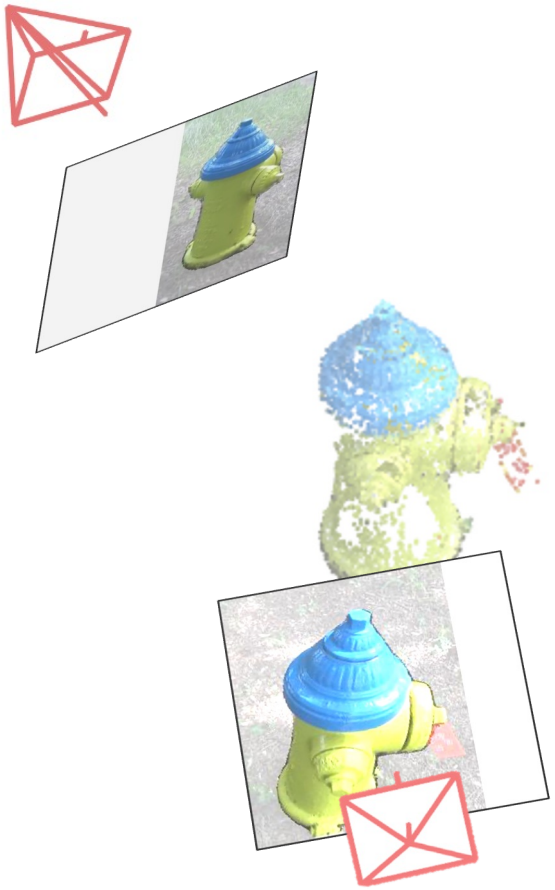
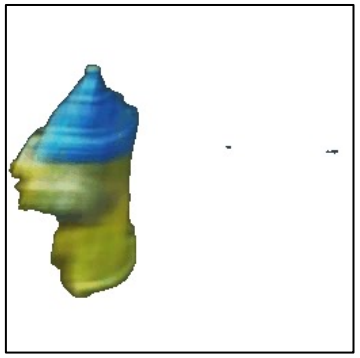
SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction

Zhizhuo (Z) Zhou, Shubham Tulsiani

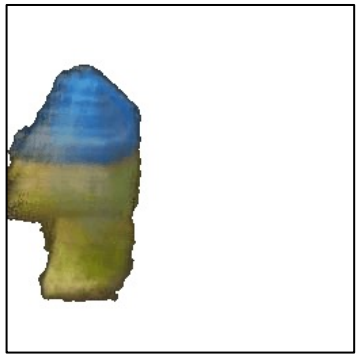
Task: Sparse-view Reconstruction



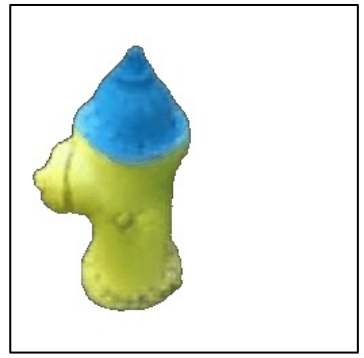
PixelNeRF



NerFormer



ViewFormer



SparseFusion

SparseFusion Overview

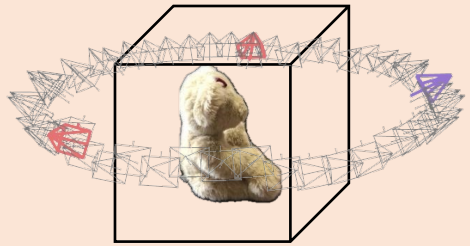
View-conditioned Latent Diffusion

$p_\phi($  $|$  $)$

$\boldsymbol{x}$ $\boldsymbol{\pi}, C$

Diffusion Distillation

$(x, y, z) \rightarrow [f_\theta] \rightarrow (rgb, \sigma)$



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

SparseFusion Overview

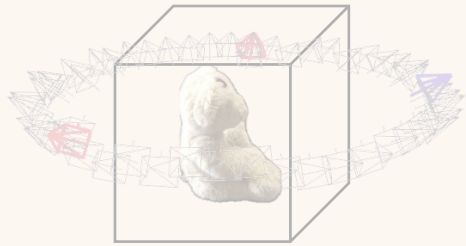
View-conditioned Latent Diffusion

$p_\phi($  $|$  $)$

$\boldsymbol{x}$ $\boldsymbol{\pi}, C$

Diffusion Distillation

$(x, y, z) \rightarrow [f_\theta] \rightarrow (rgb, \sigma)$



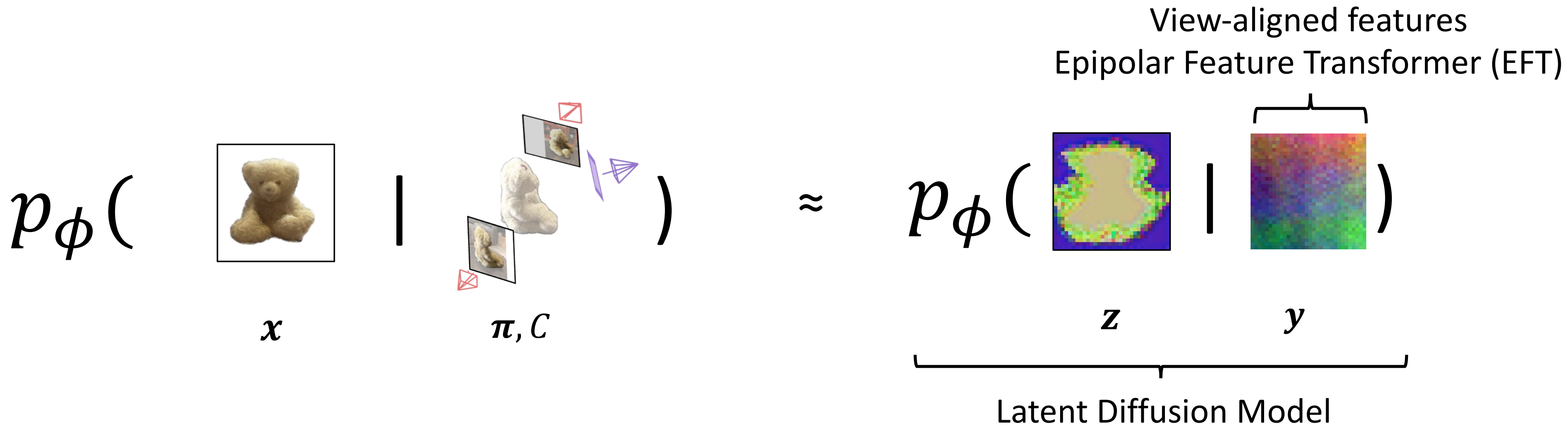
$$\min_\theta \mathbb{E}_\pi[-\log p_\phi(f_\theta(\boldsymbol{\pi})|\boldsymbol{\pi}, C)]$$

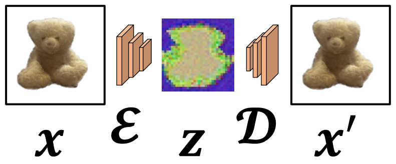
View-conditioned Latent Diffusion (VLDM)

$$p_\phi(\ \boxed{\text{  }}\ \mid \text{  }\ ) \quad \approx \quad p_\phi(\ \boxed{\text{  }}\ \mid \text{  }\ )$$

$\boldsymbol{x}$ $\quad$ $\boldsymbol{\pi}, \boldsymbol{C}$ $\quad$ $\boldsymbol{x}$ $\quad$ $\boldsymbol{y}$

View-conditioned Latent Diffusion (VLDM)





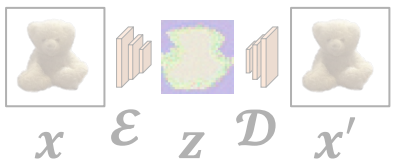
VAE from High-Resolution Image Synthesis
with Latent Diffusion Model
Rombach et al. CVPR 2022.

View-conditioned Latent Diffusion (VLDM)

$$p_\phi(\ x\ |\ \pi, C\ ) \approx p_\phi(\ z\ |\ y\ )$$

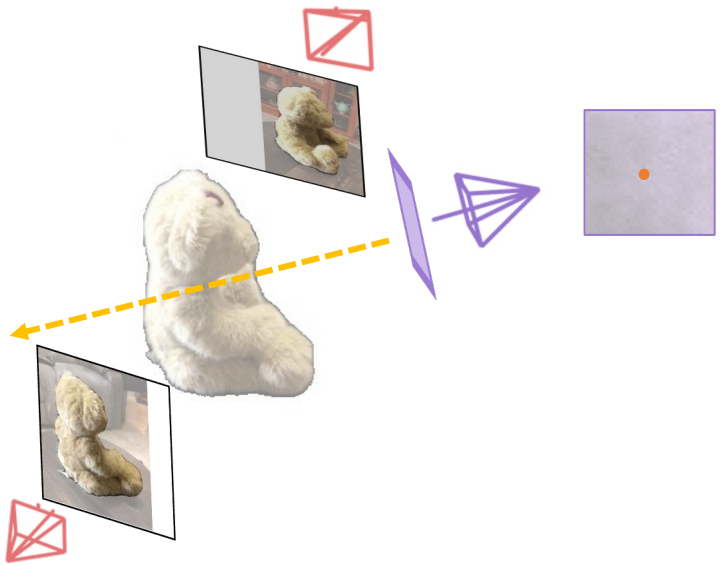
View-aligned features
Epipolar Feature Transformer (EFT)

