NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild

Jason Y. Zhang Gengshan Yang Shubham Tulsiani* Deva Ramanan*

NeurIPS 2021

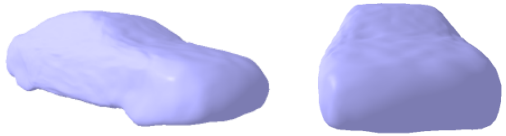
(* denotes equal coding)

Goal: 3D from Sparse Views

Given:

Several Images + Masks
of Same Instance

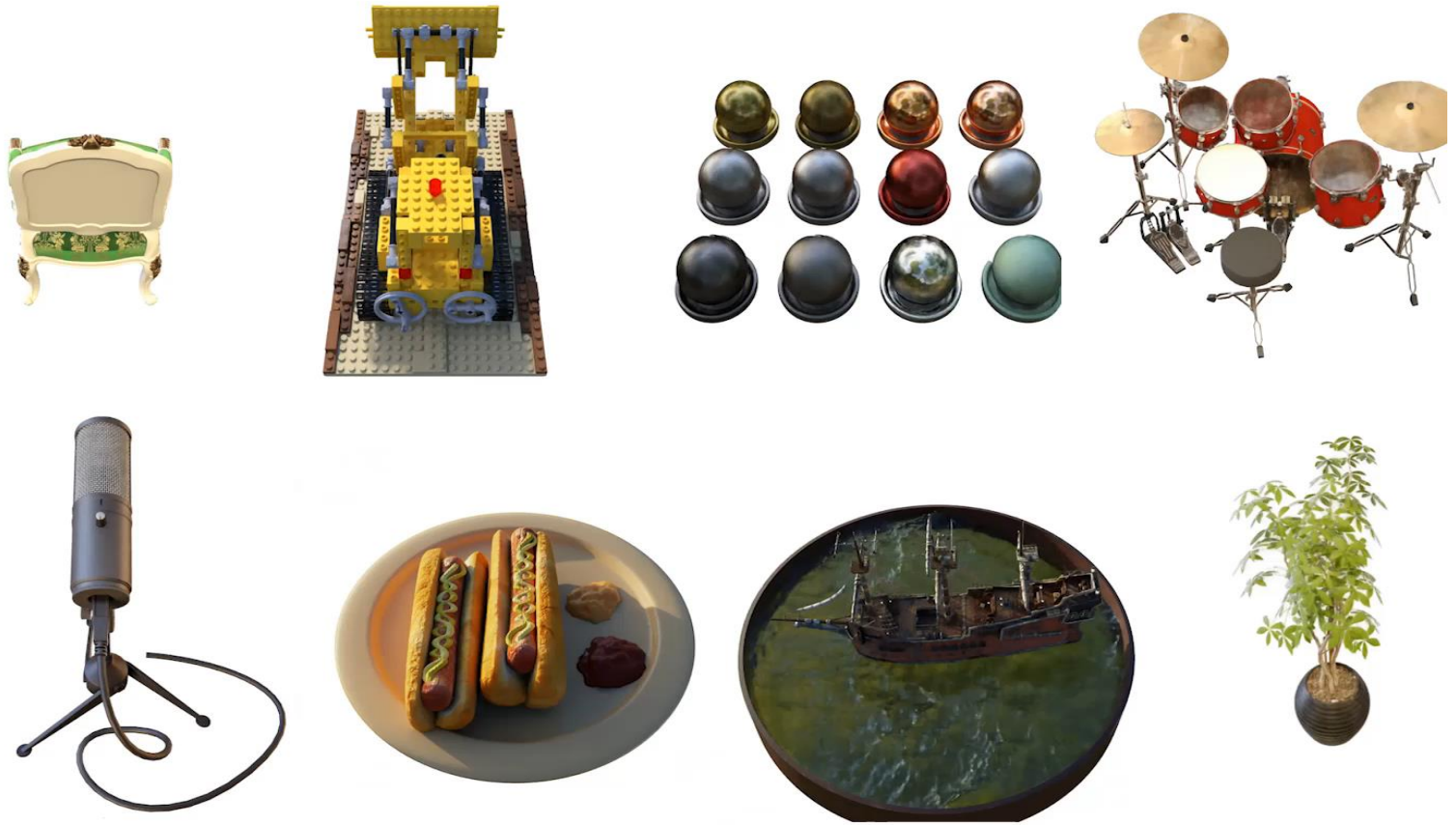
Coarse Initial Mesh +
Coarse Off-the-shelf Poses



