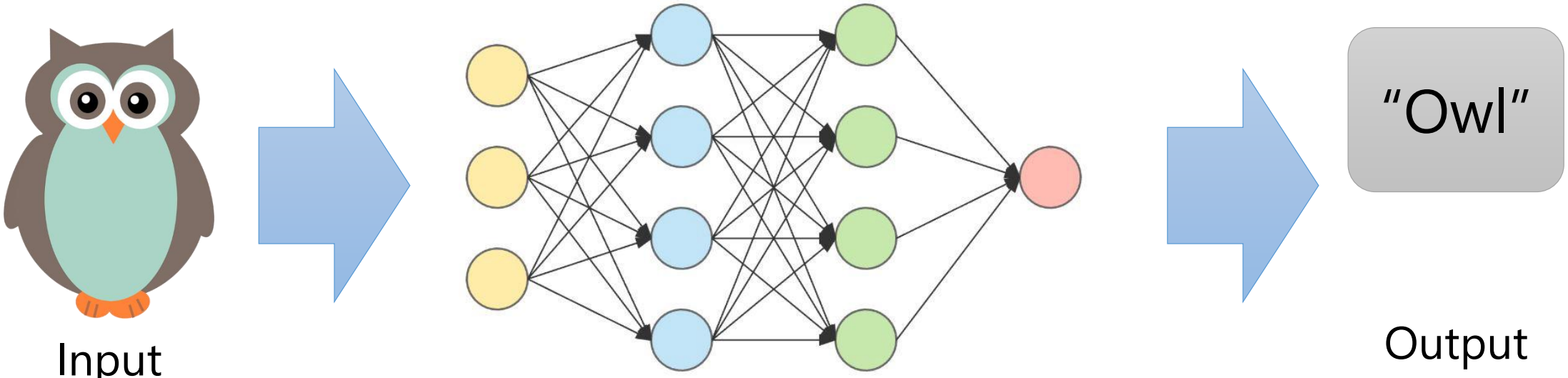
Weak Supervision, n-shot Learning, Data Efficiency

Prof. Angela Dai

Brief Recap

Deep Learning



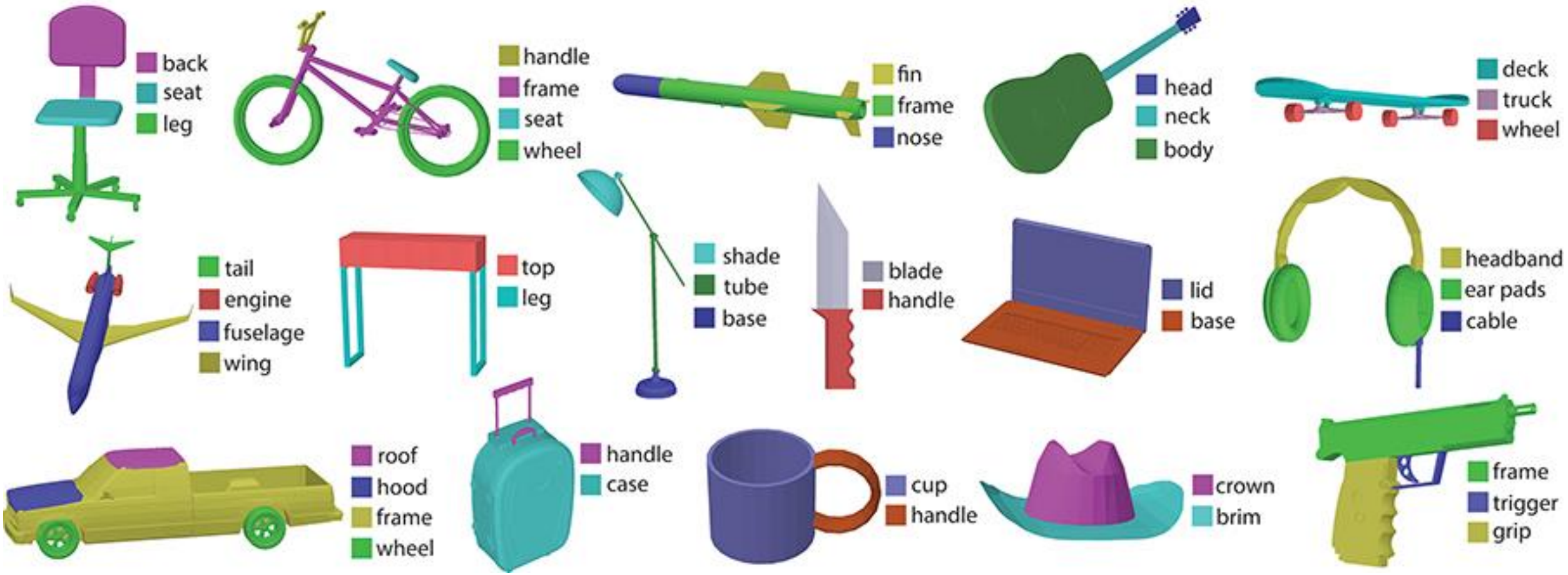
Input

Learned Feature Extraction
and Classification

Output

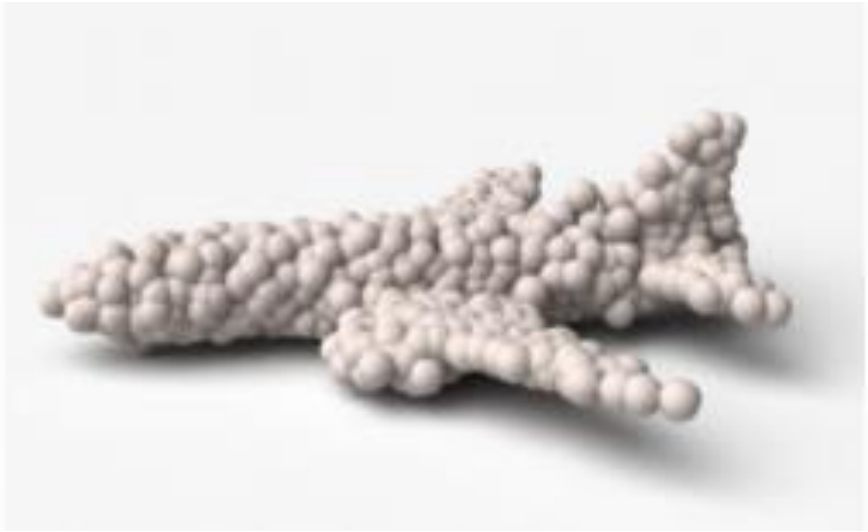
Want to automatically learn good feature representations for the task

Shape segmentation into parts

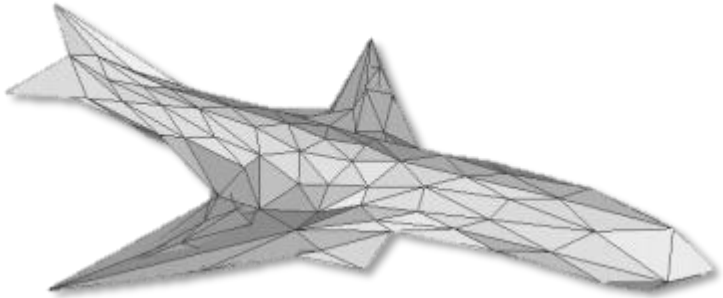


Generating Shapes

Signed Distance Fields

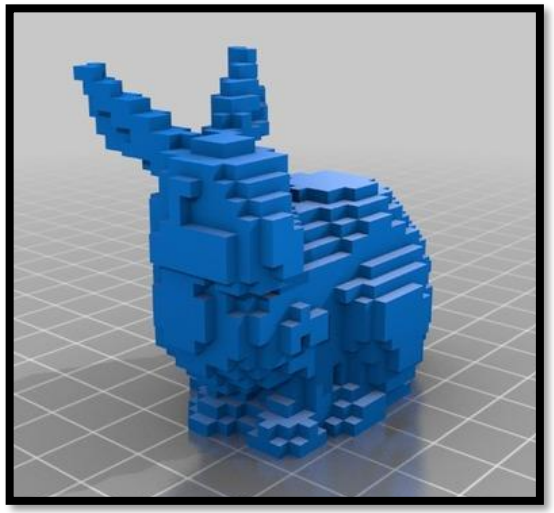


Point Clouds

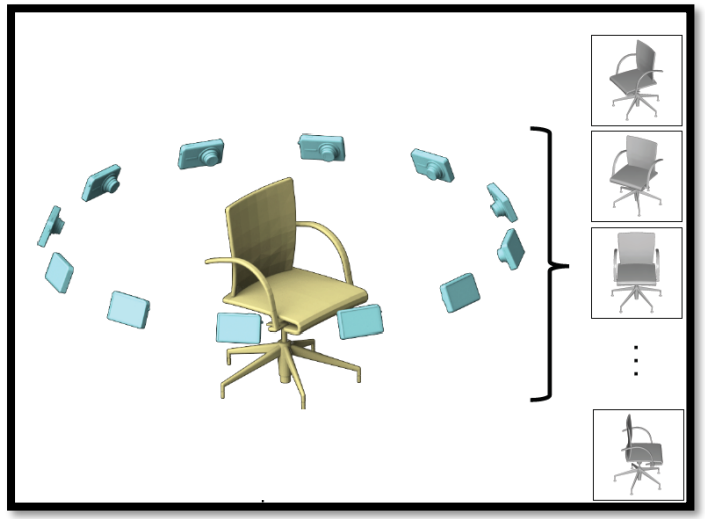


Meshes

3D Deep Learning by Representations

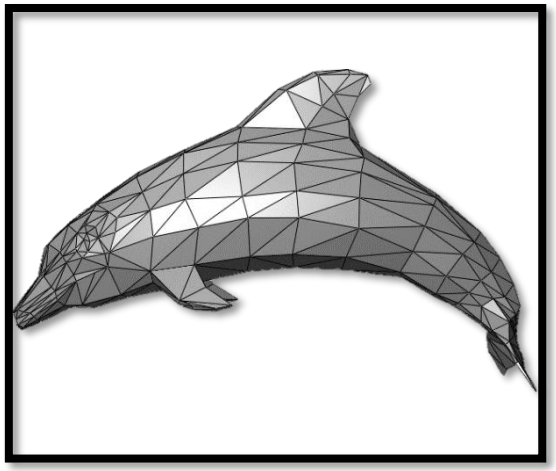


Volumetric
3D CNNs: Dense, Hierarchical, Sparse



Multi-View
(also: multi-view + volumetric/point/mesh)