Latent Diffusion Model



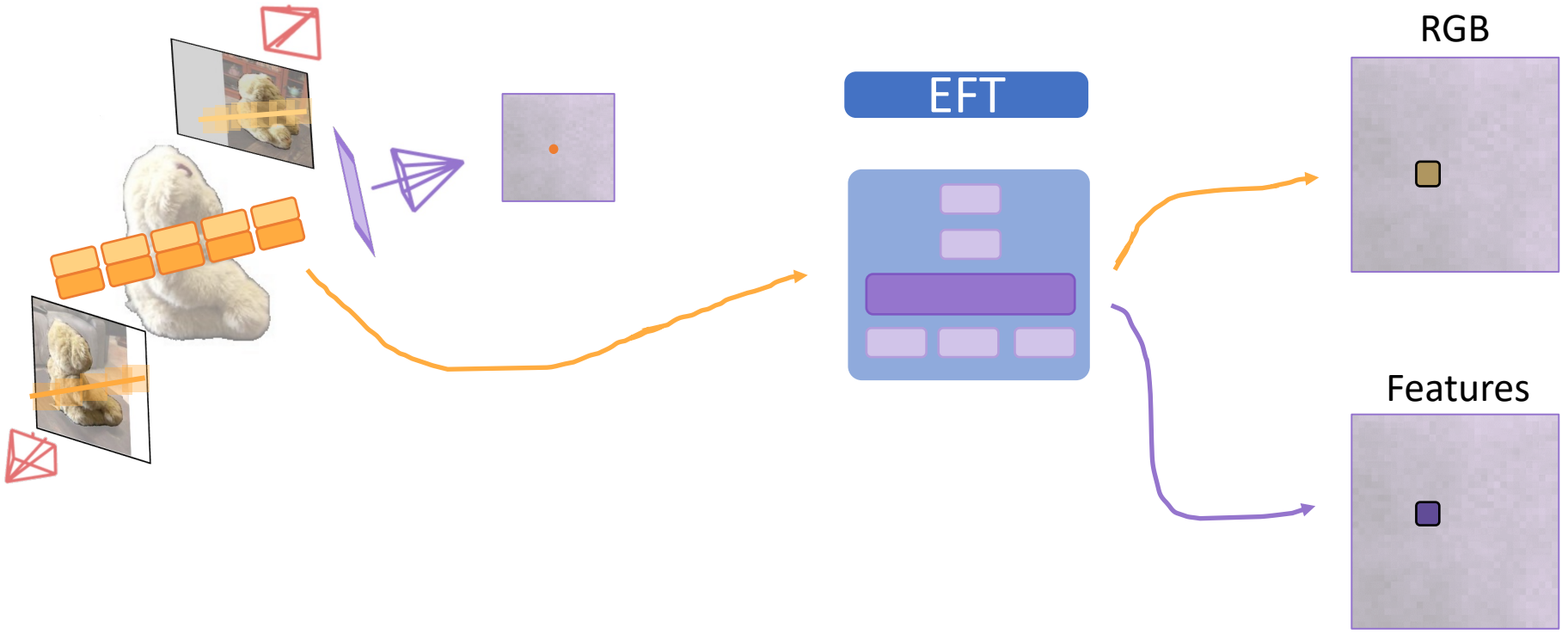
VAE from High-Resolution Image Synthesis with Latent Diffusion Model
Rombach et al. CVPR 2022.

Epipolar Feature Transformer (EFT)



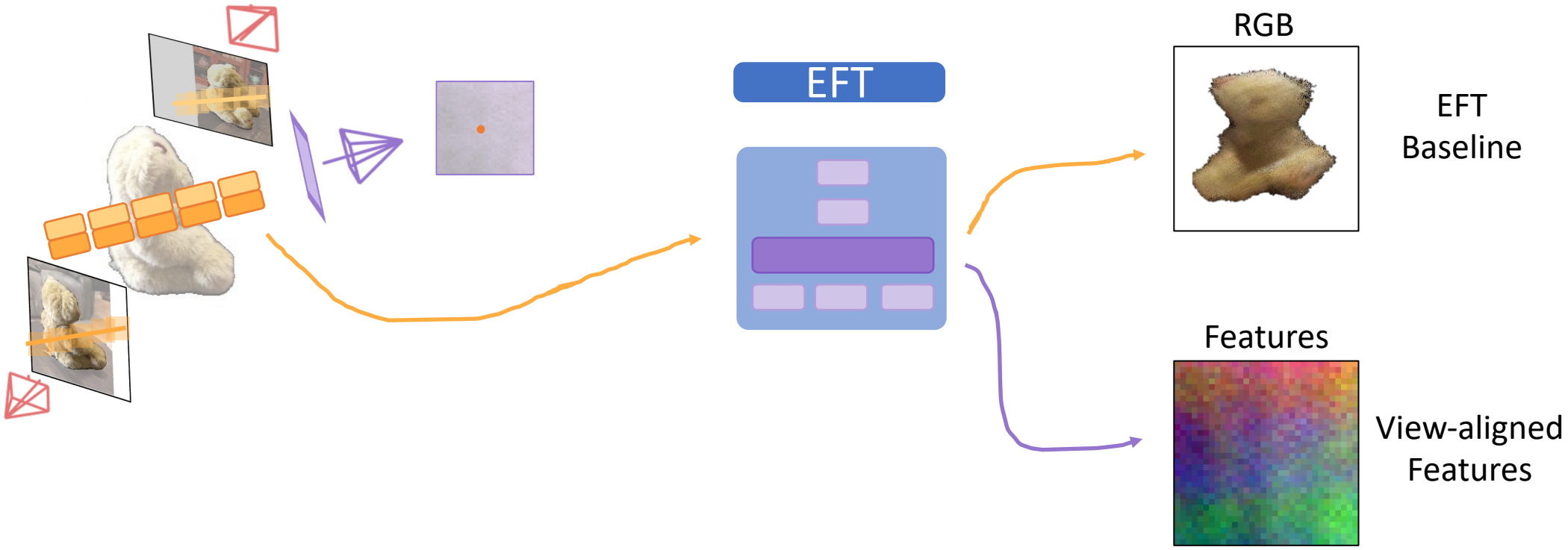
Modified from Generalizable Patch-Based Neural Rendering
Suhail et al. ECCV 2022.

Epipolar Feature Transformer (EFT)



Modified from Generalizable Patch-Based Neural Rendering
Suhail et al. ECCV 2022.

Epipolar Feature Transformer (EFT)



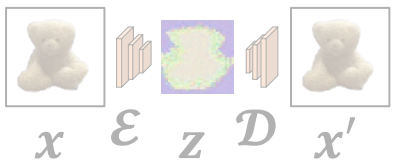
Modified from Generalizable Patch-Based Neural Rendering
Suhail et al. ECCV 2022.

View-conditioned Latent Diffusion (VLDM)

$$p_\phi(\ x\ |\ \pi, C\ ) \approx p_\phi(\ z\ |\ y\ )$$

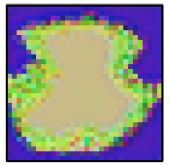
View-aligned features
Epipolar Feature Transformer (EFT)

Latent Diffusion Model



VAE from High-Resolution Image Synthesis with Latent Diffusion Model
Rombach et al. CVPR 2022.

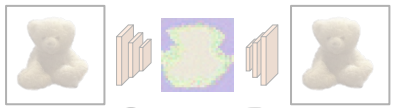
View-conditioned Latent Diffusion (VLDM)

$p_\phi($  $|$  $)$ $\approx$ $p_\phi($  $|$  $)$

x π, C z y

View-aligned features
Epipolar Feature Transformer (EFT)

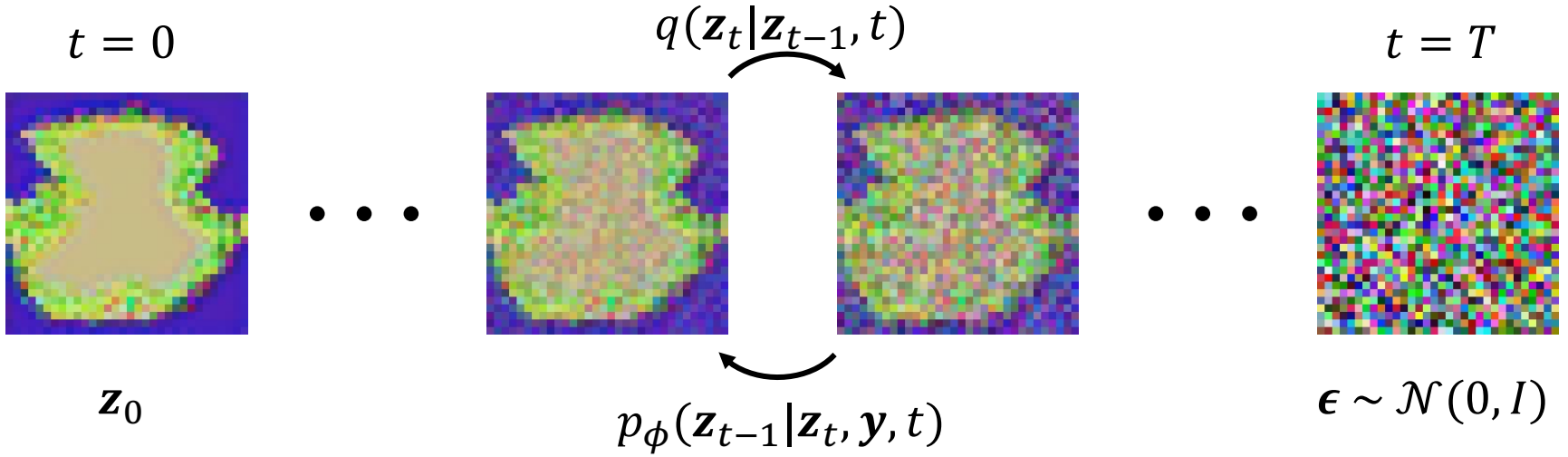
Latent Diffusion Model

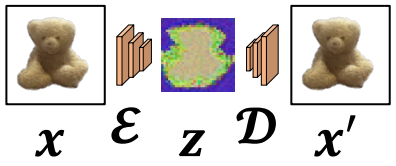


x $\mathcal{E}$ z $\mathcal{D}$ x'

VAE from High-Resolution Image Synthesis with Latent Diffusion Model
Rombach et al. CVPR 2022.

