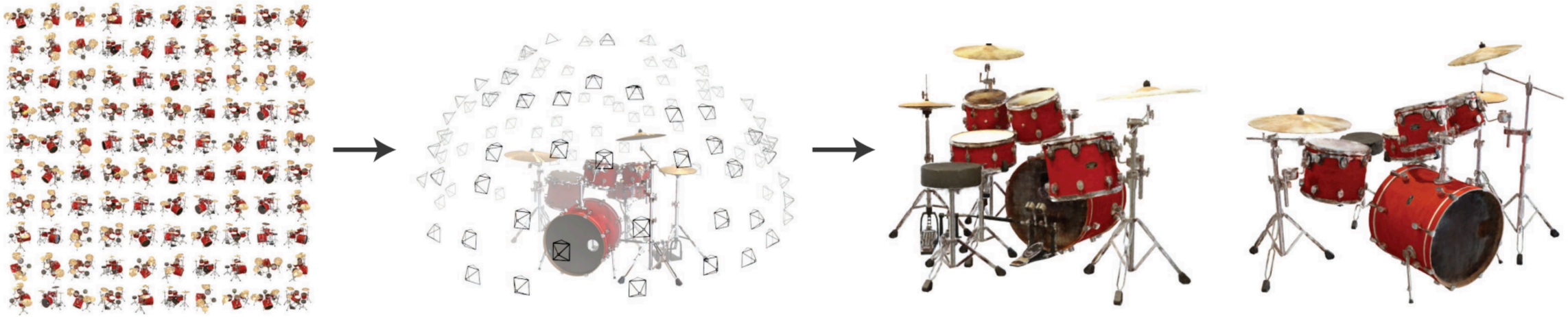
Lecture 26: Recent directions in 3D

Recall: 3D view synthesis



Input views

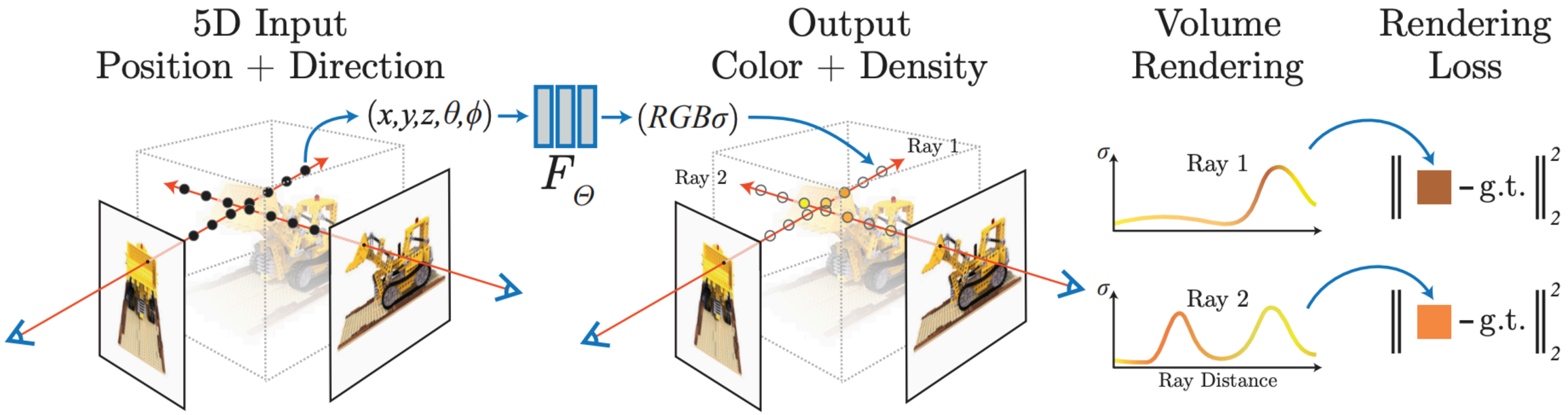
Create model

Render new views

What representation should we use?

[Source: Mildenhall et al., “NeRF”, 2020]

Recall: Fitting a NeRF to a scene



[Source: Mildenhall et al., “NeRF”, 2020]

Recall: Results

Uses hundreds of photos

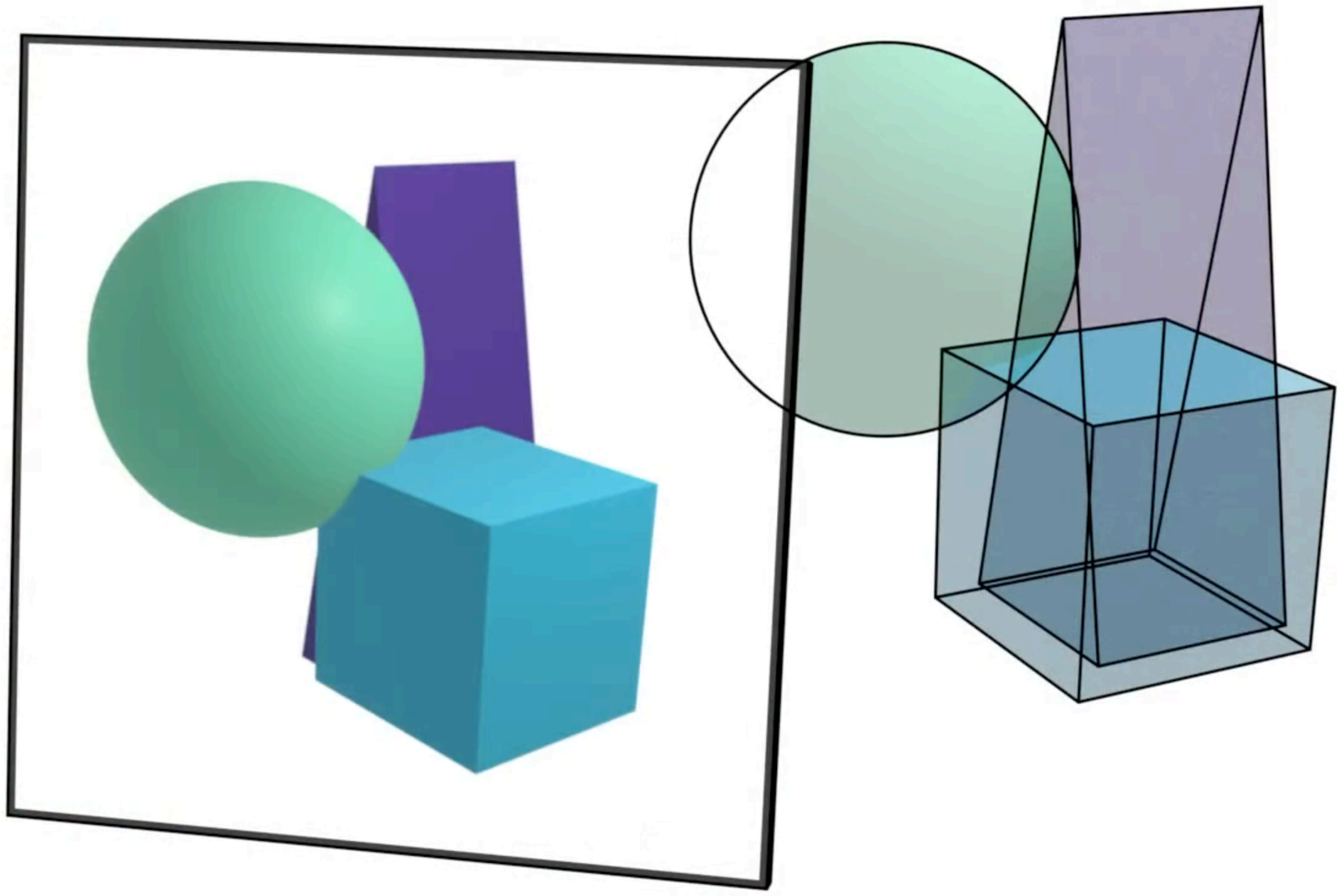
[Mildenhall*, Srinivasan*, Tanick*, et al. 2020]

What happens if we use fewer views?

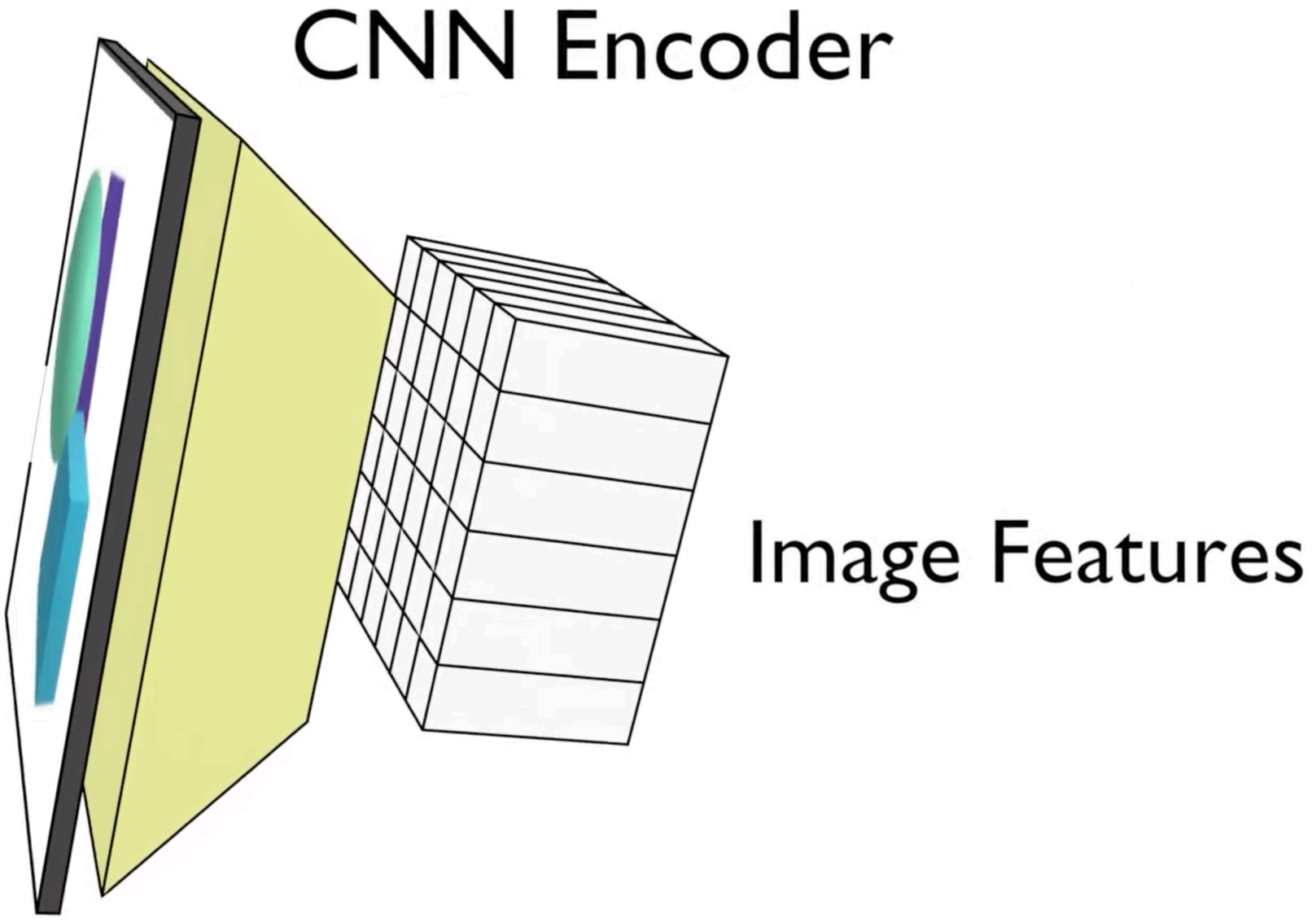
NeRF

Why does this happen?

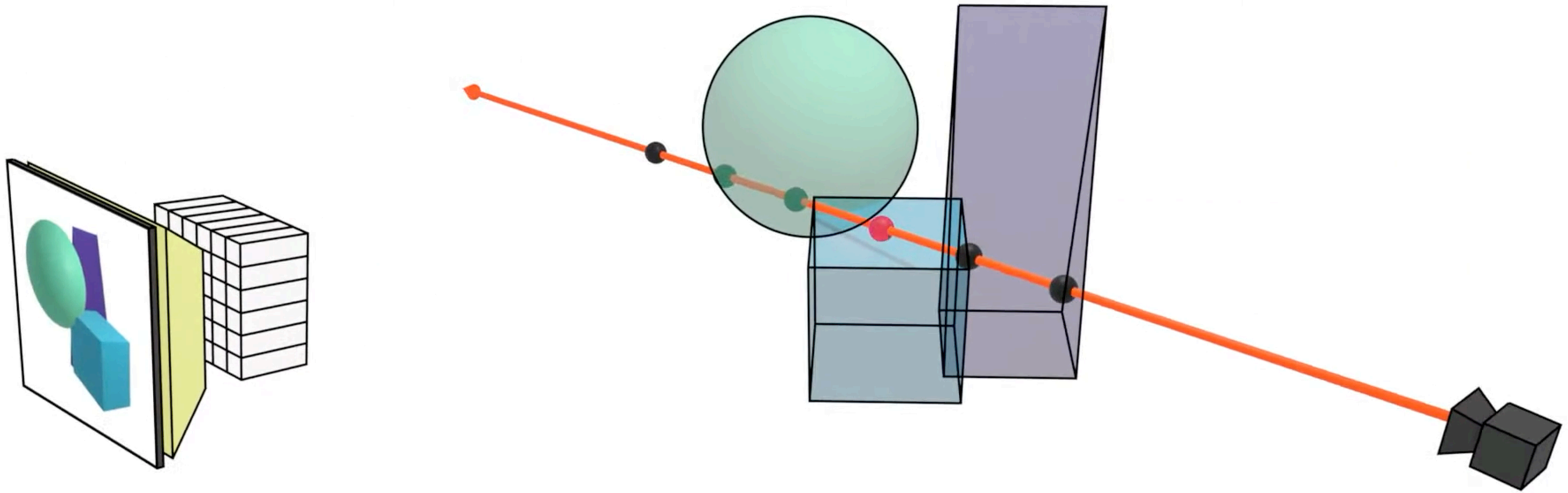
[Yu et al., “pixelNeRF: Neural Radiance Fields from One or Few Images”, 2021]



