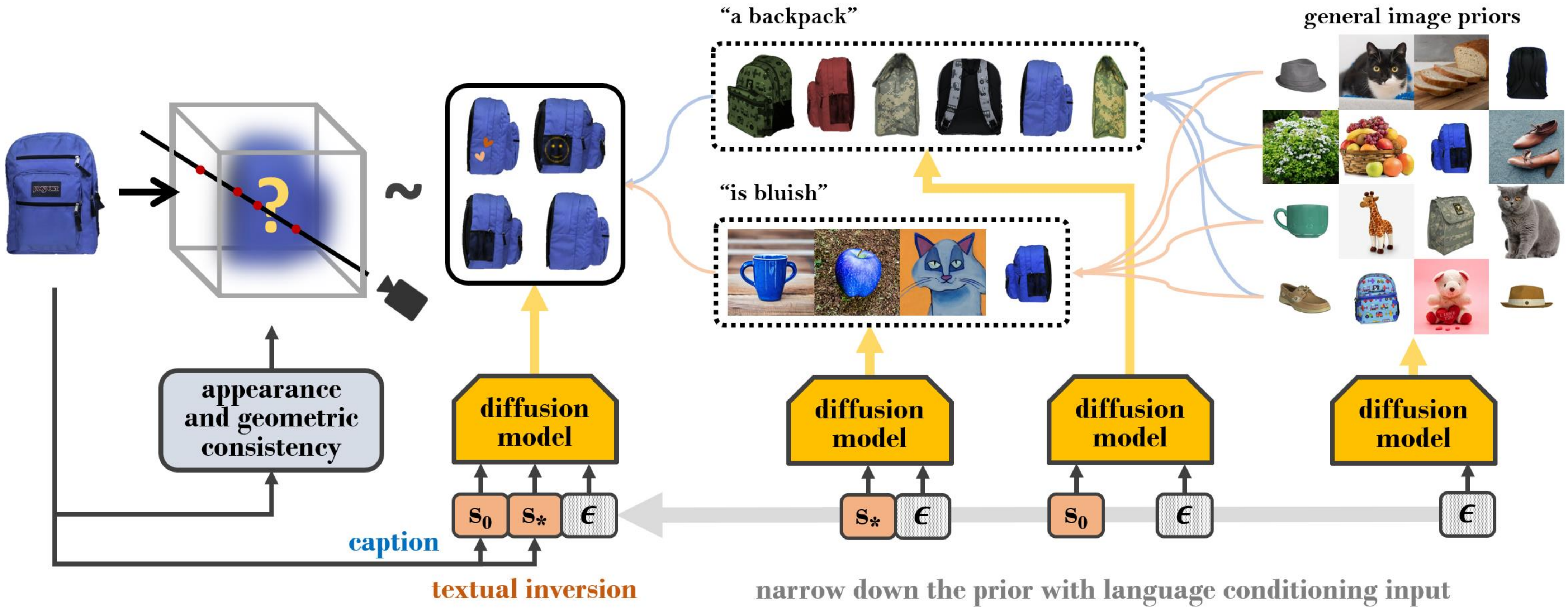
NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors

Congyue Deng[2], Max “Chiyu” Jiang[1], Charles R. Qi[1], Xinchen Yan[1], Yin Zhou[1], Leonidas Guibas[2,3], Dragomir Anguelov[1]

[1]Waymo **[2]Stanford University** **[3]Google Research**

Overview

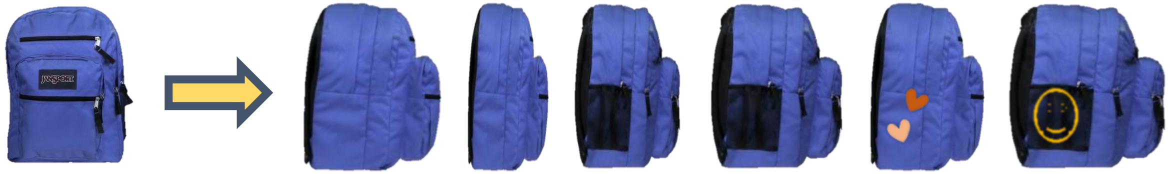

"a backpack"
general image priors
"is bluish"
~
?
appearance and geometric consistency
diffusion model
diffusion model
diffusion model
diffusion model
s_0
s_*
ϵ
s_*
ϵ
s_0
ϵ
ϵ
caption
textual inversion
narrow down the prior with language conditioning input

Intuitions

How does the **side** of this **backpack** look like?

A standard single-view 2D-to-3D reconstruction task

Supervised learning?



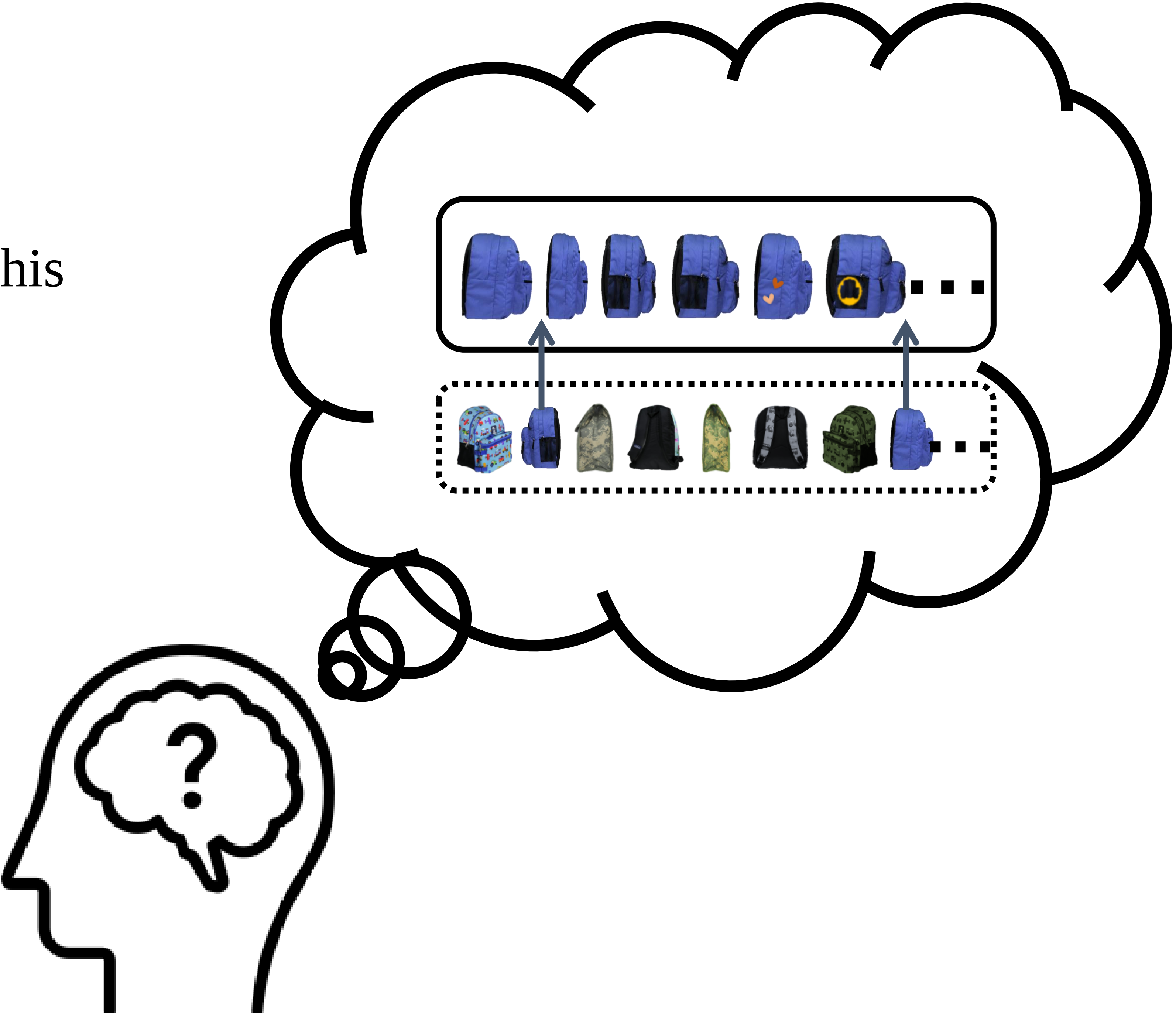
How do we train a network to answer such questions?