Latent Diffusion Model

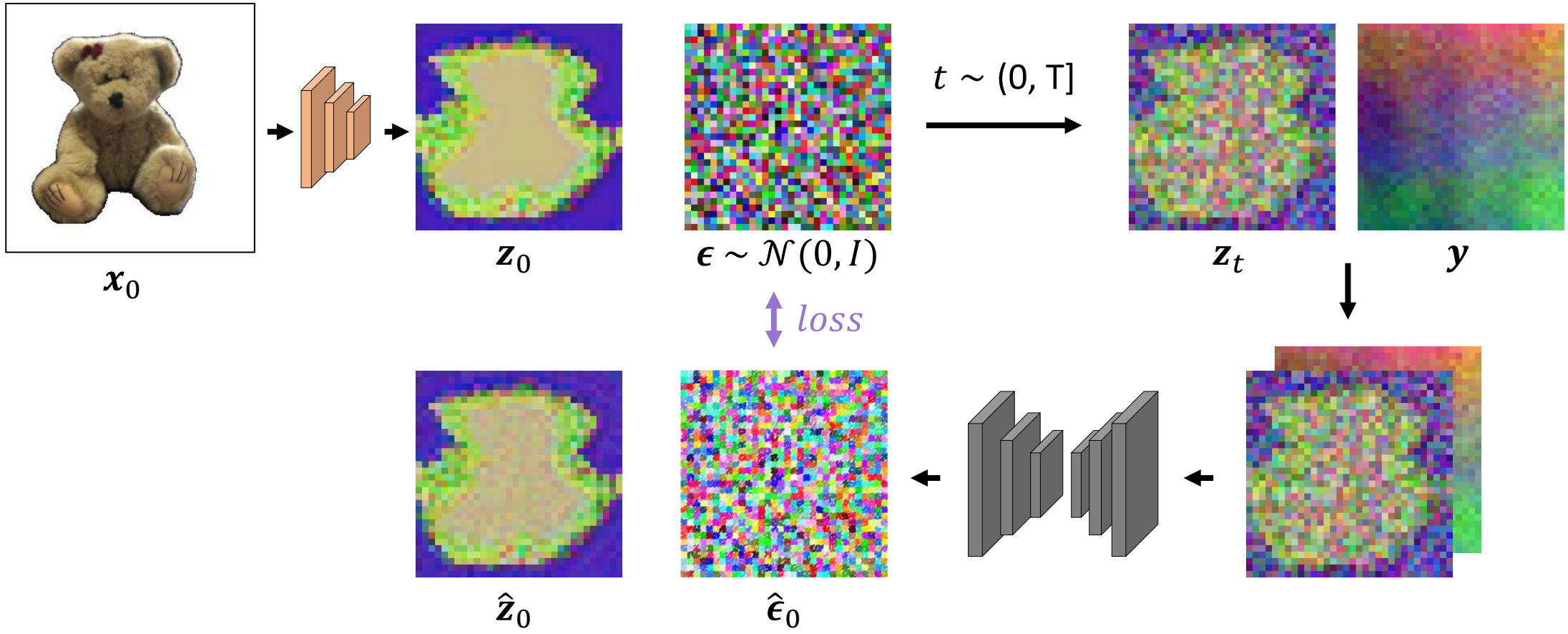




Denoising Diffusion Probabilistic Models
Ho et al. NeurIPS 2020.

VAE from High-Resolution Image Synthesis with Latent Diffusion Model
Rombach et al. CVPR 2022.

View-conditioned Latent Diffusion (VLDM)



$p_\phi($ z | y $)$

Denoising Diffusion Probabilistic Models
Ho et al. NeurIPS 2020.

VAE from High-Resolution Image Synthesis with Latent Diffusion Model
Rombach et al. CVPR 2022.

View-conditioned Latent Diffusion (VLDM)

$$p_\phi(\ \boldsymbol{x}\ \mid\ \boldsymbol{\pi}, C\ )$$

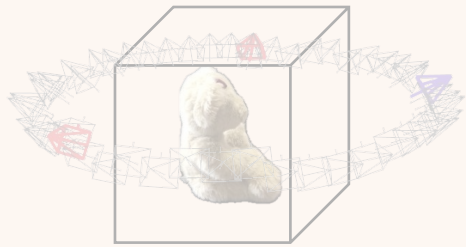
SparseFusion Overview

View-conditioned Latent Diffusion

$$p_\phi(\ \boldsymbol{x}\ |\ \boldsymbol{\pi}, C\ )$$

Diffusion Distillation

$$(x, y, z) \rightarrow f_\theta \rightarrow (rgb, \sigma)$$



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

SparseFusion Overview

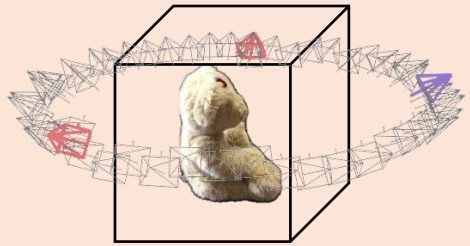
View-conditioned Latent Diffusion

$p_\phi($  $|$  $)$

$\boldsymbol{x}$ $\quad$ $\boldsymbol{\pi}, C$

Diffusion Distillation

$(x, y, z) \rightarrow [f_\theta] \rightarrow (rgb, \sigma)$



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

SparseFusion Overview

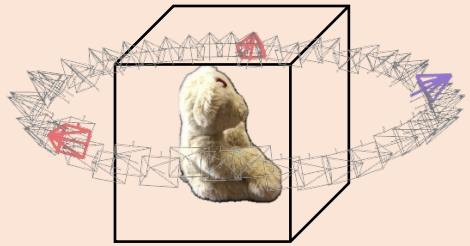
View-conditioned Latent Diffusion

$p_\phi($  $|$  $)$

x π, C

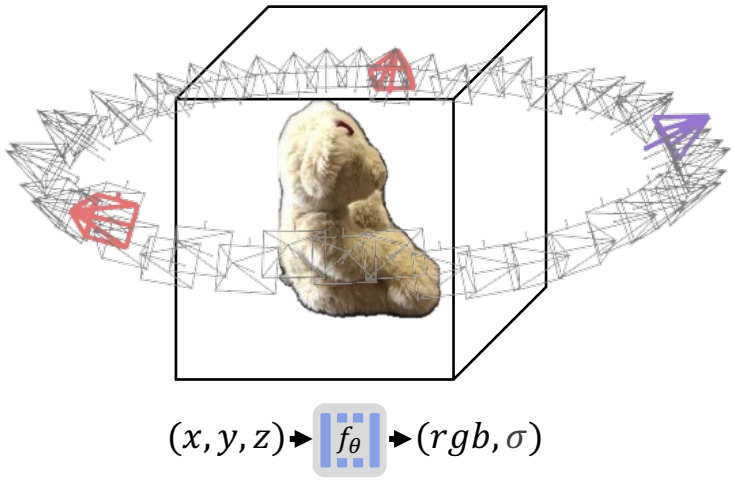
Diffusion Distillation

$(x, y, z) \rightarrow [f_\theta] \rightarrow (rgb, \sigma)$



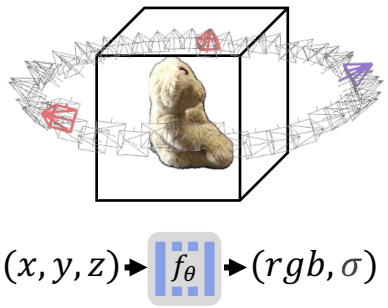
$$\min_\theta \mathbb{E}_\pi[-\log p_\phi(f_\theta(\boldsymbol{\pi})|\boldsymbol{\pi}, C)]$$

Diffusion Distillation



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

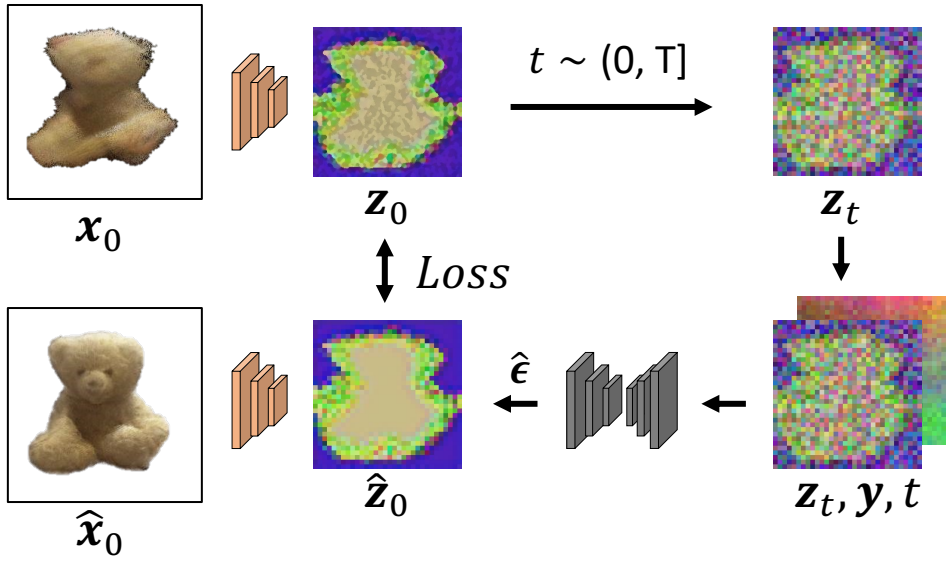
Diffusion Distillation



$\pi \sim \Pi$

$\nabla \mathcal{L}$

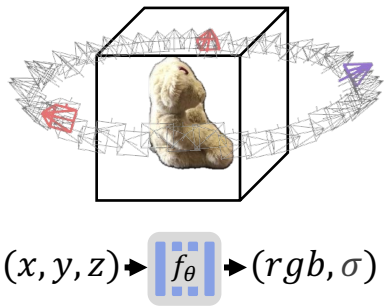
$$\log p_\phi(\ \boldsymbol{x}_0 = f_\theta(\boldsymbol{\pi})\ |\boldsymbol{\pi}, C)$$



$$\log p_\phi\ (\boldsymbol{x}_0|\boldsymbol{\pi}, C) \approx \mathbb{E}_t \|\boldsymbol{z}_0 - \hat{\boldsymbol{z}}_0\|^2$$