Recover:

Textured 3D Reconstruction
w/ Plausible Illumination

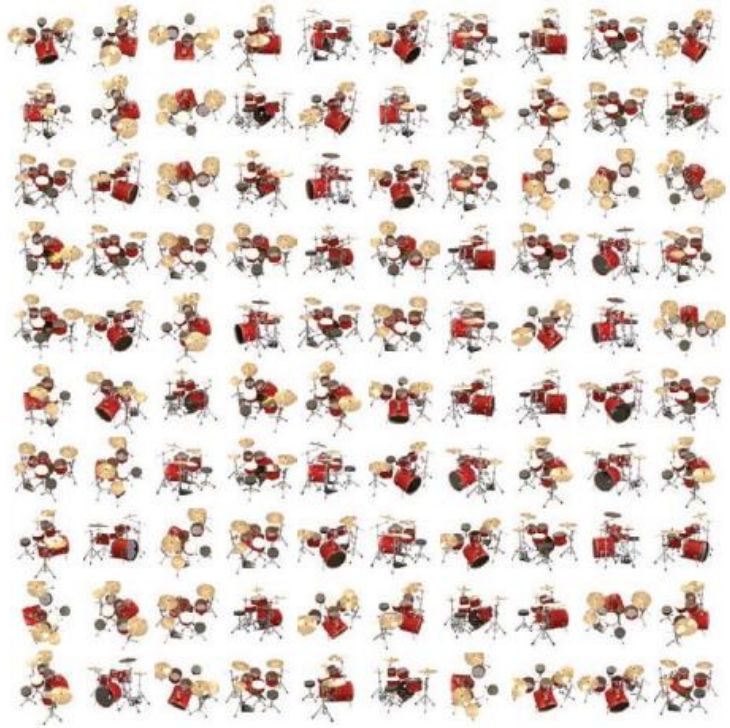
Related Work in Volumetric Rendering: High-fidelity Novel View Synthesis



Mildenhall et al. NeRF. (ECCV 2020)