In-the-wild images

Non-deterministic answer

Intuitions

How does the **side** of this **backpack** look like?



Problem Formulation

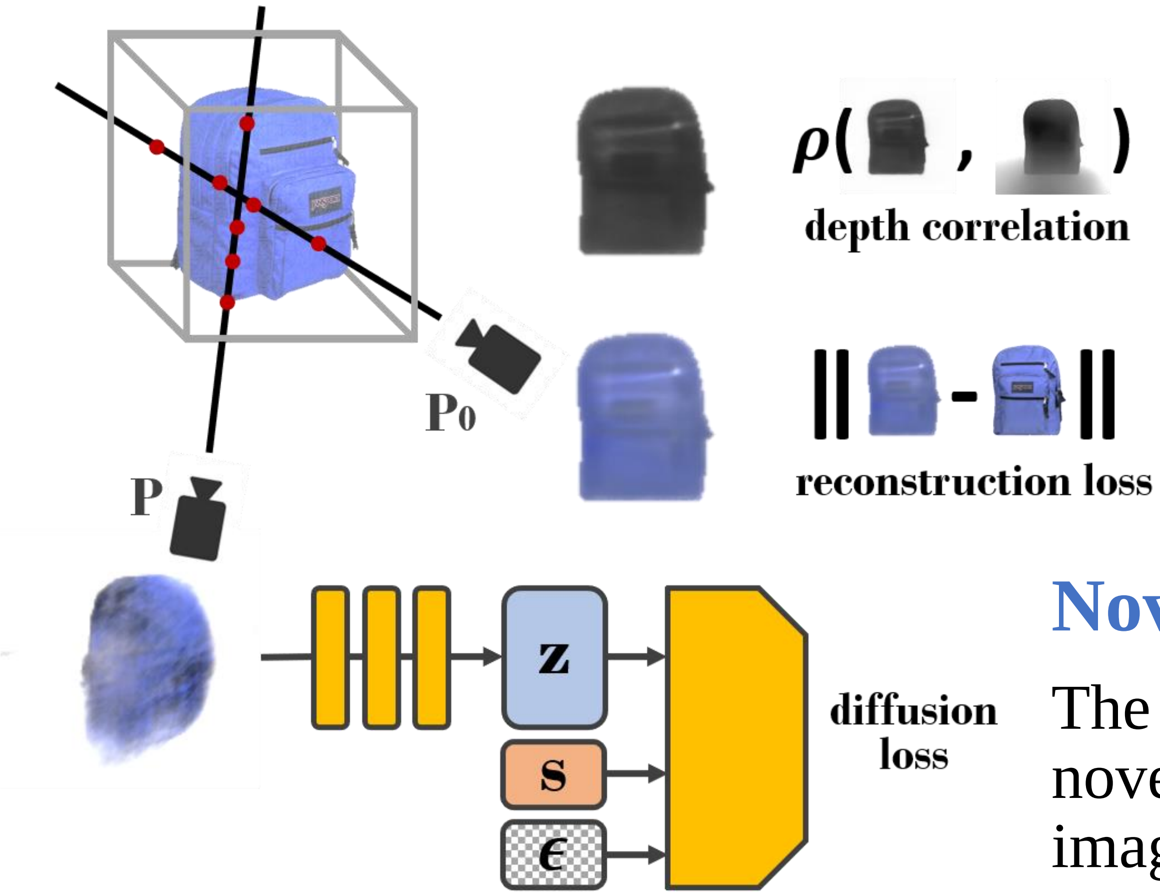
Formulate 2D-to-3D inference as conditioned generation

$$f(\cdot, \omega) \sim \text{3D scene distribution} \mid f(\mathbf{P}_0, \omega) = \mathbf{x}_0$$

3D scene distribution? 2D image distribution!

$$\forall \mathbf{P},\ f(\mathbf{P}, \omega) \sim \mathbb{P} \mid f(\mathbf{P}_0, \omega) = \mathbf{x}_0$$