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

Diffusion Distillation

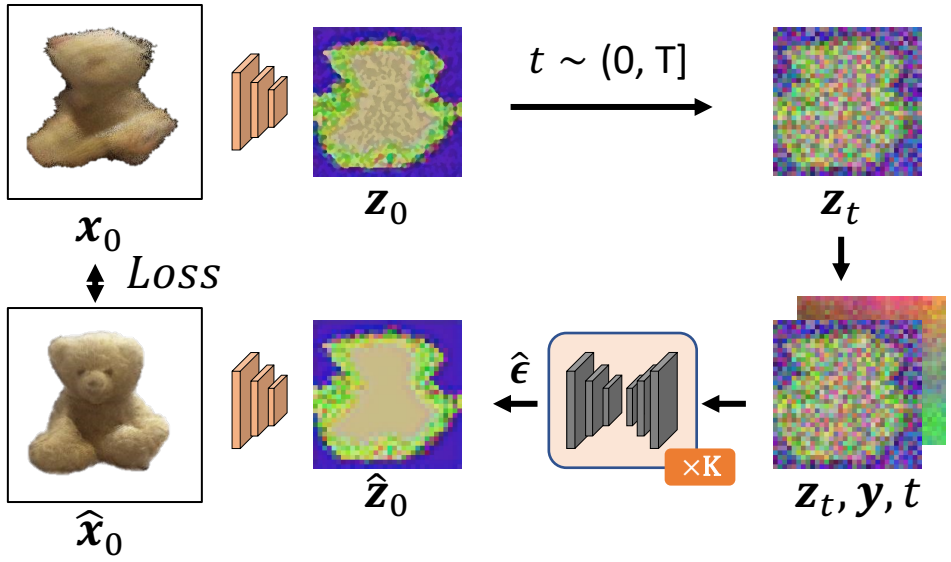


$\pi \sim \Pi$

$\nabla \mathcal{L}$

$\log p_\phi($  $|\pi, C)$

$\boldsymbol{x}_0 = f_\theta(\pi)$



$$\log p_\phi(\boldsymbol{x}_0|\boldsymbol{\pi}, C) \approx \mathbb{E}_t \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0\|^2$$

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

SparseFusion Overview

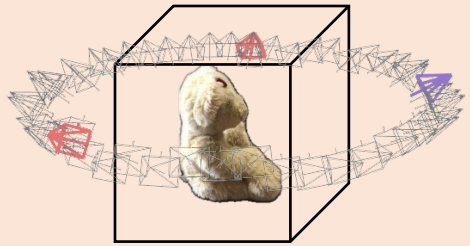
View-conditioned Latent Diffusion

$p_\phi($  $|$  $)$

$\boldsymbol{x}$ $\quad$ $\boldsymbol{\pi}, C$

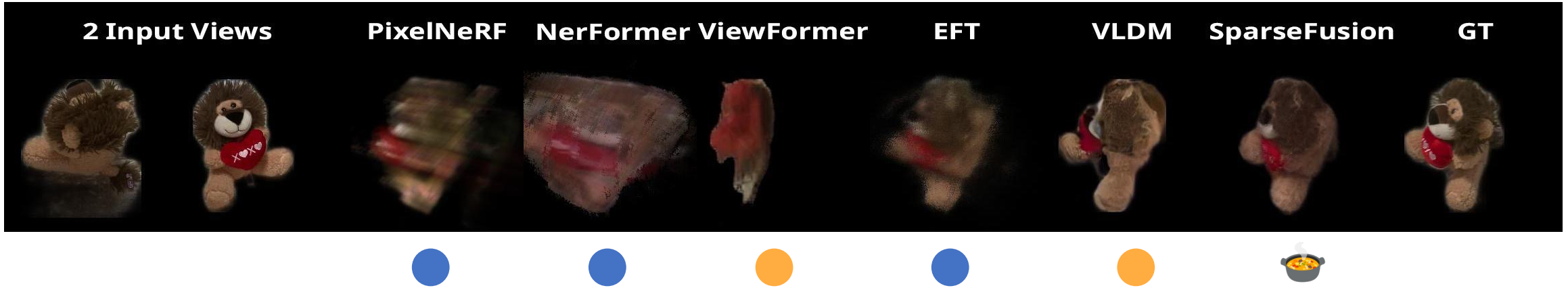
Diffusion Distillation

$(x, y, z) \rightarrow [f_\theta] \rightarrow (rgb, \sigma)$



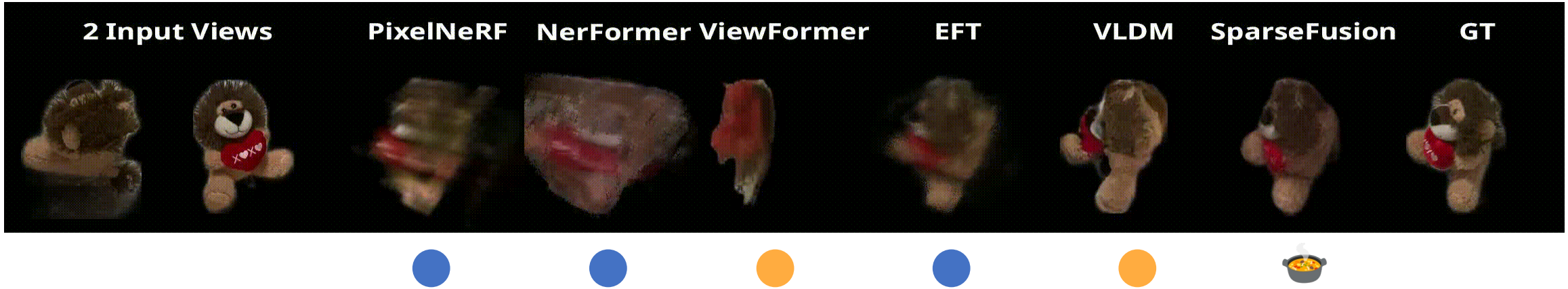
$$\min_\theta \mathbb{E}_\pi[-\log p_\phi(f_\theta(\boldsymbol{\pi})|\boldsymbol{\pi}, C)]$$

Results



- Does not hallucinate unseen regions
- Does not respect geometry
- Hallucinates unseen region + respects geometry

Results



- Does not hallucinate unseen regions
- Does not respect geometry
- Hallucinates unseen region + respects geometry

Results

CO3Dv2 Dataset

Reizenstein et al. ICCV 2021.

2-view Novel View Synthesis on 10 Categories

	PSNR↑	LPIPS↓
PixelNeRF	19.52	0.327
NerFormer	17.88	0.382
ViewFormer	18.37	0.282

Results

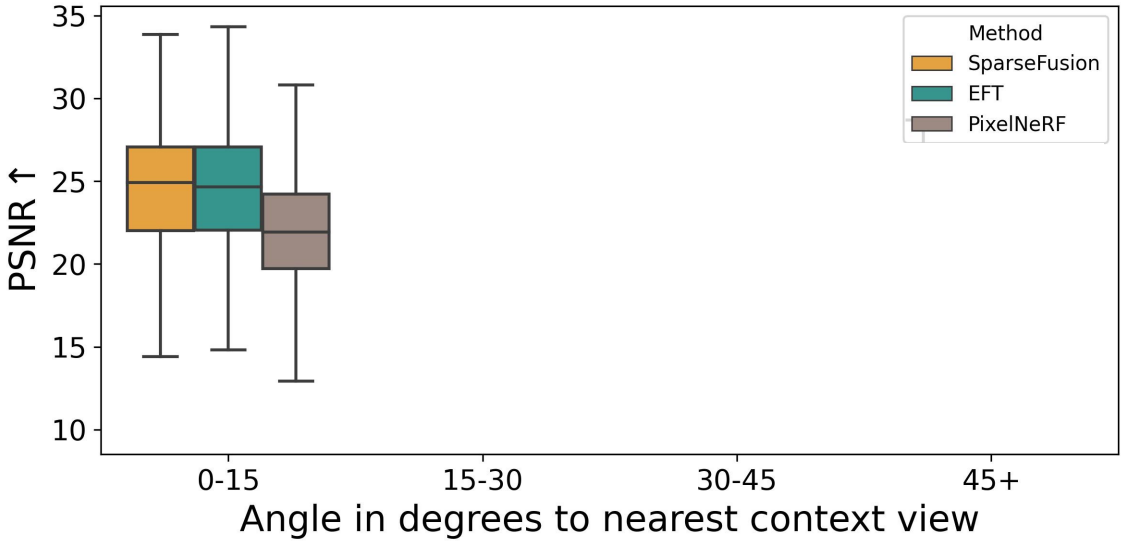
CO3Dv2 Dataset

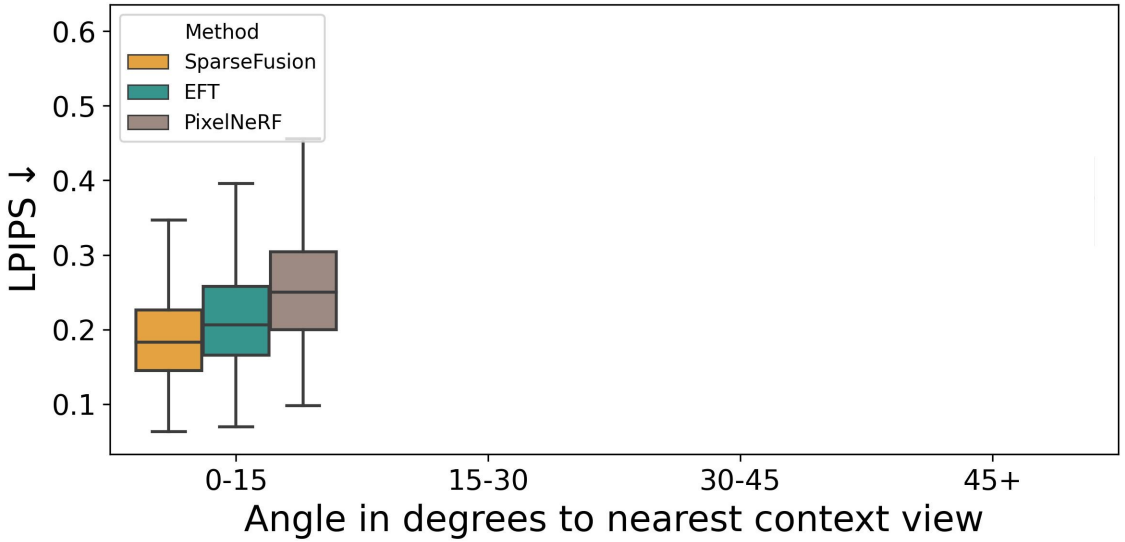
Reizenstein et al. ICCV 2021.