Bottlenecks for View Synthesis in the Wild

1. Many (50+) Views



2. Precisely Calibrated Camera Poses

NeRF Struggles to Generalize when Trained w/ Sparse Views

Training Images

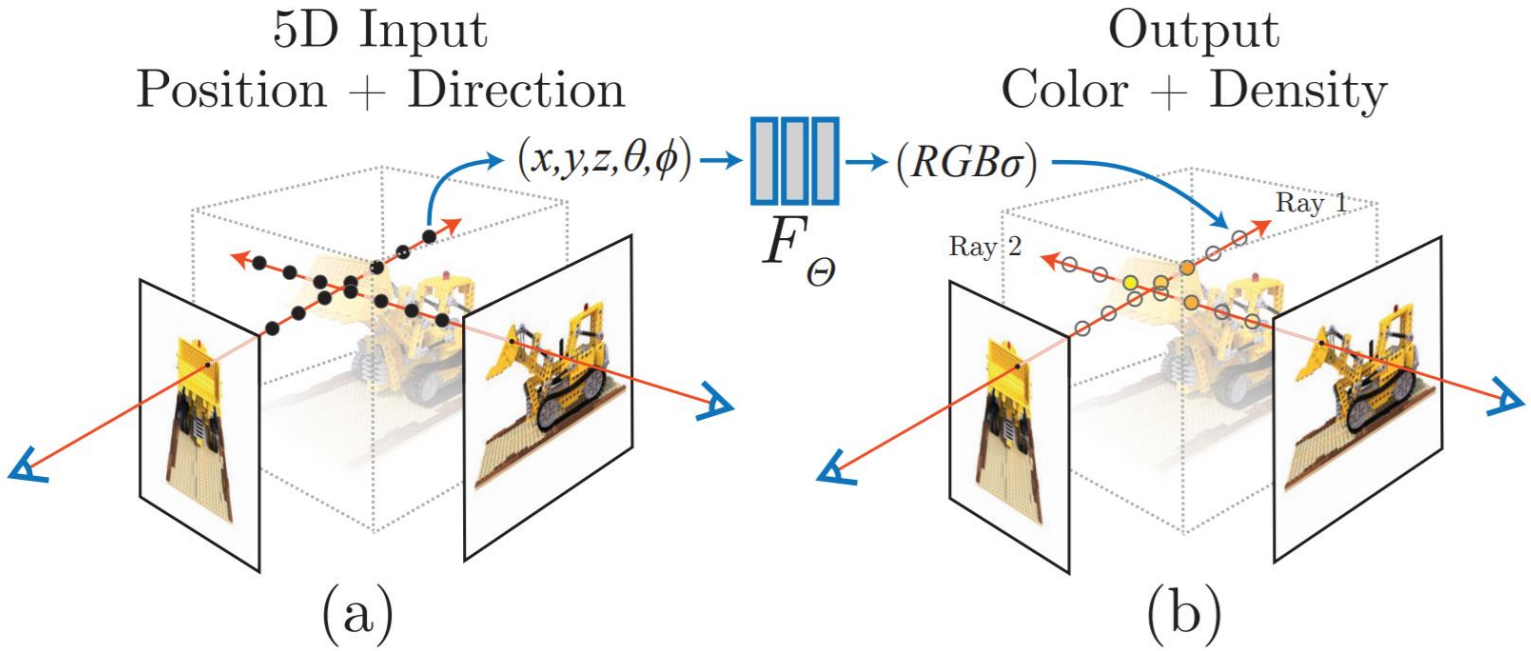
NeRF*

Training Images

NeRF*

Why does NeRF Fail with sparse views?

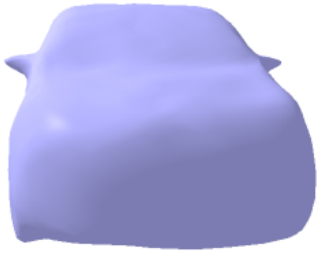
NeRF allows for arbitrary geometry and appearance