Method



Input view constraints

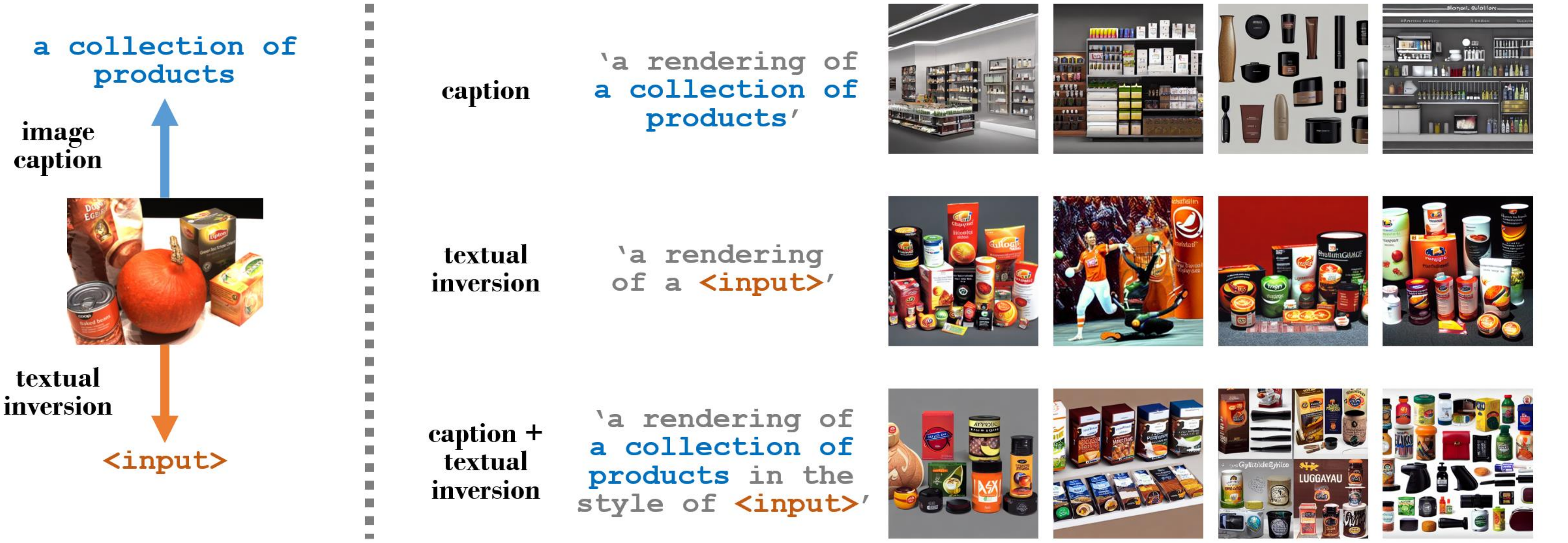
The rendering at the input view should be identical to the input image

Novel view distribution loss

The renderings at randomly sampled novel views should follow the 2D image prior

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{x}),\mathbf{s},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta(\mathbf{z}_t,t,\boxed{c_\theta(\mathbf{s})})\|_2^2\right]$$

Method: 2-Section Semantic Conditions



a collection of products

image caption

textual inversion

<input>

caption

'a rendering of a collection of products'

textual inversion

'a rendering of a <input>'

caption + textual inversion

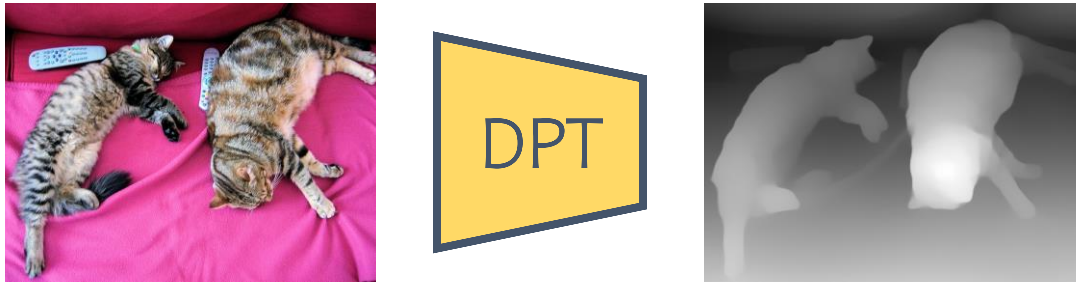
'a rendering of a collection of products in the style of <input>'

Method: Geometric Regularization

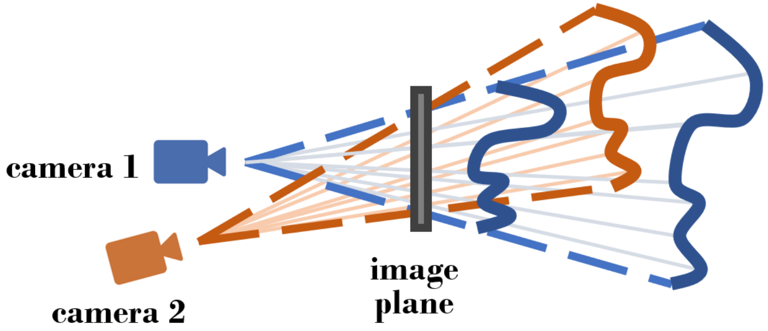
Render depth map from NeRF at the input view

$$\hat{\mathbf{d}}_0 = \int_{t_n}^{t_f} \sigma(t)\mathrm{d}t$$

Regularize it with a monocular depth estimation network



Scale uncertainties and inaccuracies of estimated depth:



Pearson correlation

$$\rho\left(\hat{\mathbf{d}}_0, \mathbf{d}_{0,\mathrm{est}}\right) = \frac{\mathrm{Cov}(\hat{\mathbf{d}}_0, \mathbf{d}_{0,\mathrm{est}})}{\sqrt{\mathrm{Var}(\hat{\mathbf{d}}_0)\mathrm{Var}(\mathbf{d}_{0,\mathrm{est}})}}$$

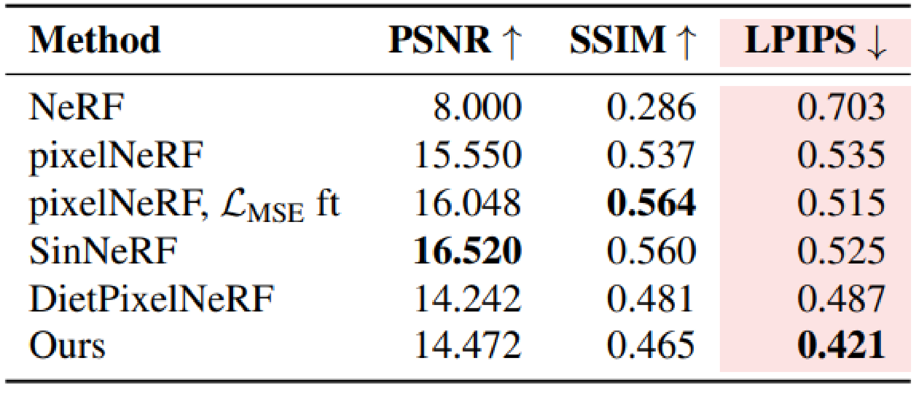
Results: DTU MVS Dataset

LPIPS (perception metric):

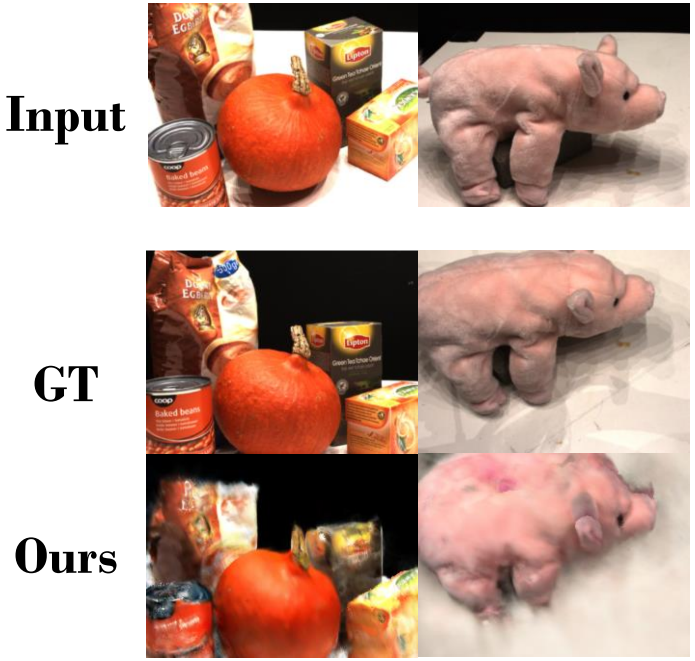
- Great improvement compared to prior methods