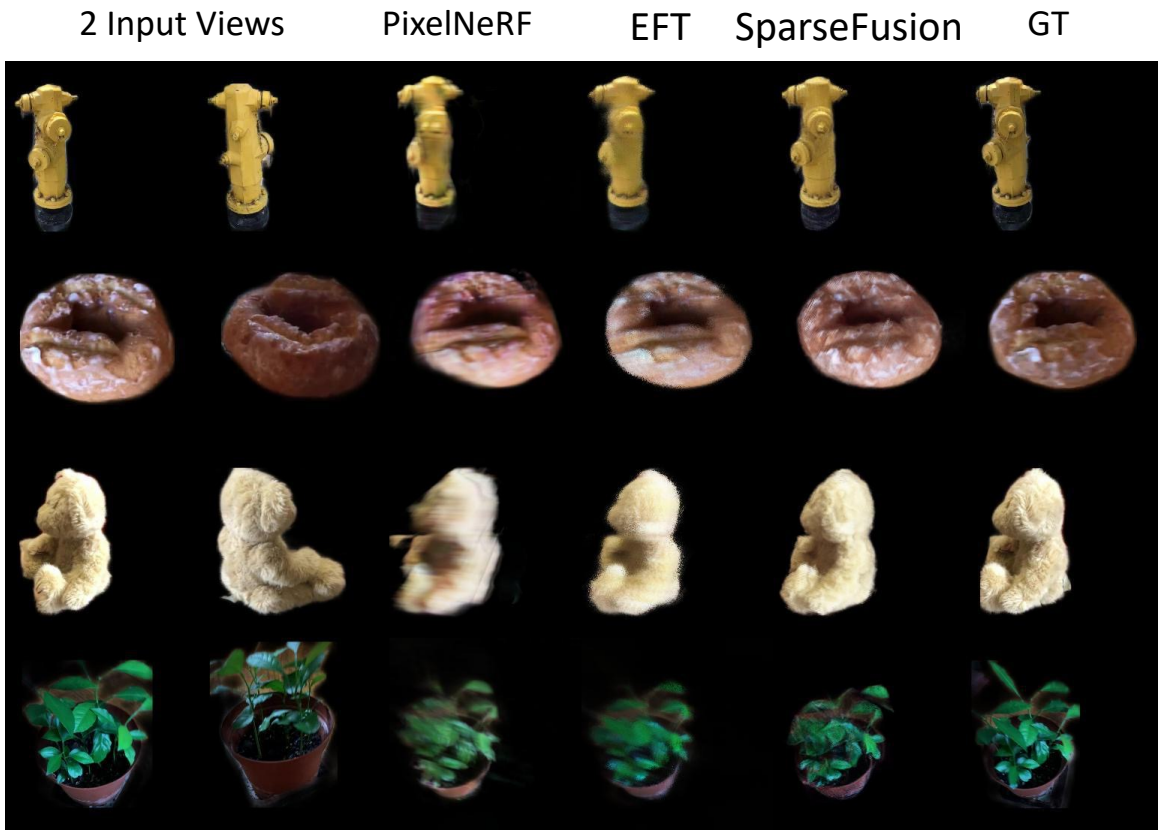
2-view Novel View Synthesis on 10 Categories

	PSNR↑	LPIPS↓
PixelNeRF	19.52	0.327
NerFormer	17.88	0.382
ViewFormer	18.37	0.282
EFT	20.85	0.289
VLDM	19.55	0.247
SparseFusion	**21.34**	**0.225**

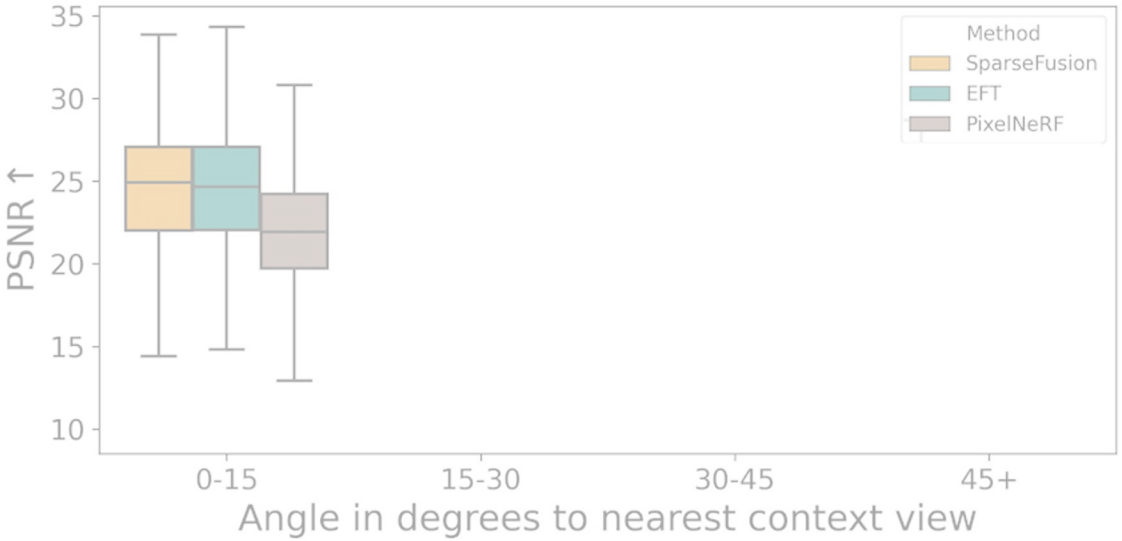
Results

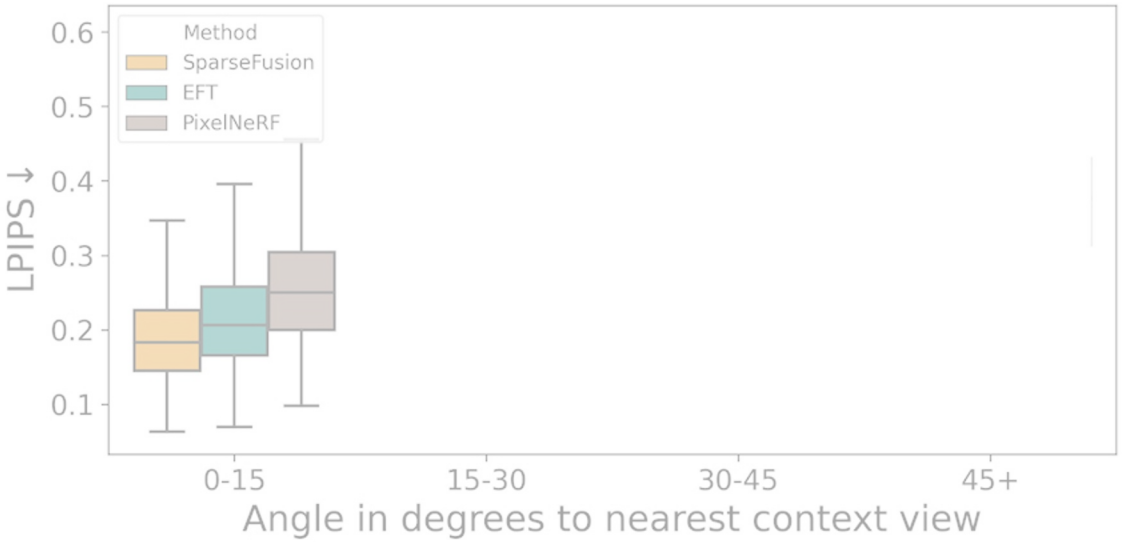

Method
SparseFusion
EFT
PixelNeRF
PSNR ↑
35
30
25
20
15
10
0-15
15-30
30-45
45+
Angle in degrees to nearest context view

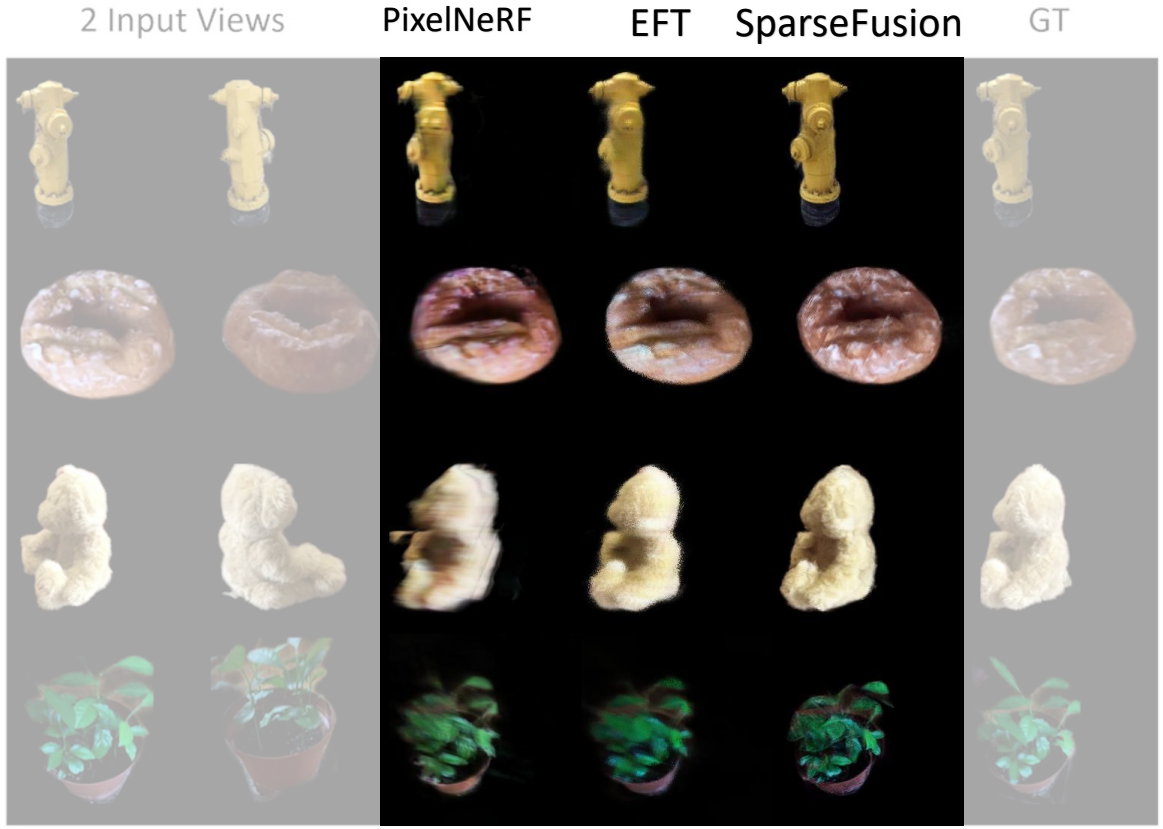

Method
SparseFusion
EFT
PixelNeRF
LPIPS ↓
0.6
0.5
0.4
0.3
0.2
0.1
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


