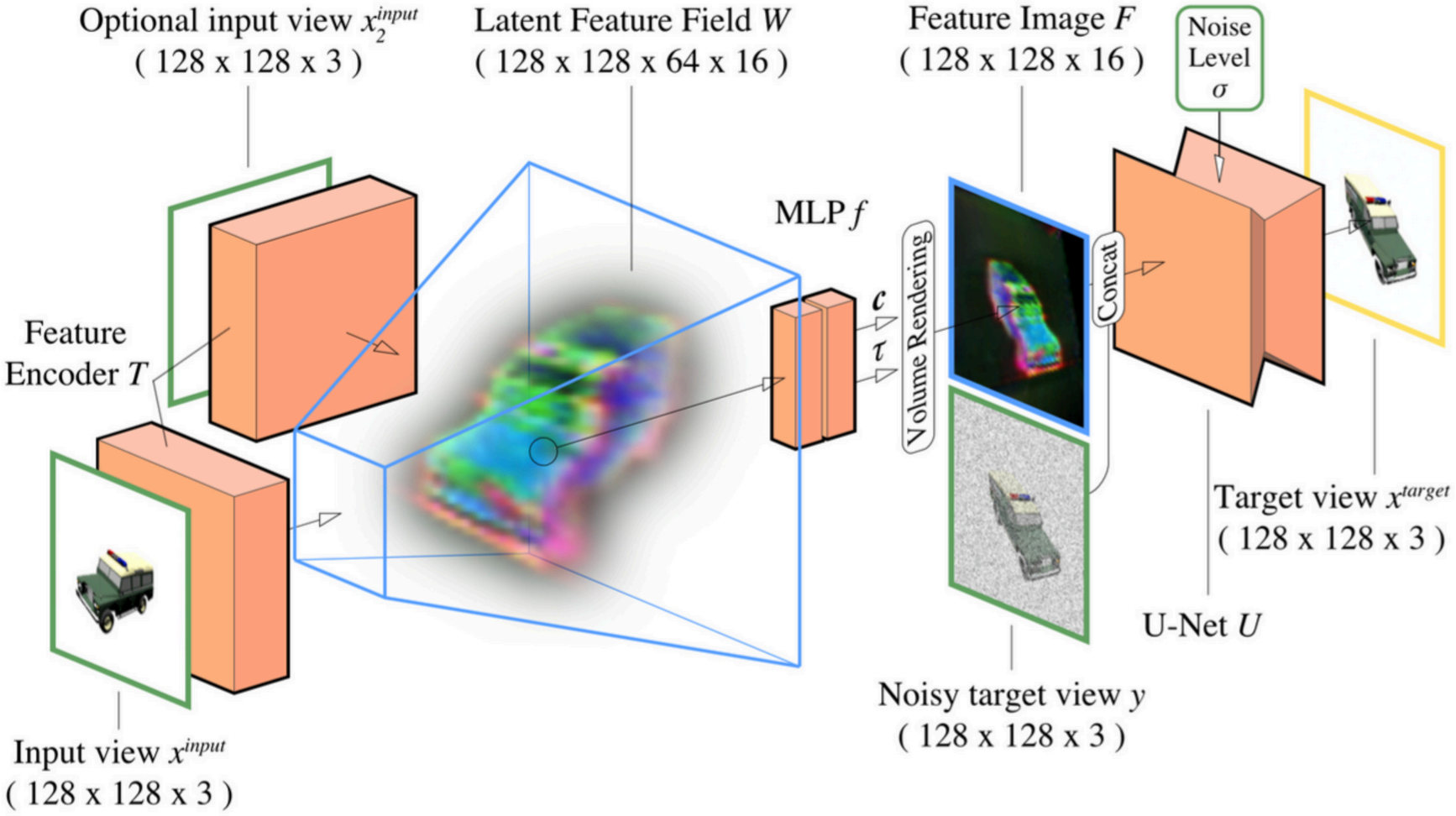
GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models

Eric R. Chan [*1,2] Koki Nagano [*2] Matthew A. Chan [*2] Alexander W. Bergman [*1] Jeong Joon Park [*1]
Axel Levy [1] Miika Aittala [2] Shalini De Mello [2] Tero Karras [2] Gordon Wetzstein [1]



GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models

Eric R. Chan [*1,2] Koki Nagano [*2] Matthew A. Chan [*2] Alexander W. Bergman [*1] Jeong Joon Park [*1]
Axel Levy [1] Miika Aittala [2] Shalini De Mello [2] Tero Karras [2] Gordon Wetzstein [1]



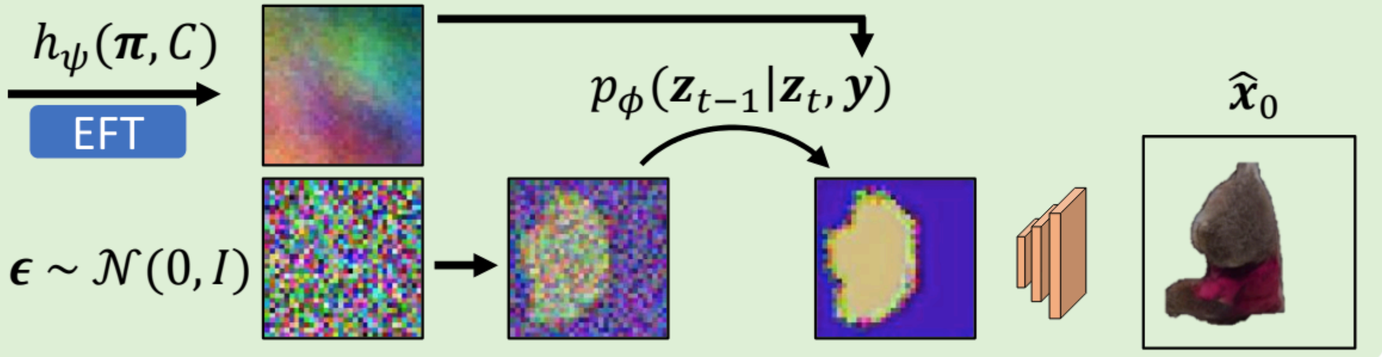
- Allows to condition on 3D model
- But generative model is still an *image* generative model.

SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction

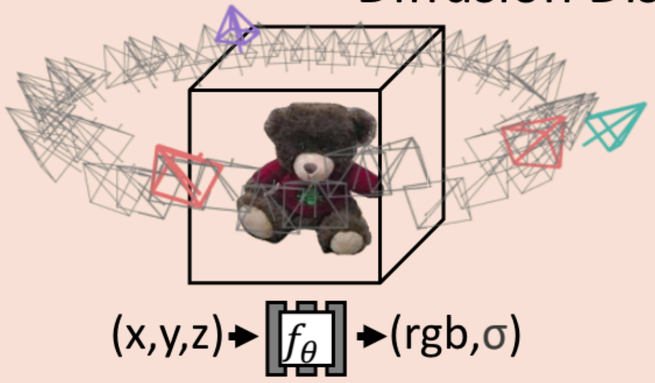
Zhizhuo Zhou Shubham Tulsiani

Input Views

View-conditioned Latent Diffusion VLDM



Diffusion Distillation



$$\min_\theta \mathbb{E}_\pi[-\log p_\phi(f_\theta(\pi)|y)]$$

But: These are all *2D* Generative Models:
The Diffusion model does not generate a 3D Scene!