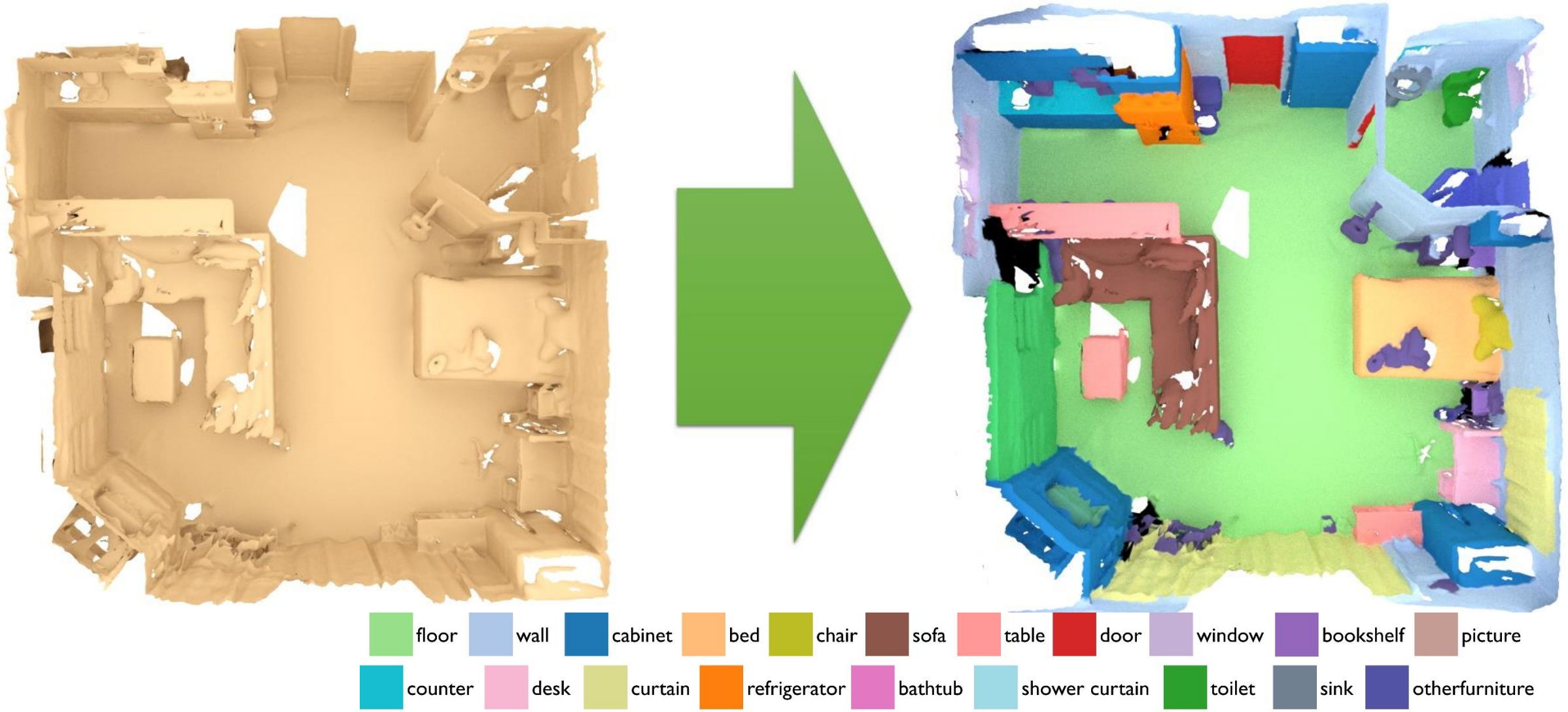
Point Cloud



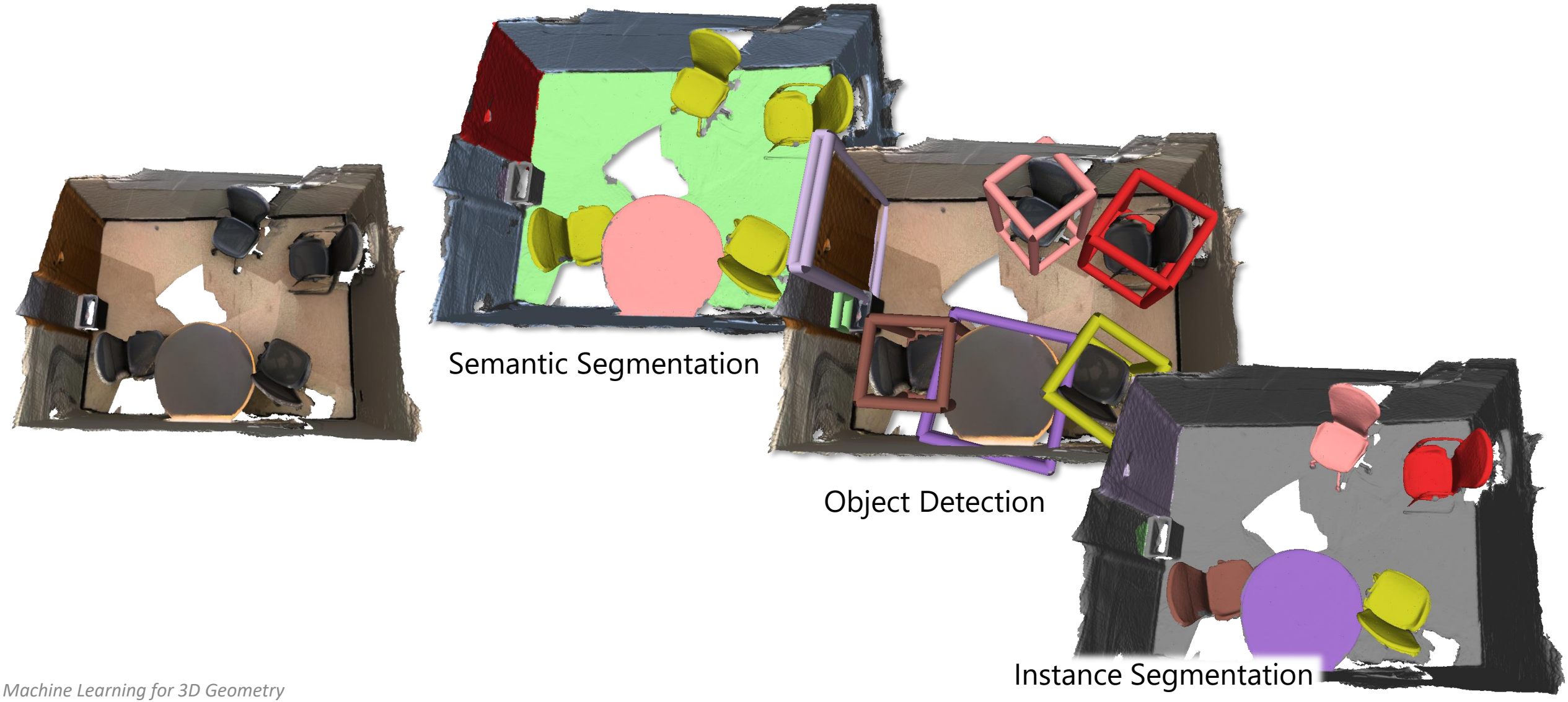
Mesh
Graph Neural Networks

and more!

3D Semantic Segmentation



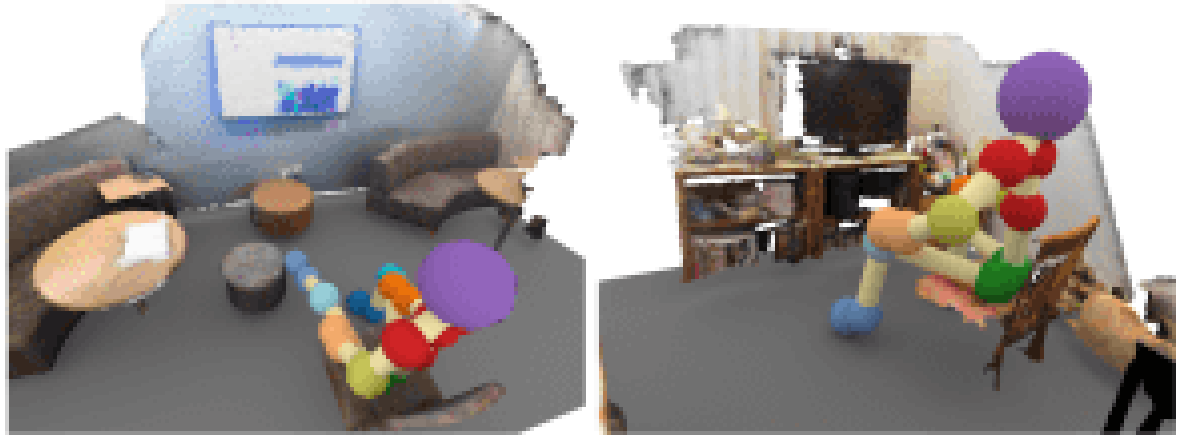
Understanding object-ness



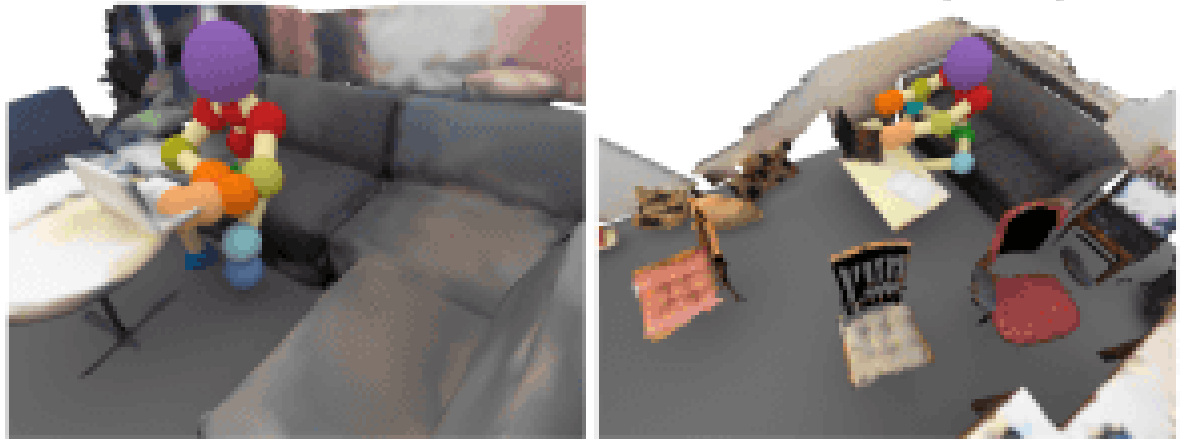
Generating 3D Scenes

Understanding interactions

"Sit on a chair and watch TV"



"Sit on a couch and use a laptop"



"Write on a whiteboard"

[Savva et al. '16]

Challenge: data for supervised learning

- Data collection is expensive
- Data annotation is expensive
- Especially for 3D / 4D data!

Training vs amount of labeled data

Traditional Supervision: Expert annotators manually label data (expensive)

Unsupervised: Learn structural patterns from unlabeled data

Semi-Supervision: Work with labeled set of data and an unlabeled set of data

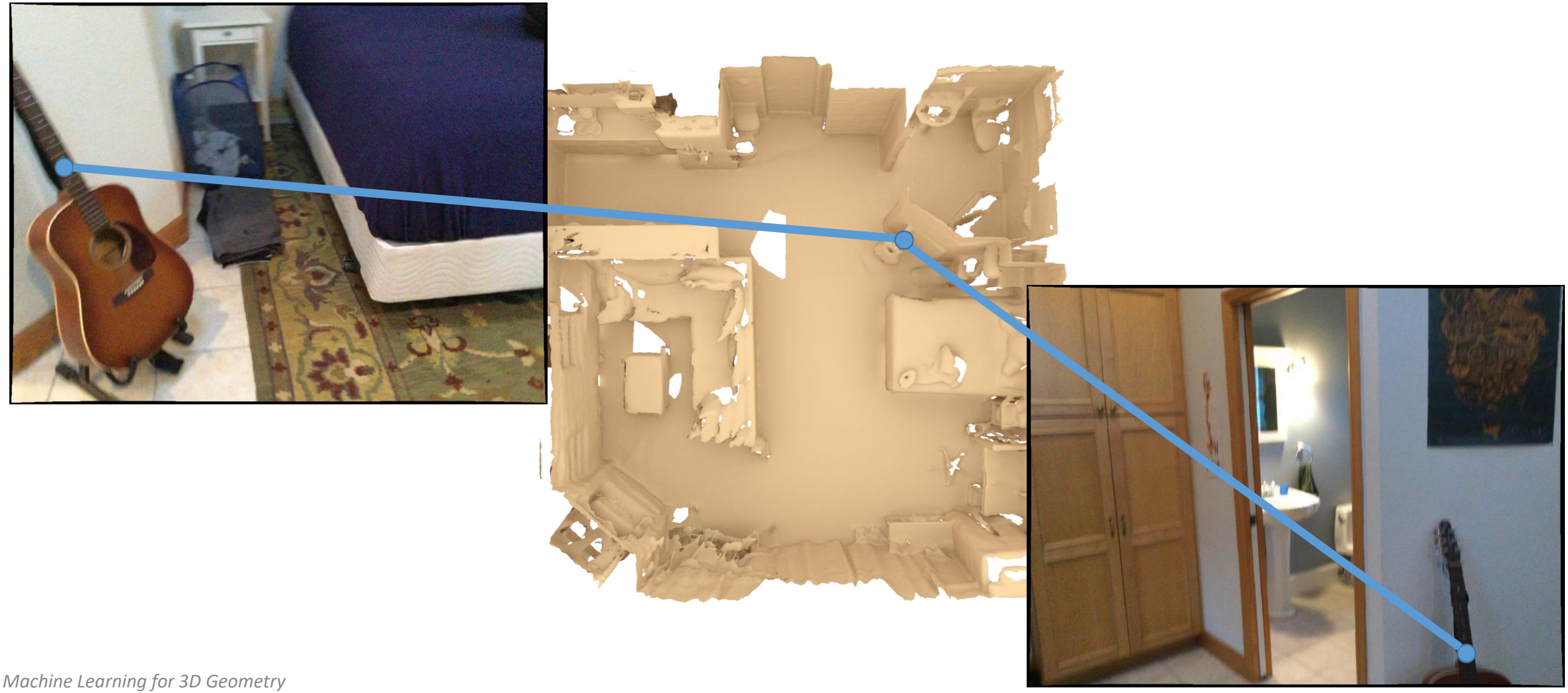
Weak Supervision: Use lower-quality or noisy labels that are easier to get

Self-Supervision: Automatically generate supervisory signal

Transfer Learning: Leverage models trained on a different task

Other hybrids: Active learning to more efficiently leverage manual annotations

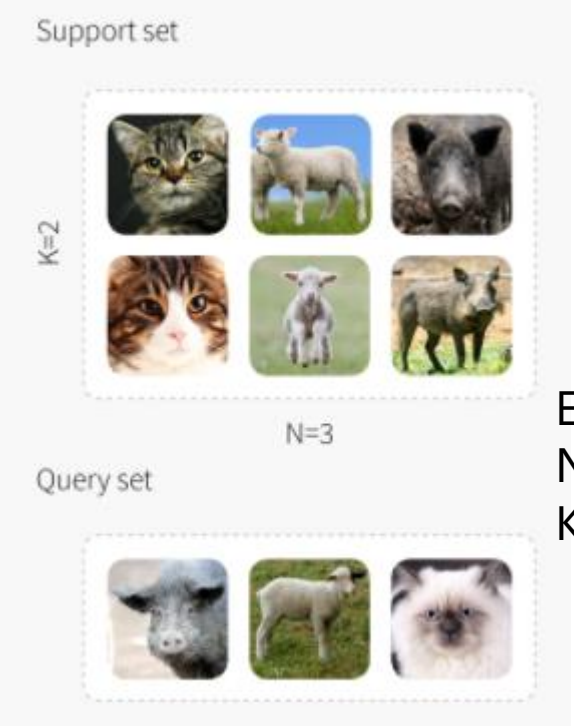
Interesting constraint for 3D: multi-view



Few-shot learning

- Challenge: collecting labeled data for every category to be handled
- Goal: handle new data having seen only a few training examples

- One-shot: only one example available



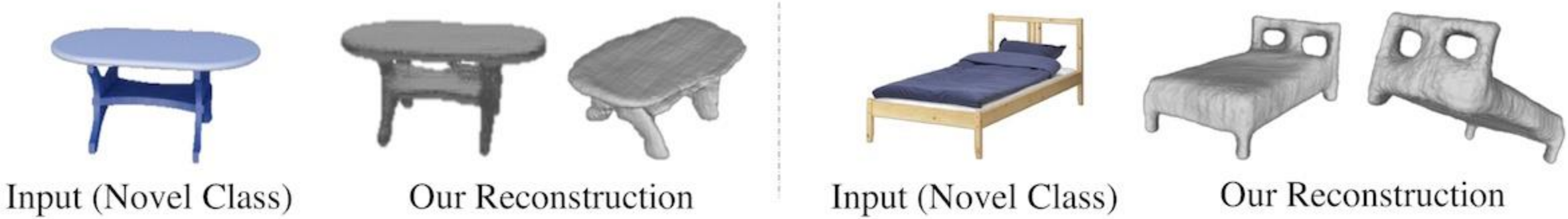
Ex: classification task
N=3 classes
K=2 known examples

Reconstruct Shapes from Unseen Classes?