PSNR & **SSIM** (pixel-aligned similarity metric):

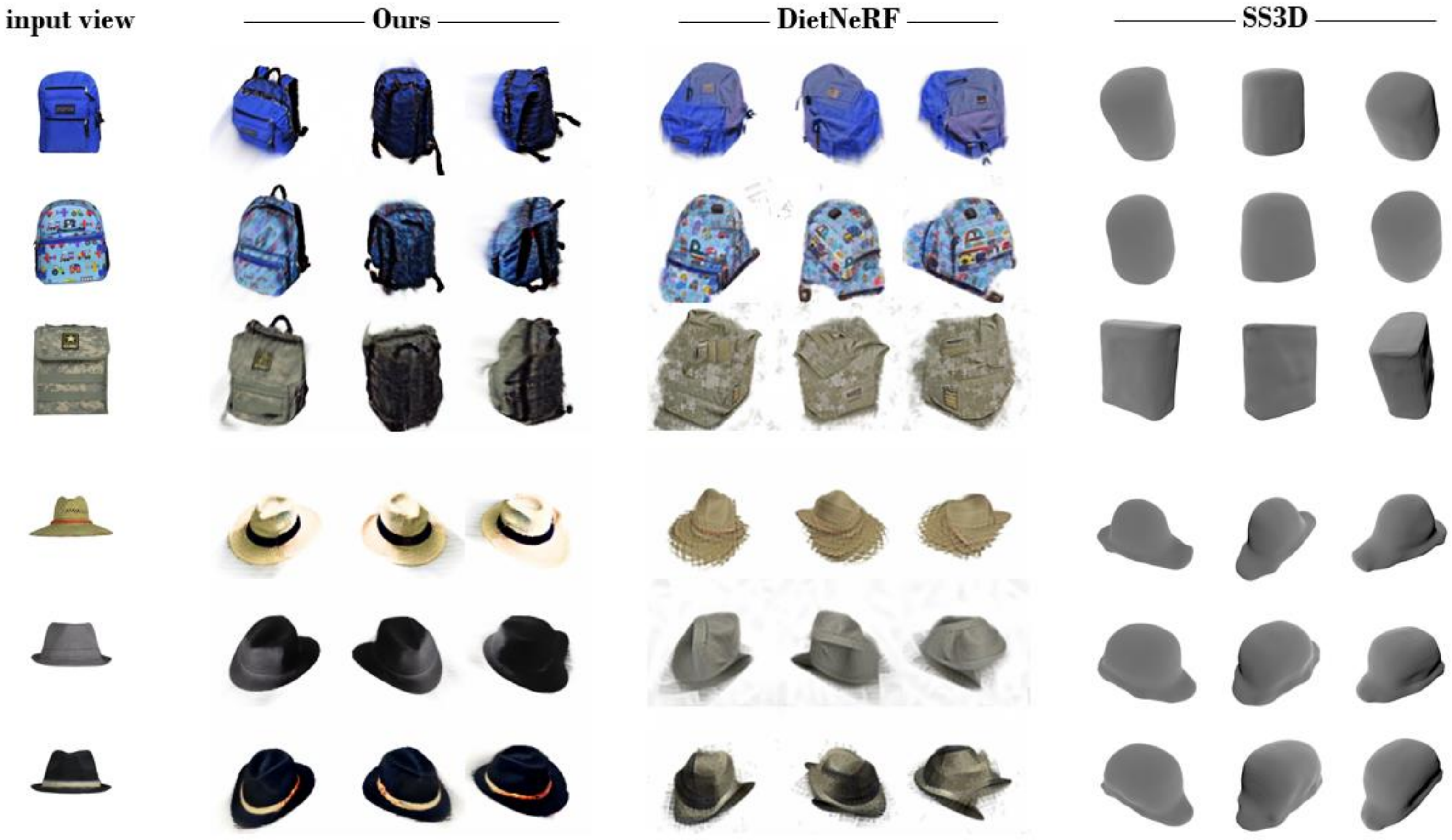
- Slightly lower than pixelNeRF
- On par with DietPixelNeRF
- Less indicative because of the 2D-3D ambiguity



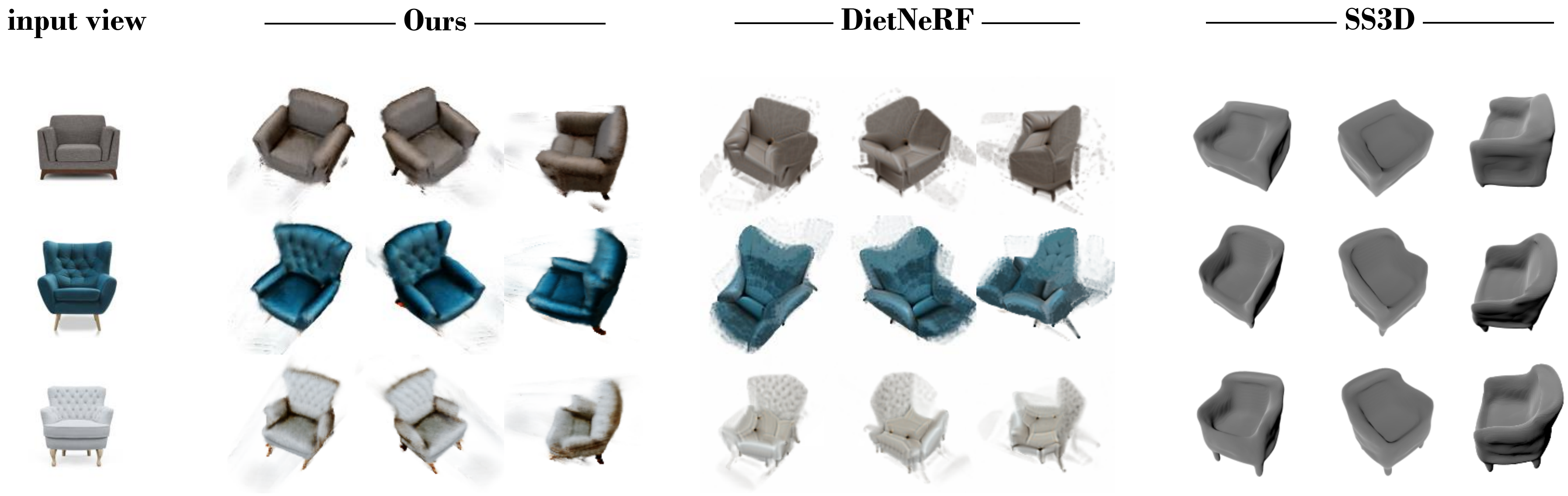
Method	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF	8.000	0.286	0.703
pixelNeRF	15.550	0.537	0.535
pixelNeRF, $\mathcal{L}_{\text{MSE}}$ ft	16.048	**0.564**	0.515
SinNeRF	**16.520**	0.560	0.525
DietPixelNeRF	14.242	0.481	0.487
Ours	14.472	0.465	**0.421**