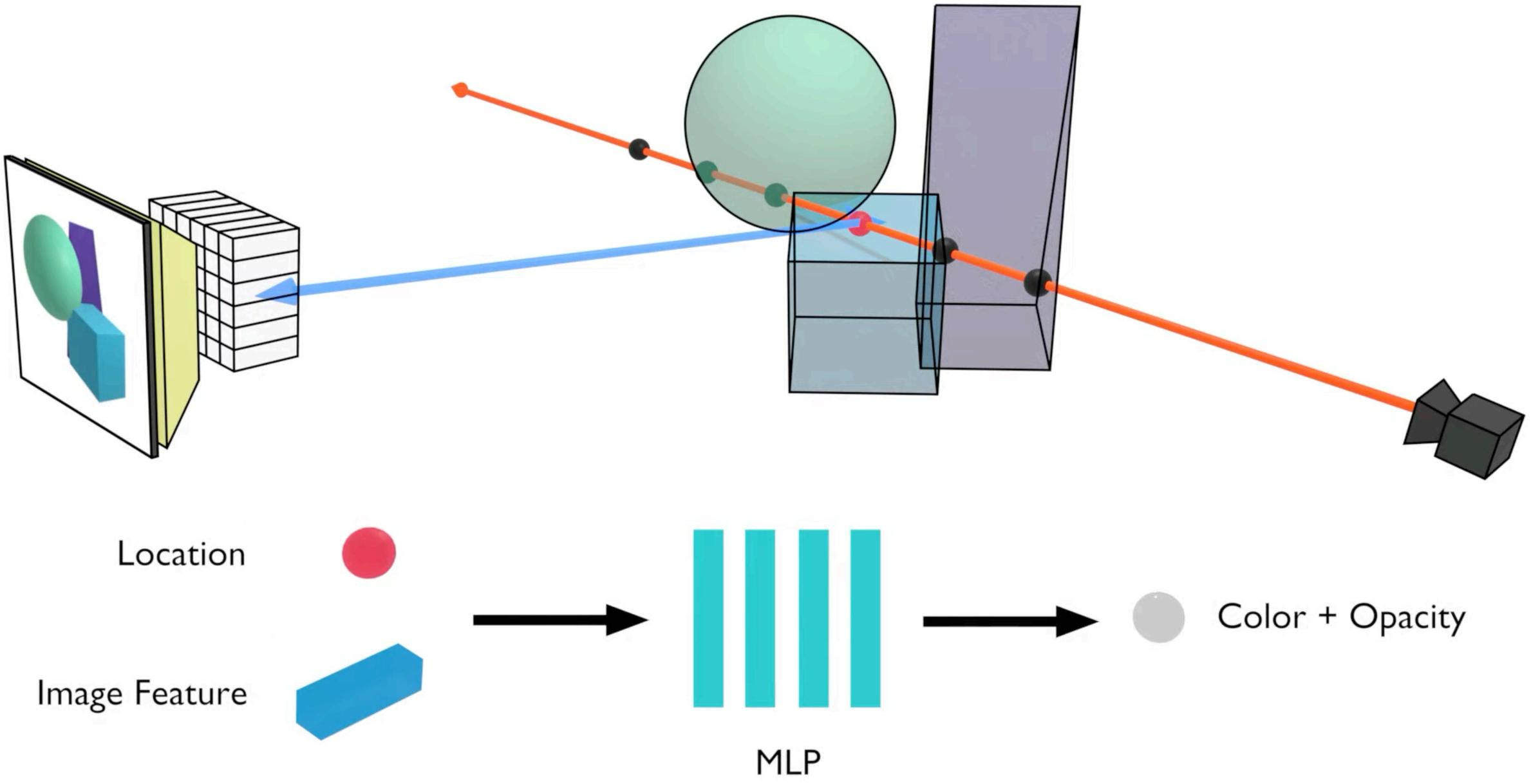
[Yu et al., “pixelNeRF: Neural Radiance Fields from One or Few Images”, 2021]



[Yu et al., “pixelNeRF: Neural Radiance Fields from One or Few Images”, 2021]



[Yu et al., “pixelNeRF: Neural Radiance Fields from One or Few Images”, 2021]



[Yu et al., “pixelNeRF: Neural Radiance Fields from One or Few Images”, 2021]

Input

Input

[Yu et al., “pixelNeRF: Neural Radiance Fields from One or Few Images”, 2021]

Camera pose estimation

Given: two RGB images with unknown relationship
Want: single, coherent reconstruction

Planar Surface Reconstruction from Sparse Views.
Linyi Jin, Shengyi Qian, Andrew Owens, David F. Fouhey. ICCV 2021. **Oral**

TOYOTA RESEARCH INSTITUTE

Source: D. Fouhey

Particular Challenge

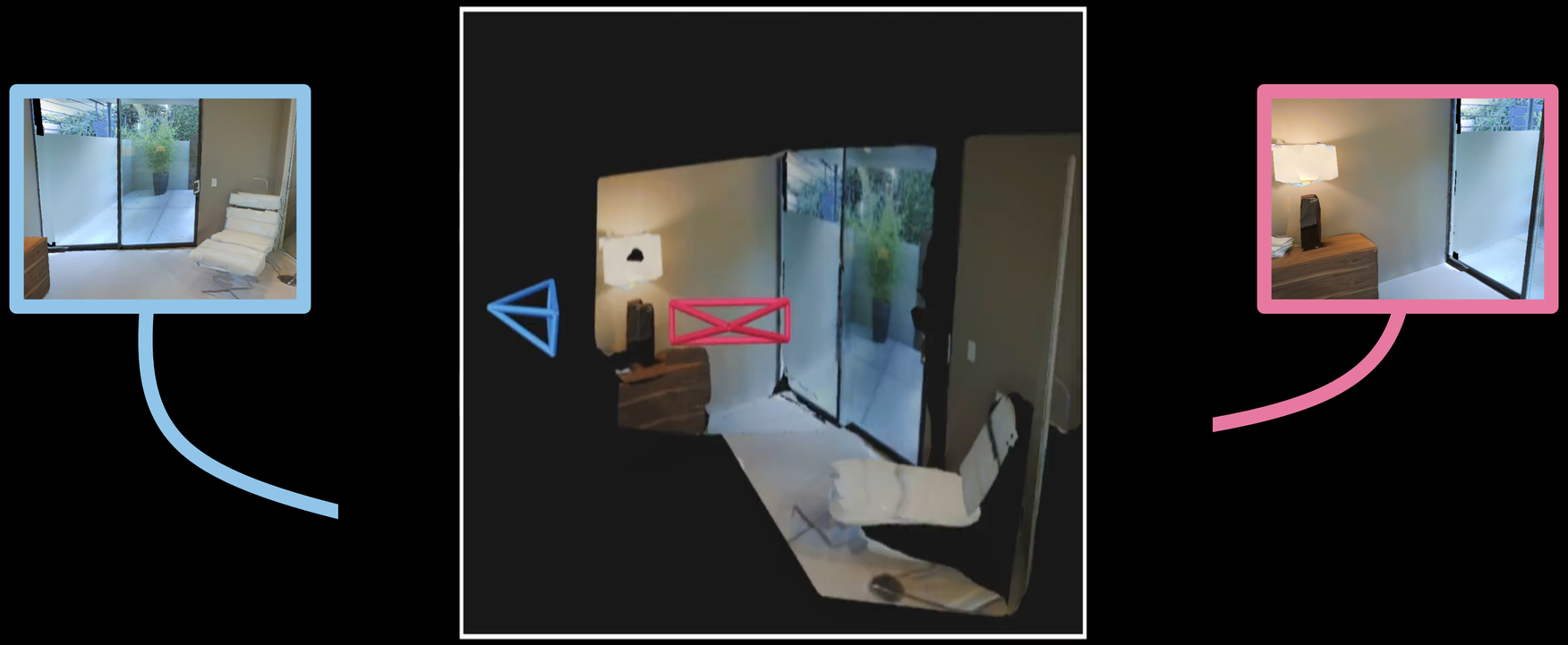
Given: two RGB images with unknown relationship
Want: single, coherent reconstruction

Output: Planes + Relative Camera Pose

Source: D. Fouhey

Particular Challenge

Given: two RGB images with unknown relationship
Want: single, coherent reconstruction



Output: Planes + Relative Camera Pose

Source: D. Fouhey

Geometric Matching

One Solution: Find Correspondence

Among many: Pritchett & Zisserman ‘98, Hartley & Zisserman ’04, Wu et al. ’08, Poursaeed et al. ‘18, Ranftl & Koltun ‘18, Dusmanu et al. ’19, Sarlin et al. ’20, etc. etc.

Source: D. Fouhey

Geometric Matching

One Solution: Find Correspondence
Problem: Where's the Overlap?

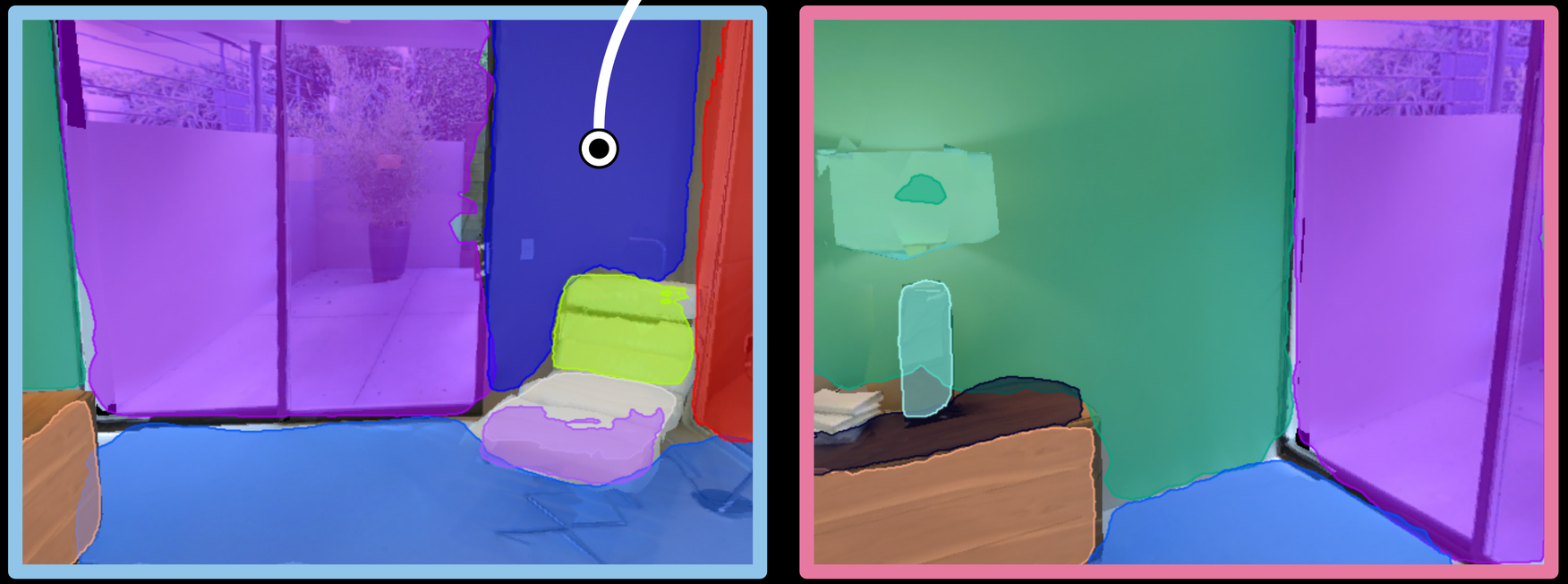
Among many: Pritchett & Zisserman '98, Hartley & Zisserman '04, Wu et al. '08, Poursaeed et al. '18, Ranftl & Koltun '18, Dusmanu et al. '19, Sarlin et al. '20, etc. etc.

Source: D. Fouhey

Single-View 3D

Lots of methods can estimate 3D from RGB

$$\pi_i^T[x, y, z, 1] = 0$$

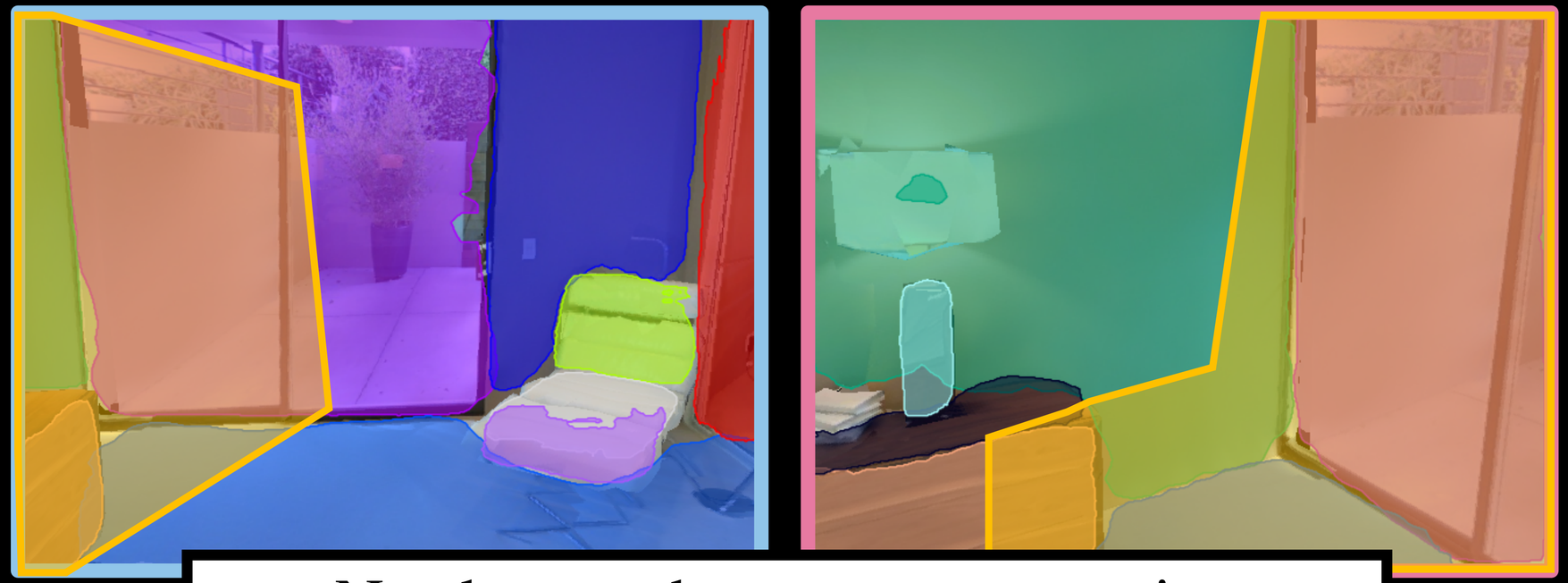


Among many: Hoiem et al. ’05, Saxena et al. ‘05, Girdhar et al. ‘16, Choy et al. ’16, Sun et al. ’18, Liu et al. ’18, Zou et al. ’18, Gkioxari et al. ’19, Jiang et al. ’20, etc. etc

Source: D. Fouhey

Single-View 3D

Lots of methods can estimate 3D from RGB
Problem: What About the Overlap?



Need one coherent reconstruction
Not soup of per-view segments

Among man 6, Sun et al. ’18, Liu et al. ’18, Zou et al. ’18, Gkioxari et al. ’19, Jiang et al. ’20, etc. etc

Source: D. Fouhey

Multiview 3D

Given: **two RGB** images with **unknown relationship**
Want: single, coherent **reconstruction**

Among many: Choy et al. ’16, Kar et al. ’17, En et al. ‘18, Huang et al. ‘18, Saito et al. ‘19, Flynn et al. ‘19, El Banani et al. ‘20, Mildenhall et al. ’20, Srinivasan et al. ’20, etc.

Source: D. Fouhey

Insight

Per-view reconstruction, *inter-view* correspondence, and *inter-view* camera pose should be solved jointly

Output: Planes + Relative Camera Pose

Source: D. Fouhey

- Results with planes on Matterport3D (re-renders)
- Single forward pass + principled optimization

Image A | Image B | Prediction

Linyi Jin, Shengyi Qian, Andrew Owens, David F. Fouhey.
Planar Surface Reconstruction from Sparse Views. ICCV 2021. ***Oral***

TOYOTA RESEARCH INSTITUTE

Source: D. Fouhey

Architecture

Input: Two Images

Source: D. Fouhey

Architecture – Plane Detection

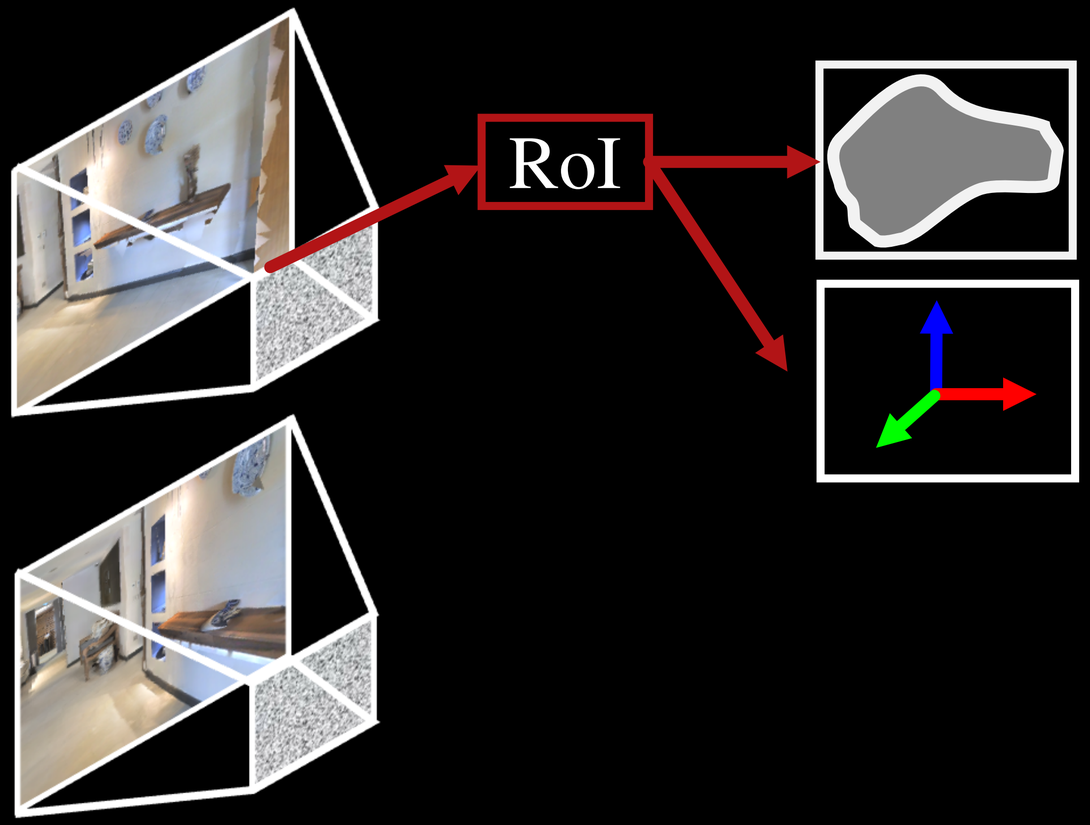
ResNet50-FPN

Tied Weights

Tsung-Yi Lin et al. Feature Pyramid Networks. CVPR 2017.