- Want to avoid overfitting to global structures
- Often: formulate lower-level reconstruct task that enables better generalization

Reconstructing Unseen Classes: Depth Inpainting

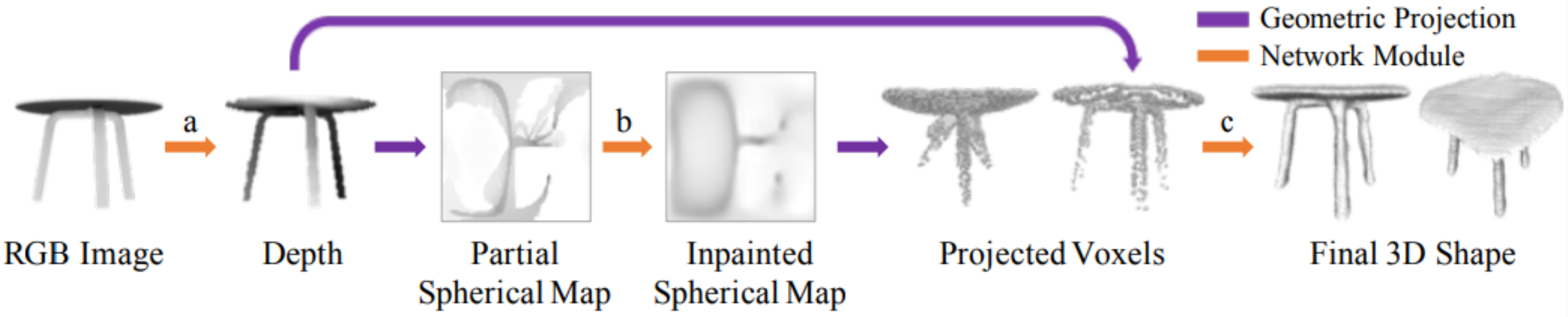
- Learn 3D shape priors from known dataset of shapes from n categories; aim to reconstruct 3D shape from images of unseen categories



[Zhang et al. ‘18]

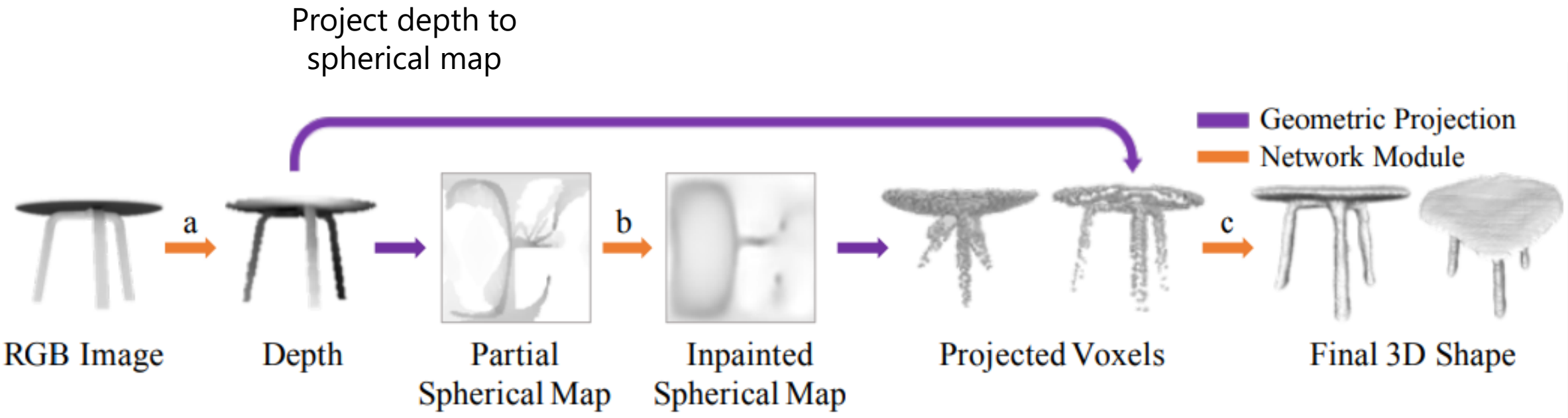
Reconstructing Unseen Classes: Depth Inpainting

Depth estimation from a single image



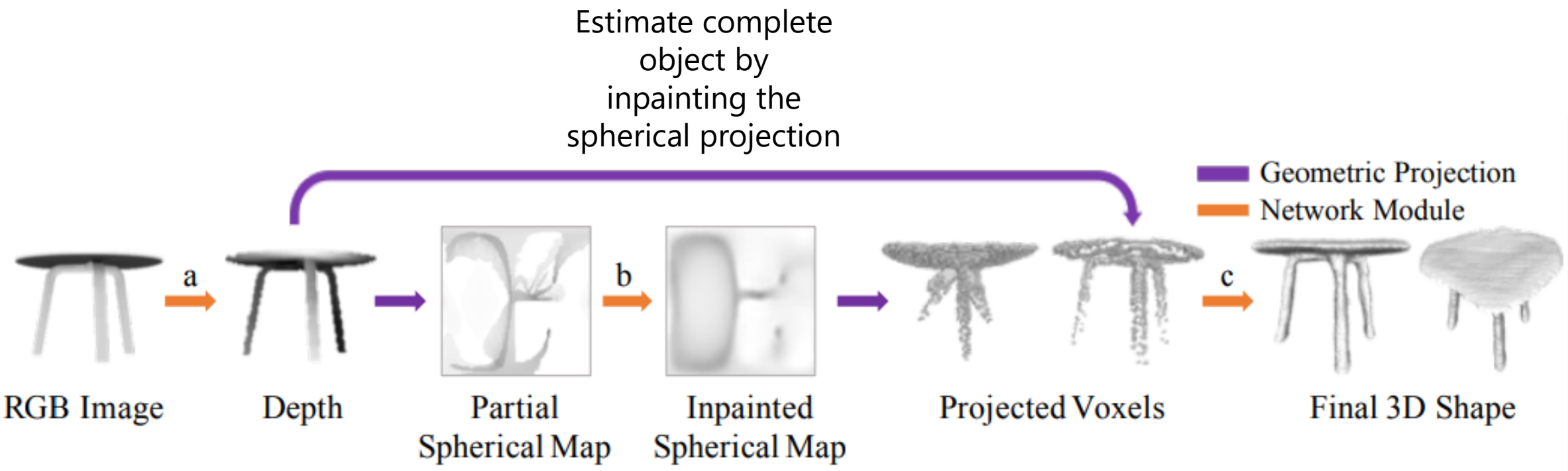
[Zhang et al. '18]

Reconstructing Unseen Classes: Depth Inpainting



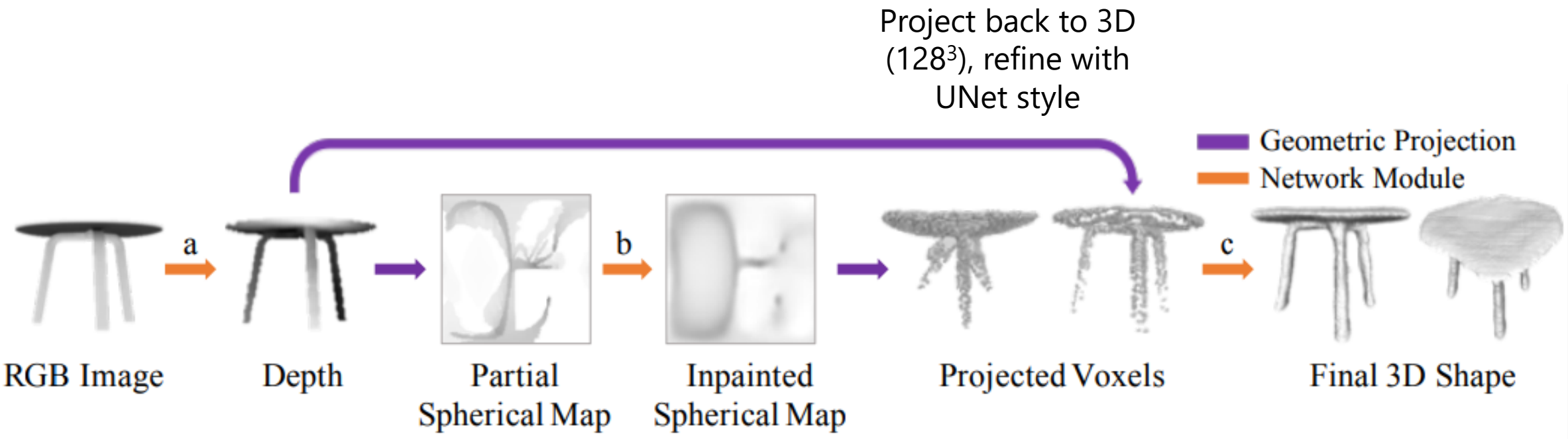
[Zhang et al. '18]

Reconstructing Unseen Classes: Depth Inpainting



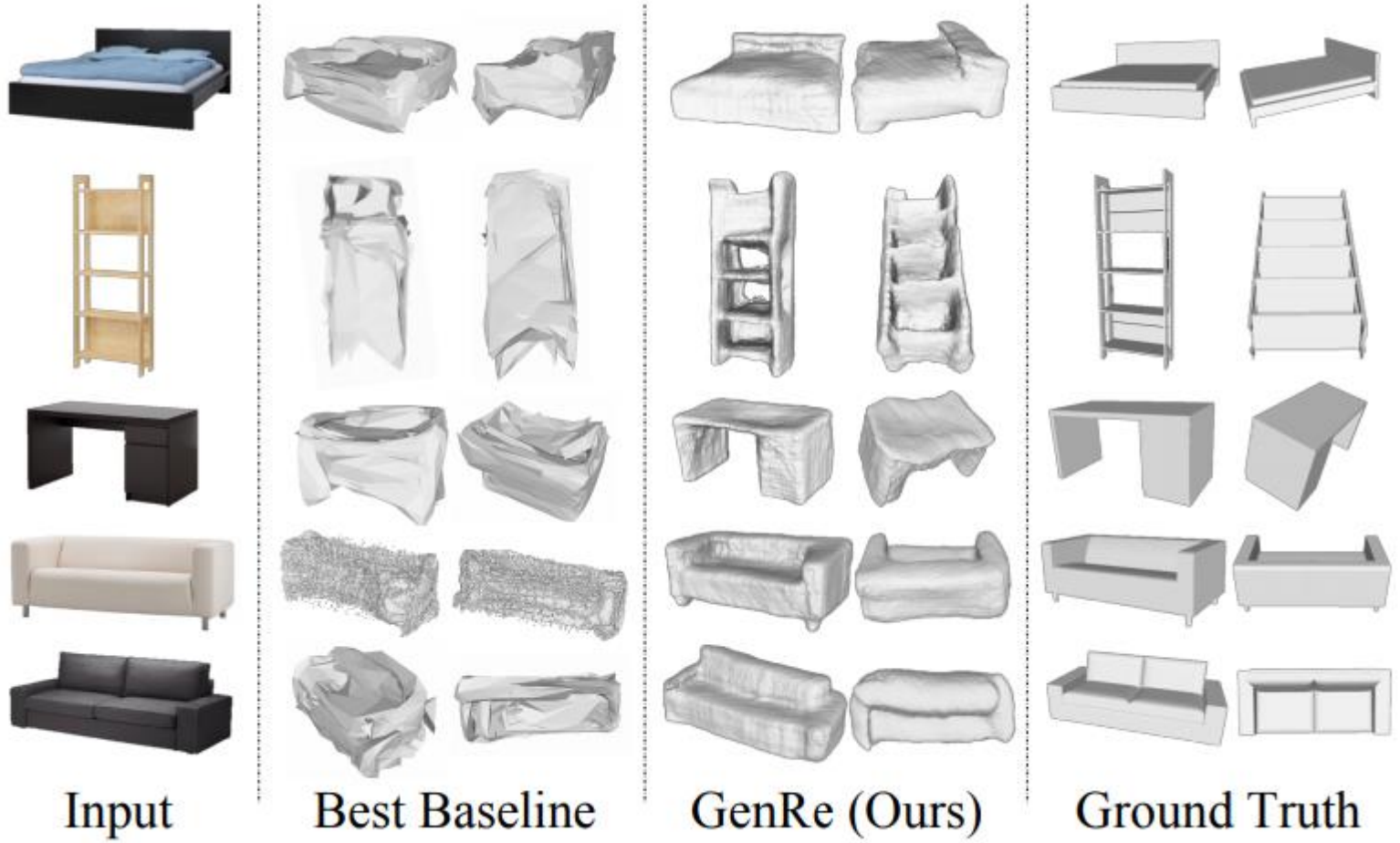
[Zhang et al. '18]

Reconstructing Unseen Classes: Depth Inpainting



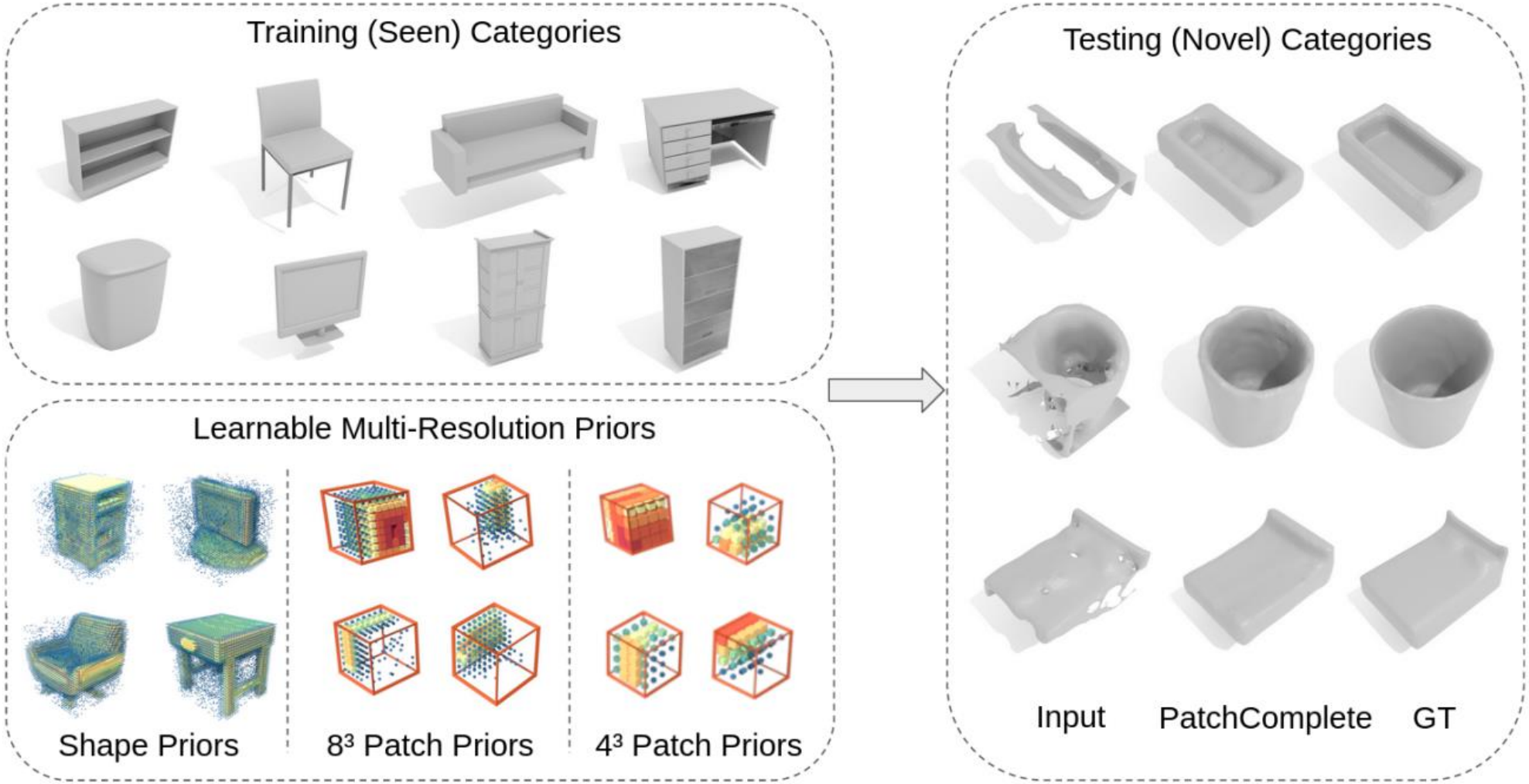
[Zhang et al. '18]