NeRS: Neural Reflectance Surfaces

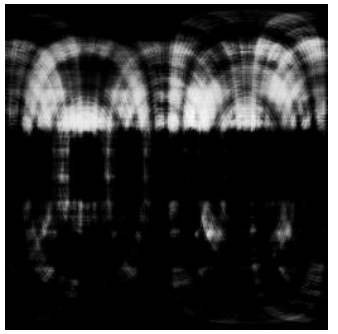
Reference

Shape



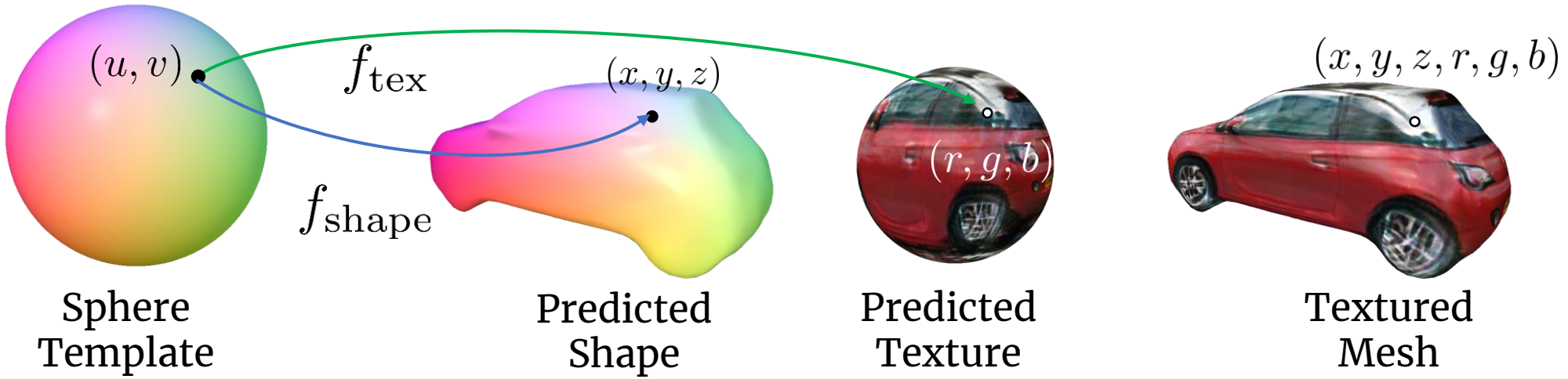
Texture

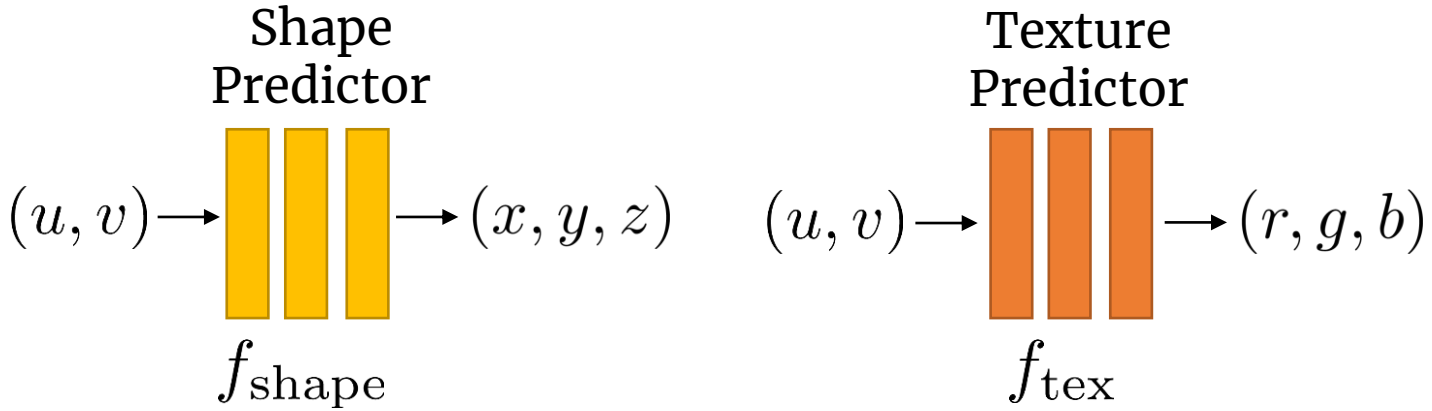
Environment Map



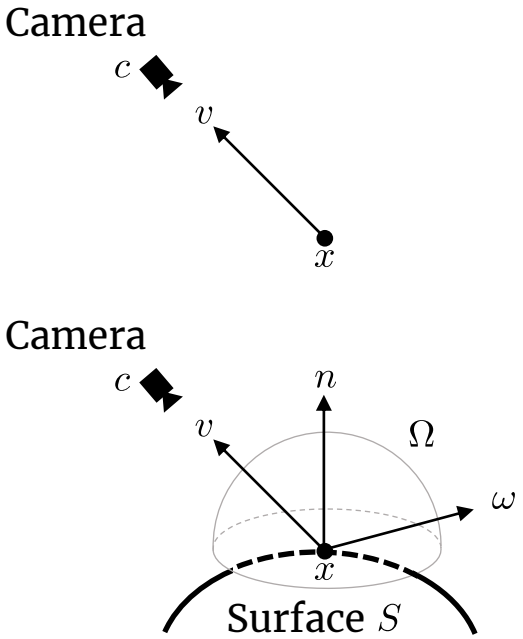
- Insight 1: Objects generally have well-defined surfaces
- Insight 2: View-dependent appearance cannot be arbitrary