Source: D. Fouhey

Architecture – Plane Detection



Plane Mask
M_i

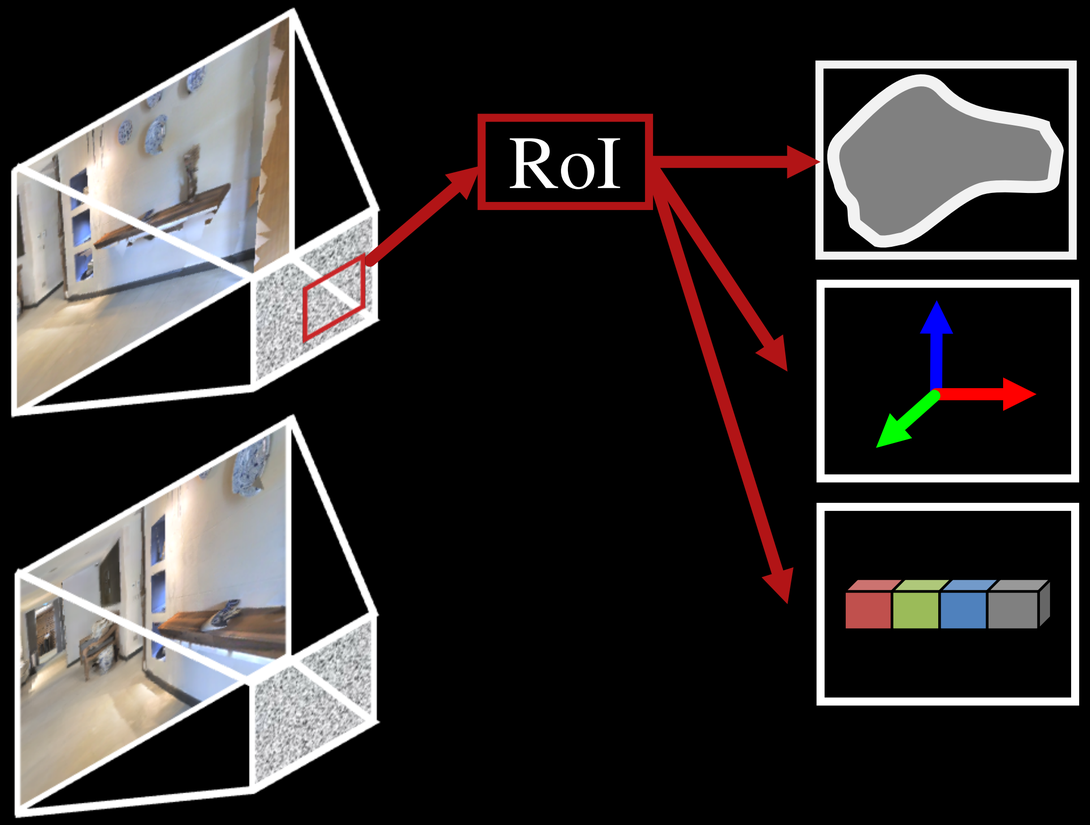
Plane Parameters
$\boldsymbol{\pi}_i = [\boldsymbol{n}_i, o_i]$

ResNet50-FPN
Tied Weights

Chen Liu et al. Planercnn: 3d plane detection and reconstruction from a single image. CVPR 2019

Source: D. Fouhey

Architecture – Plane Detection



Plane Mask
M_i

Plane Parameters
$\boldsymbol{\pi}_i = [\boldsymbol{n}_i, o_i]$

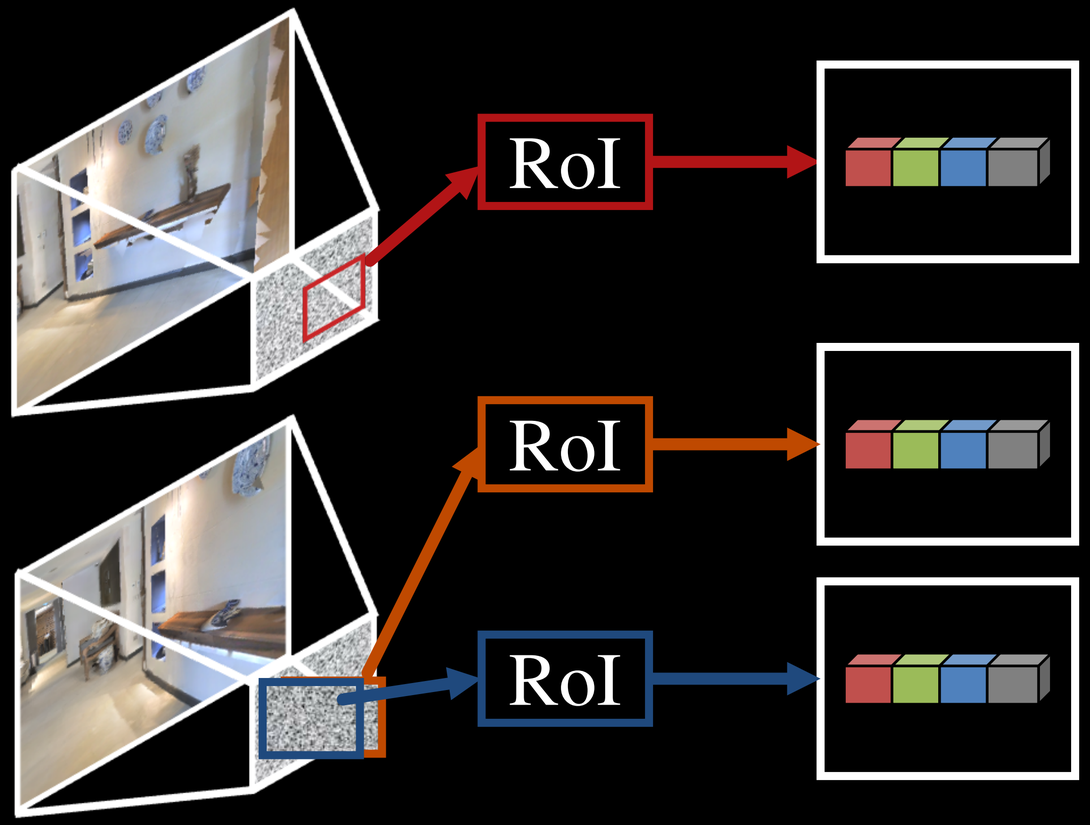
Appearance Feature
$\boldsymbol{e}_i$

ResNet50-FPN
Tied Weights

Chen Liu et al. Planercnn: 3d plane detection and reconstruction from a single image. CVPR 2019

Source: D. Fouhey

Architecture – Plane Detection



Anchor Feature
e_a

Positive Feature
e'_p

Negative Feature
e'_n

ResNet50-FPN
Tied Weights

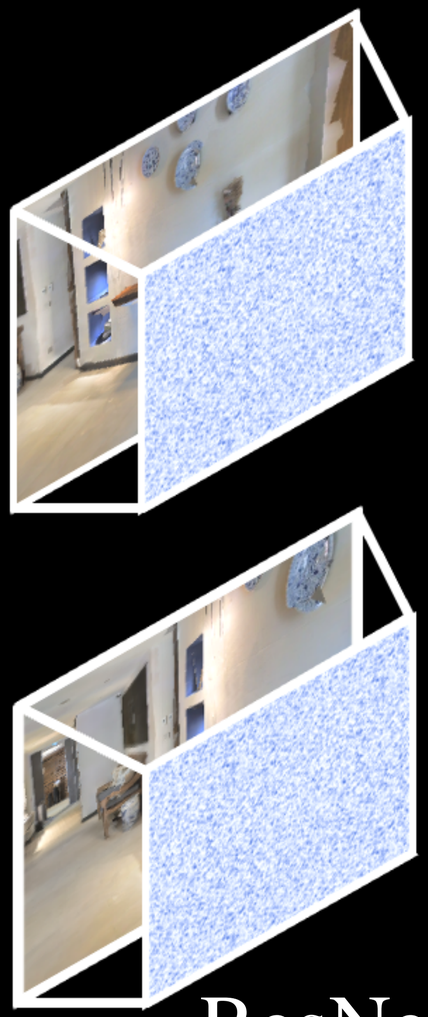
Same plane → Similar embedding

$$\max(\|e_a - e'_p\|_2 - \|e_a - e'_n\|_2 + \alpha)$$

Florian Schroff et al. Facenet: A unified embedding for face recognition and clustering. CVPR 2015. etc.

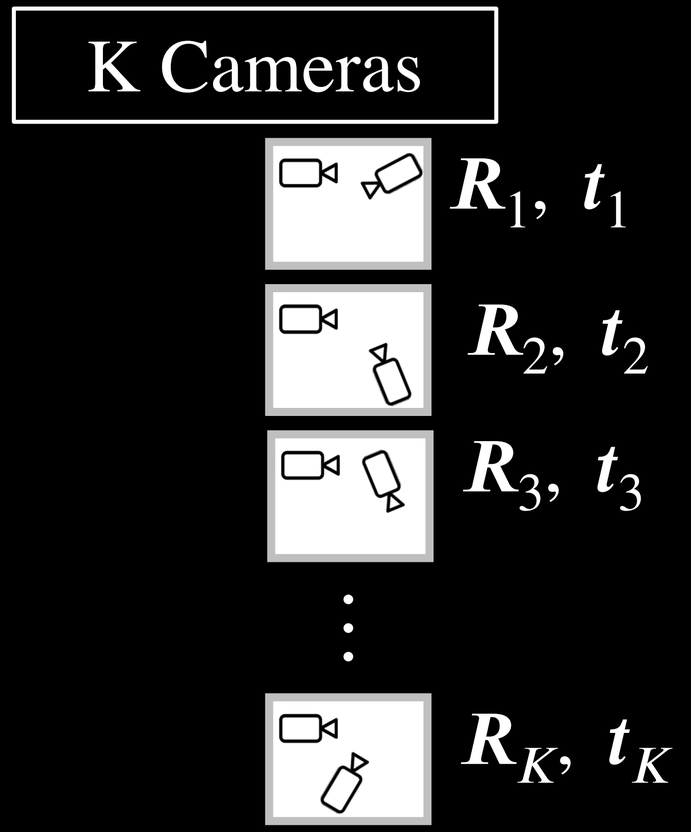
Source: D. Fouhey

Architecture – Cameras



ResNet50-P3

Tied Weights



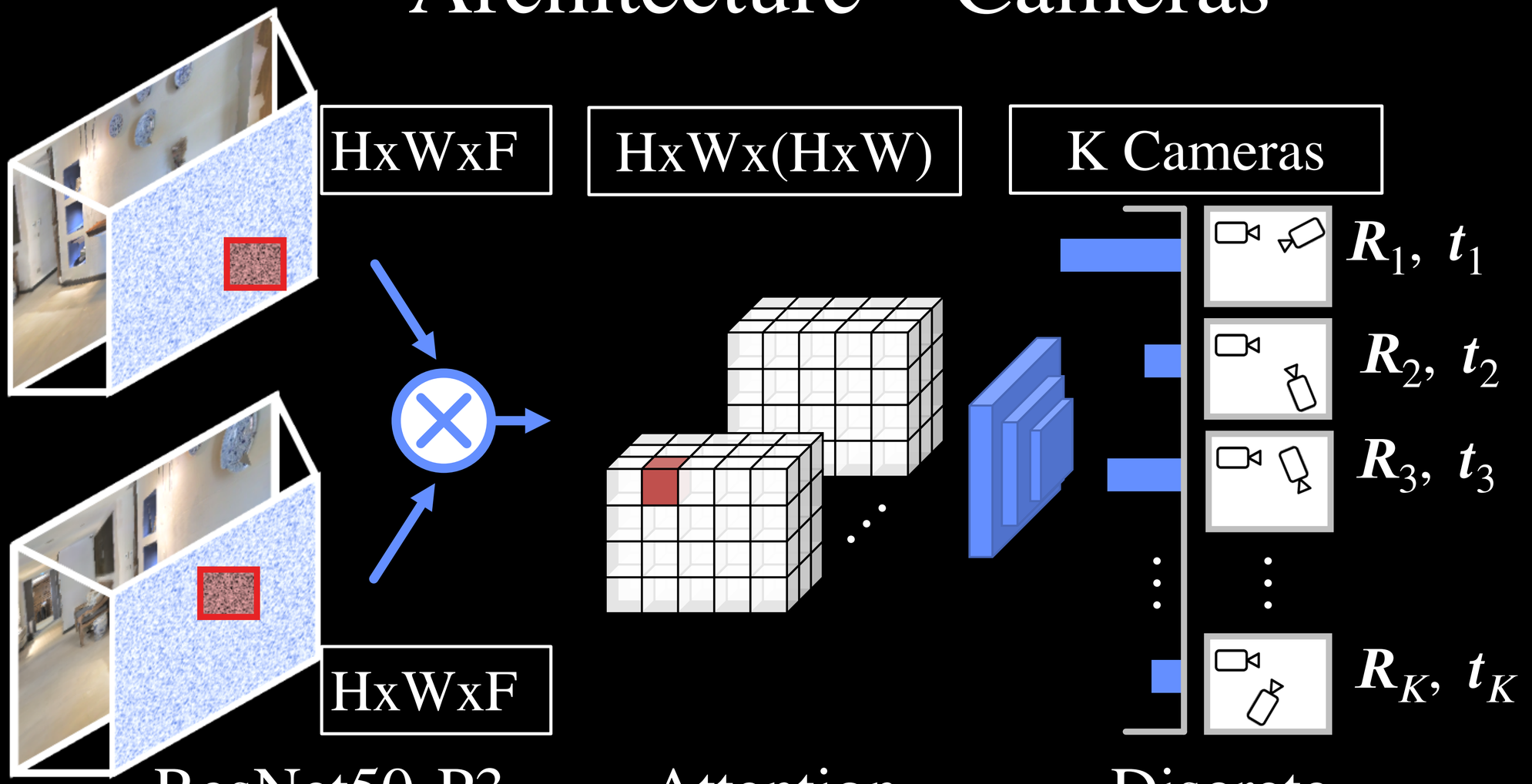
Discrete Categories

Sovann En et al. RPNet: An end-to-end network for relative camera pose estimation. ECCV 2018.

Shengyi Qian et al. Associative3D: Volumetric Reconstruction from Sparse Views. ECCV 2020.

Source: D. Fouhey

Architecture – Cameras



ResNet50-P3
Tied Weights

Attention
Features

Discrete
Distribution

Sovann En et al. RPNet: An end-to-end network for relative camera pose estimation. ECCV 2018.
Shengyi Qian et al. Associative3D: Volumetric Reconstruction from Sparse Views. ECCV 2020.

Source: D. Fouhey

Are We Done?