2 Input Views
PixelNeRF
EFT
SparseFusion
GT

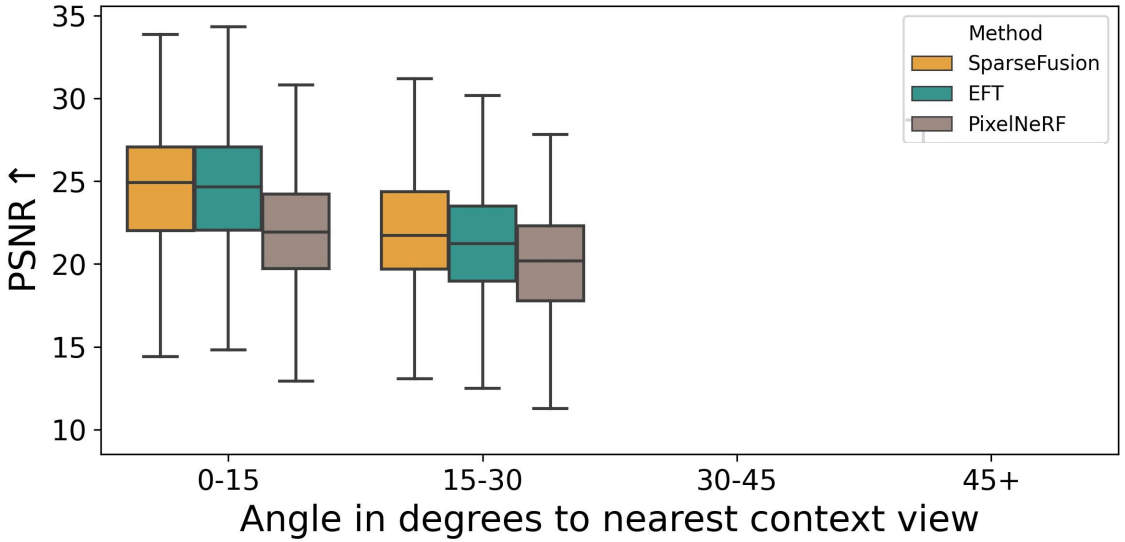
Results

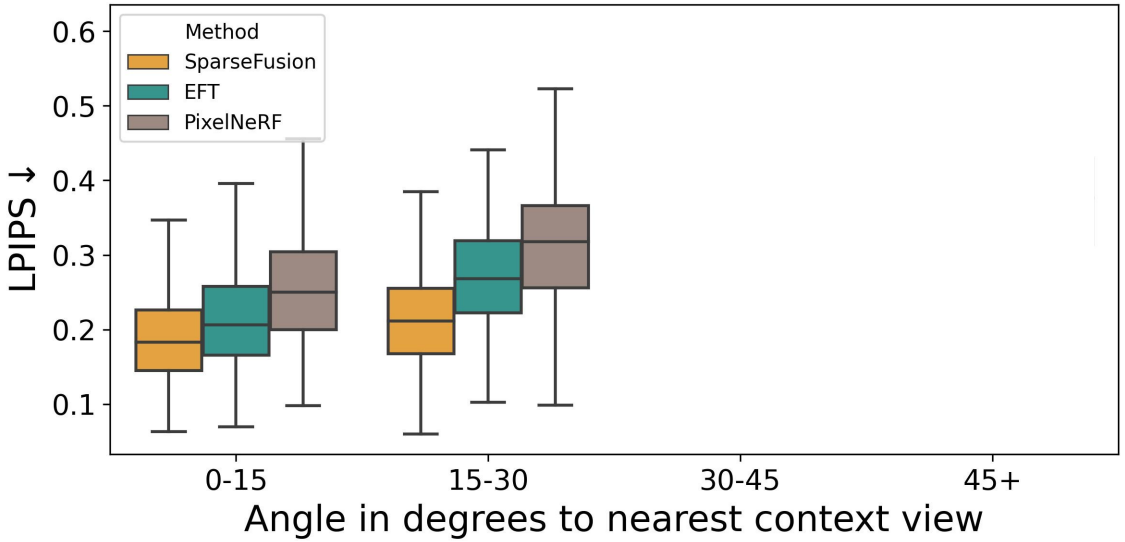

Method
SparseFusion
EFT
PixelNeRF
PSNR ↑
35
30
25
20
15
10
0-15
15-30
30-45
45+
Angle in degrees to nearest context view

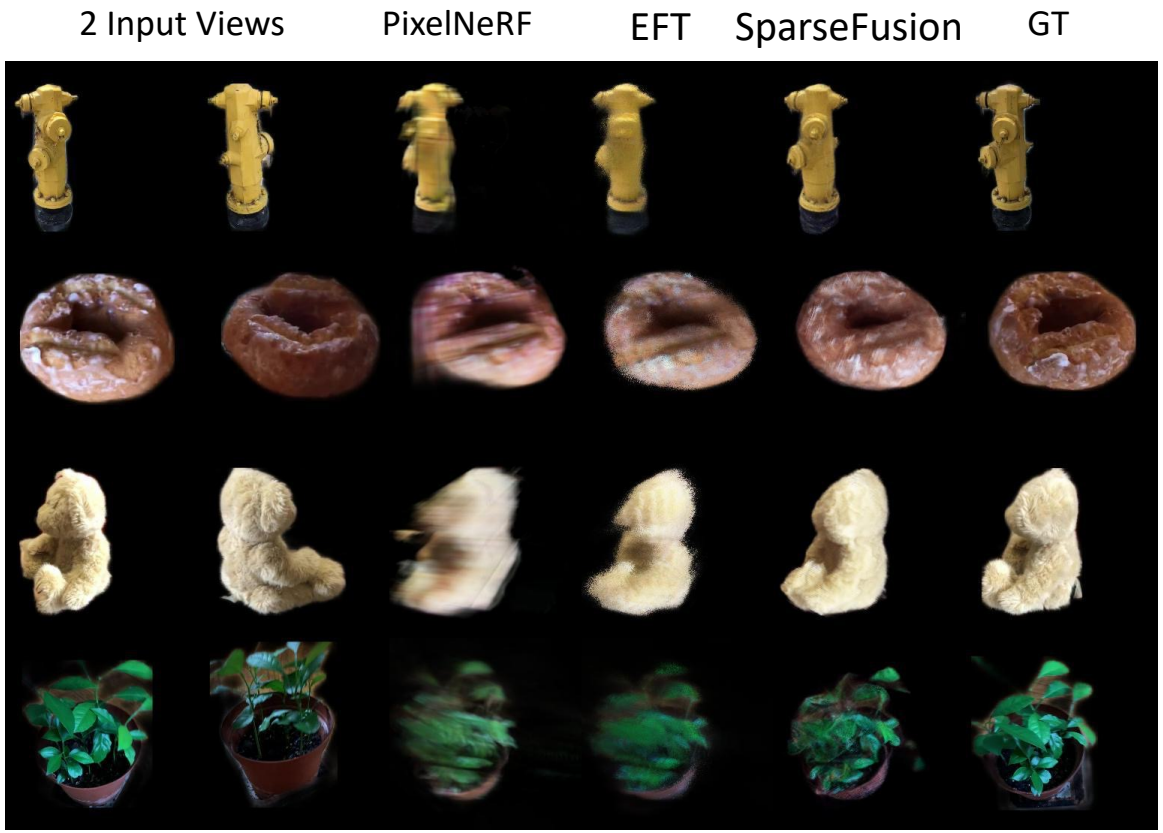

Method
SparseFusion
EFT
PixelNeRF
LPIPS ↓
0.6
0.5
0.4
0.3
0.2
0.1
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