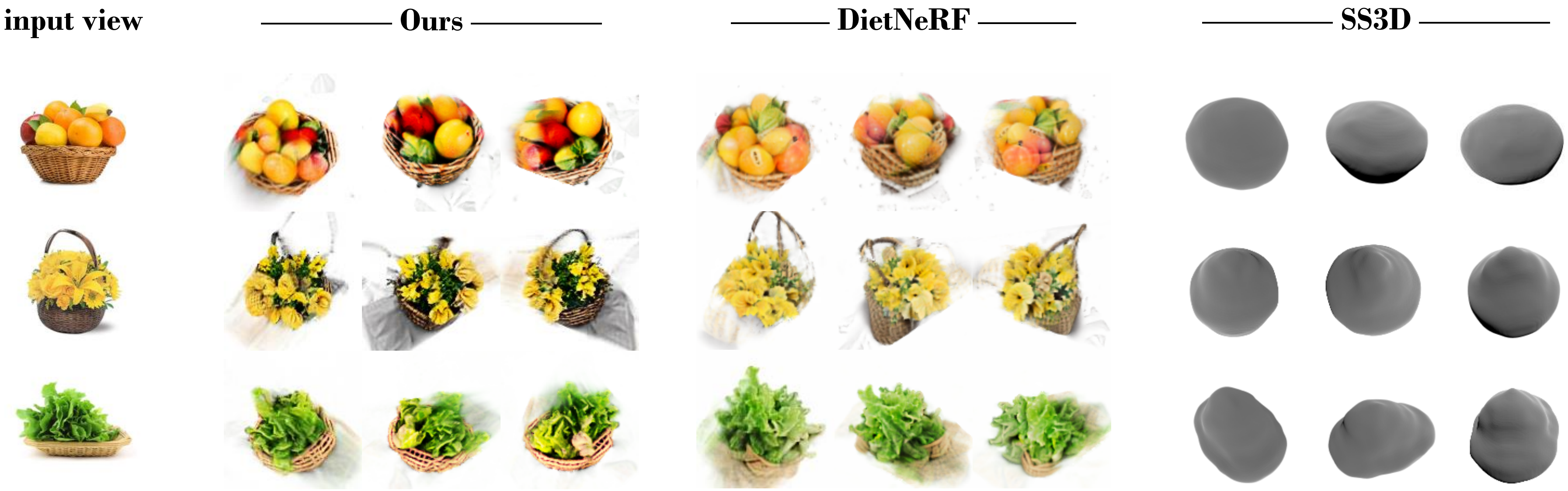
Results: Google Scanned Objects



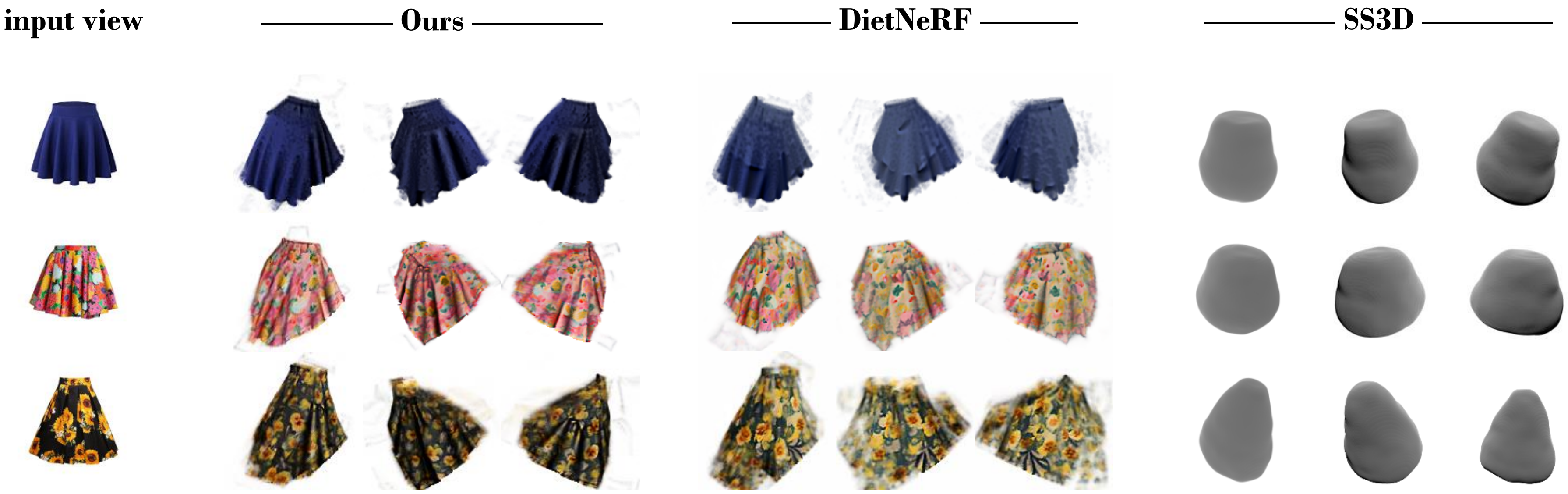
Results: Images from the Internet



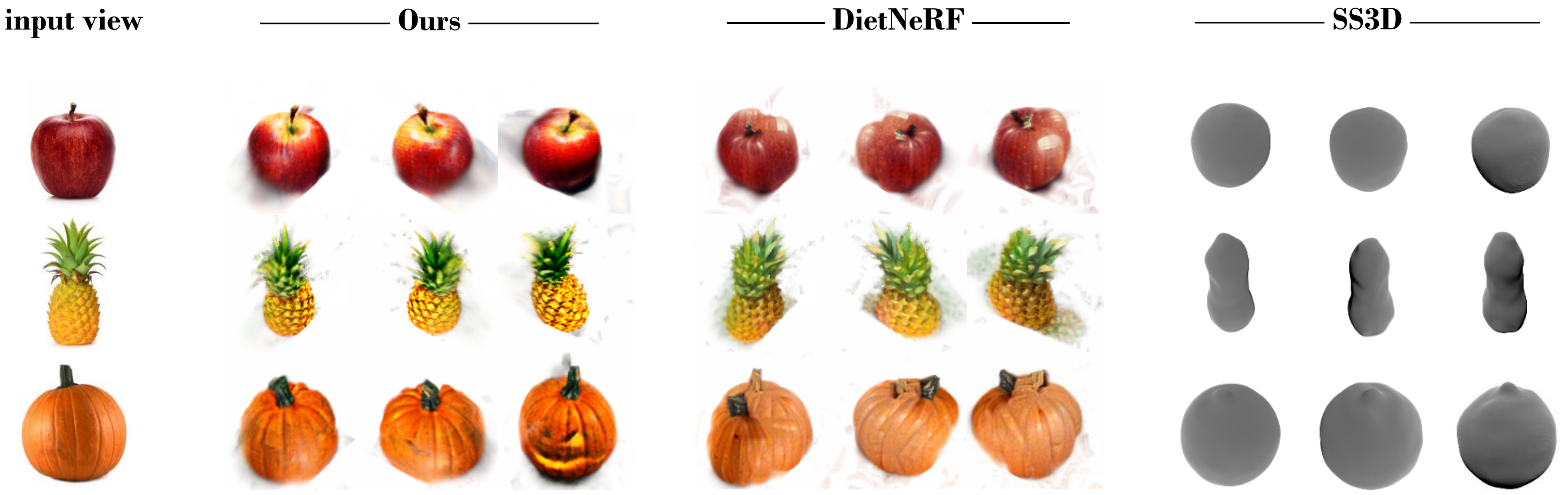
Results: Images from the Internet