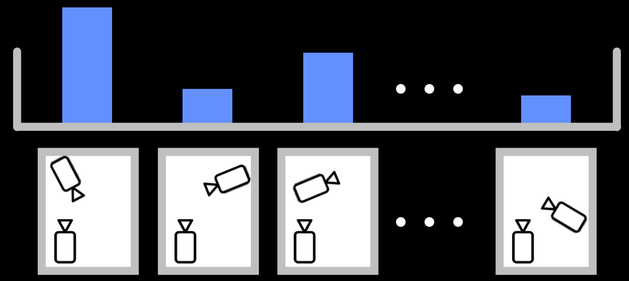
Source: D. Fouhey

Are We Done?

View 1: m planes

View 2: n planes

Cross-View: Distribution over K Cameras



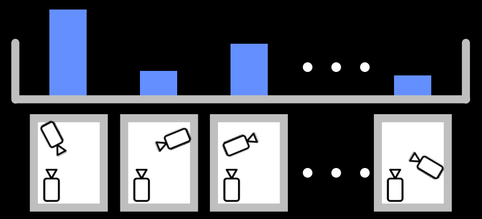
Source: D. Fouhey

Producing a Reconstruction

m planes

n planes

K cams



Goal: match **C** in $\{0,1\}^{mxn}$, camera

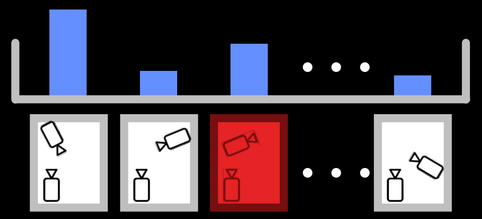
Score: plane i ↔ plane j with cam k

Source: D. Fouhey

Producing a Reconstruction

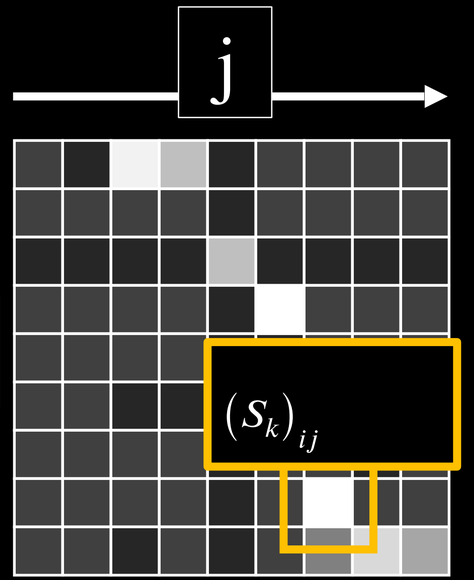
m planes

n planes

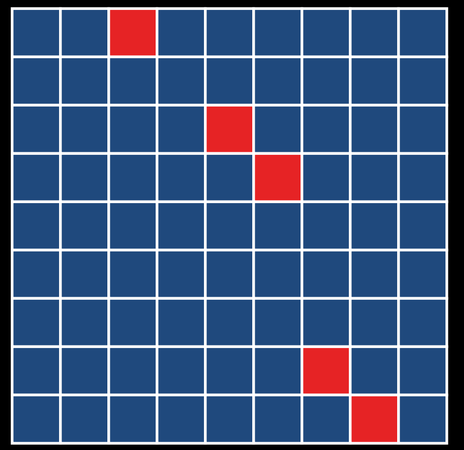


K cams

Goal: match **C** in $\{0,1\}^{mxn}$, camera
Score: plane i ↔ plane j with cam k



Score Matrix S_k for camera k

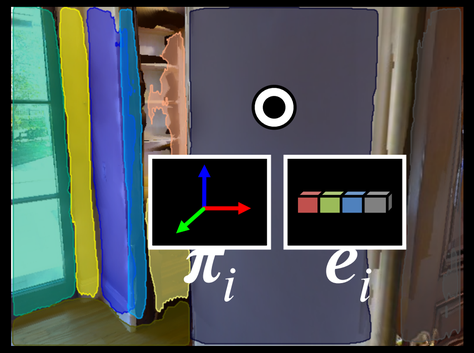


Match Matrix **C**

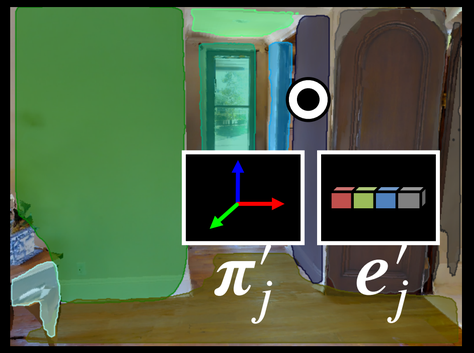
Source: D. Fouhey

Producing a Reconstruction

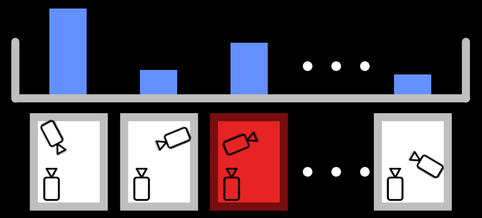
m planes



n planes



K cams



Goal: match **C** in $\{0,1\}^{mxn}$, camera

Score: plane i ↔ plane j with cam k

$$(S_k)_{ij} =$$

$$\left\| e_i - e'_j \right\| + \quad \textit{(Appearance)}$$

$$\lambda_n \operatorname{acos}\left(\left| n_i^T n'_j \right| \right) + \quad \textit{(Normals)}$$

$$\lambda_o \left| o_i - o'_j \right| \quad \textit{(Offset)}$$

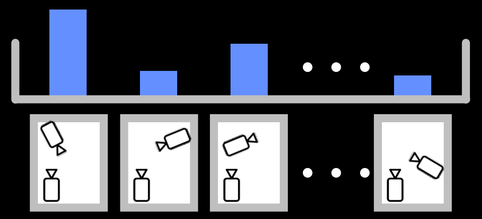
Source: D. Fouhey

Producing a Reconstruction

m planes

n planes

K cams



Goal: match $\mathbf{C}$ in $\{0,1\}^{m \times n}$, camera

Score: plane i ↔ plane j with cam k

$$\underset{k,\mathbf{C}}{\operatorname{argmin}}$$

$$\sum_{i,j} \left(\boldsymbol{C} \circ \boldsymbol{S}_k\right)_{ij} \qquad \textit{(Match Well)}$$

$$-\sum_{i,j} \boldsymbol{C}_{ij} \qquad \textit{(Match A Lot)}$$

$$-\lambda_c \log\left(p_k\right) \qquad \textit{(Pick Good Camera)}$$

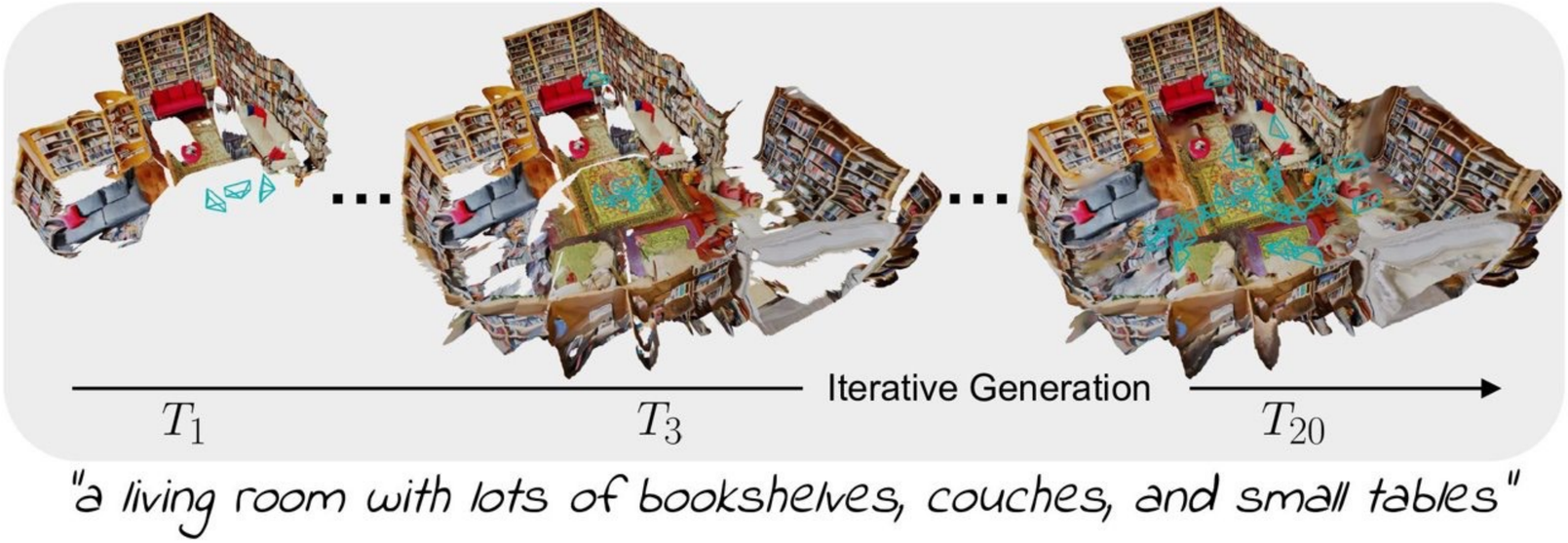
Source: D. Fouhey

Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models

Lukas Höllein[1*], Ang Cao[2*], Andrew Owens[2], Justin Johnson[2], Matthias Nießner[1]

[1]Technical University of Munich, [2]University of Michigan

*joint first authorship



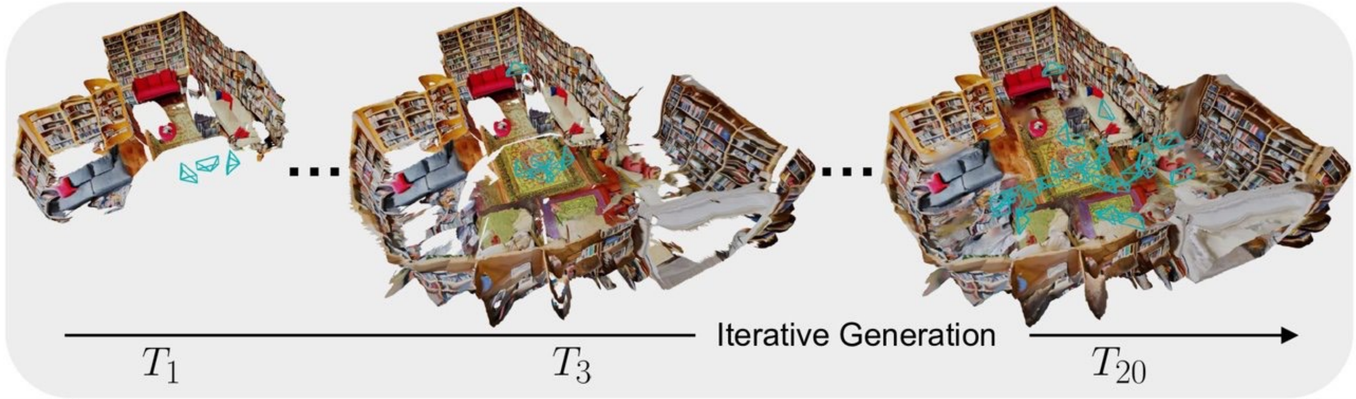
"a living room with a lit furnace, couch, and cozy curtains, bright lamps that make the room look well-lit"