Representing Neural Surfaces





Rendering View-dependent Effects

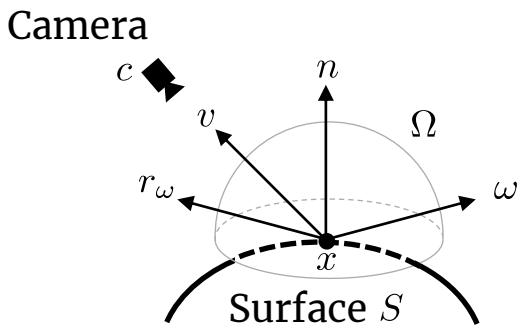


NeRF Solution: $L_o(x, v) = F_\Theta(x, v)$

Rendering Equation:

$$L_o(x, v) = \int_\Omega \underbrace{f_r(x, v, \omega)}_{\text{BRDF}} \underbrace{L_i(x, \omega)}_{\text{Incoming Radiance}} \underbrace{(\omega \cdot n)}_{\text{Cosine Reduction}} d\omega$$

Rendering using Phong Shading



Rendering Equation:

$$L_o(x, v) = \int_\Omega f_r(x, v, \omega) L_i(x, \omega)(\omega \cdot n) d\omega$$

$$\approx T(x) I_{\text{diffuse}}(x) + k_s I_{\text{specular}}(x, v)$$

Specular Coefficient

$$I_{\text{diffuse}}(x) = \sum_{\omega \in \Omega} (\omega \cdot n) L_i(\omega)$$

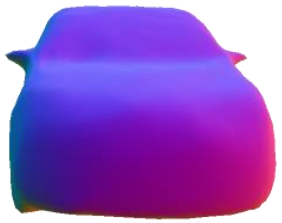
Shininess

$$I_{\text{specular}}(x, v) = \sum_{\omega \in \Omega} (r_{\omega,n} \cdot v)^\alpha L_i(\omega)$$

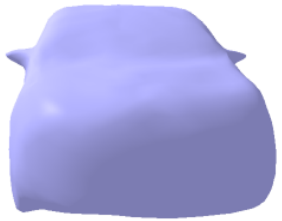
Environment Map:

$$L_i(x, \omega) \equiv L_i(\omega) = f_{\text{env}}(\omega)$$

Surface-based Illumination



Normals
(n)

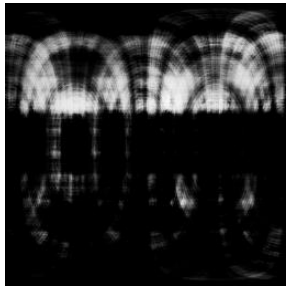


Shape
(f_{shape})

Diffuse Lighting
(I_{diffuse})

Texture
(f_{tex})

View Indep.
$(T \odot I_{\text{diffuse}})$



Env. Map
(f_{env})

Specular Lighting
(I_{specular})

Reference View
(I)

Output Radiance
(L_o)

$$L_{\text{perceptual}} + L_{\text{mask}} + L_{\text{regularize}}$$

Qualitative Results

NeRS on Everyday Objects

Training View	Initial Mesh	Output	Training View	Initial Mesh	Output
					
					

NeRS Recovers 3D at Scale (from Online Marketplace Images)