Diffusion with Forward Models:
Solving Stochastic Inverse Problems Without Direct Supervision

Input: Single Image

Deterministic Reconstruction

Diffusion with Forward Models:
Solving Stochastic Inverse Problems Without Direct Supervision

Input: Single Image

Deterministic Reconstruction

Diffusion with Forward Models:
Solving Stochastic Inverse Problems Without Direct Supervision

Input: Single Image

Deterministic Reconstruction

Ours

Diffusion with Forward Models:
Solving Stochastic Inverse Problems Without Direct Supervision

Input: Single Image

Deterministic Reconstruction

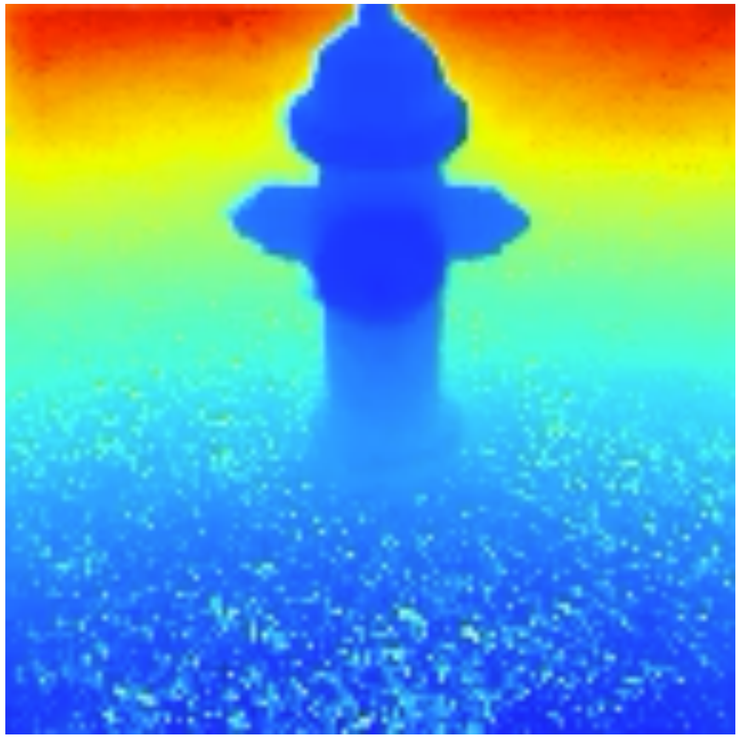
Ours

This is a **conditional generative model** that learns to **directly sample 3D scenes, trained only on images.**

Diffusion with Forward Models:
Solving Stochastic Inverse Problems Without Direct Supervision

Input: Single Image

Depth



Ours

This is a **conditional generative model** that learns to
directly sample 3D scenes, trained only on images.

Single Input Image

Our 3D
Generative Model

Sampled 3D Scene 1

Sampled 3D Scene 2

Sampled 3D Scene 3

Sampled 3D Scene 4

Single Input Image

Our 3D
Generative Model

Sampled 3D Scene 1

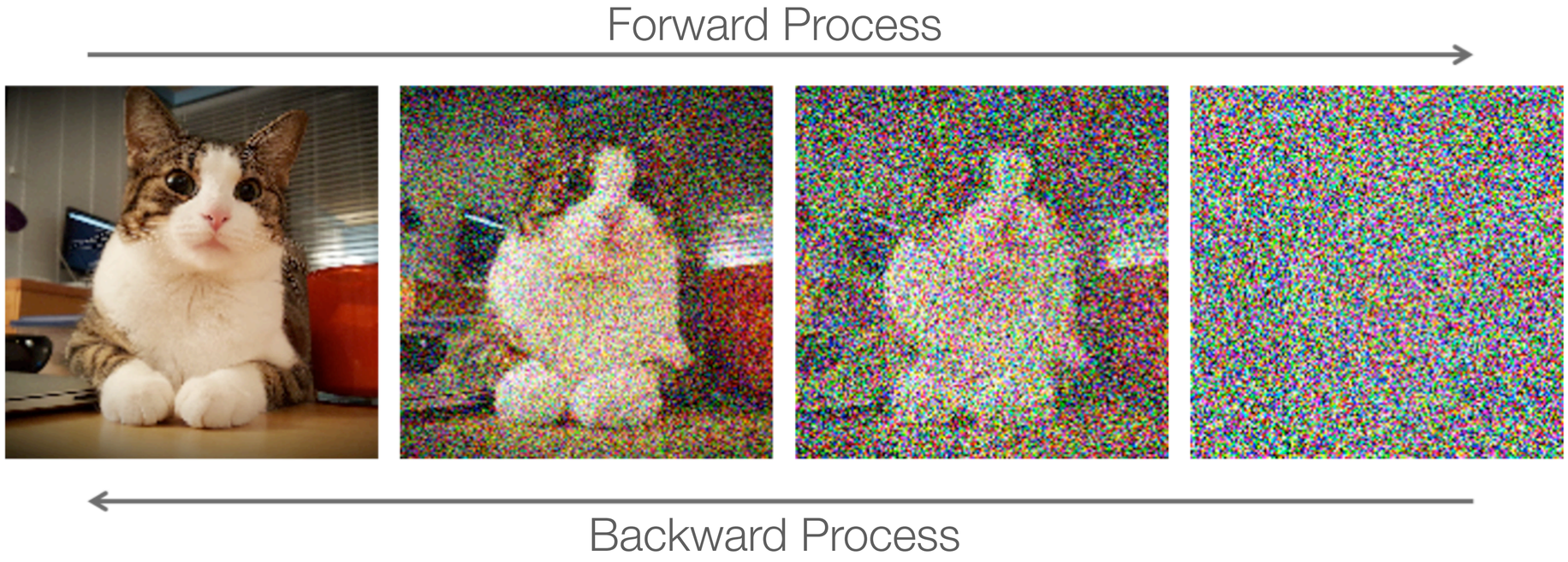
Sampled 3D Scene 2

Sampled 3D Scene 3

Sampled 3D Scene 4

Conventional diffusion models:

We always have access to the distribution that we want to sample from.