Reconstructing Unseen Classes: Depth Inpainting



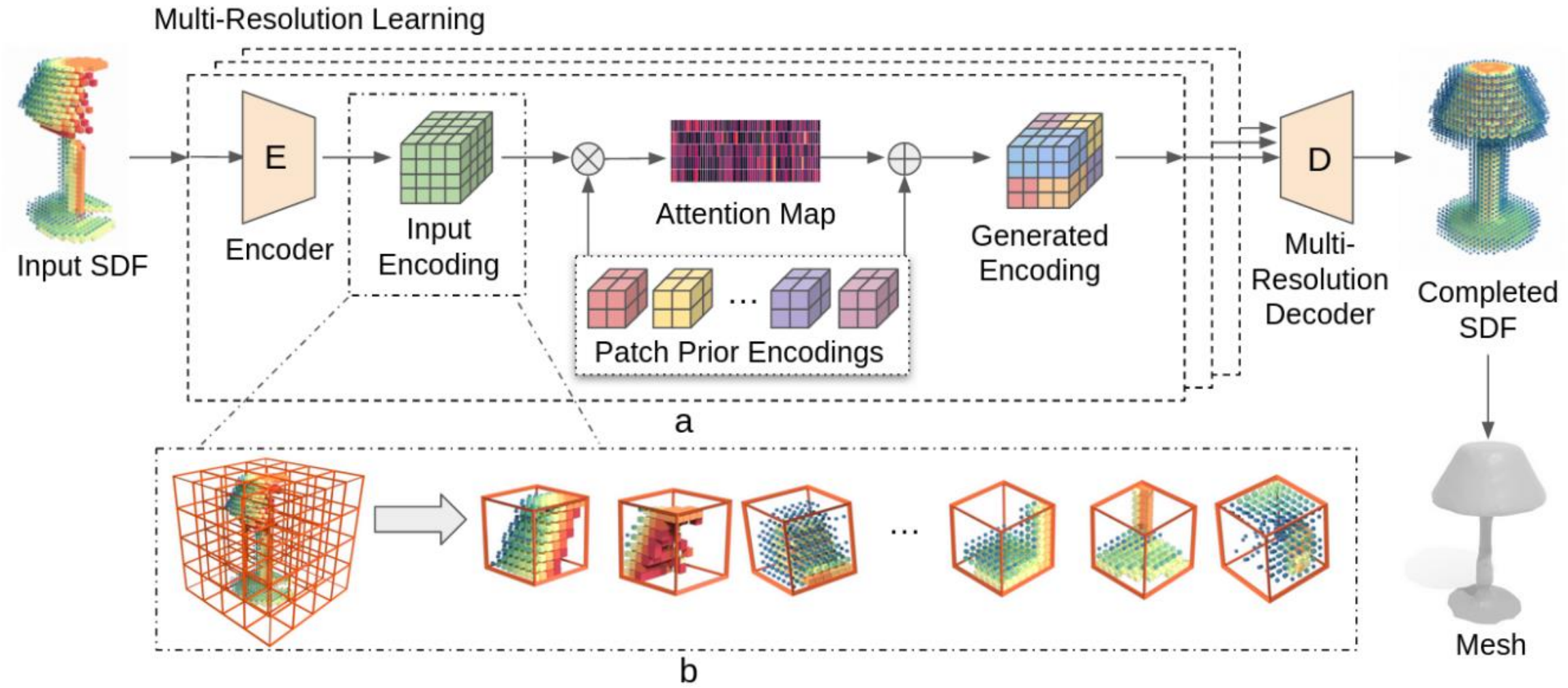
[Zhang et al. ‘18]

Reconstructing Unseen Classes: Patch Priors



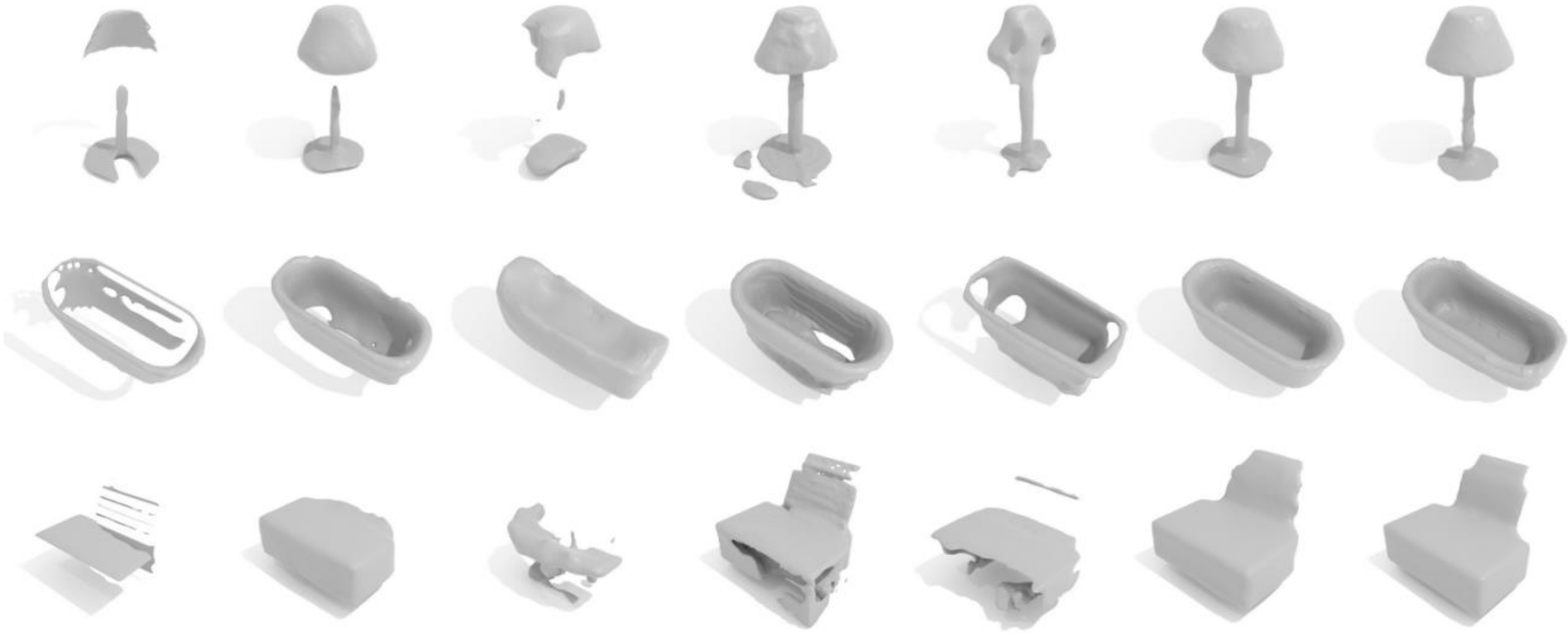
[Rao et al. '22]

Reconstructing Unseen Classes: Patch Priors



[Rao et al. '22]

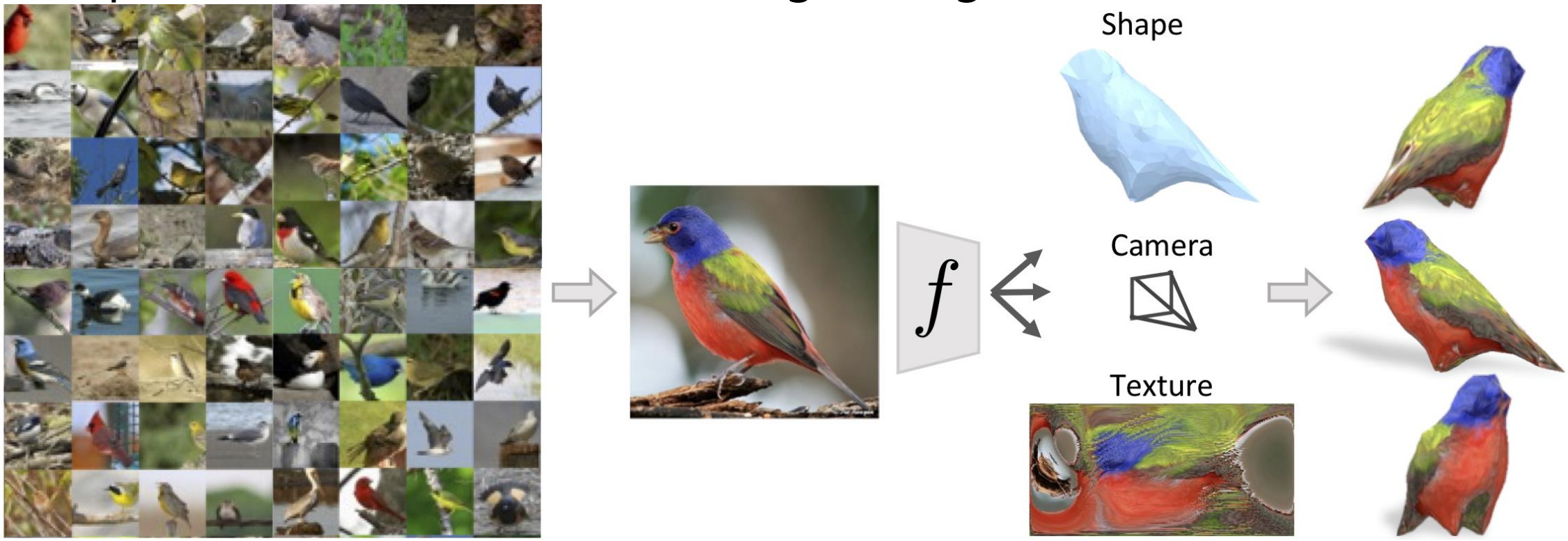
Reconstructing Unseen Classes: Patch Priors



[Rao et al. ‘22]

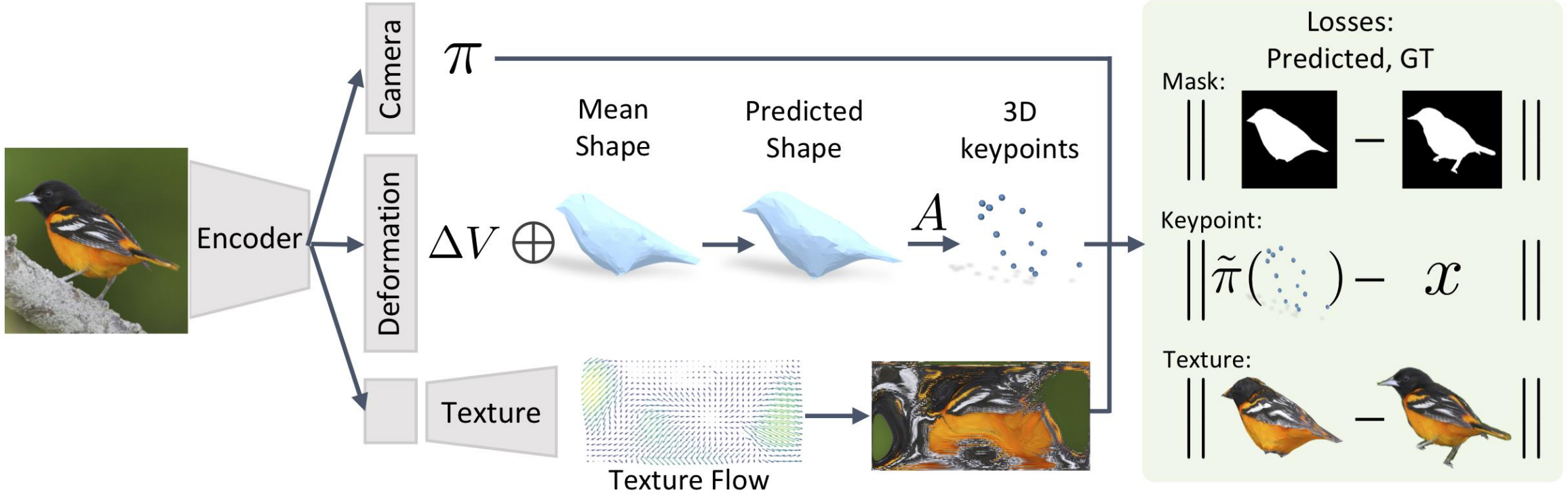
Learning Category-Specific Mesh Reconstruction

- Given annotated images from an object category, estimate 3D shape, camera, texture from a single image



[Kanazawa et al. ‘18]

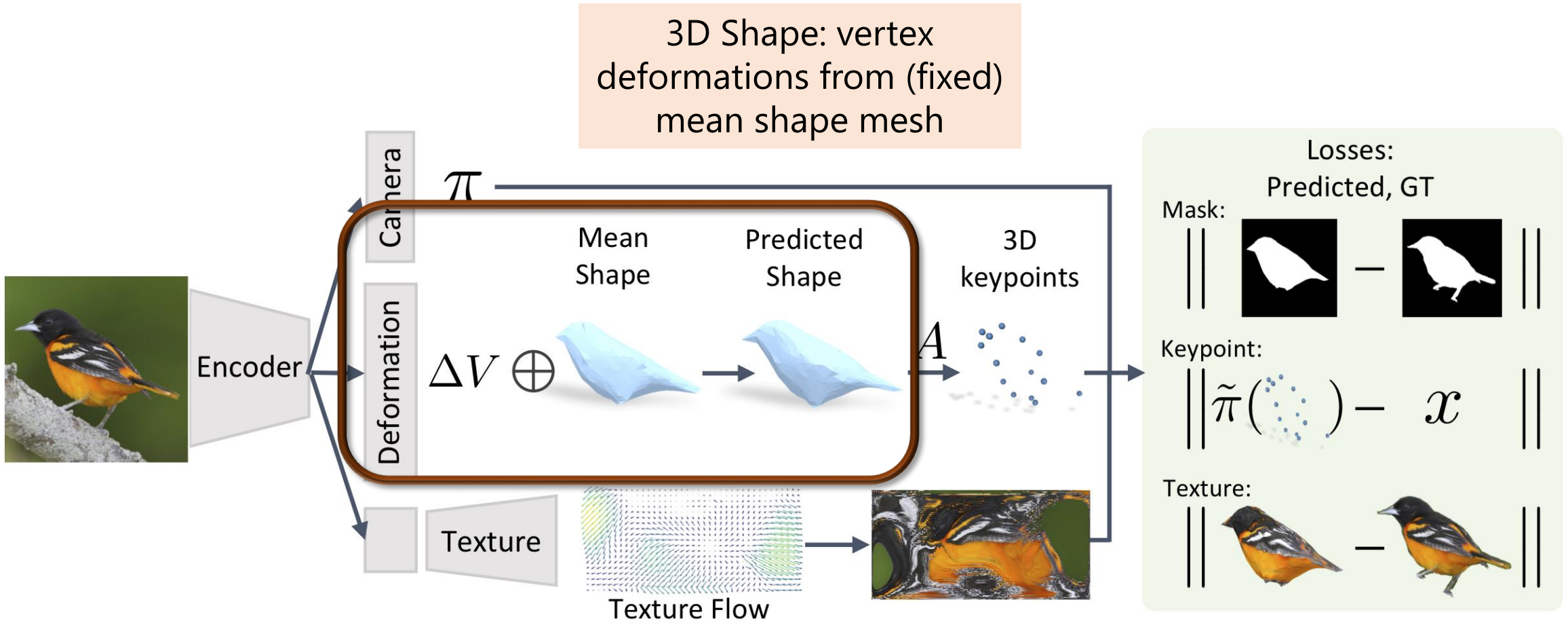
Learning Category-Specific Mesh Reconstruction