Results: Images from the Internet

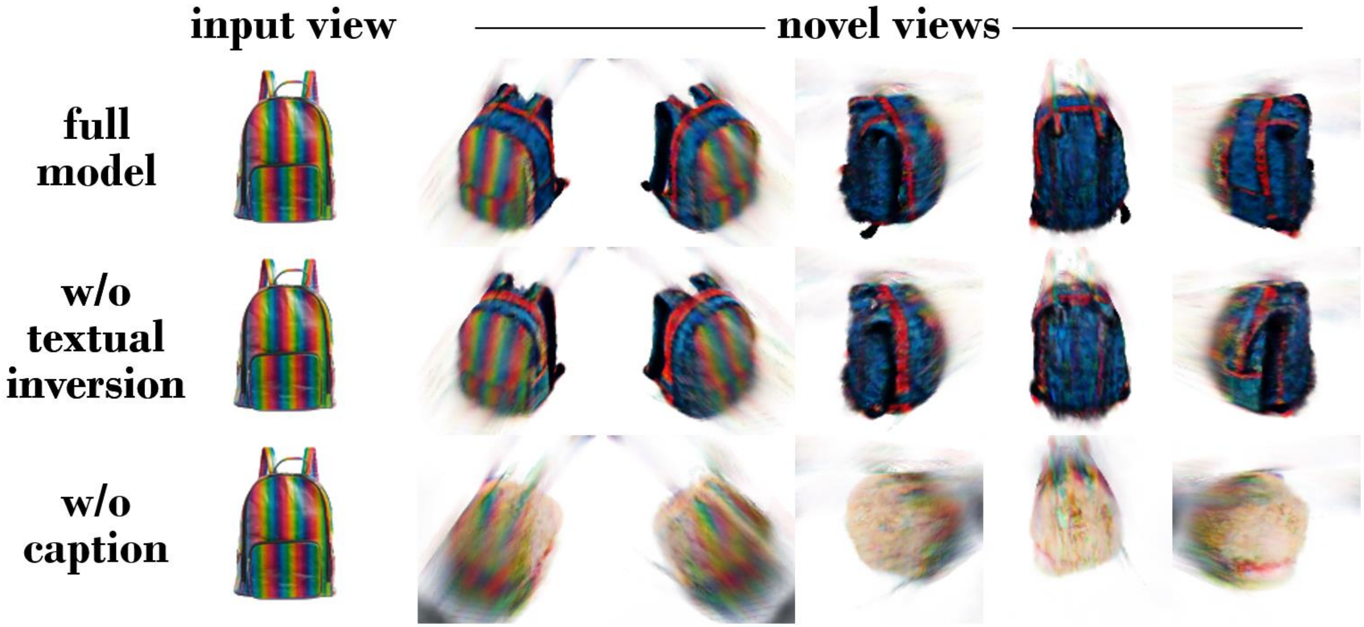


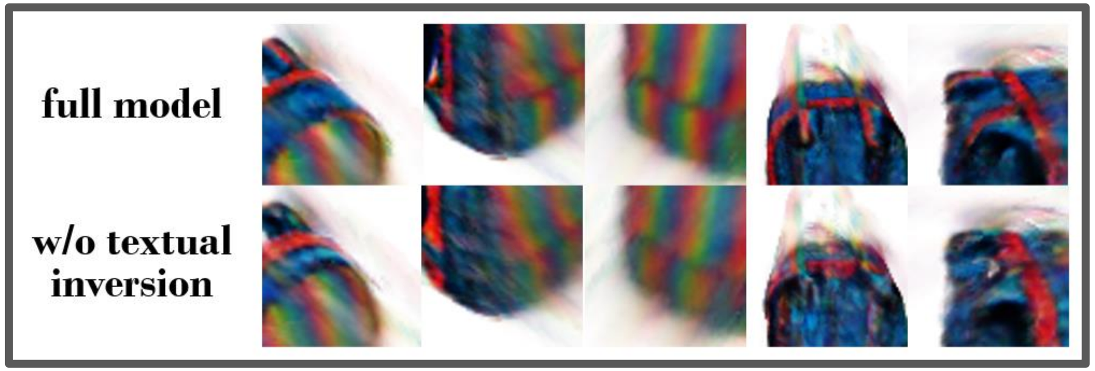
Results: Images from the Internet



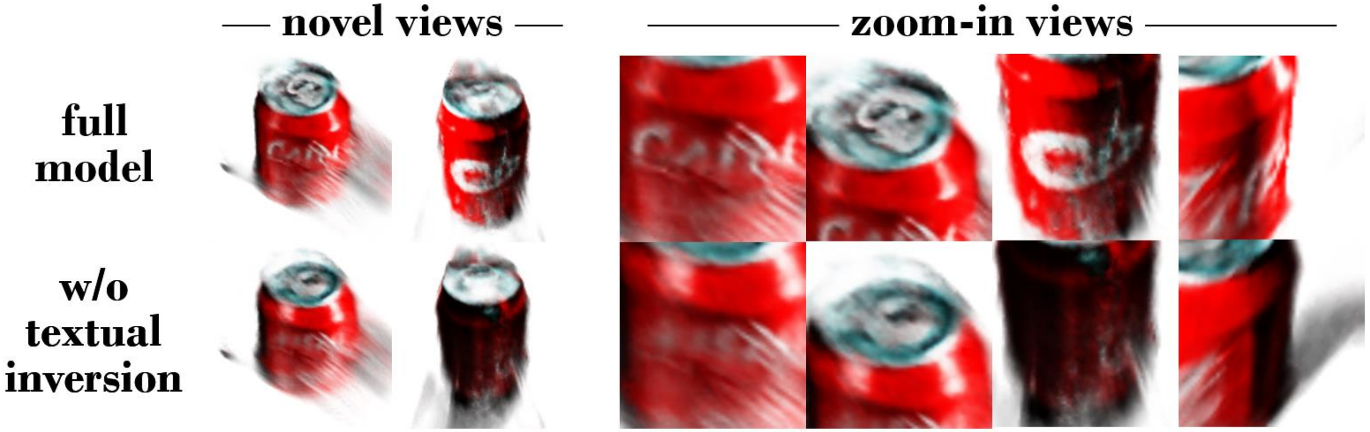
“a pumpkin”