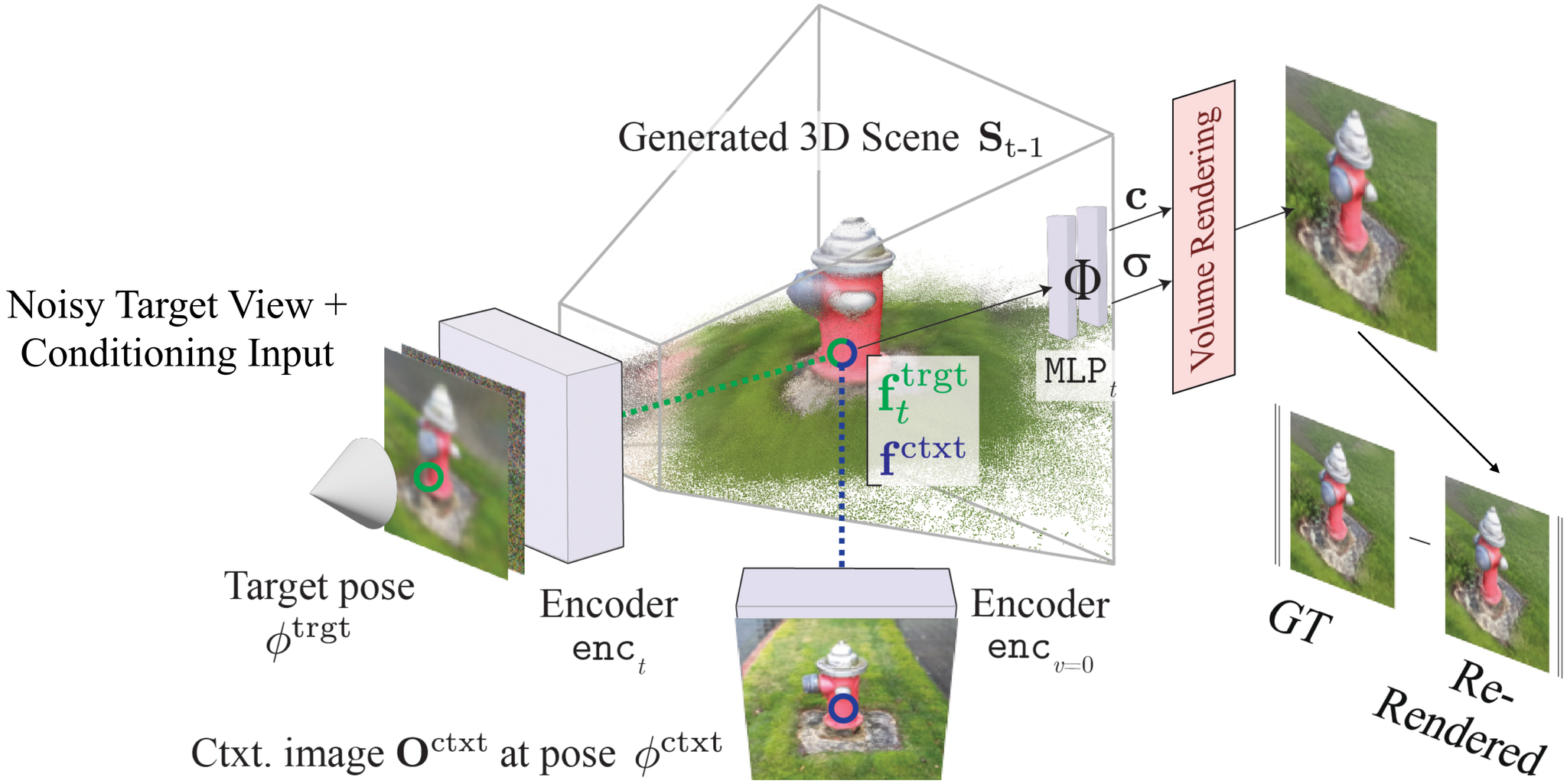
We don’t have 3D Scenes, just images!!

We don’t have 3D Scenes, just images!!

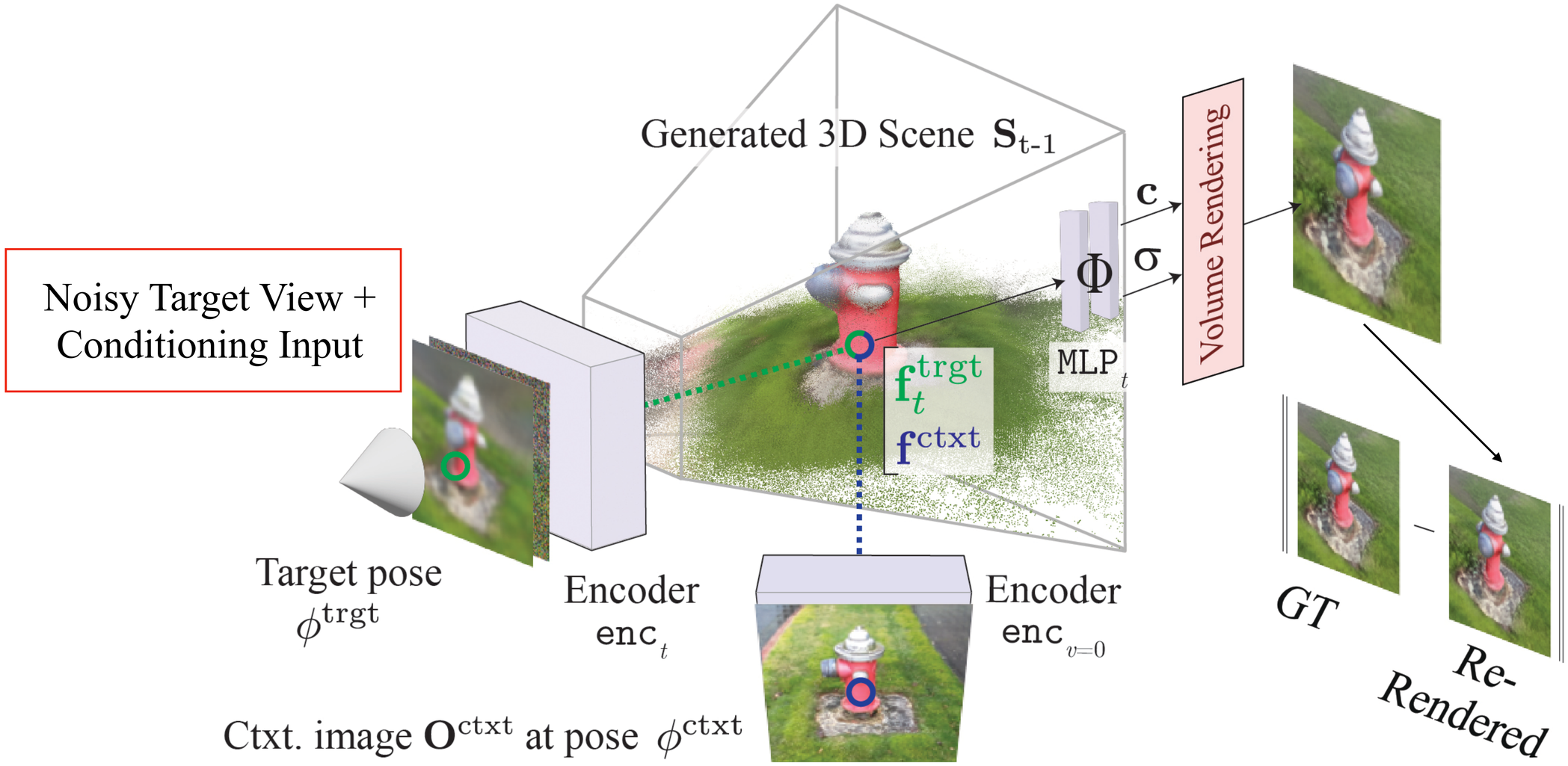
Unobserved: *Cannot* Train neural network to denoise…

Our model: Train on Images, learn to generate 3D Scenes!