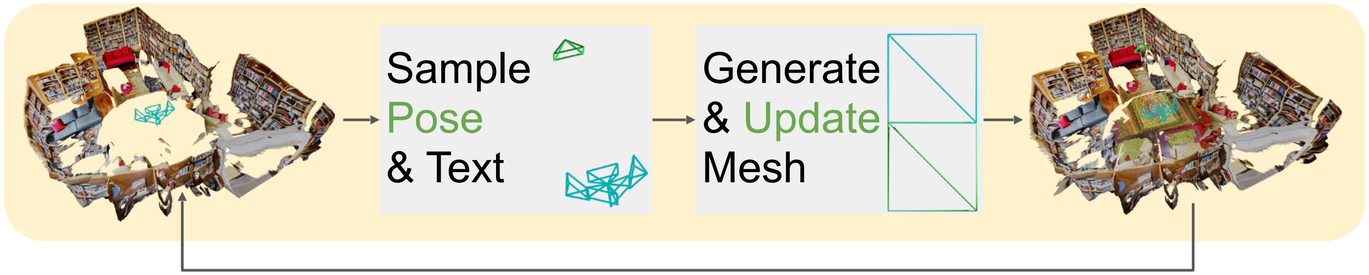
Scene Generation Stage

iterative scene generation

generated images

Two-Stage Scene Generation

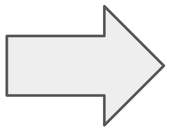




Scene Generation Stage

Completion Stage

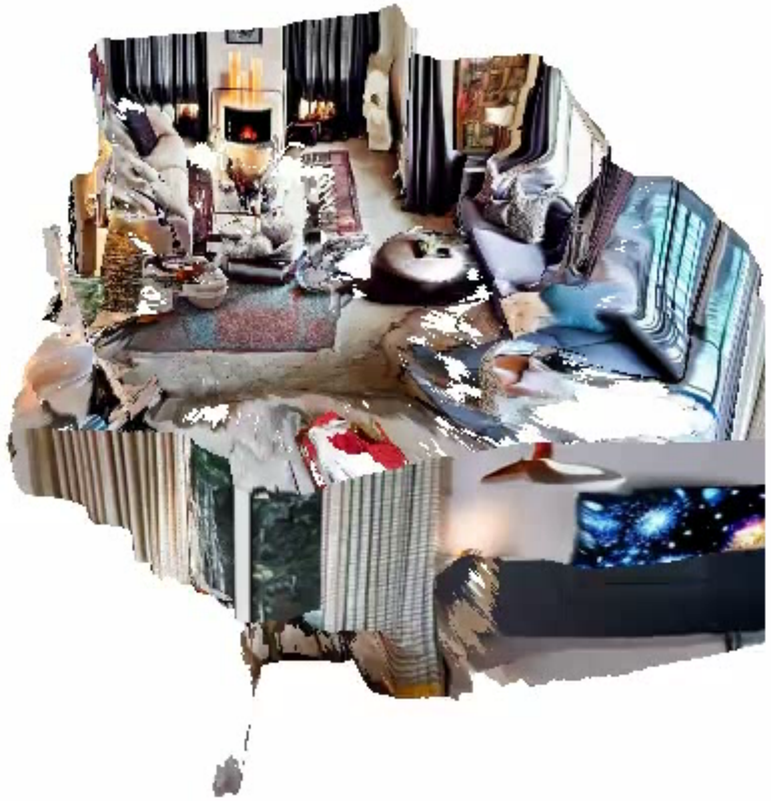
Completion Stage



after first stage: mesh contains holes

completion fills-in holes

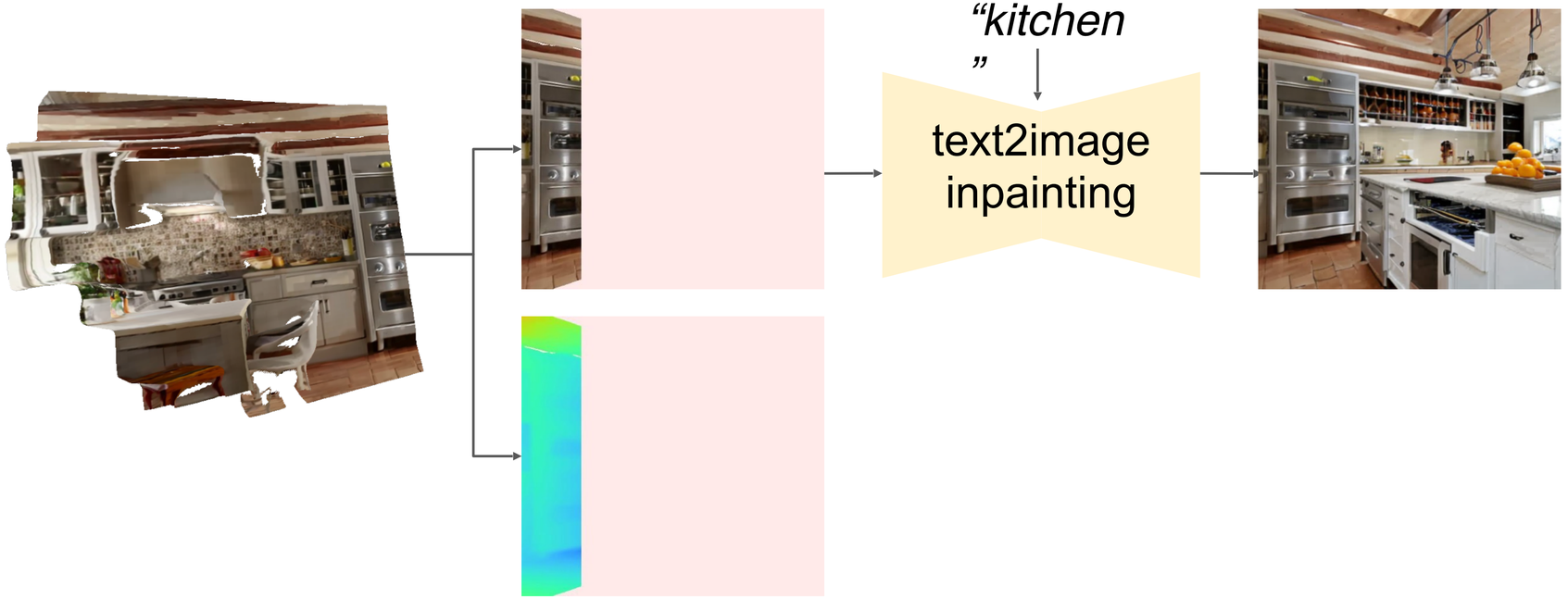
Completion Stage



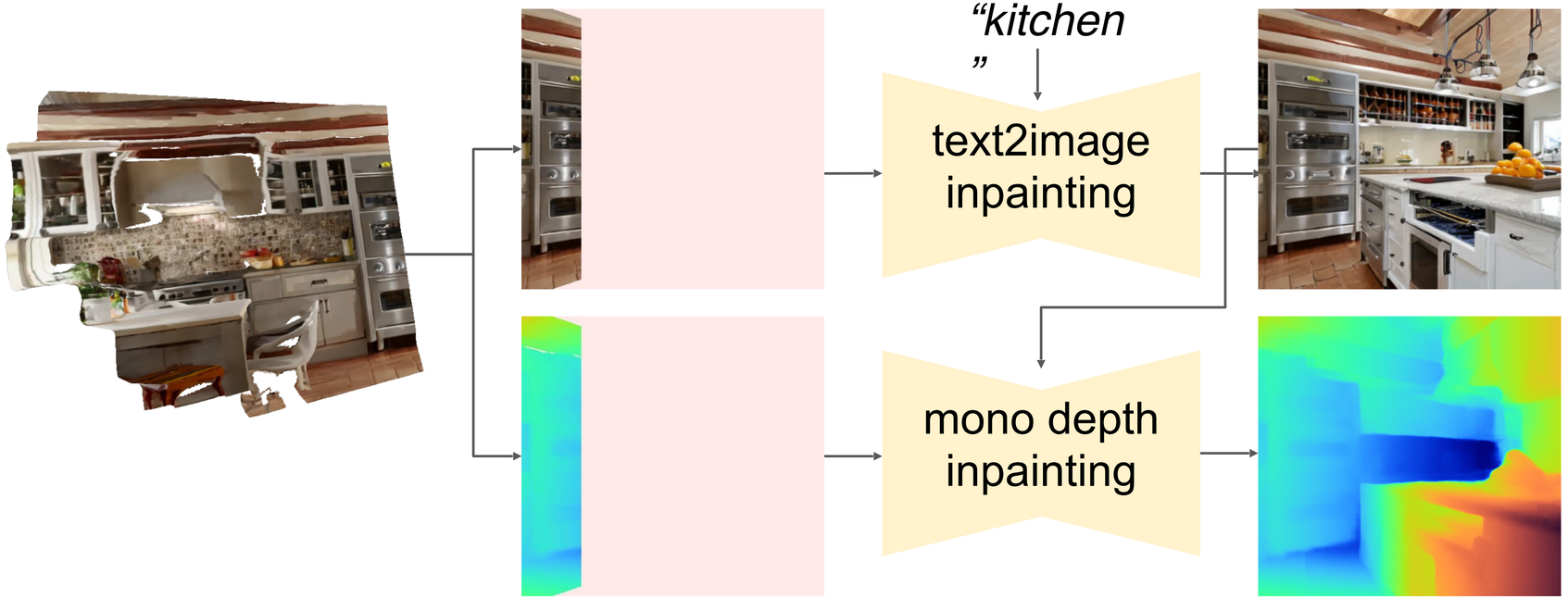
after first stage: mesh contains holes

completion fills-in holes

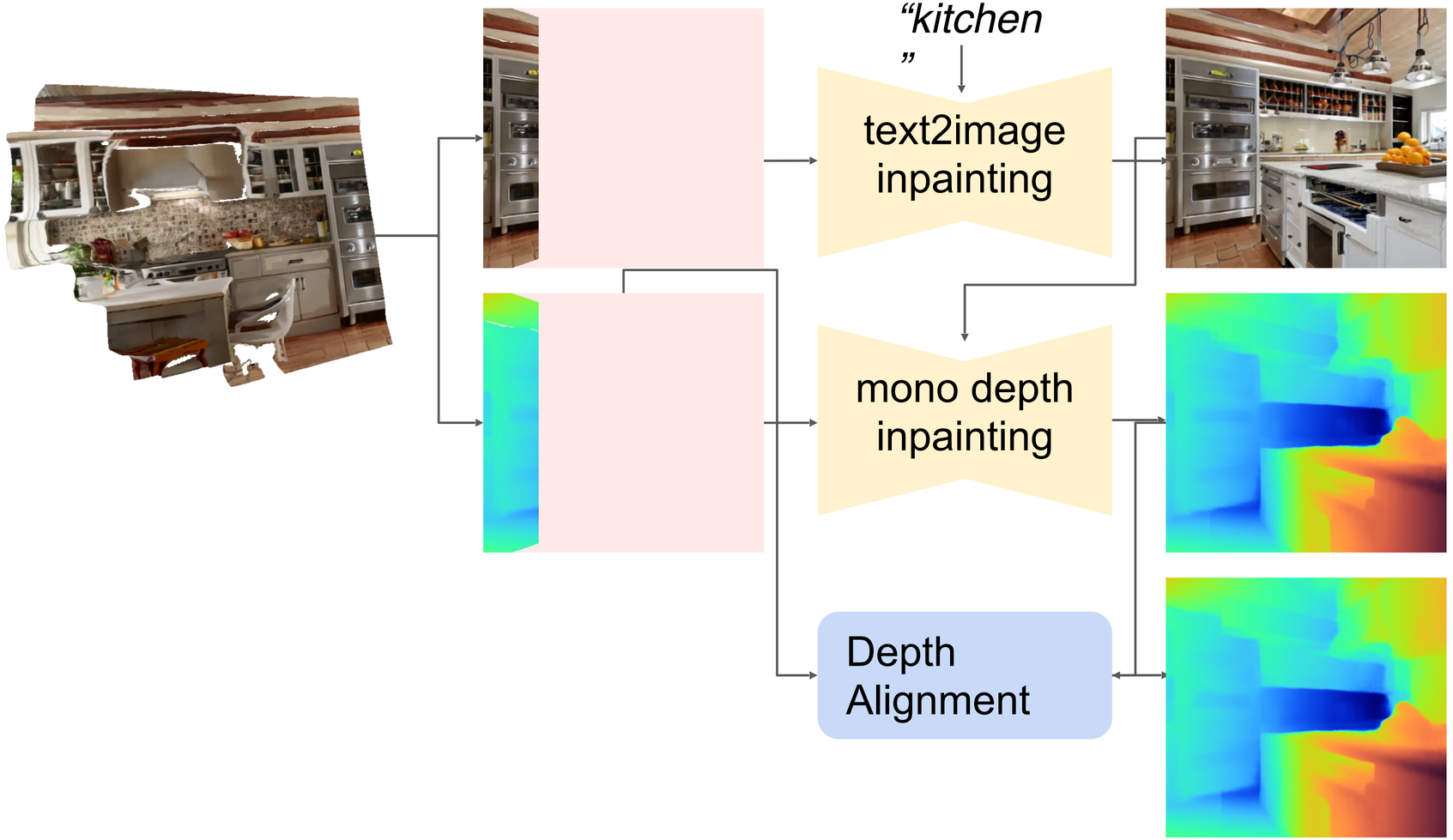
Iterative Scene Generation



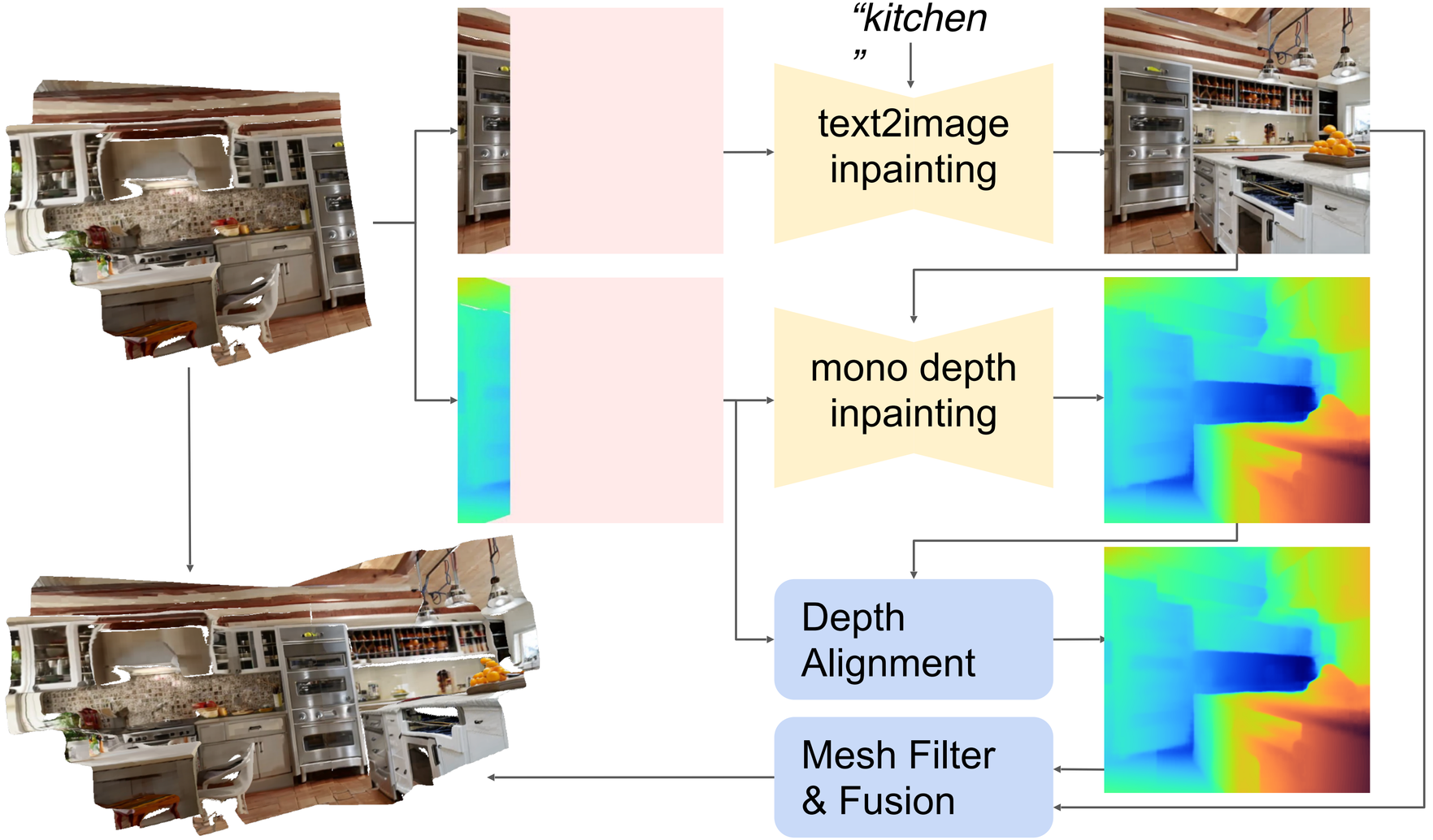
Iterative Scene Generation



Iterative Scene Generation



Iterative Scene Generation



a living room with a lit furnace, couch, and cozy curtains, bright lamps that make the room look well-lit

Editorial Style Photo, Coastal Bathroom, Clawfoot Tub, Seashell, wicker, Mosaic Tile, Blue and white

Generating 3D models

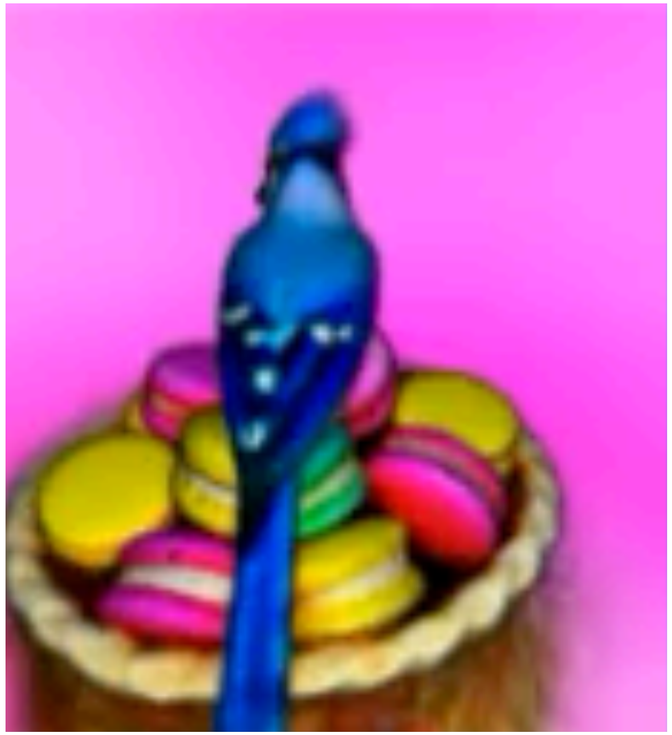
Do we actually need 3D supervision?