[Kanazawa et al. '18]

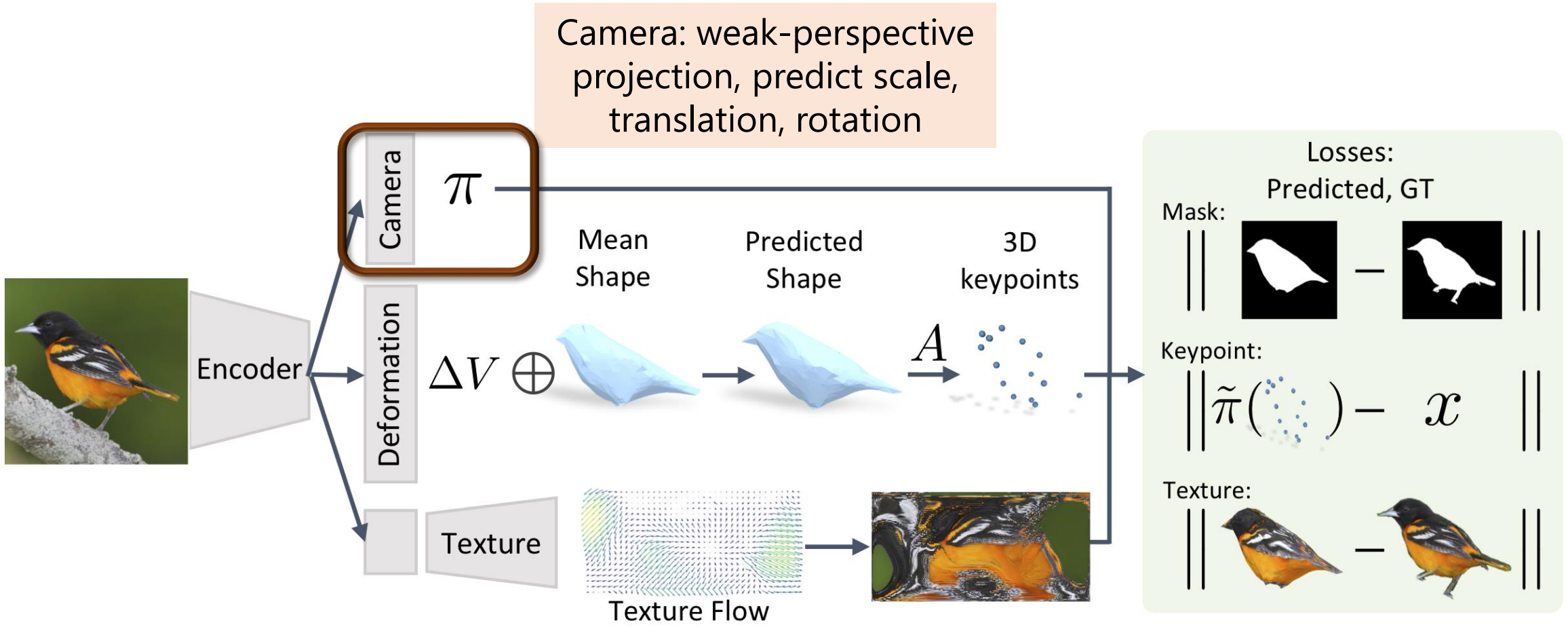
Learning Category-Specific Mesh Reconstruction

3D Shape: vertex deformations from (fixed) mean shape mesh



[Kanazawa et al. '18]

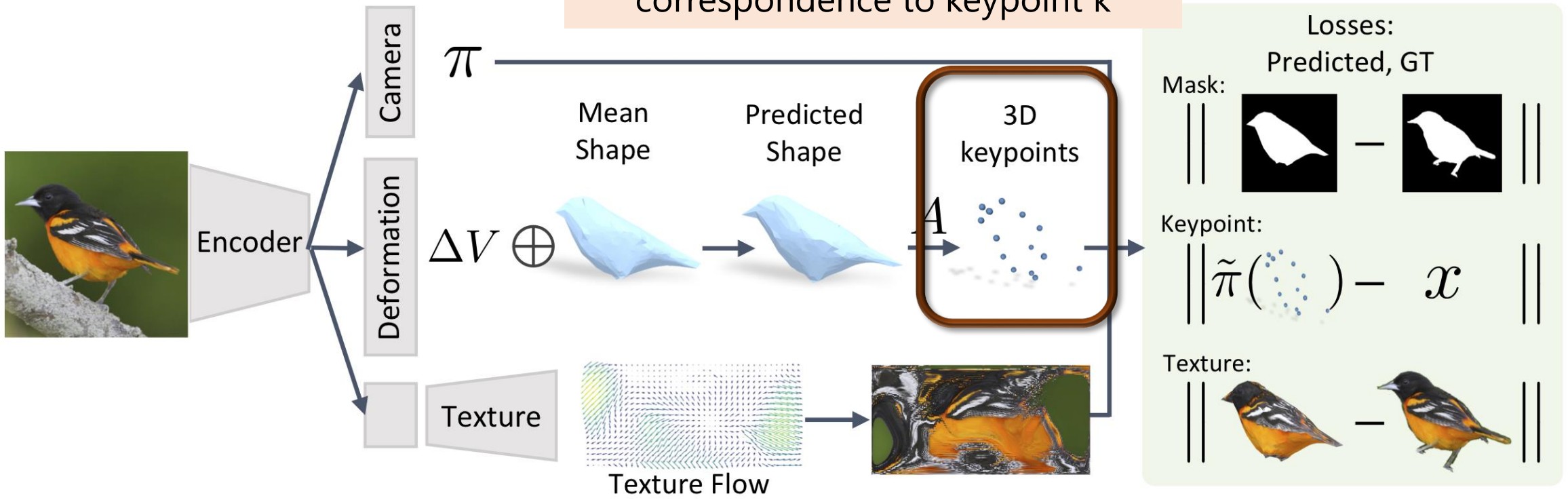
Learning Category-Specific Mesh Reconstruction



[Kanazawa et al. '18]

Learning Category-Specific Mesh Reconstruction

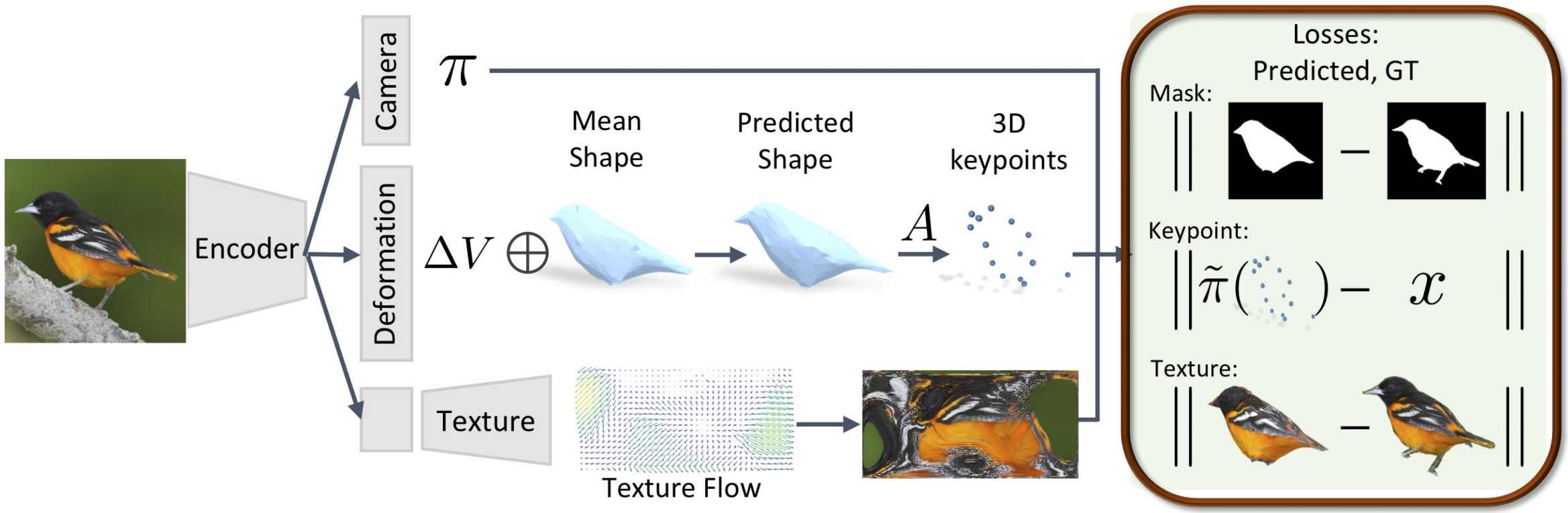
Keypoint supervision: learn to associate semantic keypoints; predict probability for each vertex its correspondence to keypoint k



[Kanazawa et al. '18]

Learning Category-Specific Mesh Reconstruction

Losses: based on 2D annotations from images (segmentation mask, keypoints) and texture (mask*image)



[Kanazawa et al. '18]

Learning Category-Specific Mesh Reconstruction

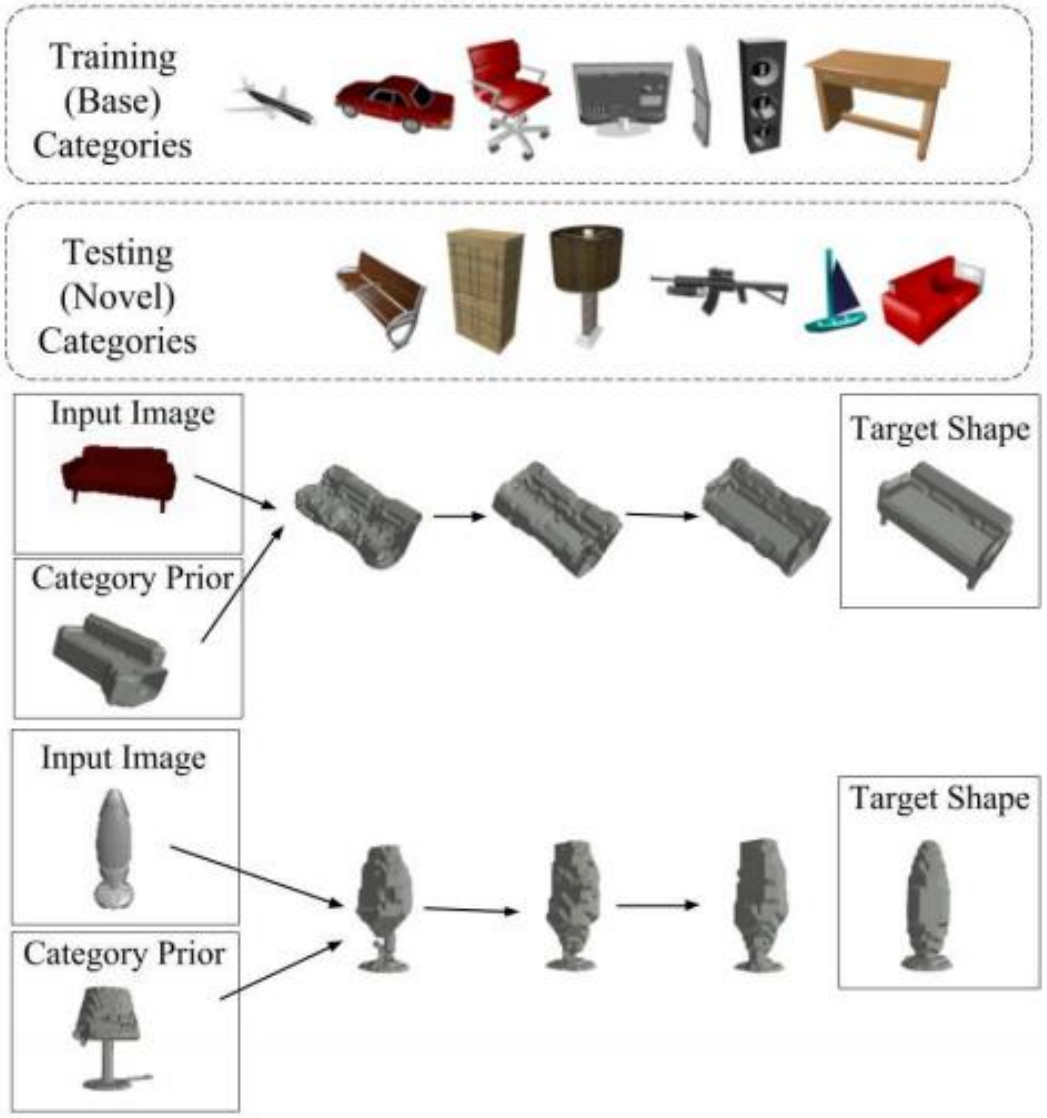
[Kanazawa et al. ‘18]

Learning Category-Specific Mesh Reconstruction

[Kanazawa et al. ‘18]

Few-shot single image reconstruction

Image to voxel reconstruction using a category-specific prior shape (average of category examples)