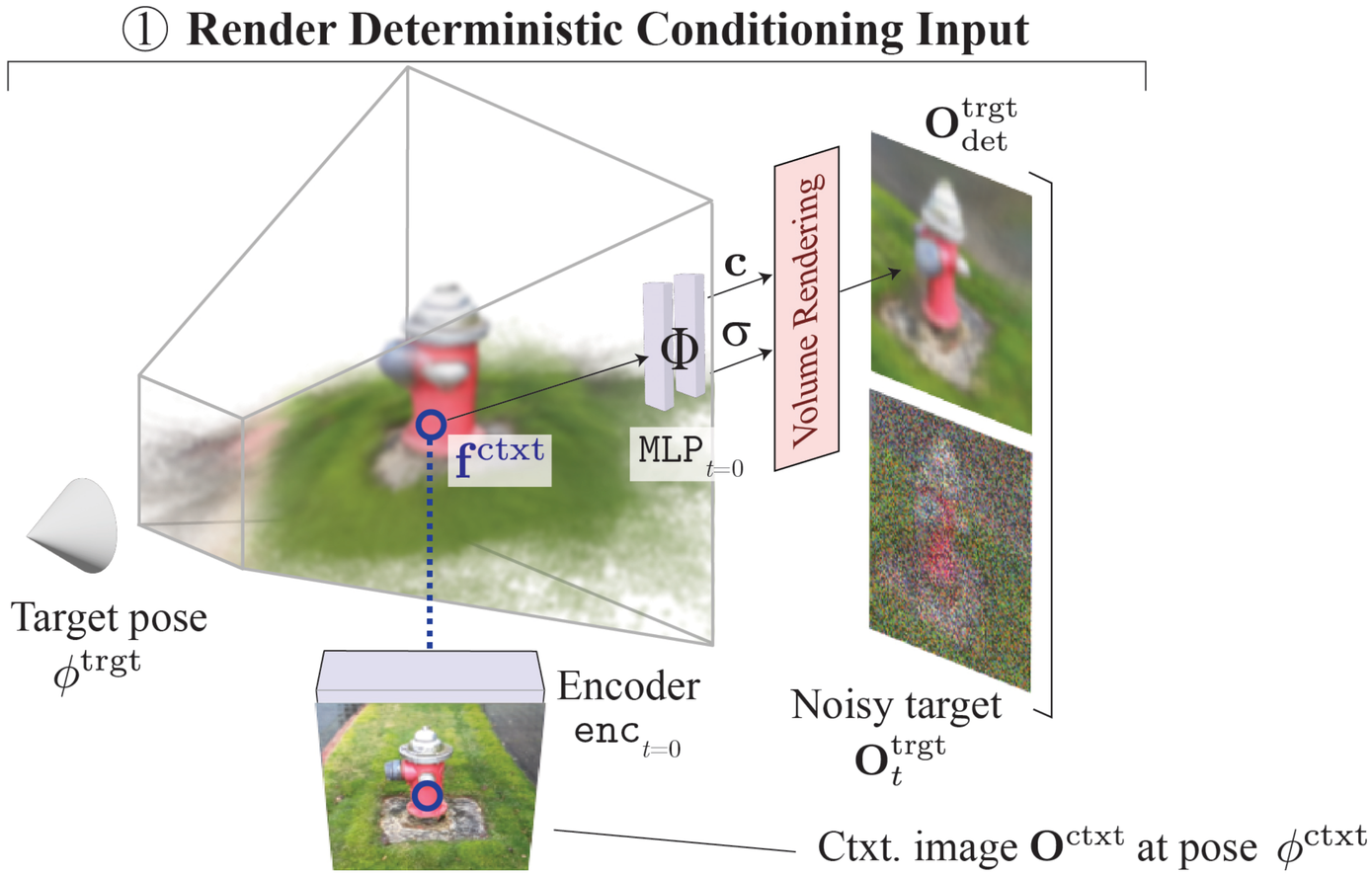
Key Idea:
Denoise Target View, but make rendering part of denoising operator!



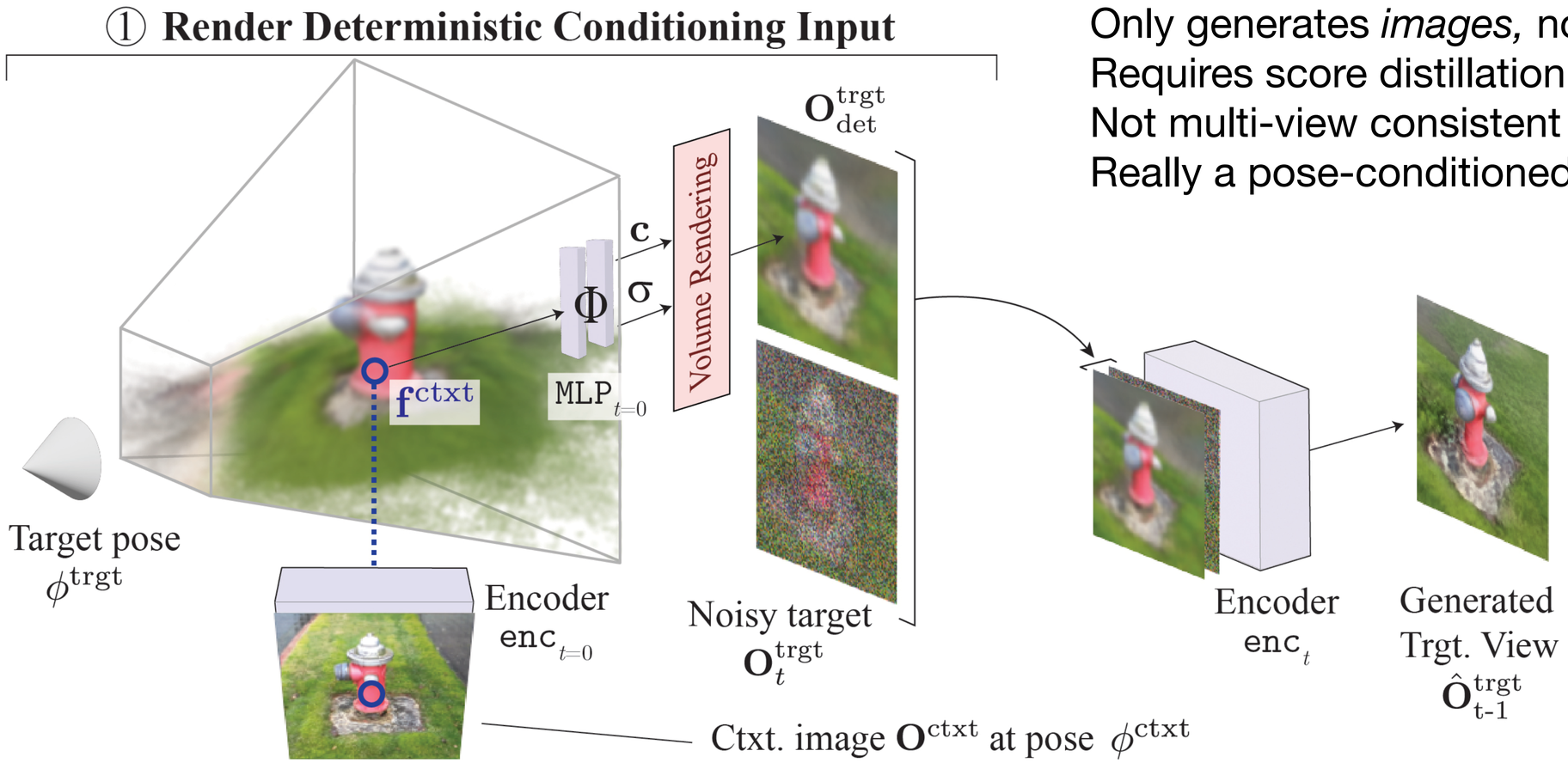
Key Idea:
Denoise Target View, but make rendering part of denoising operator!