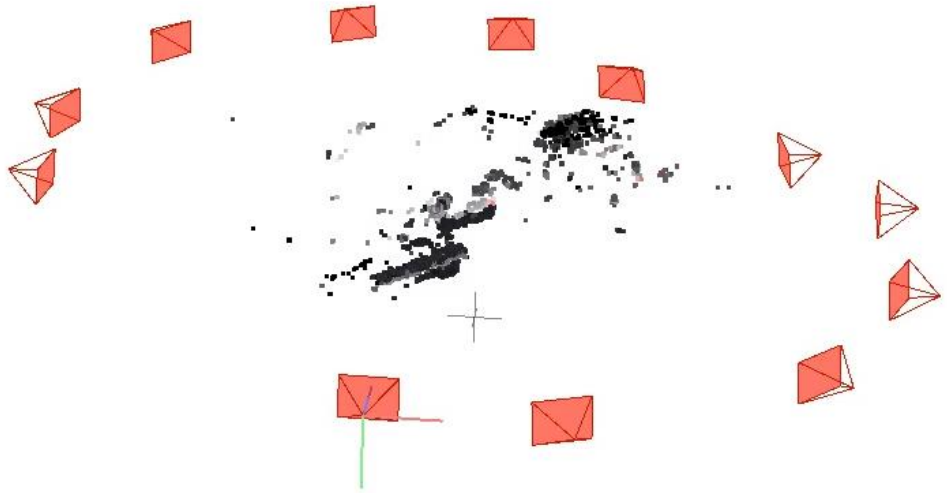




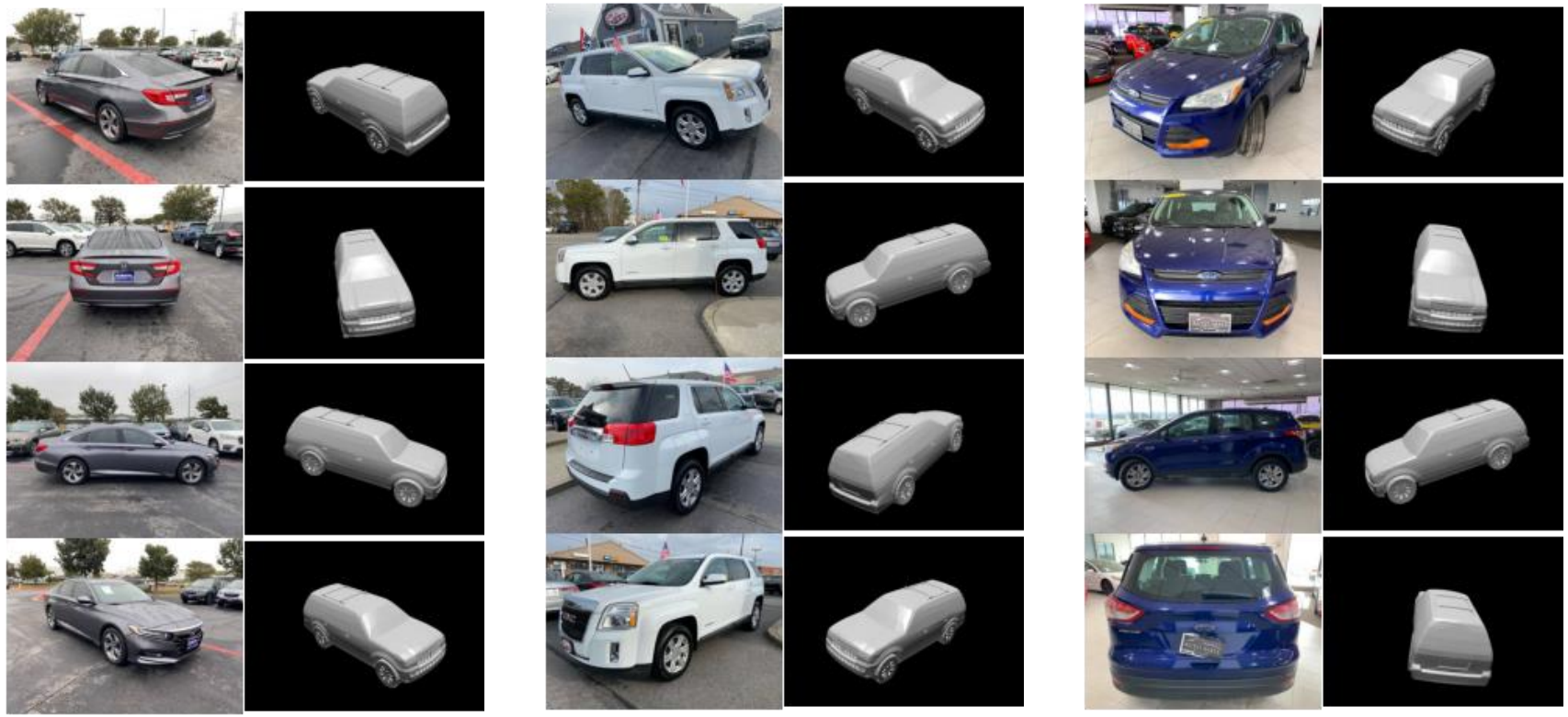

Evaluation

Challenge: Evaluation w/out GT Cameras

- Novel View Synthesis evaluation requires GT poses
- COLMAP fails to recover meaningful poses and reconstructions given wide-baseline inputs:



Approx. Off-the-Shelf Camera Poses



Xiao et al. PoseFromShape. (BMVC 2019)

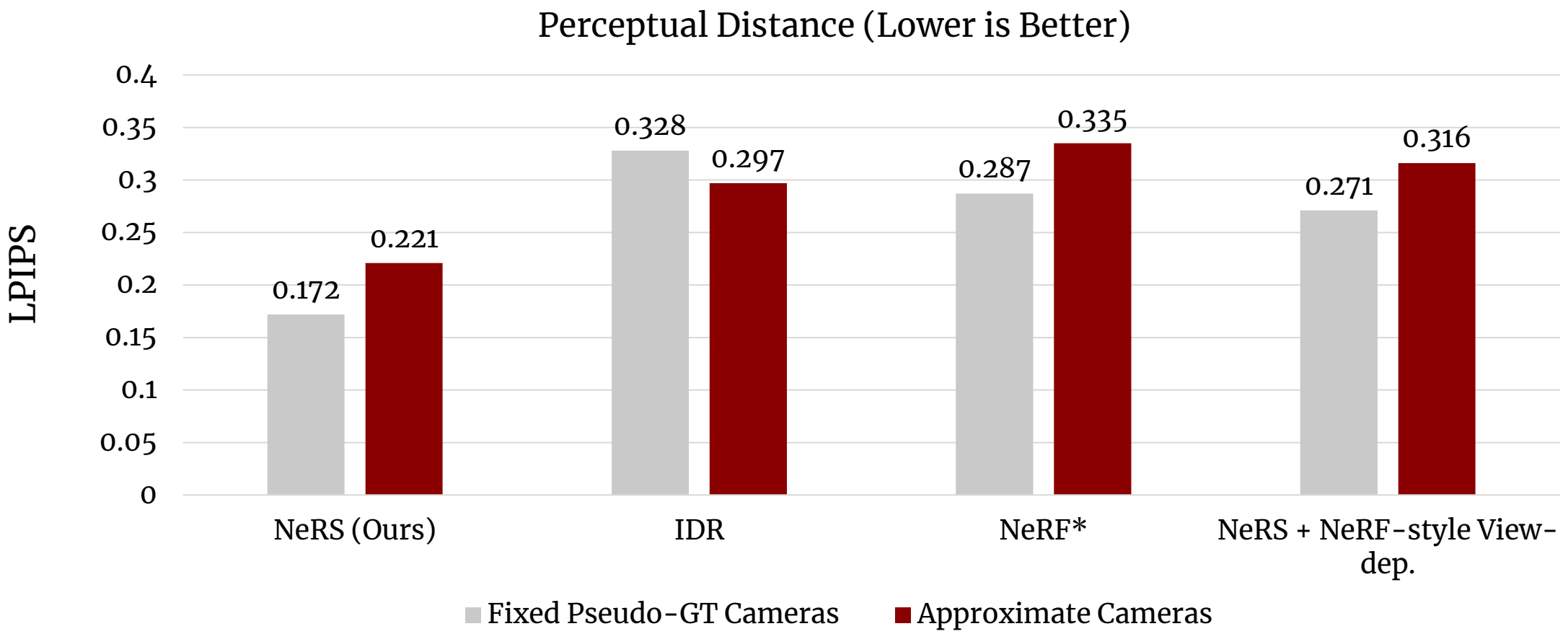
Evaluation with Fixed Pseudo-GT Poses

- Manually correct camera poses jointly optimized over all images in an instance

Evaluation with Approximate OTS Poses

- Cameras can be *refined* during both training and testing

Quantitative Evaluation







Thanks for Watching!

Project webpage: jasonyzhang.com/ners