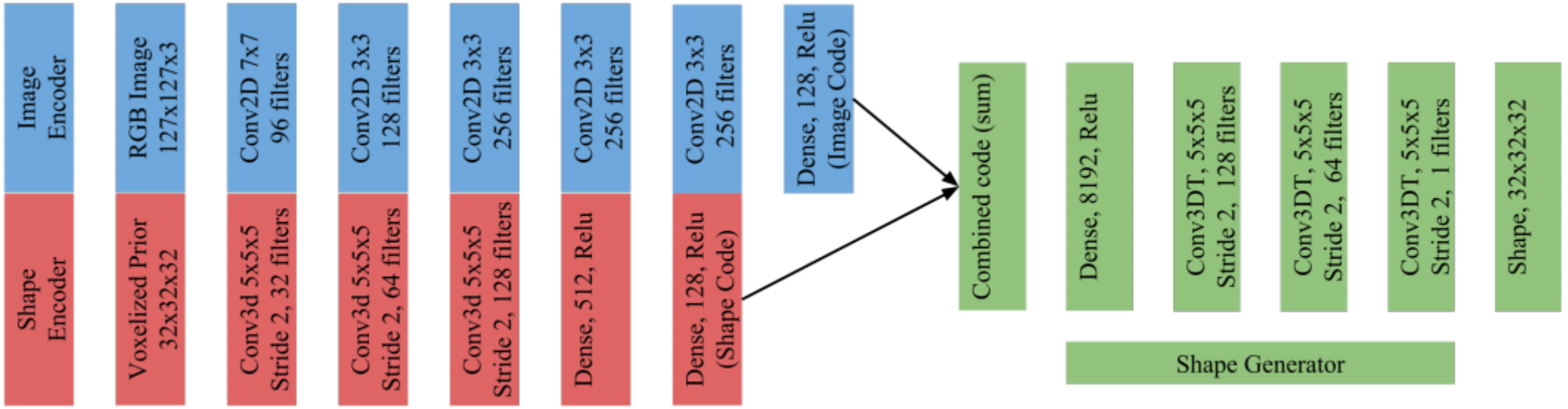


Train categories have image/3D pairs. New categories at test time see limited number of 3D shapes (≤ 25)

[Wallace and Hariharan '19]

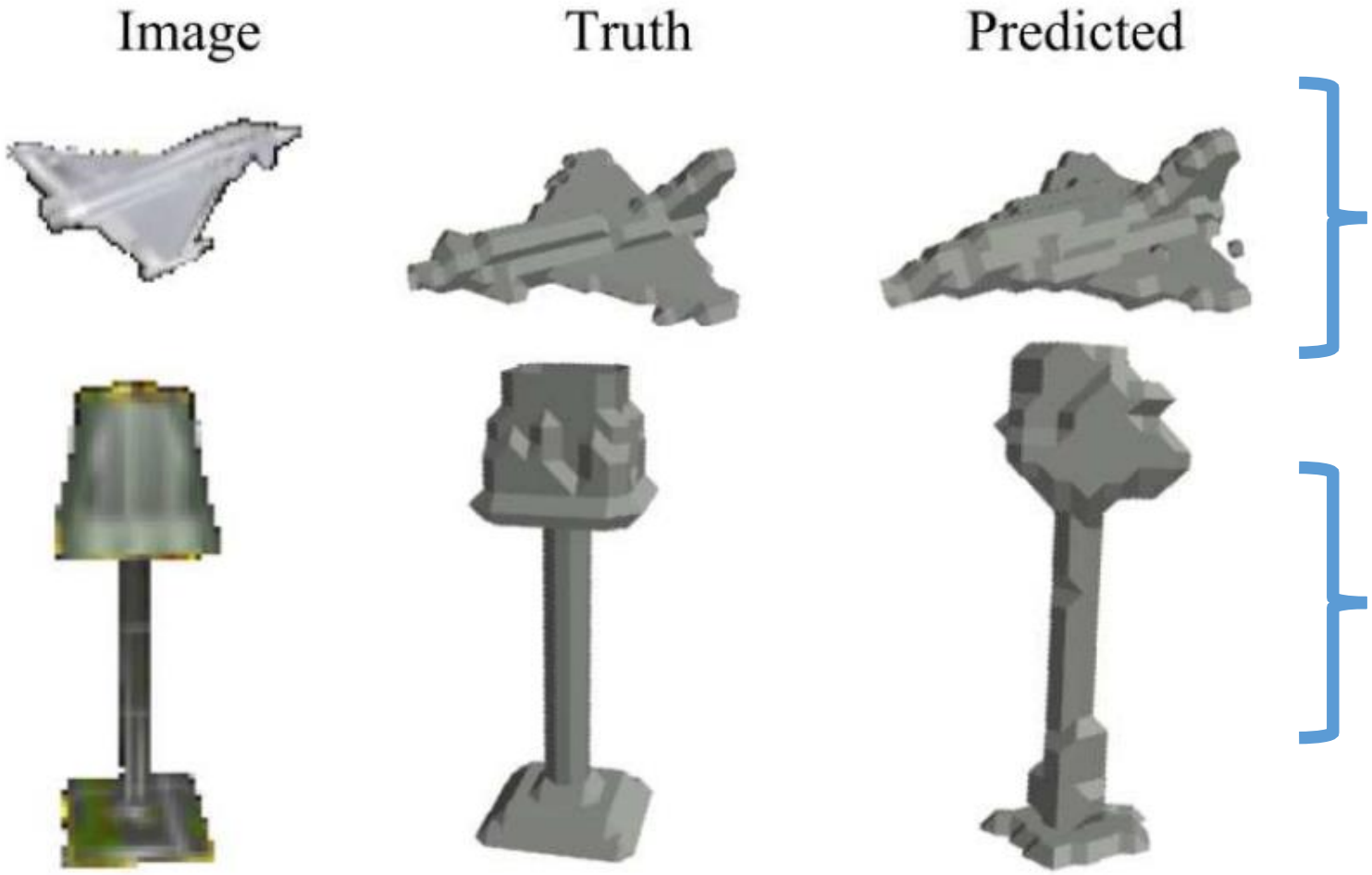
Few-shot single image reconstruction

- Encode input image and 32^3 shape prior; decode 32^3 shape



[Wallace and Hariharan '19]

Few-shot single image reconstruction



Reconstruction of new image from a train category

Reconstruction of new image from a unseen category

[Wallace and Hariharan '19]

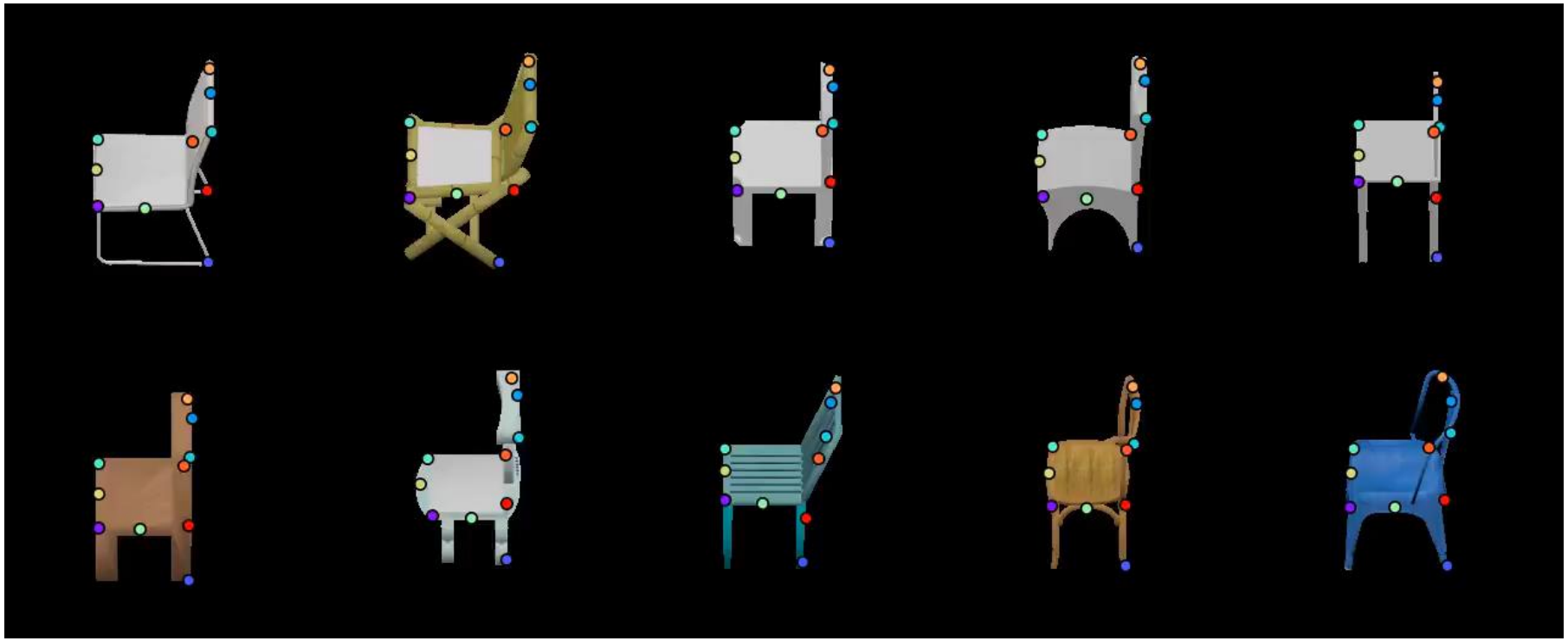
Train in Germany, Test in The USA

- Generalizing 3D object detectors across datasets
- LIDAR measurements from different different LIDAR models
- Applying state-of-the-art point-based detectors -> perf. gap
- Observation: main factor in car detection is varying car sizes (more mislocalization than misdetection)
- Domain adaptation by data normalization

[Wang et al. '20]

Discovery of Latent 3D Keypoints

- End-to-end learning of category-specific 3D keypoints and detectors

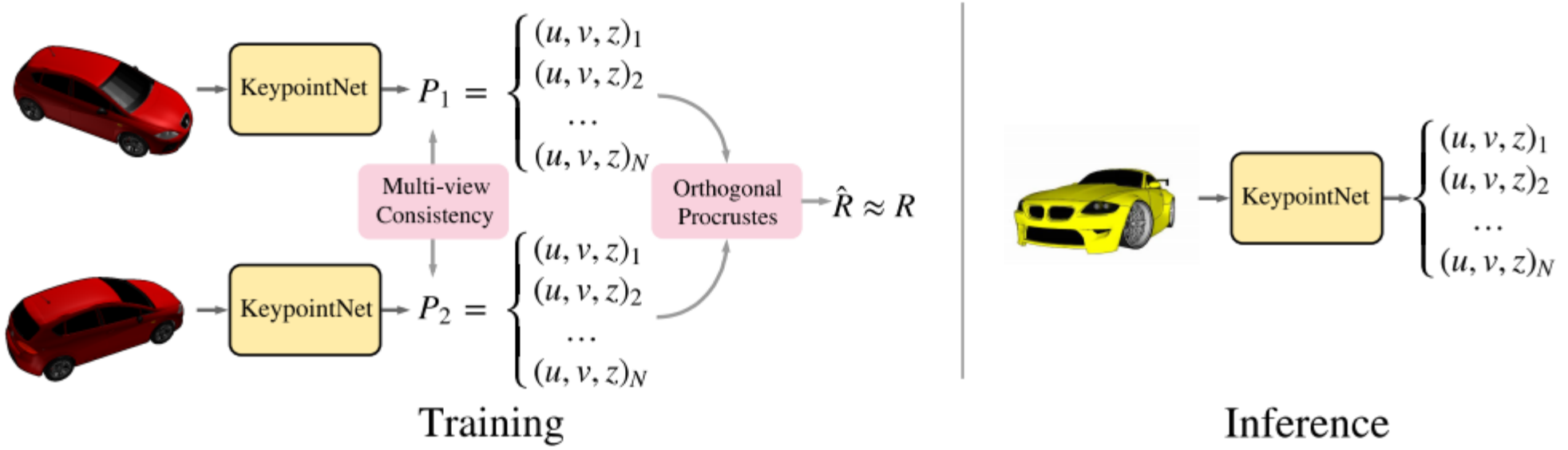


No manual annotation of keypoints required

[Suwajanakorn et al. ’18]

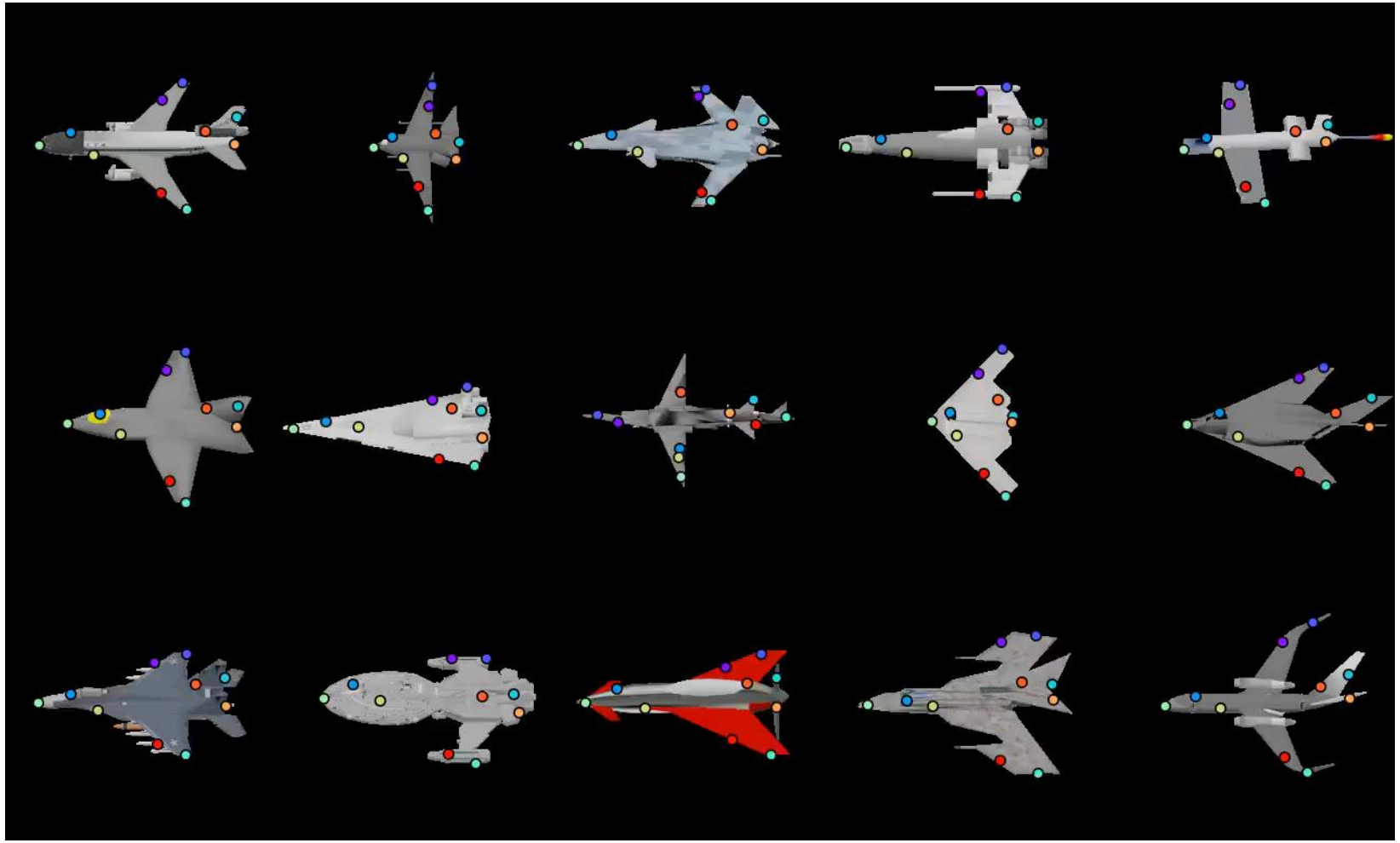
Discovery of Latent 3D Keypoints

- End-to-end learning of category-specific 3D keypoints and detectors
- End task of 3D pose estimation of an object from an image