Generating a Conditioning Input



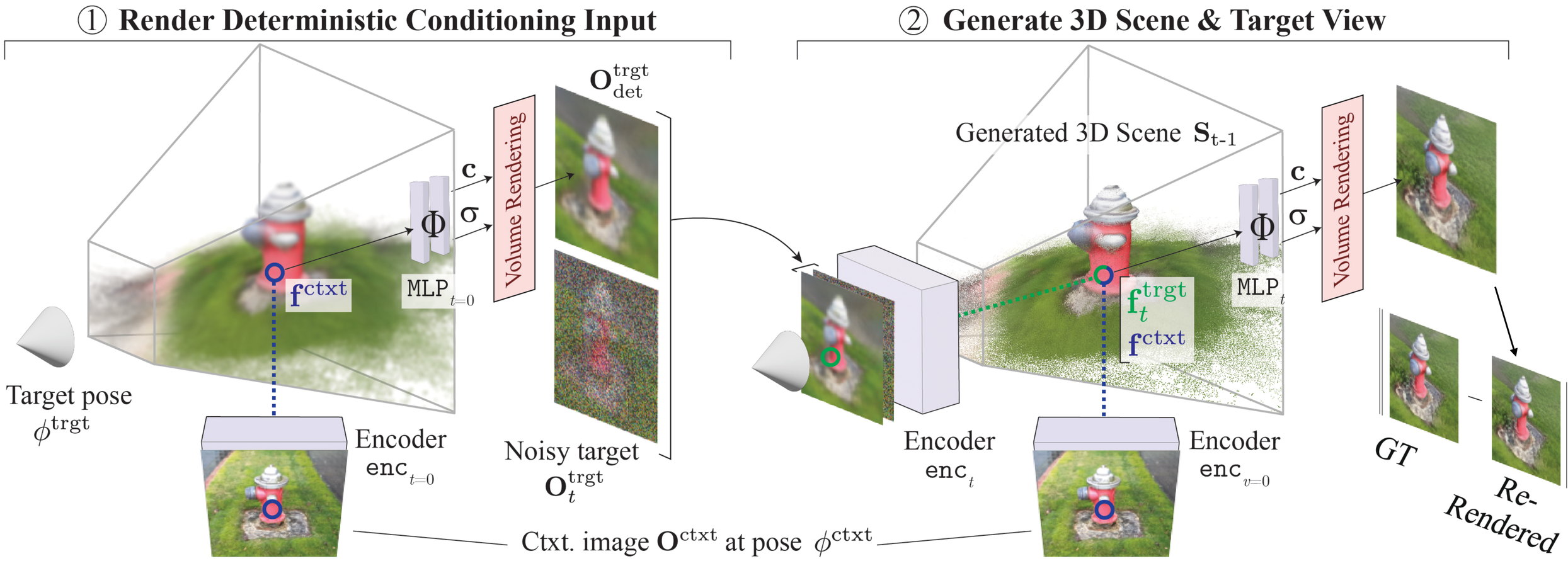
Non 3D-structured Baseline



Only generates *images,* not *3D scenes*
Requires score distillation to sample 3D scenes
Not multi-view consistent
Really a pose-conditioned *image* generative model

Generative Novel View Synthesis with 3D-Aware Diffusion Models, Chan et al. 2023
SparseFusion, Zhou et al. 2023

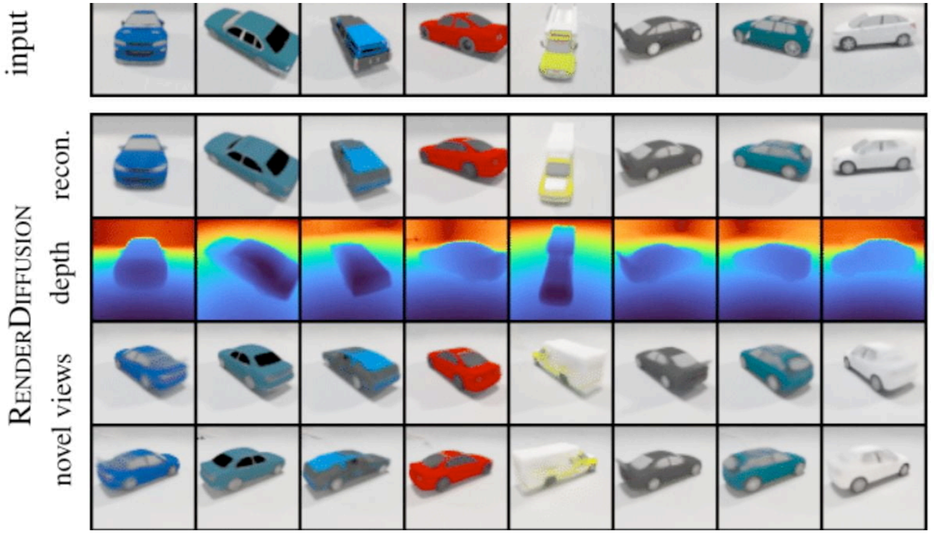
Conditional Diffusion with Rendering in-the-loop



Trained end-to-end. Directly samples 3D scenes (geometry & appearance). Multi-view consistent novel view synth.

Related & Concurrent Work

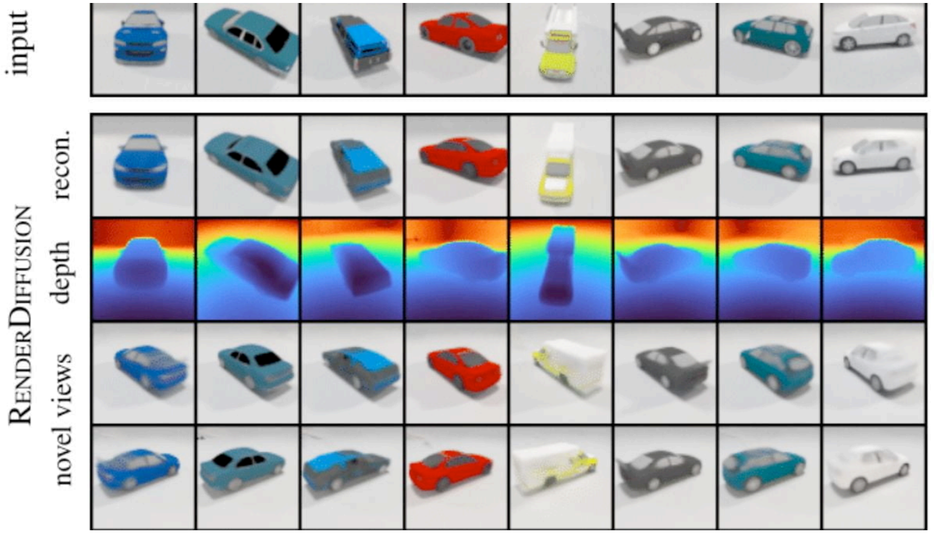
Generative Novel View Synthesis with 3D-Aware Diffusion Models, Chan et al. 2023
SparseFusion, Zhou et al. 2023
Not 3D, not multi-view consistent, requires score distillation.