2 Input Views
PixelNeRF
EFT
SparseFusion
GT

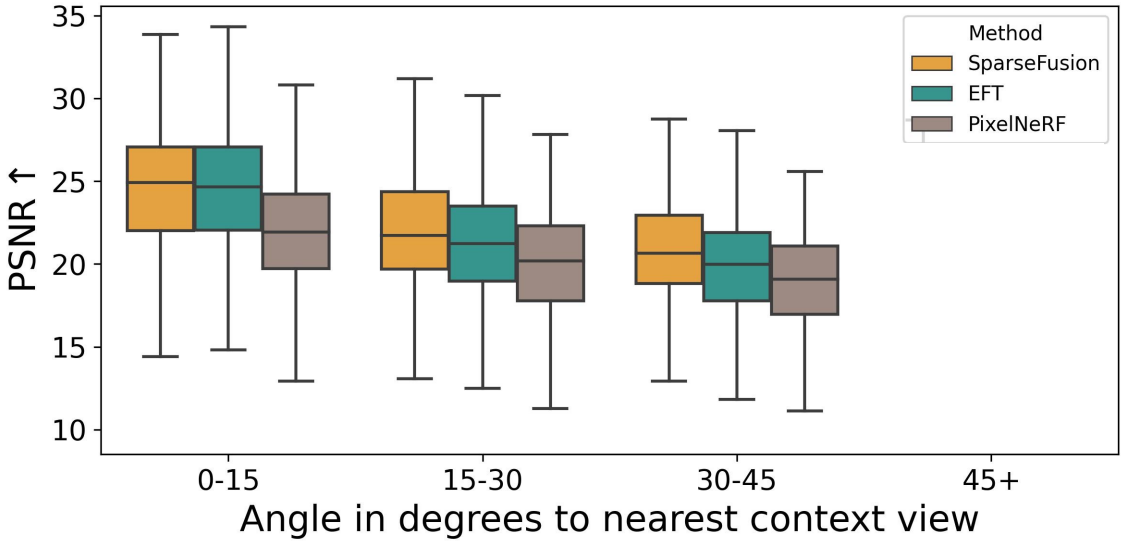
Results

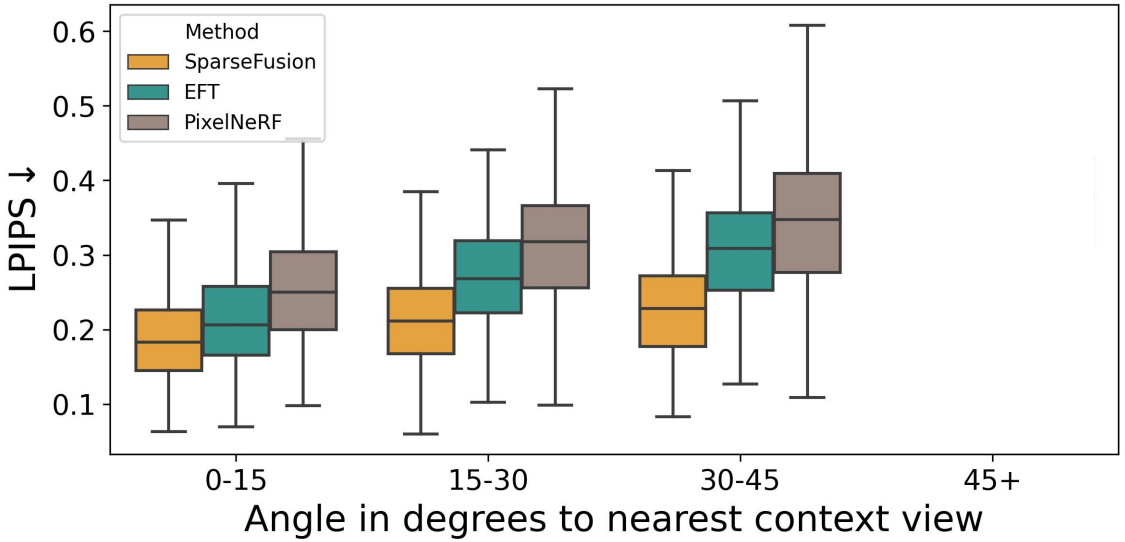

Method
SparseFusion
EFT
PixelNeRF
PSNR ↑
35
30
25
20
15
10
0-15
15-30
30-45
45+
Angle in degrees to nearest context view

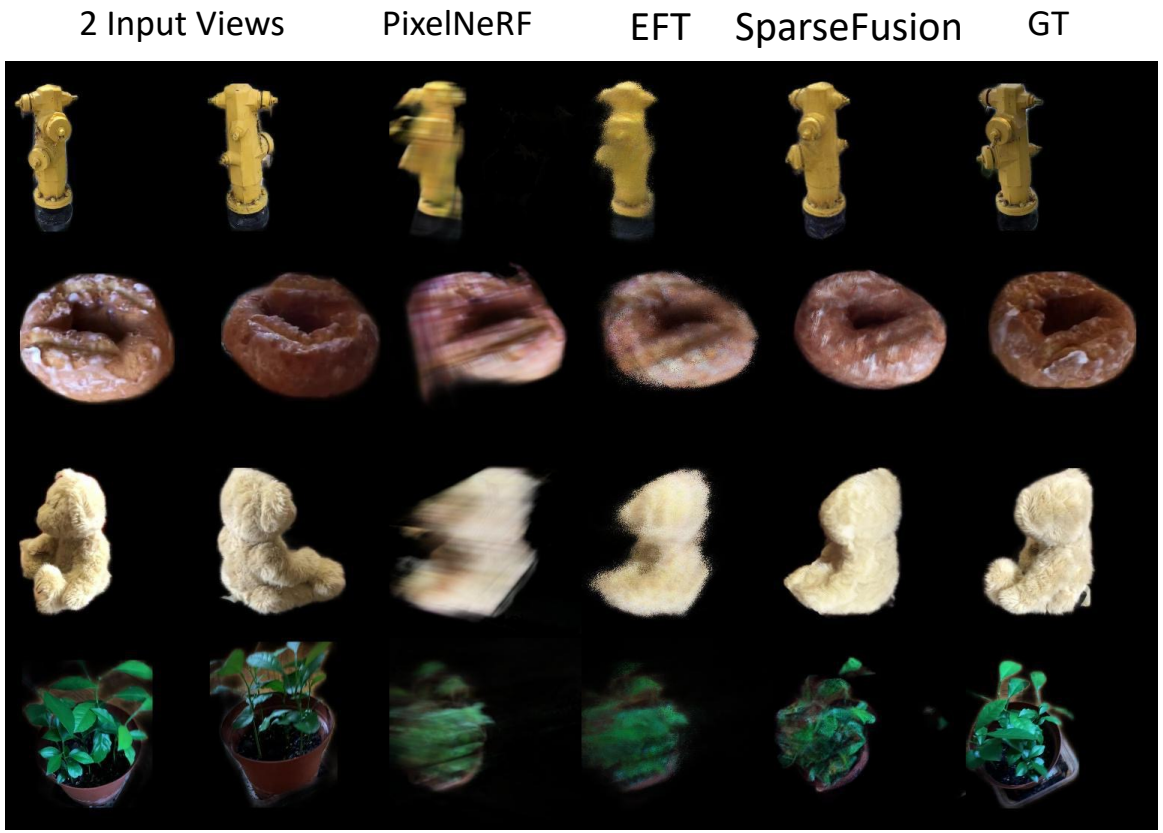

Method
SparseFusion
EFT
PixelNeRF
LPIPS ↓
0.6
0.5
0.4
0.3
0.2
0.1
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


2 Input Views
PixelNeRF
EFT
SparseFusion
GT

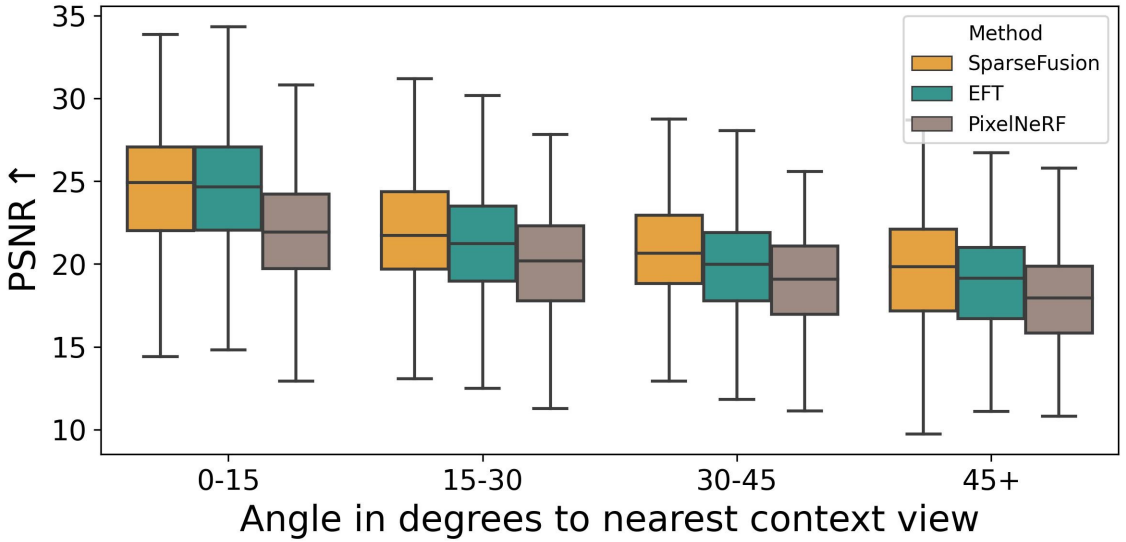
Results

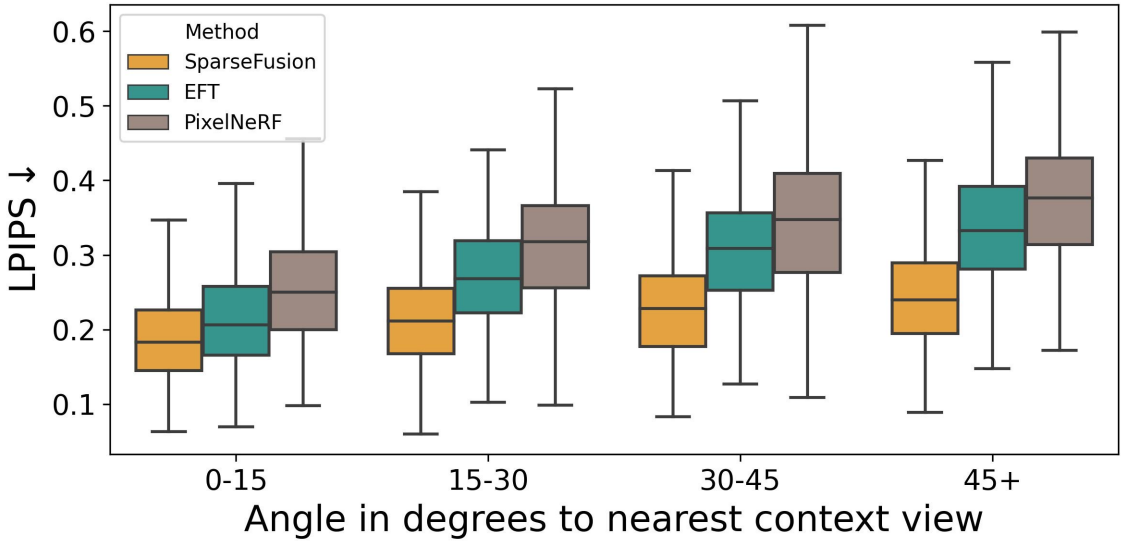

Method
SparseFusion
EFT
PixelNeRF
PSNR ↑
35
30
25
20
15
10
0-15
15-30
30-45
45+
Angle in degrees to nearest context view