[Suwajanakorn et al. ’18]

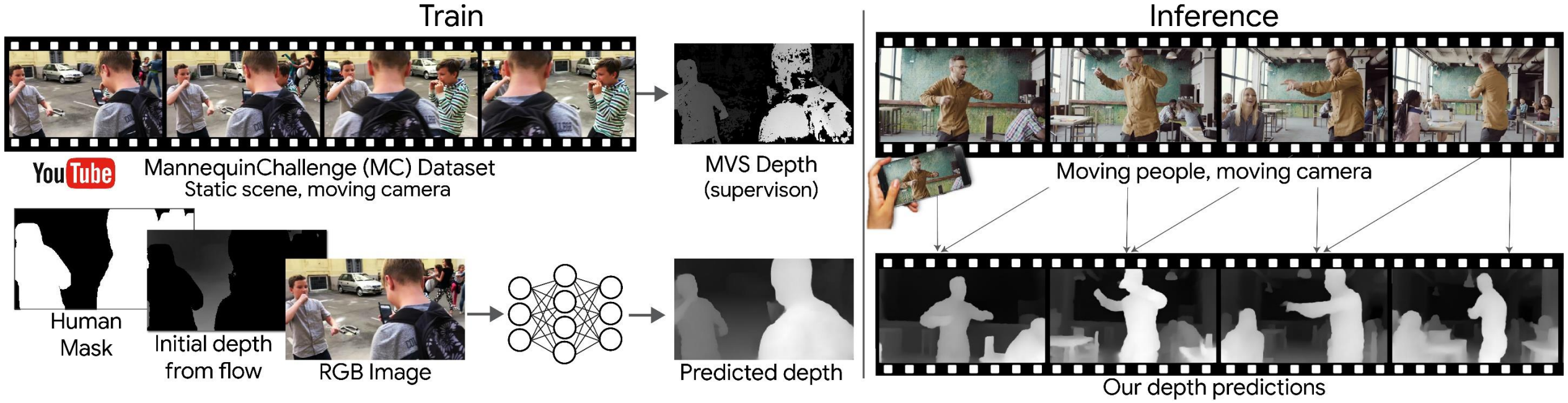
Discovery of Latent 3D Keypoints



[Suwajanakorn et al. ’18]

Learning the Depths of Moving People

- Dense depth prediction from monocular videos of people
  - Dynamic camera and person movements
- Train on internet videos of people imitating mannequins



[Li et al. '19]

Learning the Depths of Moving People

- MannequinChallenge Dataset
  - 2000 YouTube Videos
  - People all staying still while camera moving

https://youtu.be/qFaUhLkdRPg

[Li et al. ‘19]

Learning the Depths of Moving People

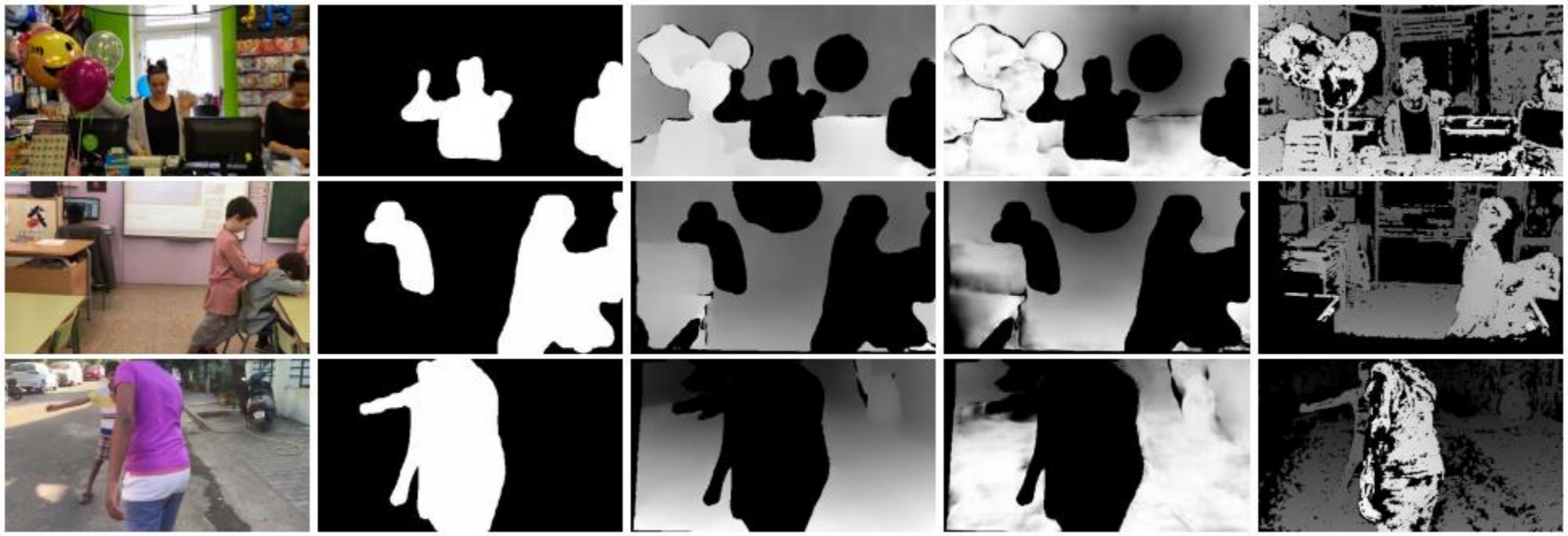
- MannequinChallenge Dataset
  - 2000 YouTube Videos
  - People all staying still while camera moving
  - Stationary people -> rigid scene -> use structure from motion and multi-view stereo to get depth estimates to use for supervision
  - Semi-dense depth supervision from SFM

https://youtu.be/qFaUhLkdRPg

[Li et al. '19]

Learning the Depths of Moving People

- Input: RGB image, human mask, masked depth (computed from motion parallax w.r.t source image), masked confidence



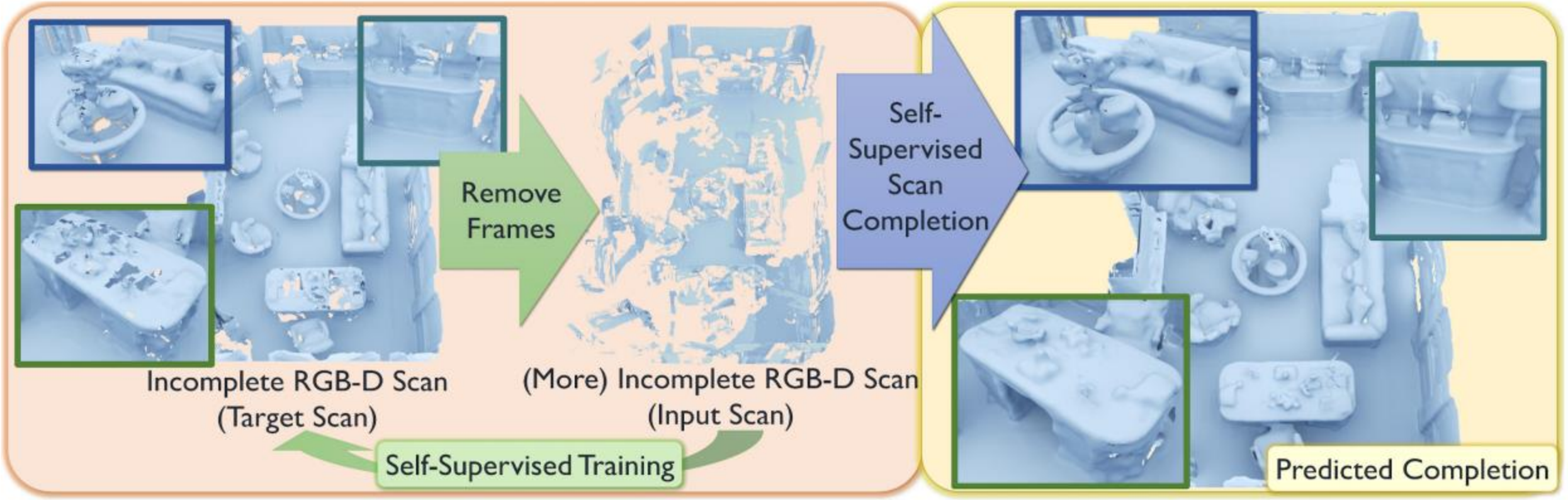
(a) Reference image I^r (b) Human mask M (c) Input depth D_{pp} (d) Input confidence C (e) MVS depth D_{MVS}

https://youtu.be/qFaUhLkdRPg

[Li et al. '19]

SG-NN: Self-supervised scan completion

- Generating completion data by removing frames
- Loss formulation to avoid penalizing incomplete targets



[Dai et al. '20]

PointContrast

- Unsupervised pre-training for point cloud semantic understanding

[Xie et al. ‘20]

2D Representation Learning

- SimCLR [Chen et al. ’20], MoCo [He et al. ’20]
- SimCLR:
- Minibatch of N examples -> contrastive prediction on augmented examples: $2N$ data points
- Contrastive loss function

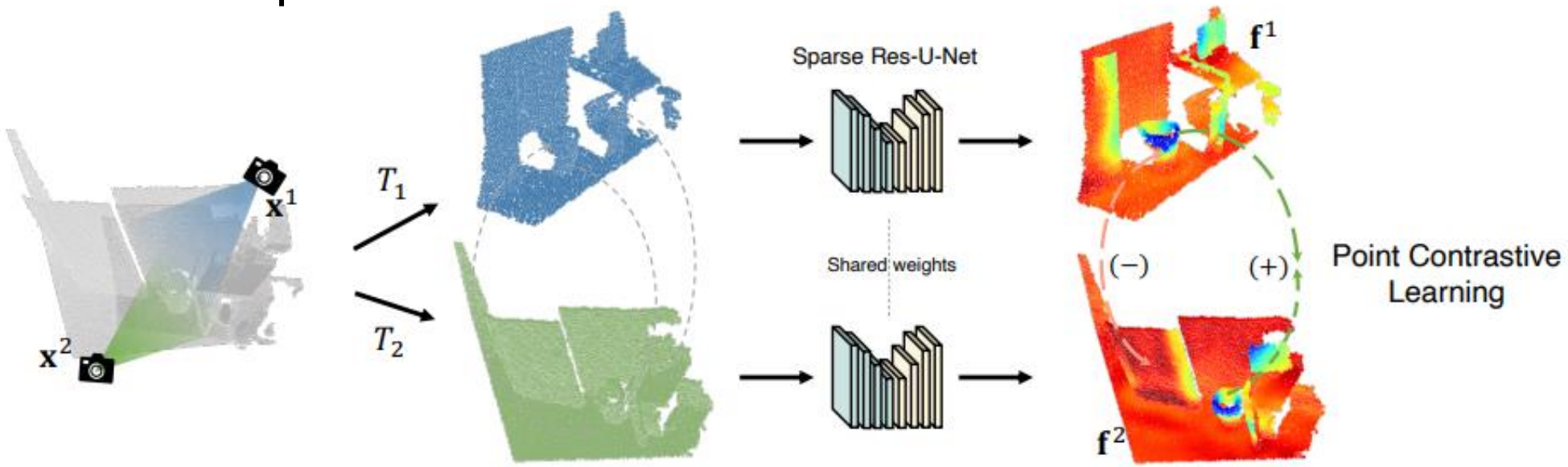
$$L = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$

- *Effect of data augmentation*: random cropping+resizing, random color distortion

[Chen et al. ’20]

PointContrast

- Unsupervised pre-training for point cloud semantic understanding
- Pretraining: from two views of the same point geometry, use point correspondences



[Xie et al. '20]

PointContrast

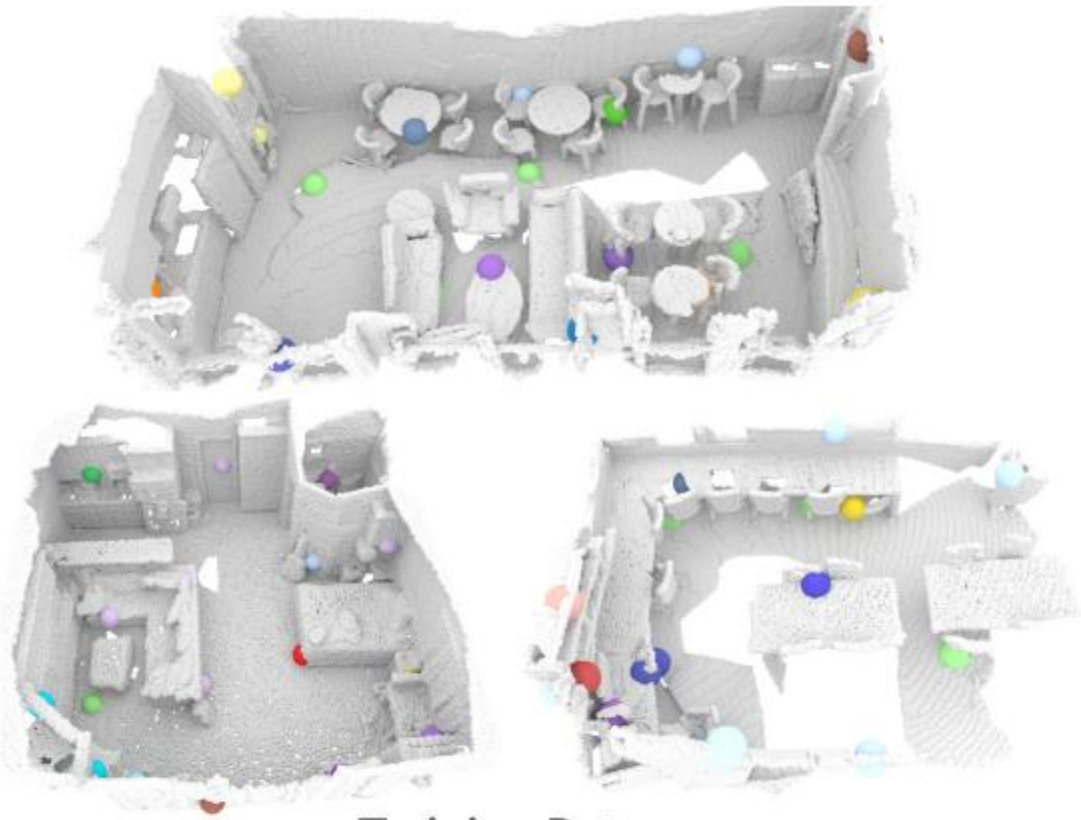
- Boost performance across various ng on the respective dataset/task)