RenderDiffusion, Anciukevicius et al. 2022
DiffRF, Müller et al. 2023
HoloDiffusion, Karnewar et al. 2023
…
3D, but unconditional, thus limited to simple distributions.

Related & Concurrent Work

Generative Novel View Synthesis with 3D-Aware Diffusion Models, Chan et al. 2023
SparseFusion, Zhou et al. 2023
Not 3D, not multi-view consistent, requires score distillation.



RenderDiffusion, Anciukevicius et al. 2022
DiffRF, Müller et al. 2023
HoloDiffusion, Karnewar et al. 2023
…
3D, but unconditional, thus limited to simple distributions.

Single Input Image

Our 3D
Generative Model

Sampled 3D Scene 1

Sampled 3D Scene 2

Sampled 3D Scene 3

Sampled 3D Scene 4

Single Input Image

Our 3D
Generative Model

Sampled 3D Scene 1

Sampled 3D Scene 2

Sampled 3D Scene 3

Sampled 3D Scene 4

Results - Co3D Hydrants

Input | Novel Views | Depth

Results - Co3D Hydrants

Input | Novel Views | Depth

Comparison - Co3D Hydrants

Input

Comparison - Co3D Hydrants

Input

Comparison - Co3D Hydrants

Input

PixelNeRF

Comparison - Co3D Hydrants

Input