[Poole et al., “DreamFusion”, 2023]

We already have good text-to-image models

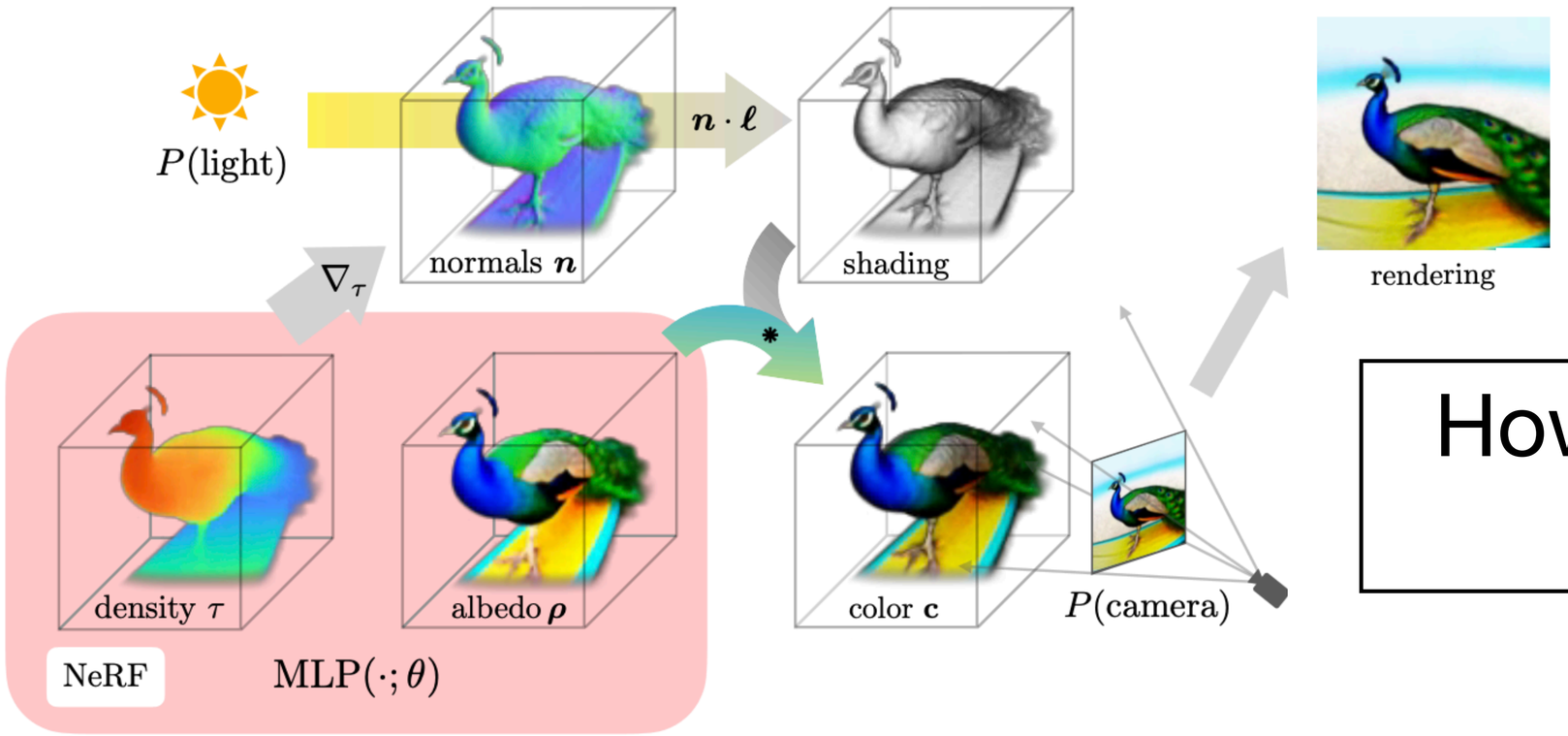
“A blue jay standing on a large basket of rainbow macarons.”

Could we fit a 3D model to this image?



Single image could “supervise” single viewpoint.

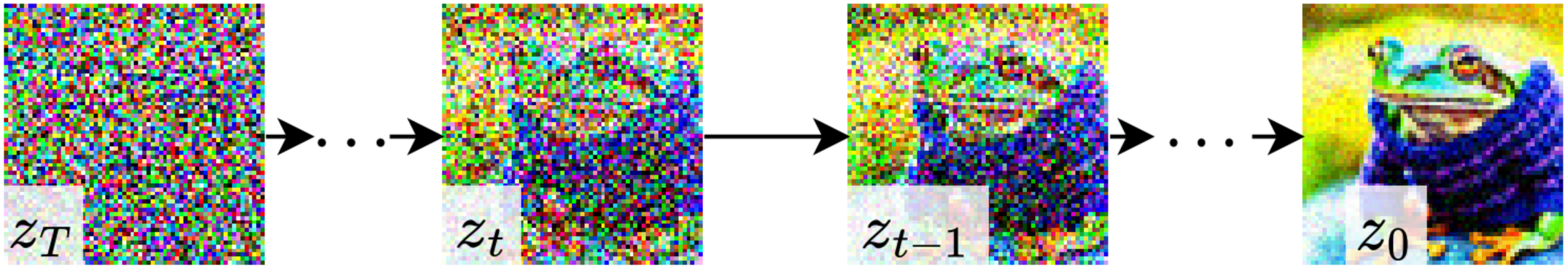
Solve for a NeRF that “looks right” from every viewpoint



How do we measure “looks right”?

$p(\text{rendered image} \mid \text{a DSLR photo of a peacock on a surfboard}'')$

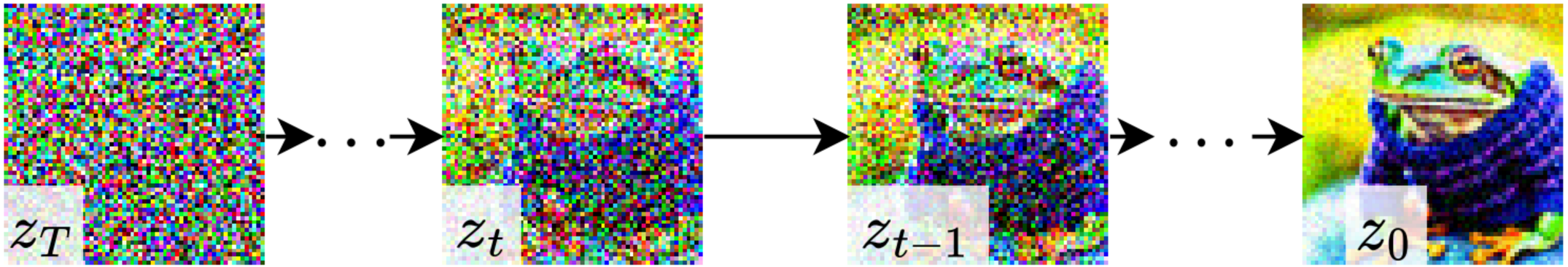
Can we use a pretrained diffusion model?



Diffusion: generate image by denoising

… but also it models $p(\text{image} \mid \text{text})$

Score distillation sampling



Noise estimator

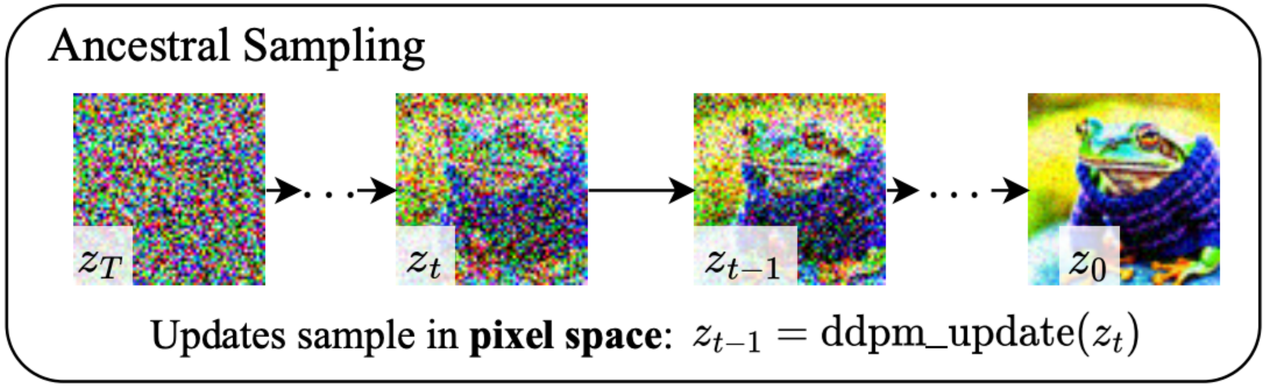
Noise

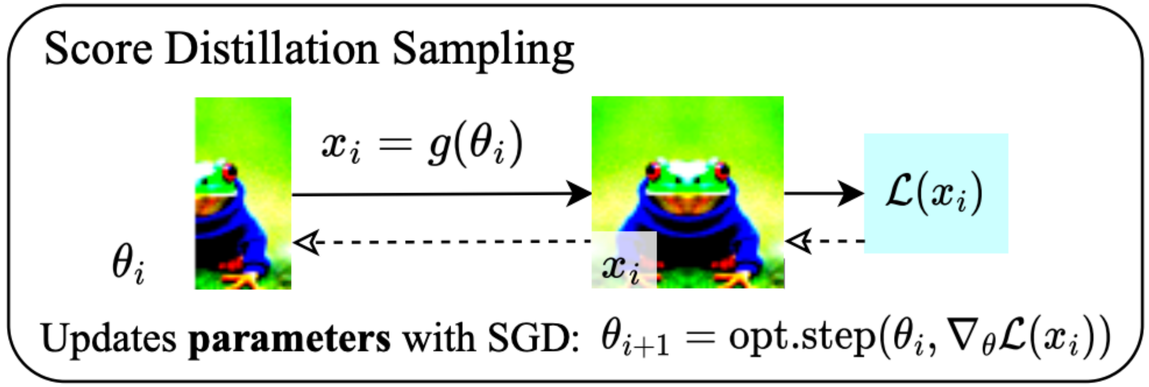
$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(t)\|\epsilon_\phi(\alpha_t\mathbf{x} + \sigma_t\epsilon; t) - \epsilon\|_2^2\right]$$

Noisy image

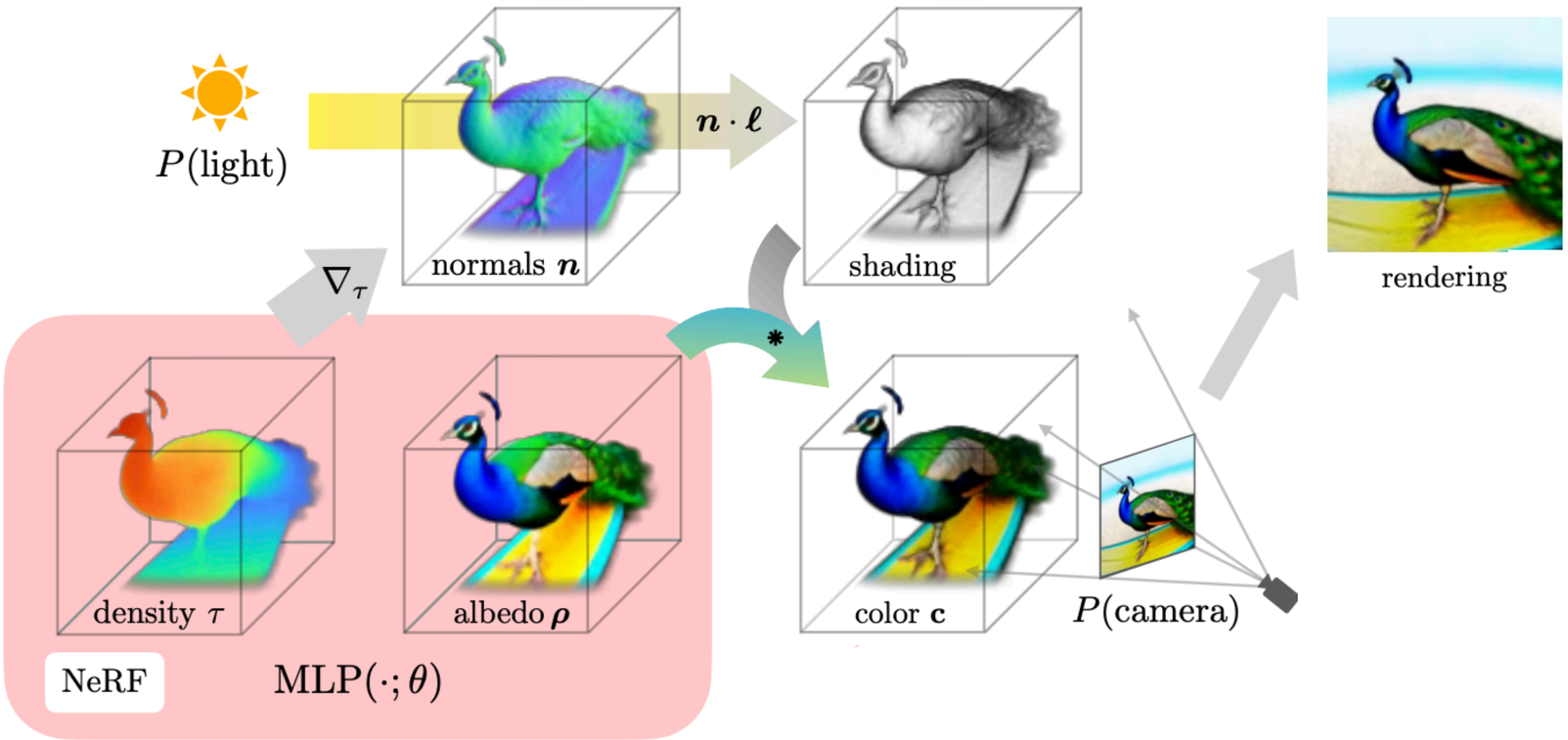
Provides a bound on $\log(p(\mathbf{x}))$ [Ho et al., 2020, Kingma et al., 2021]

Score distillation sampling





DreamFusion



Solve for a NeRF such that, when it is rendered, it has high probability under a text-to-image model.

DreamFusion

Generate 3D from text yourself!

a DSLR photo of a squirrel | an intricate wooden carving of a squirrel | a highly detailed metal sculpture of a squirrel

[...] | wearing a kimono | wearing a medieval suit of armor | wearing a purple hoodie | wearing an elegant ballgown

[...] | reading a book | riding a motorcycle | playing the saxophone | chopping vegetables | sitting at a pottery wheel shaping a clay bowl | riding a skateboard | wielding a katana | eating a hamburger | dancing

DreamFusion

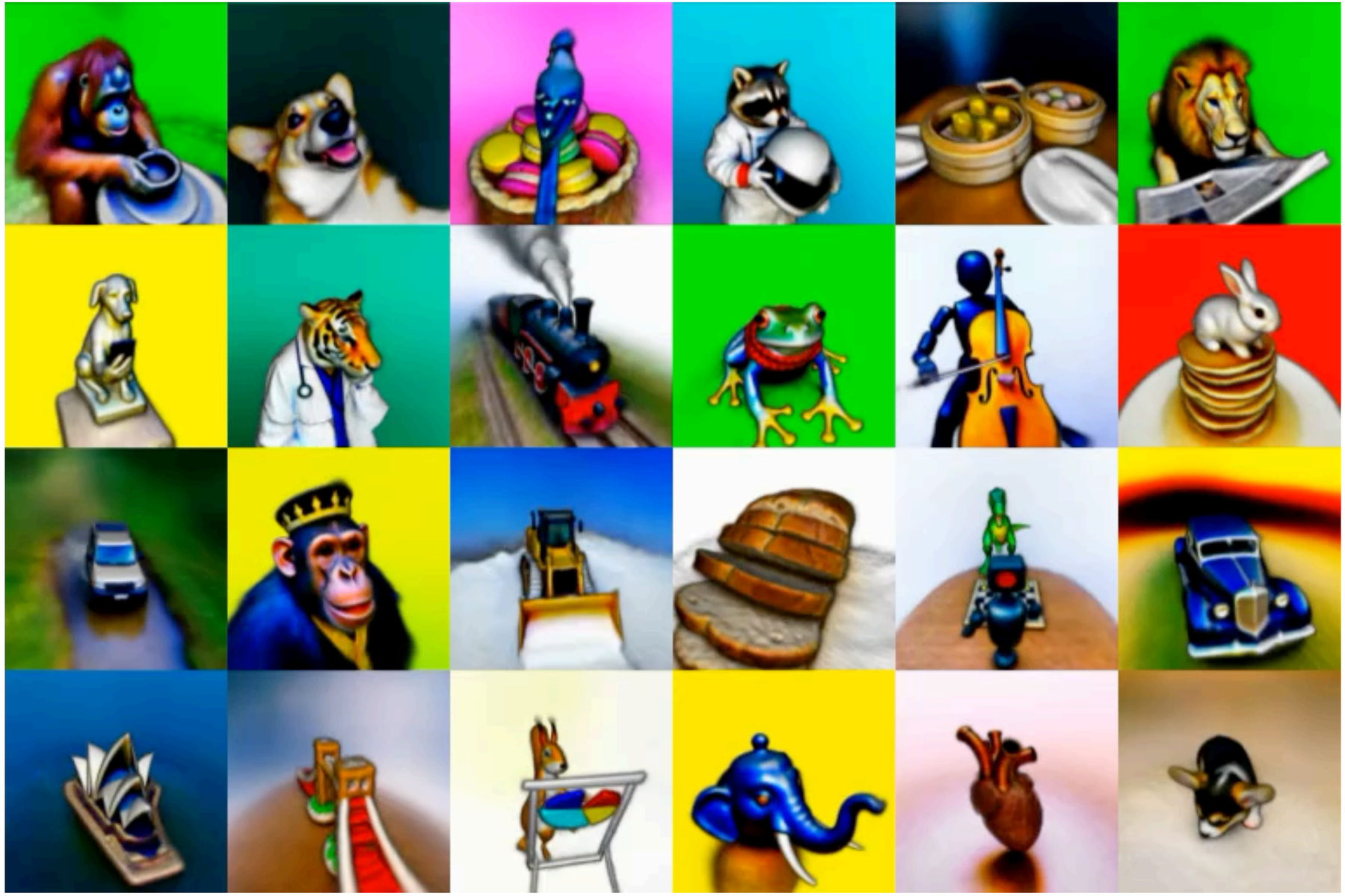
Generate 3D from text yourself!