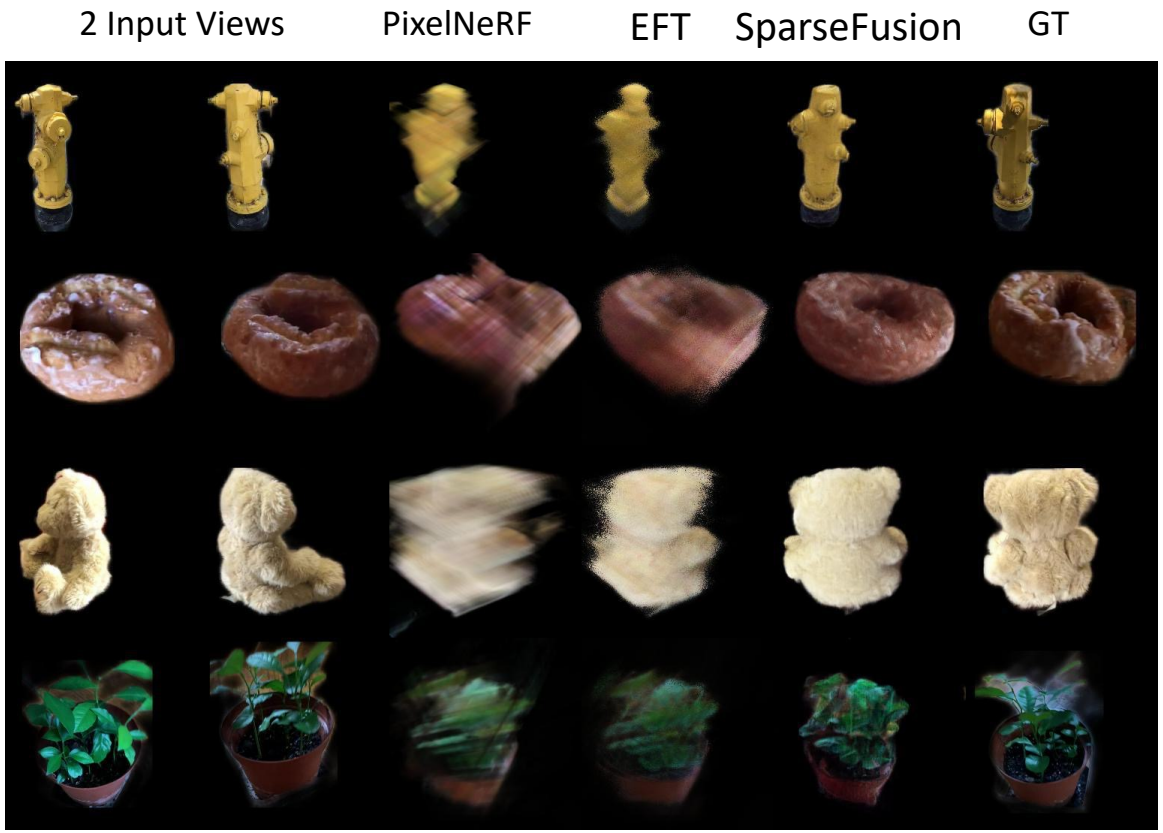

Method
SparseFusion
EFT
PixelNeRF
LPIPS ↓
0.6
0.5
0.4
0.3
0.2
0.1
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


2 Input Views
PixelNeRF
EFT
SparseFusion
GT

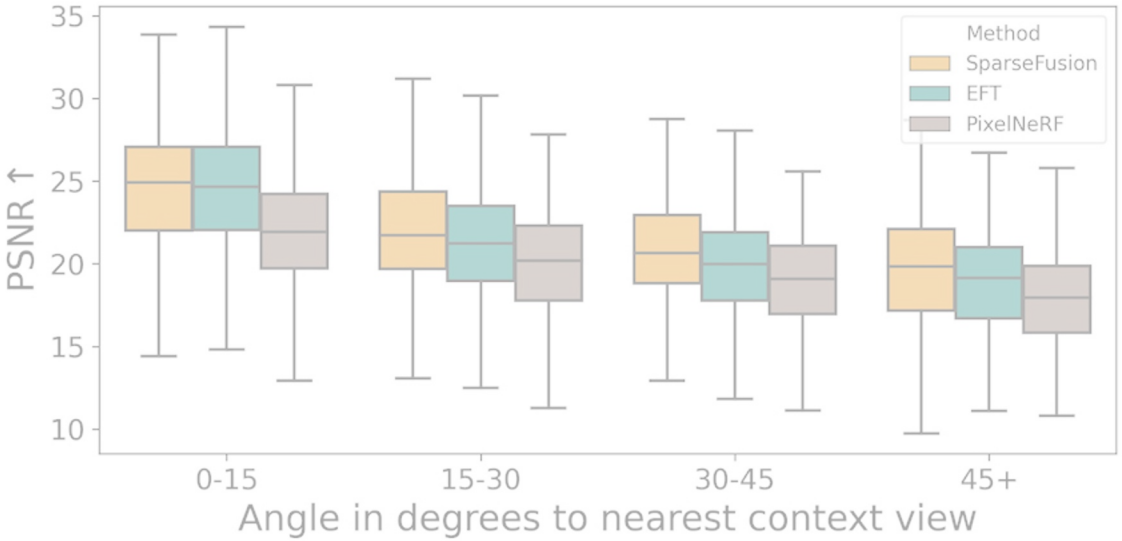
Results

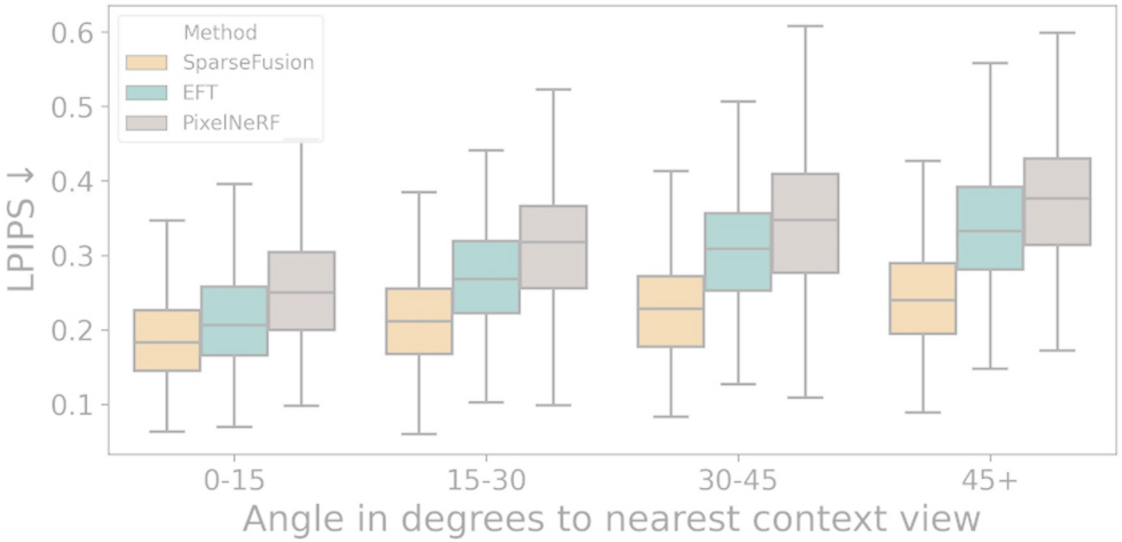

PSNR ↑
35
30
25
20
15
10
Method
SparseFusion
EFT
PixelNeRF
0-15
15-30
30-45
45+
Angle in degrees to nearest context view

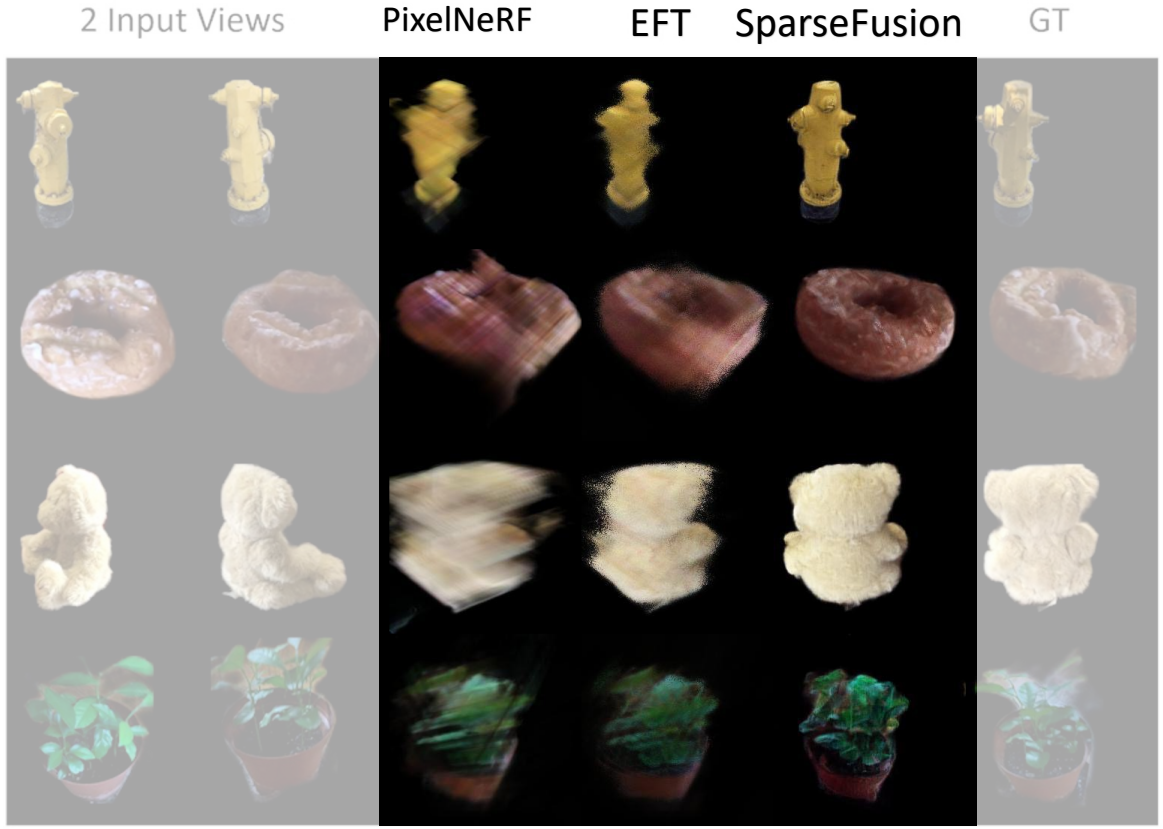

LPIPS ↓
0.6
0.5
0.4
0.3
0.2
0.1
Method
SparseFusion
EFT
PixelNeRF
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


2 Input Views
PixelNeRF
EFT
SparseFusion
GT

Results


PSNR ↑
35
30
25
20
15
10
Method
SparseFusion
EFT
PixelNeRF
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


LPIPS ↓
0.6
0.5
0.4
0.3
0.2
0.1
Method
SparseFusion
EFT
PixelNeRF
0-15
15-30
30-45
45+
Angle in degrees to nearest context view


2 Input Views
PixelNeRF
EFT
SparseFusion
GT

Carnegie Mellon University

sparsefusion.github.io

THE ROBOTICS INSTITUTE

Zhizhuo (Z) Zhou, Shubham Tulsiani