PixelNeRF

SparseFusion

Comparison - Co3D Hydrants

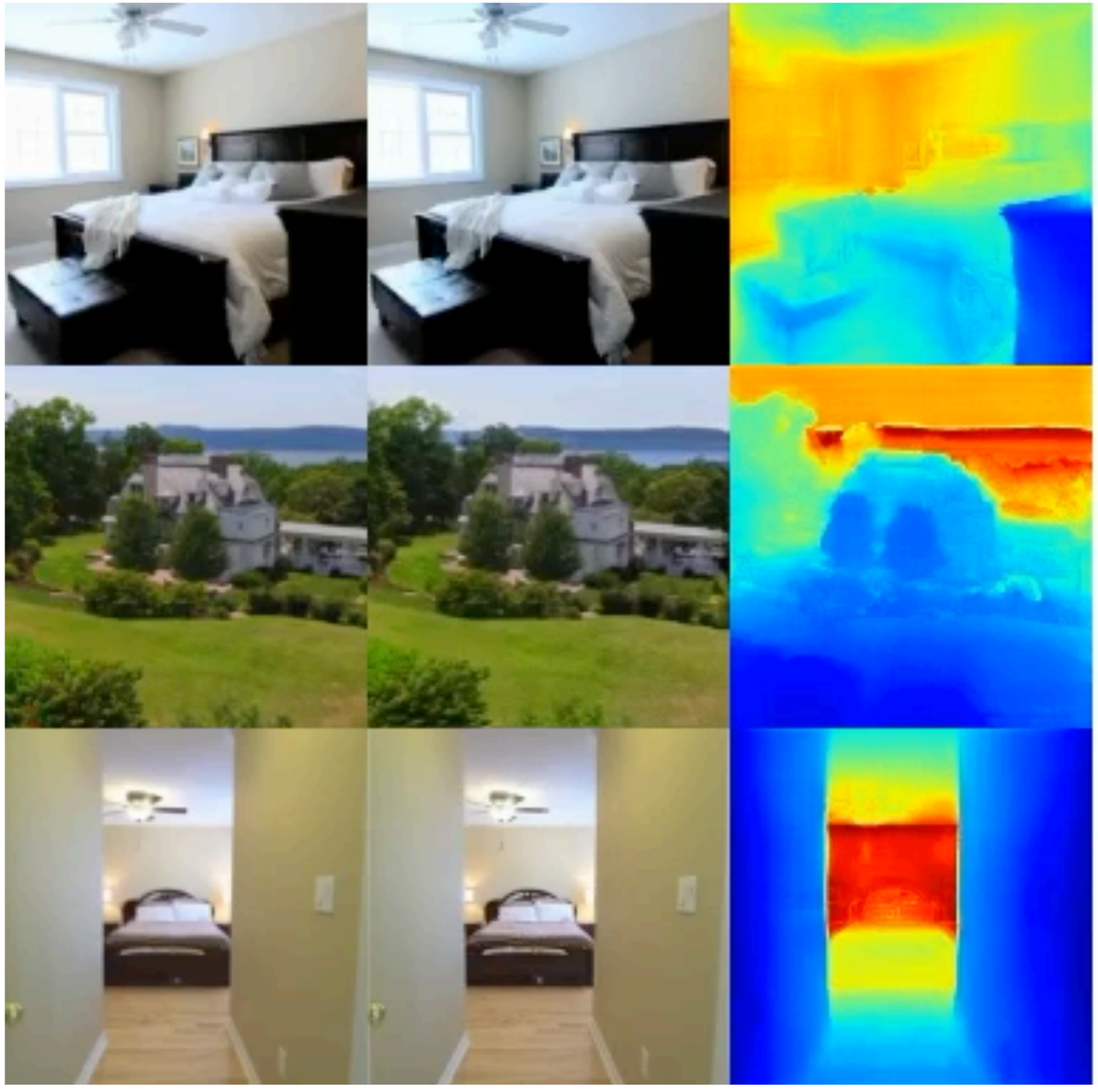
Input

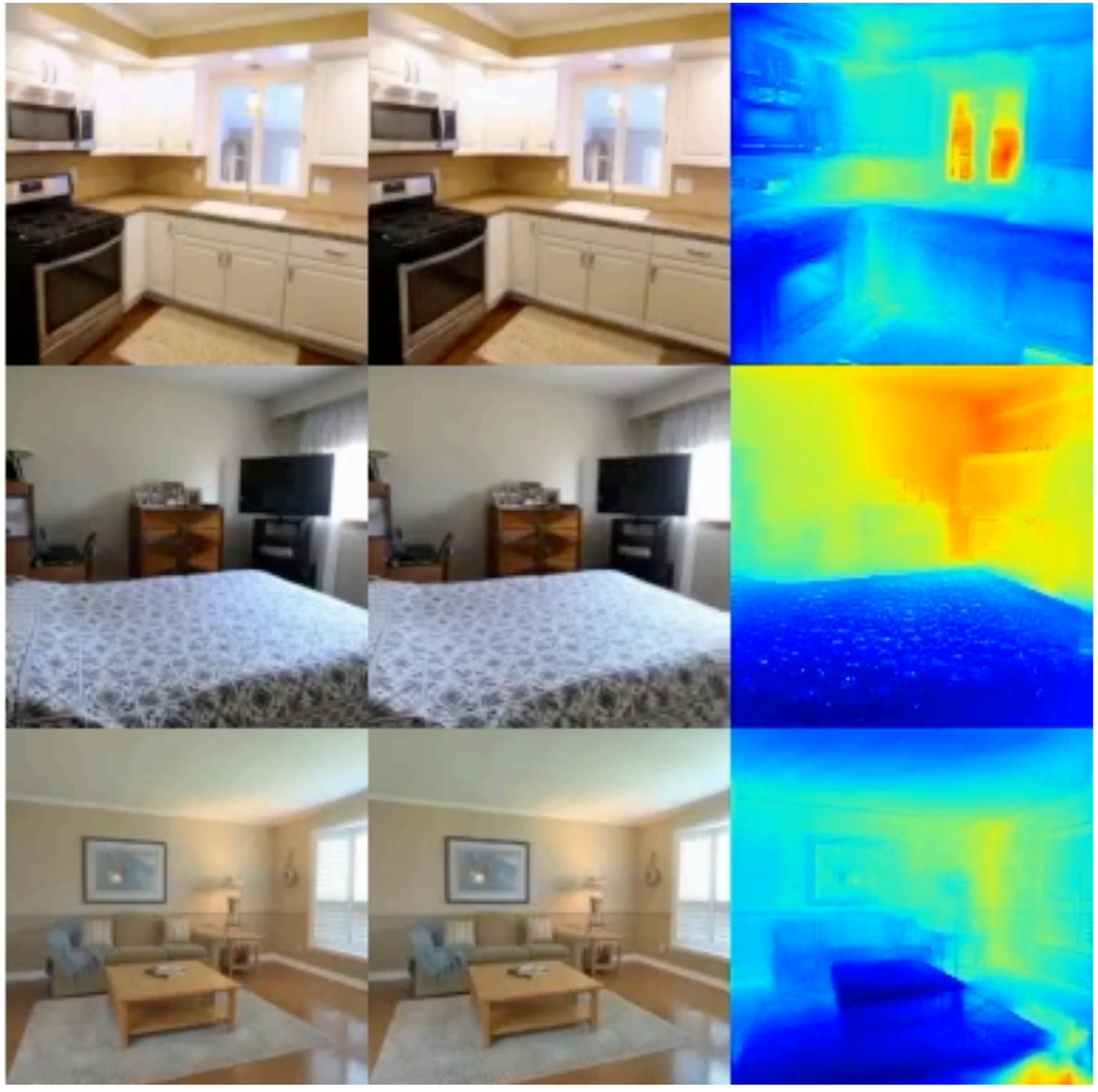
PixelNeRF

SparseFusion

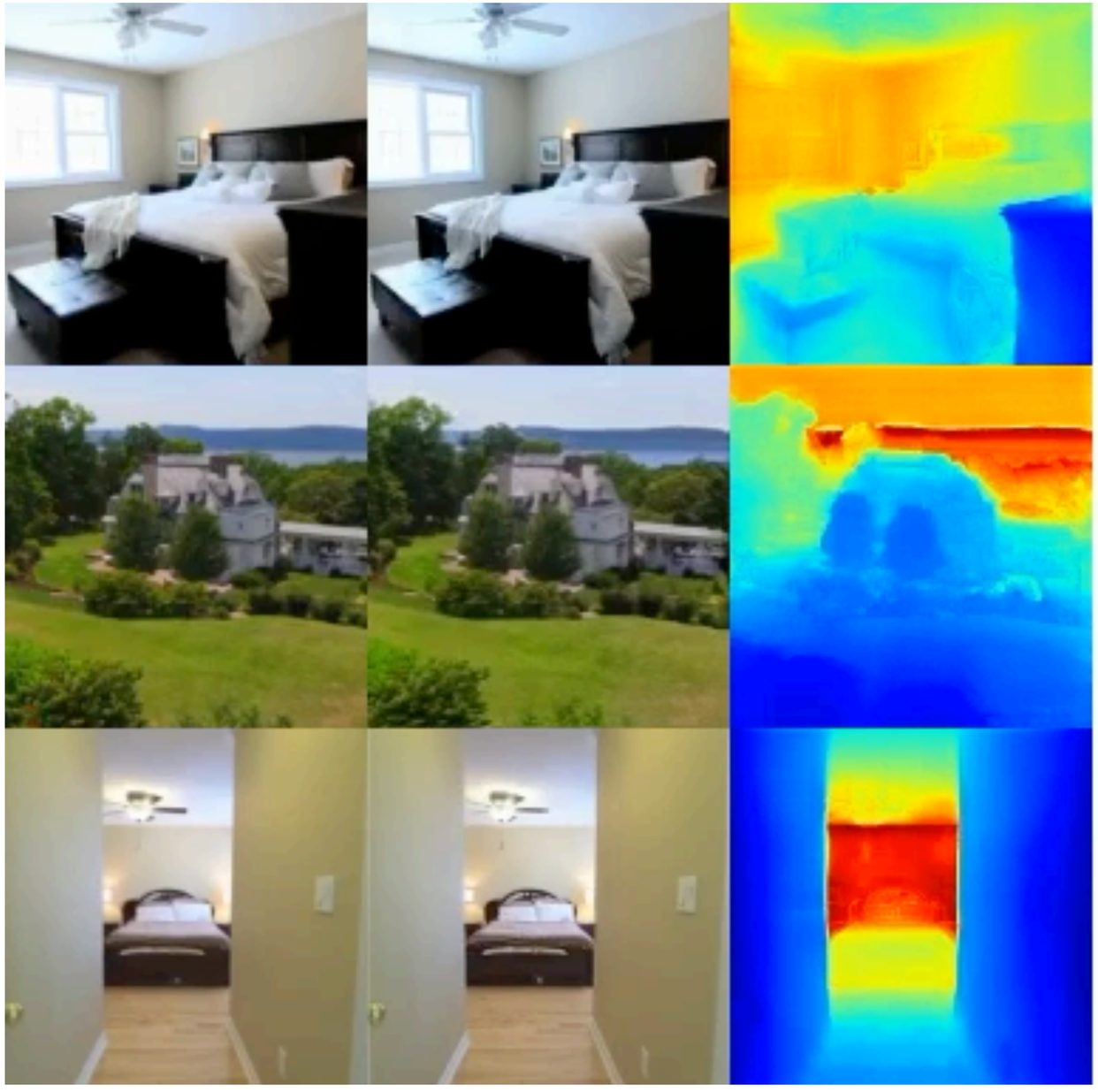
Ours

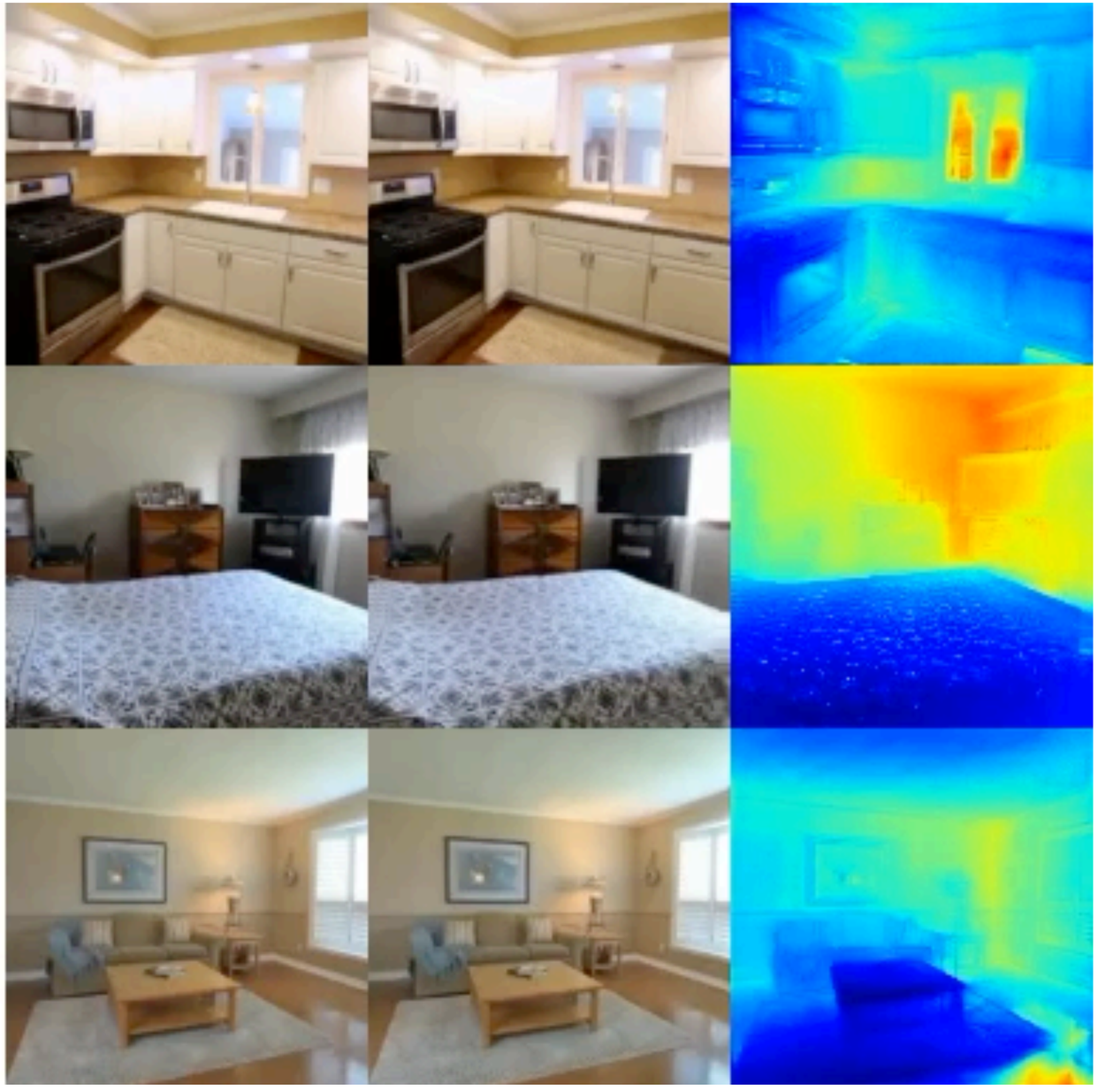
RealEstate10k - Results





RealEstate10k - Results





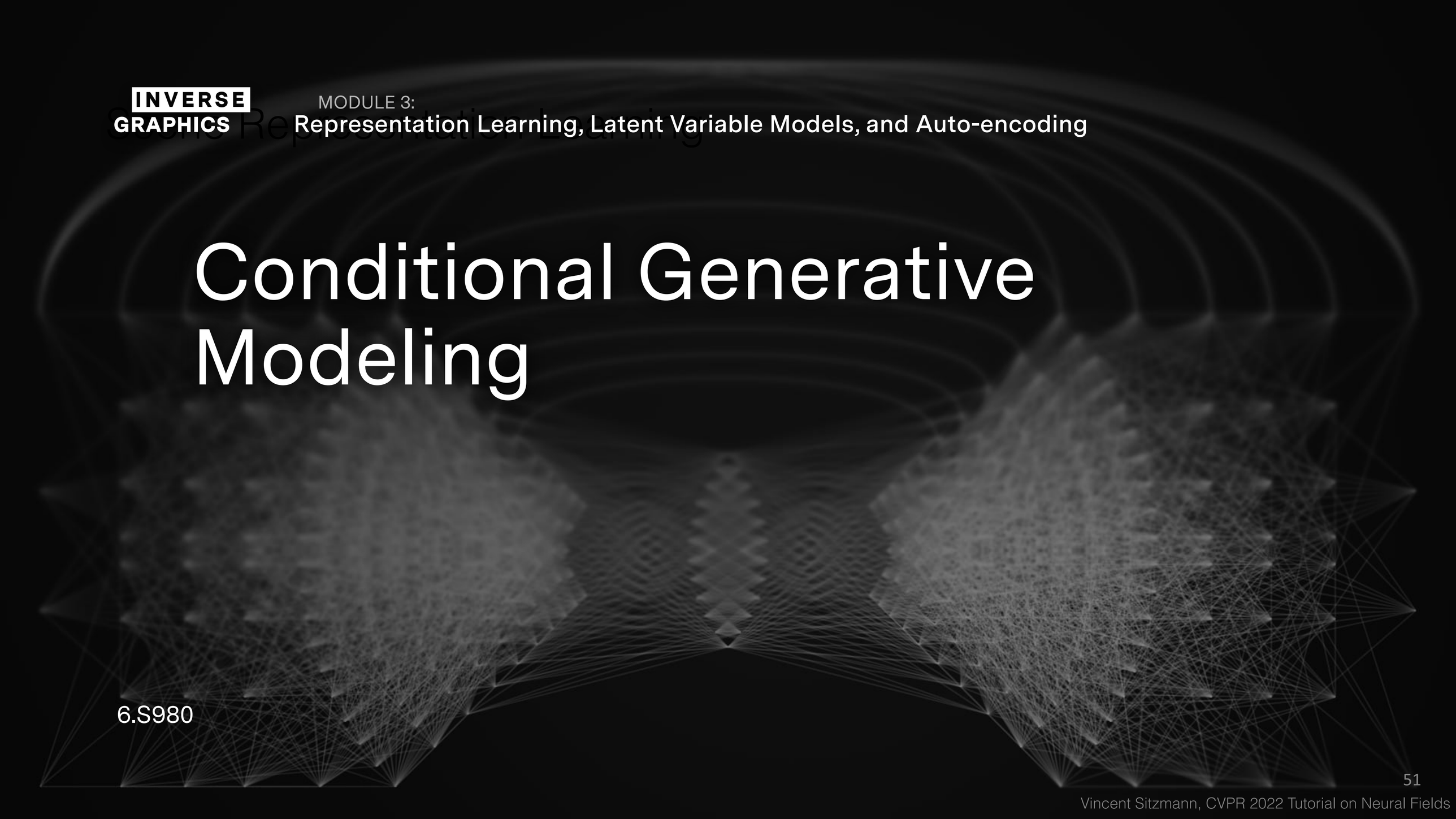
RealEstate10k - Sample Variance

RealEstate10k - Sample Variance

RealEstate10k - Sample Variance

RealEstate10k - Sample Variance

Conditional Generative Modeling



6.S980