Results: Ablation on Semantic Features

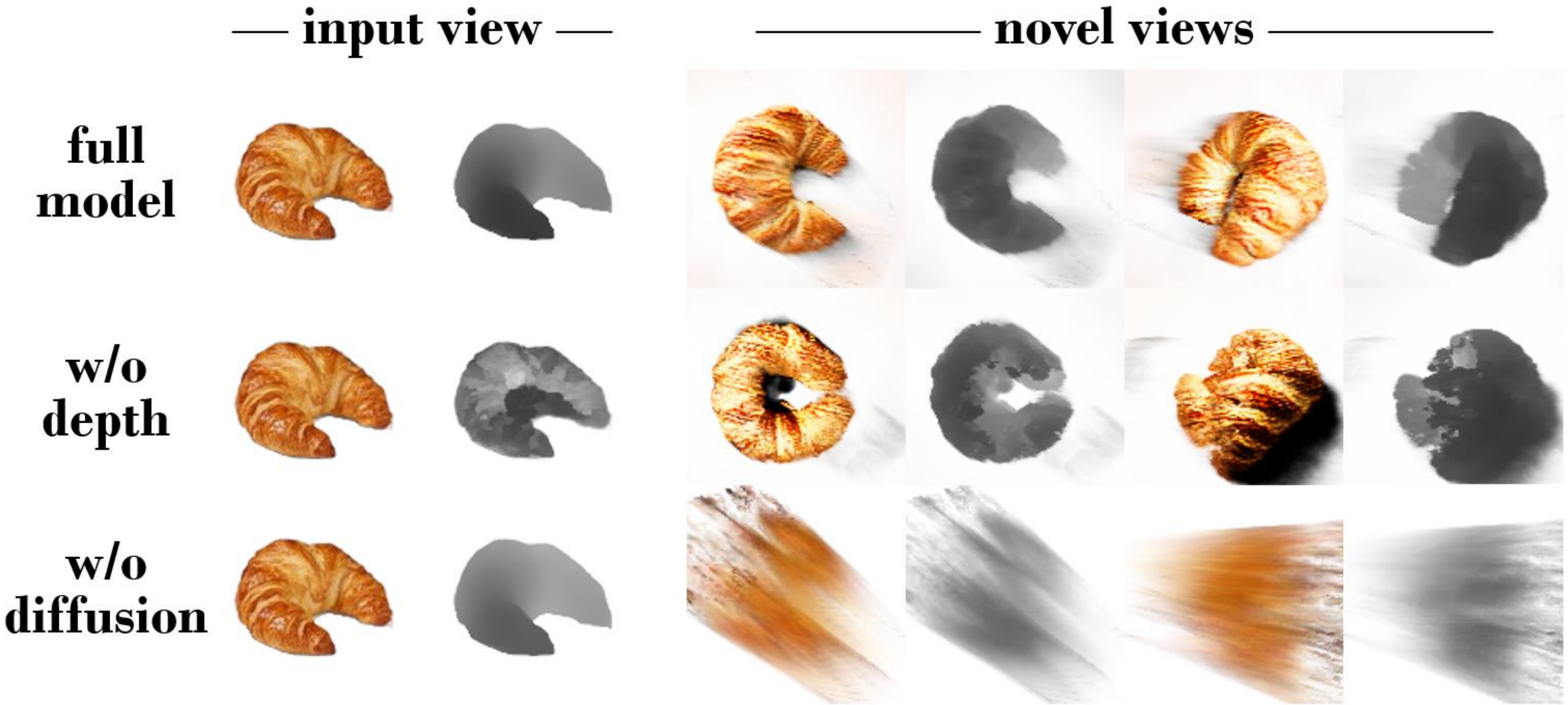




Results: Ablation on Semantic Features

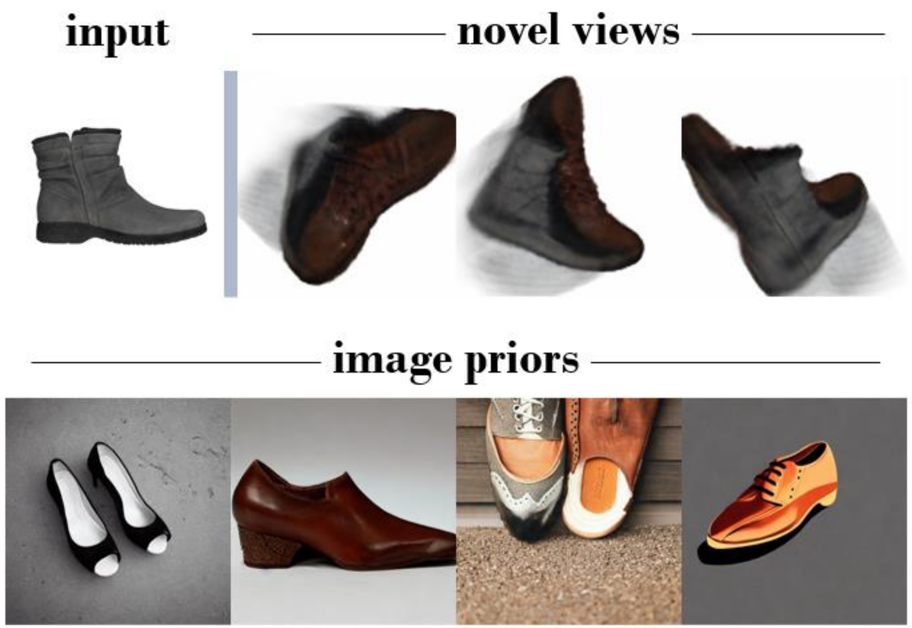


Results: Ablation on Depth Reg.



Results: Failure Cases

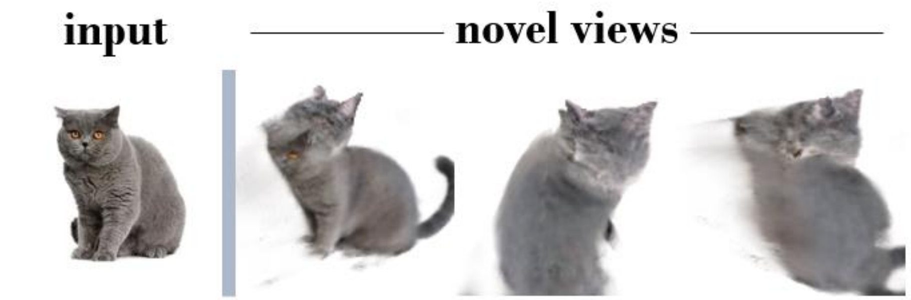
Be cautious of the biases in large models!