a DSLR photo of a squirrel | an intricate wooden carving of a squirrel | a highly detailed metal sculpture of a squirrel

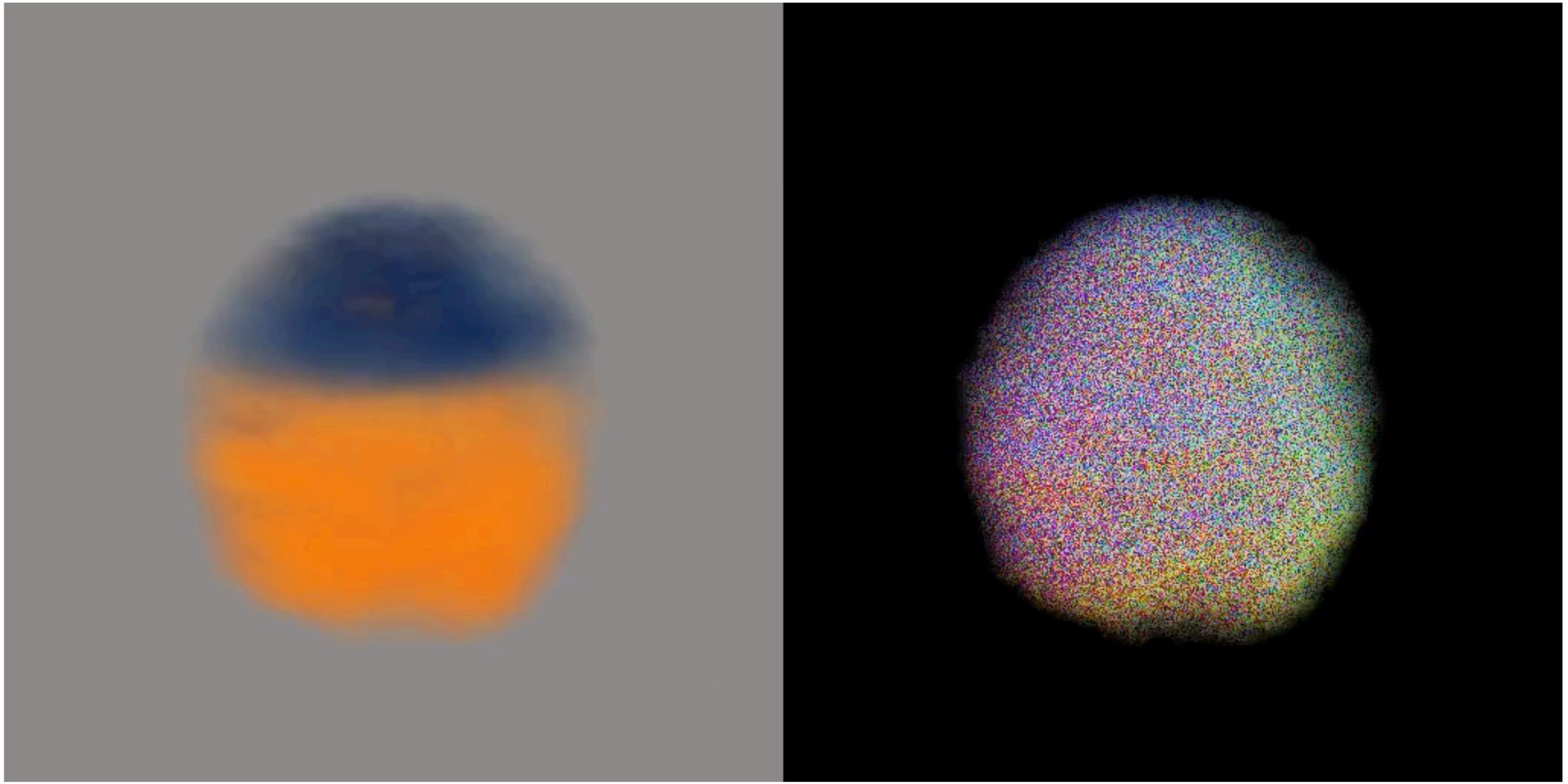
[...] | wearing a kimono | wearing a medieval suit of armor | wearing a purple hoodie | wearing an elegant ballgown

[...] | reading a book | riding a motorcycle | playing the saxophone | chopping vegetables | sitting at a pottery wheel shaping a clay bowl | riding a skateboard | wielding a katana | eating a hamburger | dancing

DreamFusion



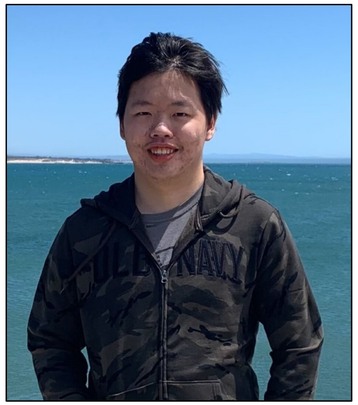
More recent text-to-3D



[Shi et al., MVDream, 2023]

GANmouflage: 3D Object Nondetection with Texture Fields

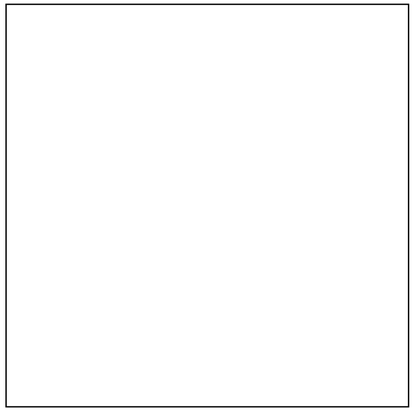
CVPR 2023



Rui Guo

Jasmine Collins

Oscar de Lima



Andrew Owens


Roger T. Hanlon

3D camouflage problem

A. Owens, C. Barnes, A. Flint, H. Singh, W. T. Freeman.
Camouflaging an Object from Many Viewpoints. CVPR 2014.

3D camouflage problem

A. Owens, C. Barnes, A. Flint, H. Singh, W. T. Freeman.
Camouflaging an Object from Many Viewpoints. CVPR 2014.

3D camouflage problem

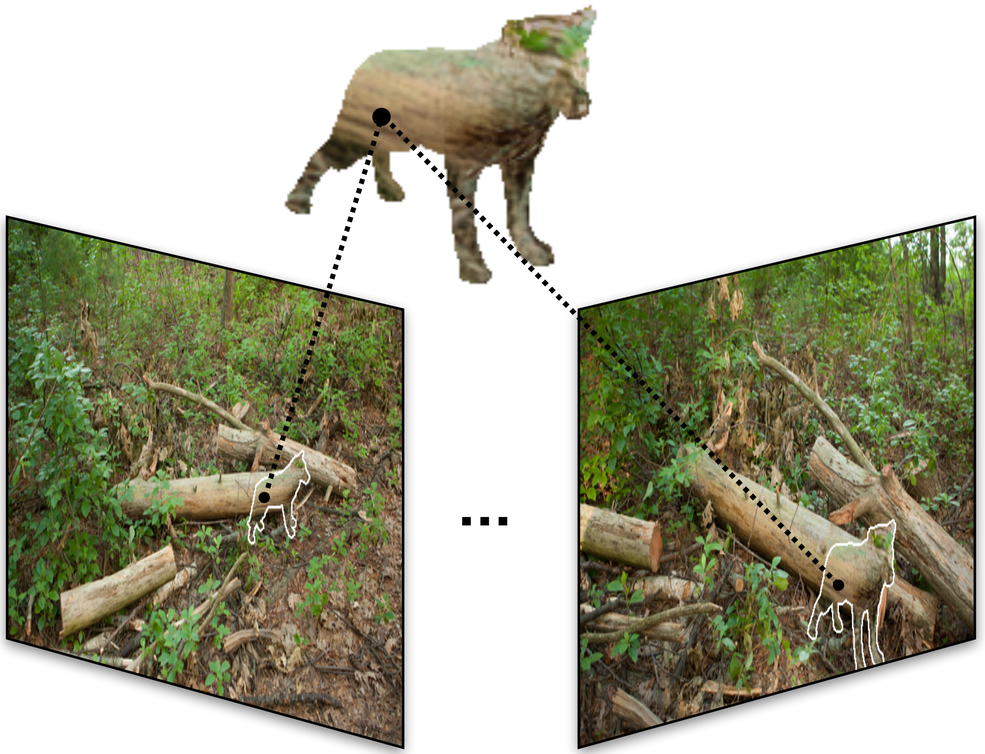


A. Owens, C. Barnes, A. Flint, H. Singh, W. T. Freeman.
Camouflaging an Object from Many Viewpoints. CVPR 2014.

Applications

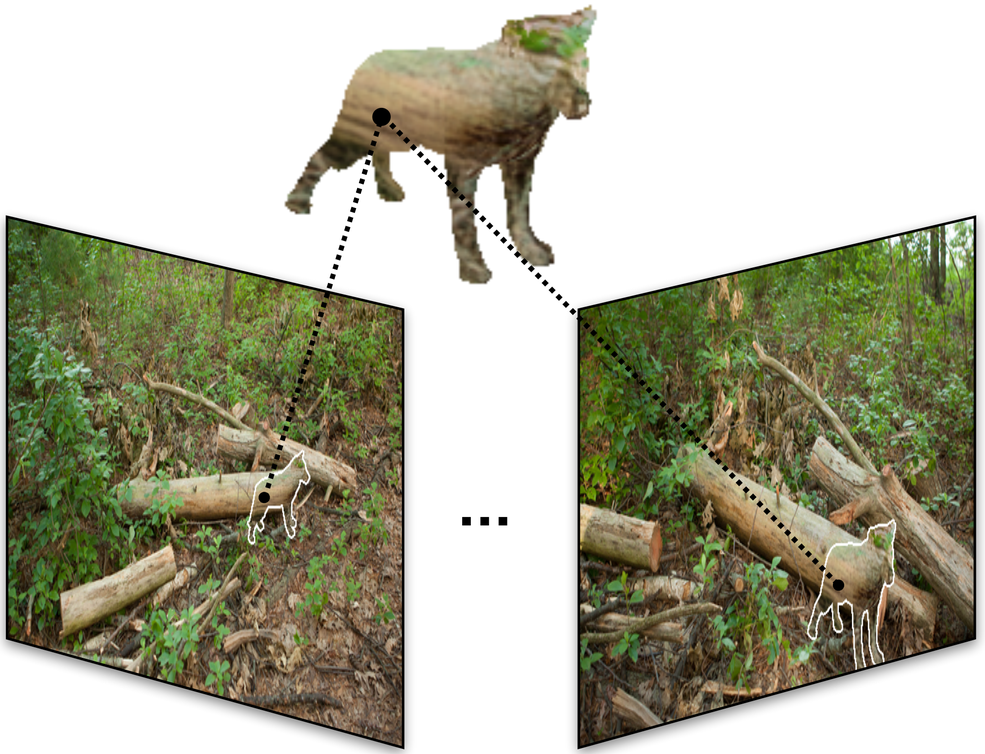
Hiding unsightly objects

GANmouflage model

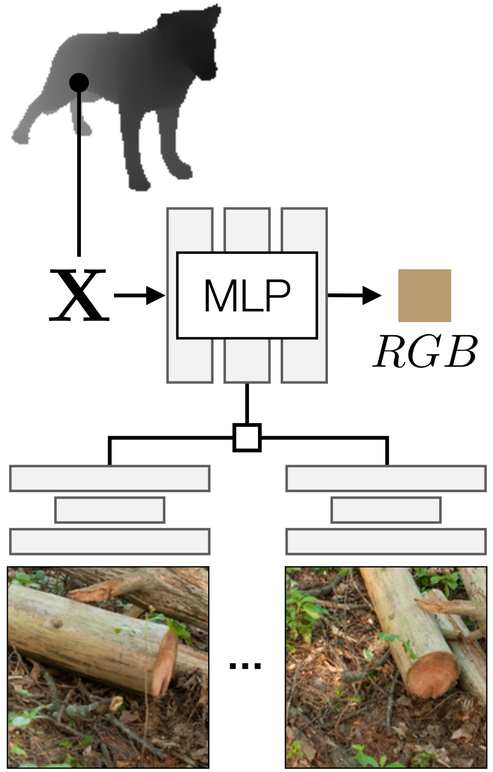


Multi-view camouflage

GANmouflage model

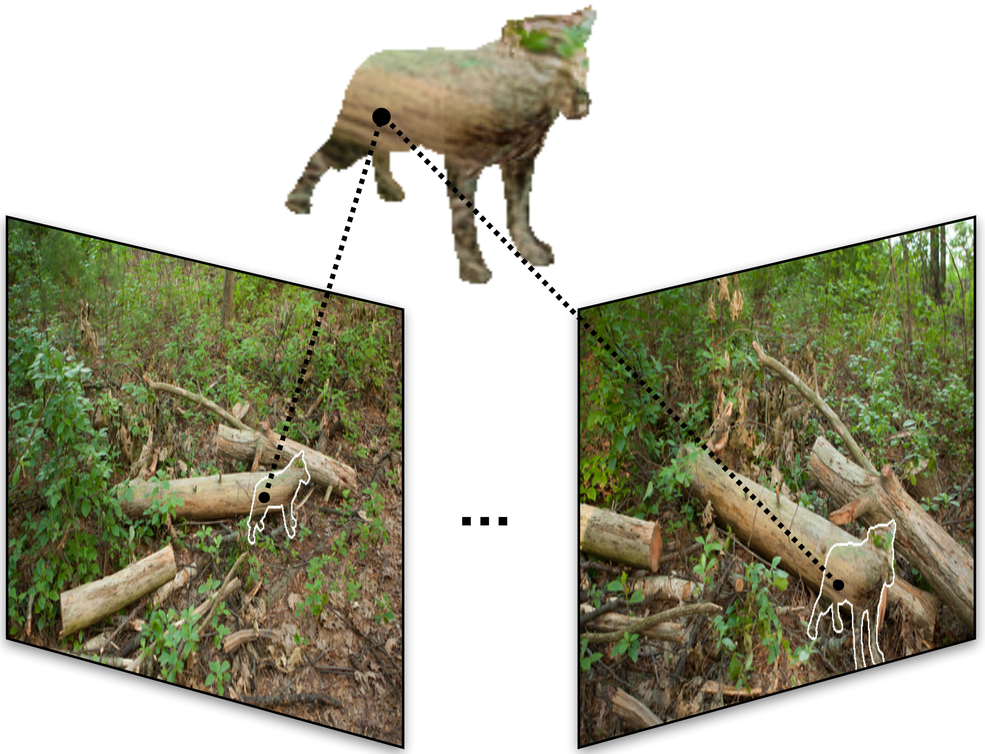


Multi-view camouflage

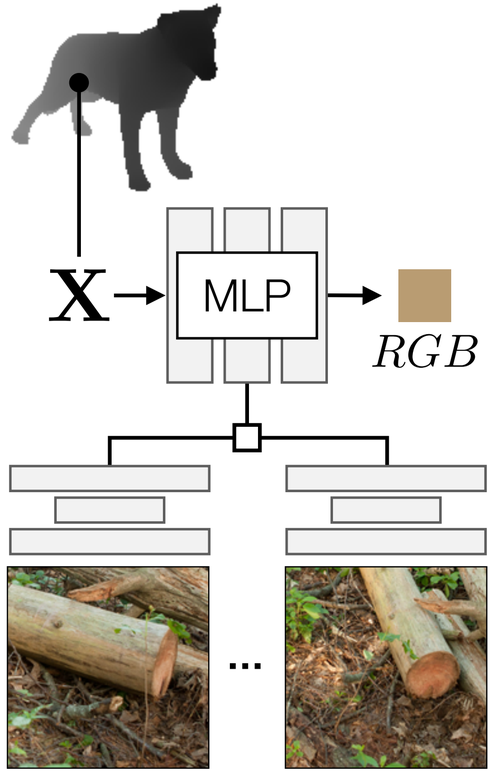


Texture model

GANmouflage model



Multi-view camouflage

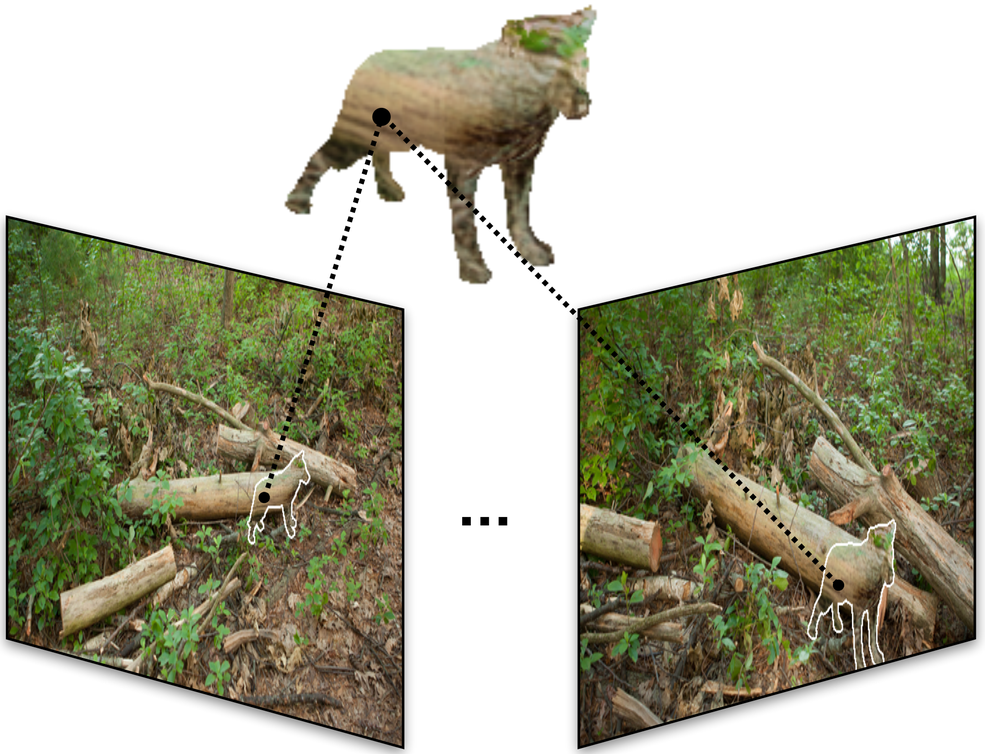


Texture model

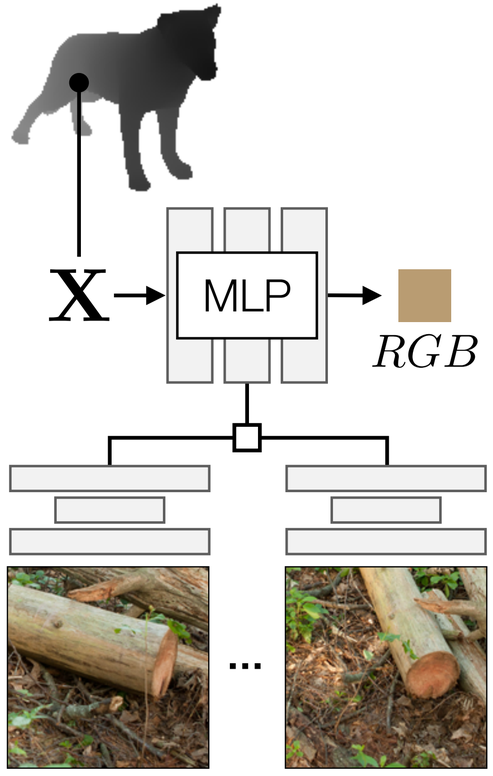
$$\mathcal{L}_p\left(\quad,\quad\right)$$

Photoconsistency

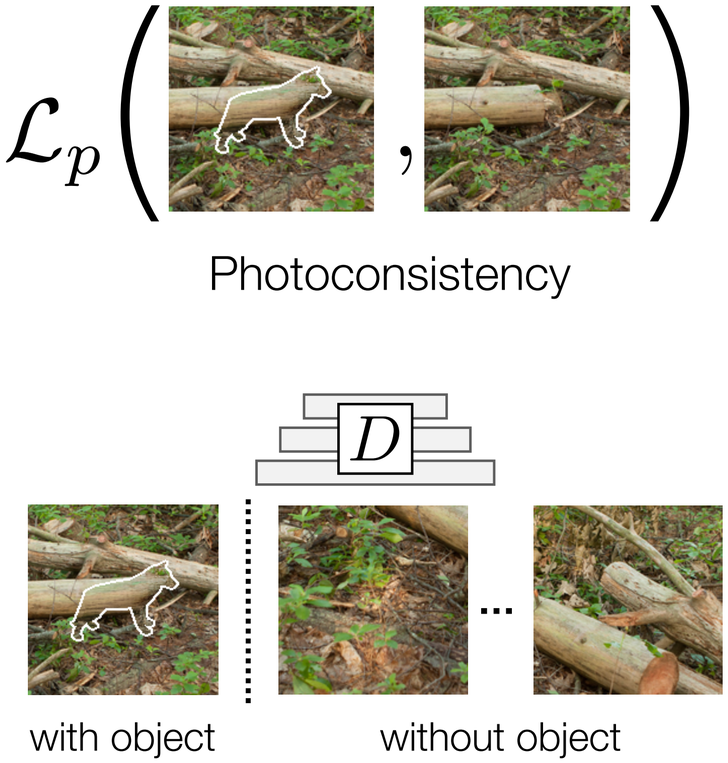
GANmouflage model



Multi-view camouflage

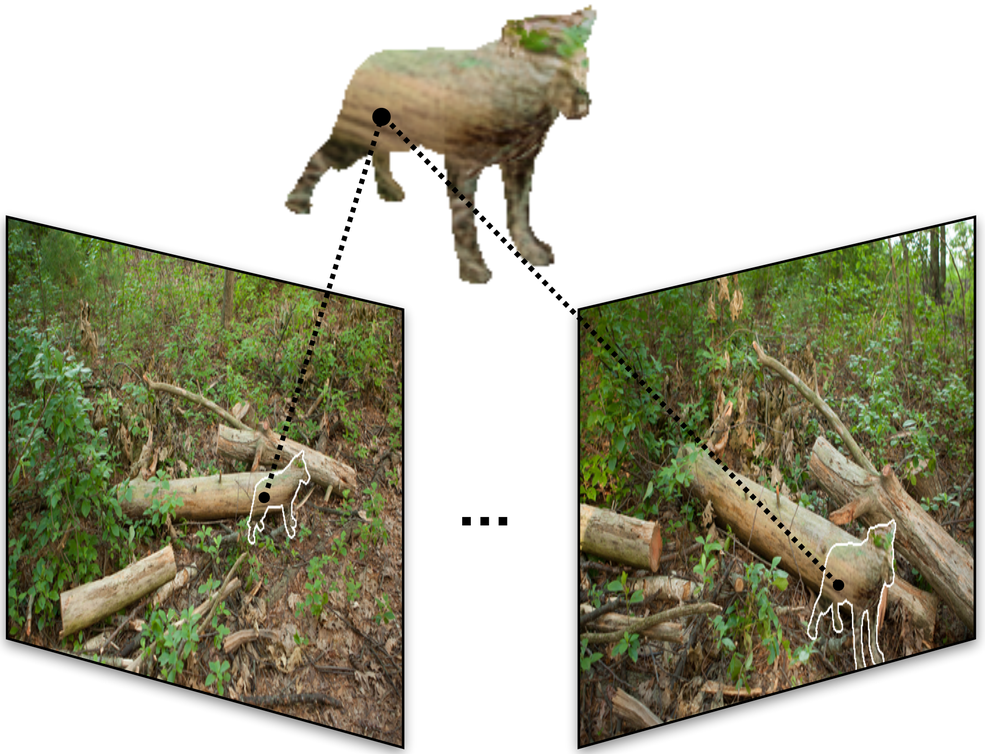


Texture model

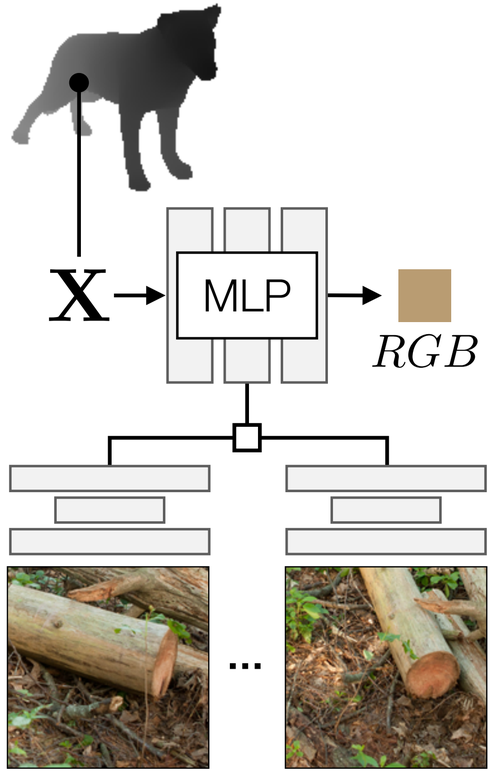


Adversarial loss

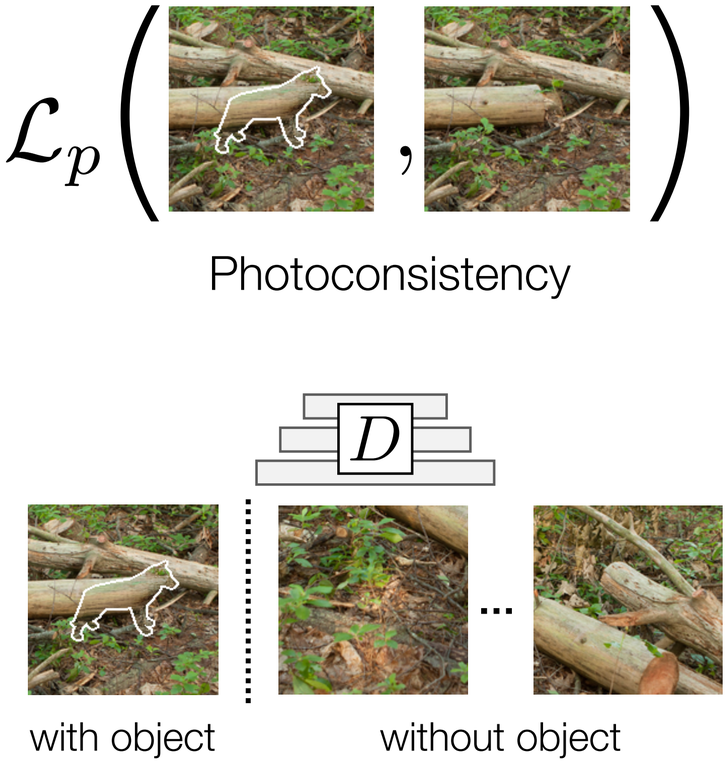
GANmouflage model



Multi-view camouflage

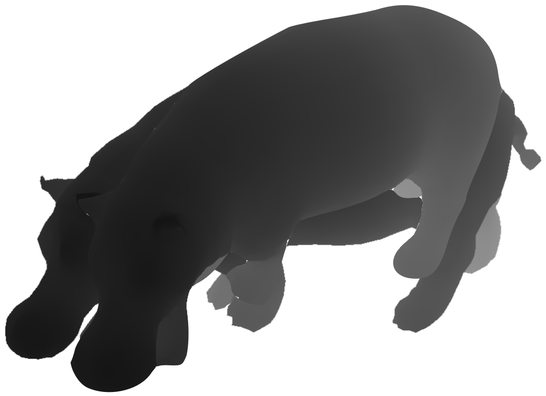


Texture model



Adversarial loss

Multi-view camouflage



Virtual object

Photo of a scene

Object inserted into same pose

Multi-view camouflage

Camouflaged object

Multiple viewpoints of camouflaged object

Object inserted into stage


More viewpoints

Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models

Daniel Geng, Inbum Park, Andrew Owens

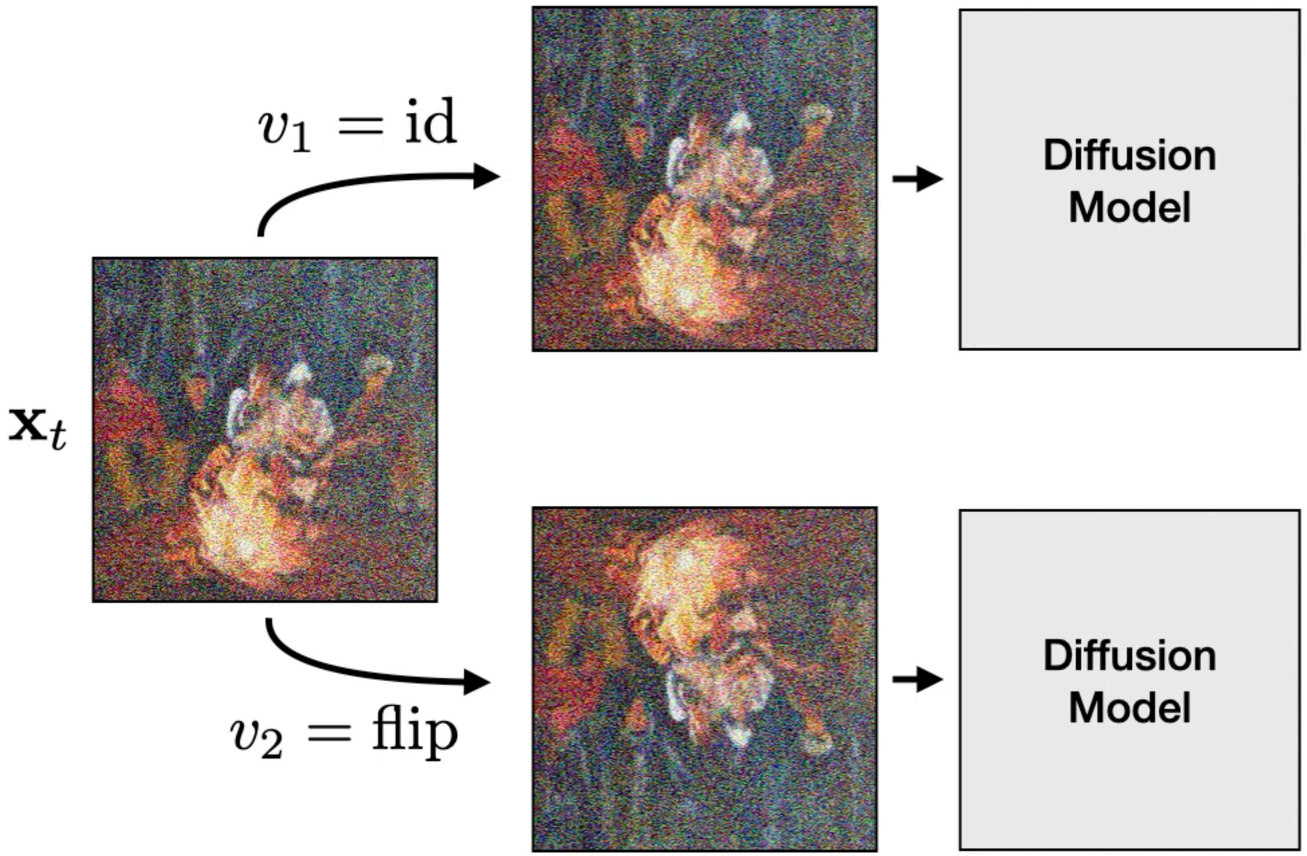
arXiv 2023

"The Fruit Basket." Giuseppe Arcimboldo, 1590.

Creating illusions

$\mathbf{x}_t$ 

Creating illusions



Multi-View Optical Illusions

Visual Anagrams