“a shoe”

Highly deformable instances

—— varying states cannot be easily captured with a simple language embedding



More Results

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

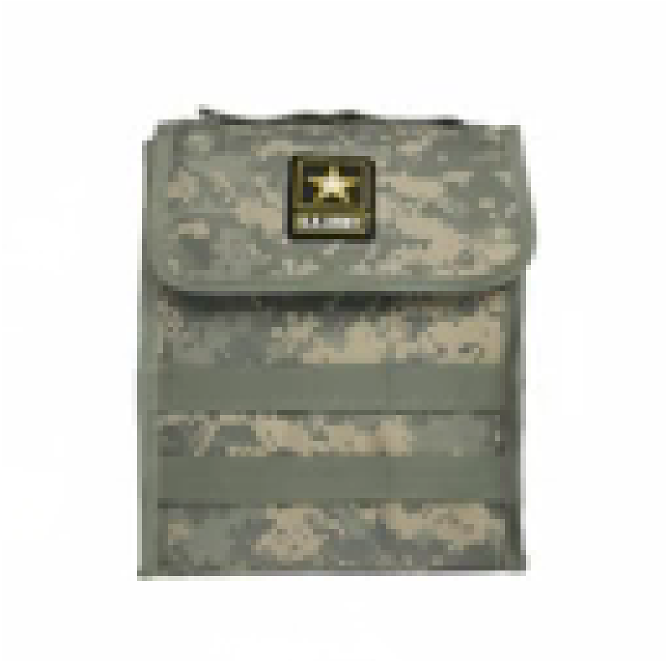
DietNeRF
(+depth)

Input view

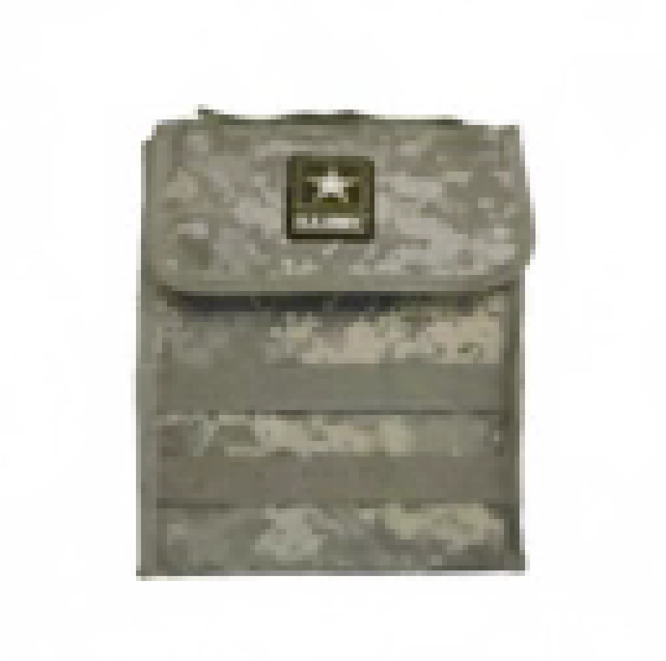
Ours

DietNeRF (+depth)

Novel views



Ours



DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

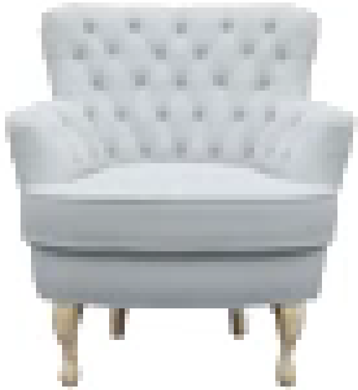
DietNeRF (+depth)

Input view

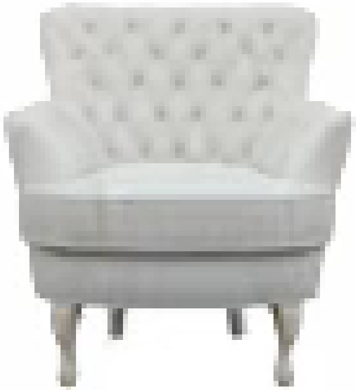
Ours

DietNeRF (+depth)

Novel views



Ours



DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

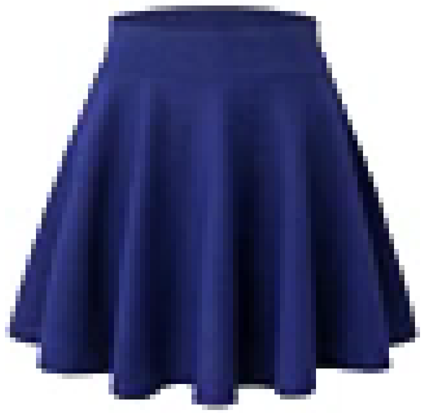
DietNeRF (+depth)

Input view

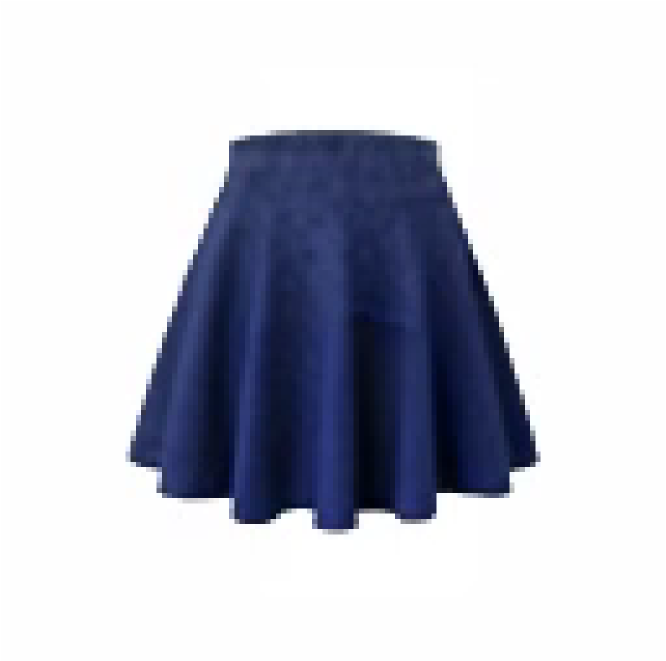
Ours

DietNeRF (+depth)

Novel views



Ours



DietNeRF (+depth)

Input view

Ours

DietNeRF (+depth)

Novel views

Ours

DietNeRF (+depth)

Thanks for watching!