PointContrast: Downstream Tasks for Fine-tuning

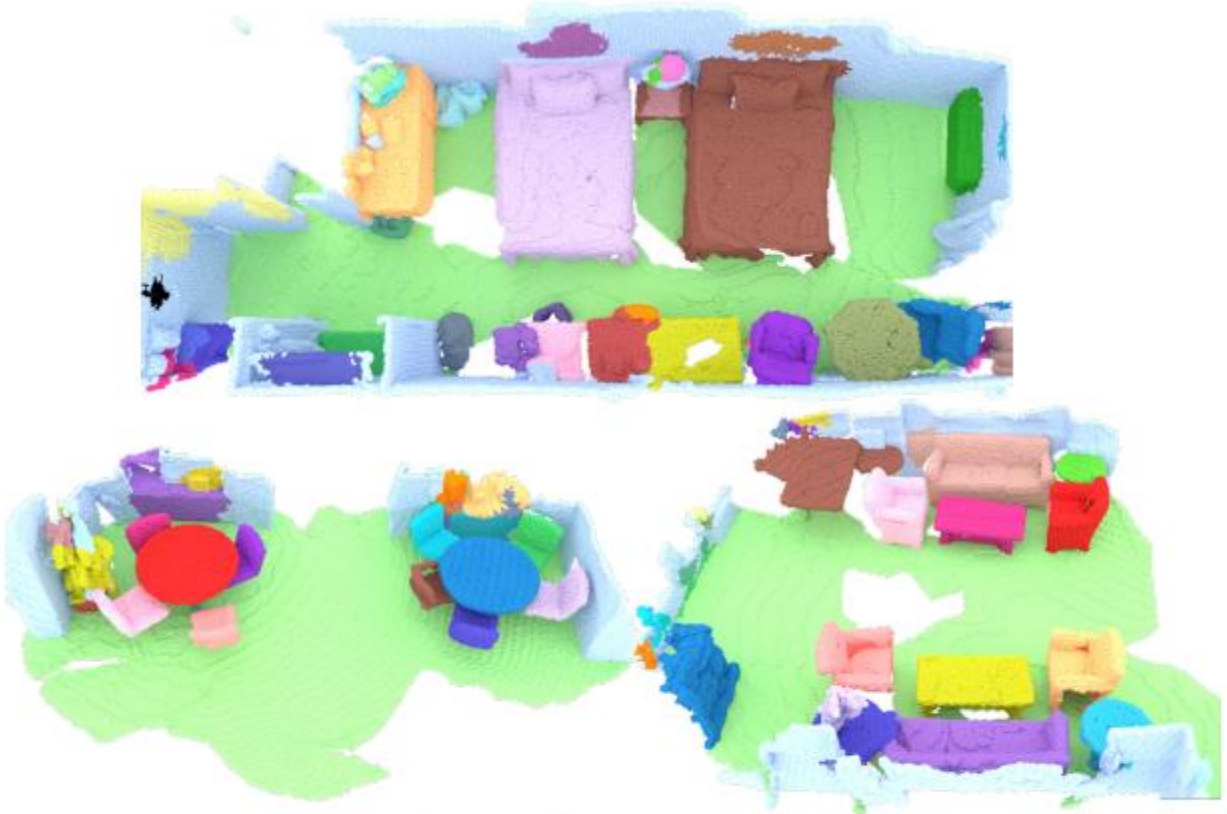
Datasets	Real / Synth.	Complexity	Env.	Task	Rel. gain
S3DIS	Real	Entire floor, office	Indoor	Segmentation	(+2.7%) mIoU
SUN RGB-D	Real	Medium-sized cluttered rooms	Indoor	Detection	(+3.1%) mAP0.5
ScanNetV2	Real	Large rooms	Indoor	Segmentation Detection	(+1.9%) mIoU (+2.6%) mAP0.5
ShapeNet	Synth.	Single objects	Indoor & outdoor	Classification	(+4.0%) Acc.*
ShapeNetPart	Synth.	Object parts	Indoor & outdoor	Segmentation	(+2.2%) mIoU*
Synthia 4D	Synth.	Street scenes, driving envs.	Outdoor	Segmentation	(+3.3%) mIoU

[Xie et al. '20]

Data-Efficient 3D Scene Understanding



Training Data
with Limited Annotations

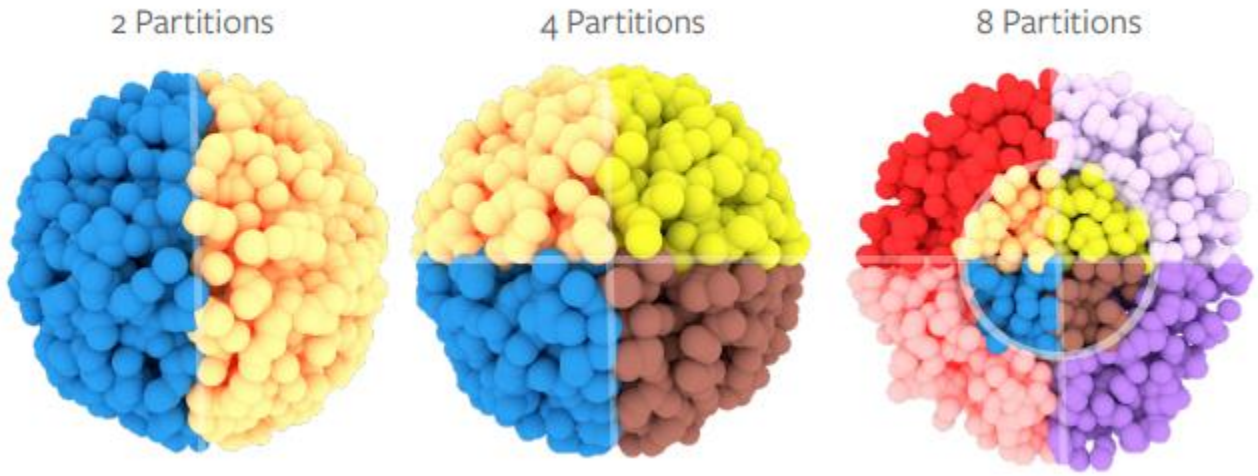


Instance Segmentation
Predictions

[Hou et al. ‘21]

Data-Efficient 3D Scene Understanding

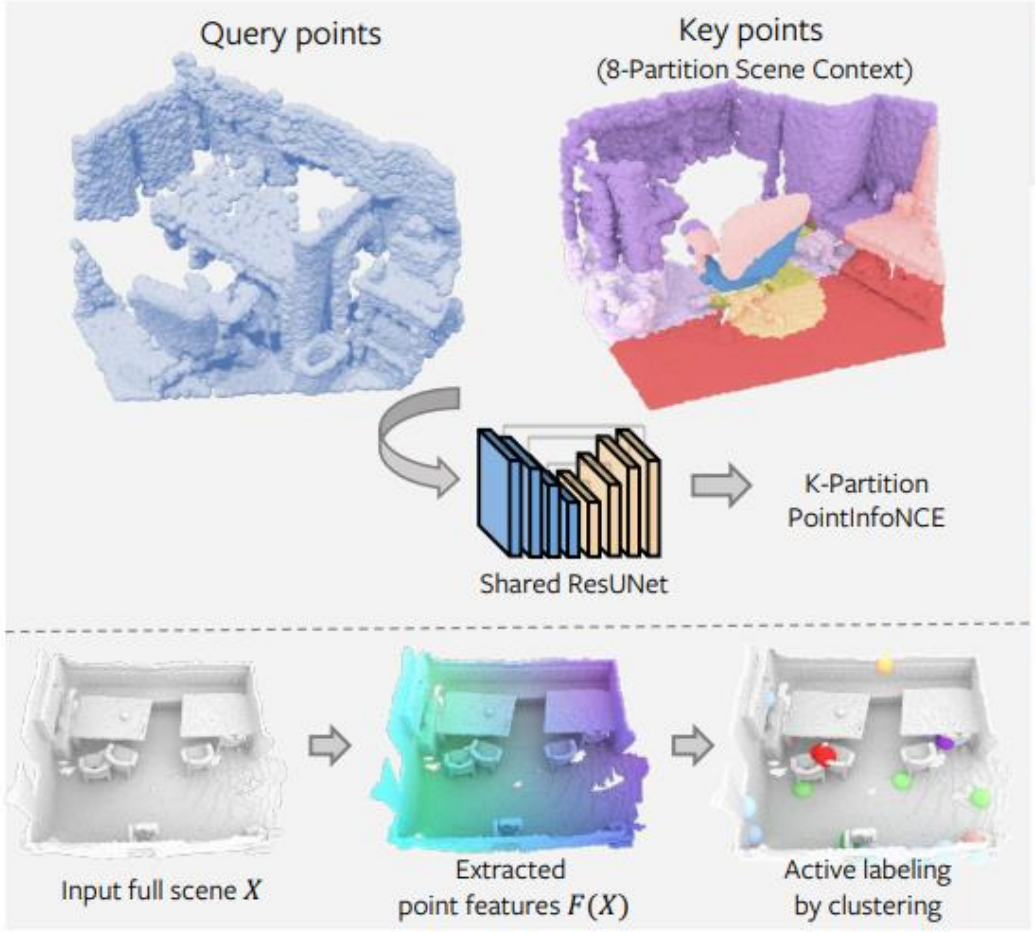
- Incorporate scene context into point contrastive pre-training
- For a point location, compute histogram of neighboring points in each cell from partitioning



- Performance improvement from sampling more points

[Hou et al. '21]

Data-Efficient 3D Scene Understanding



Unsupervised Pre-training
Large-scale partial 3D scans, no label

Pre-trained weights W
Point feature extractor F

Limited Scene Reconstructions
Limited number of scenes
100% points labeled each scene

Fine-tuning W

Limited Annotations
100% scene reconstructions
Limited points labelling budget

Active labeling with $F(X)$
and / or
Fine-tuning W

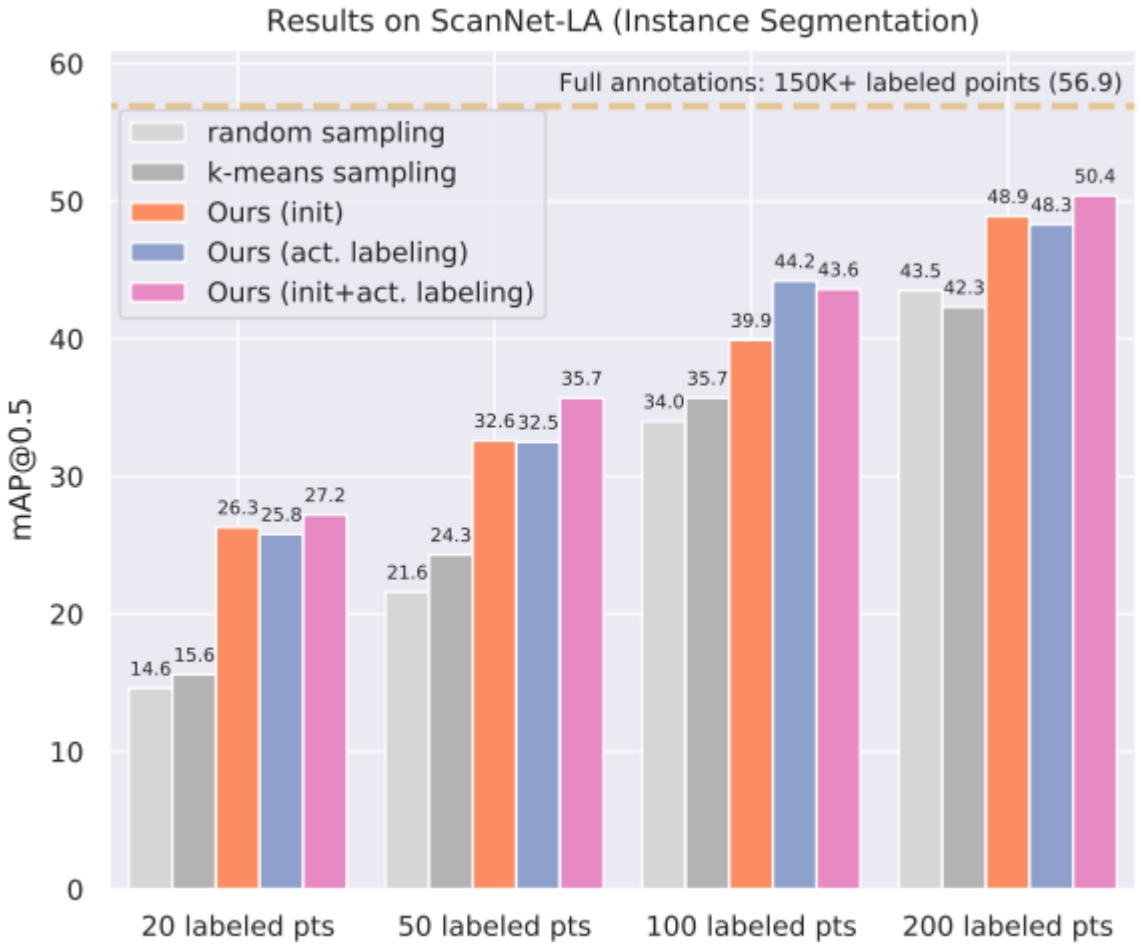
Supervised Training
0%-100% scene reconstructions
0%-100% points labeled each scene

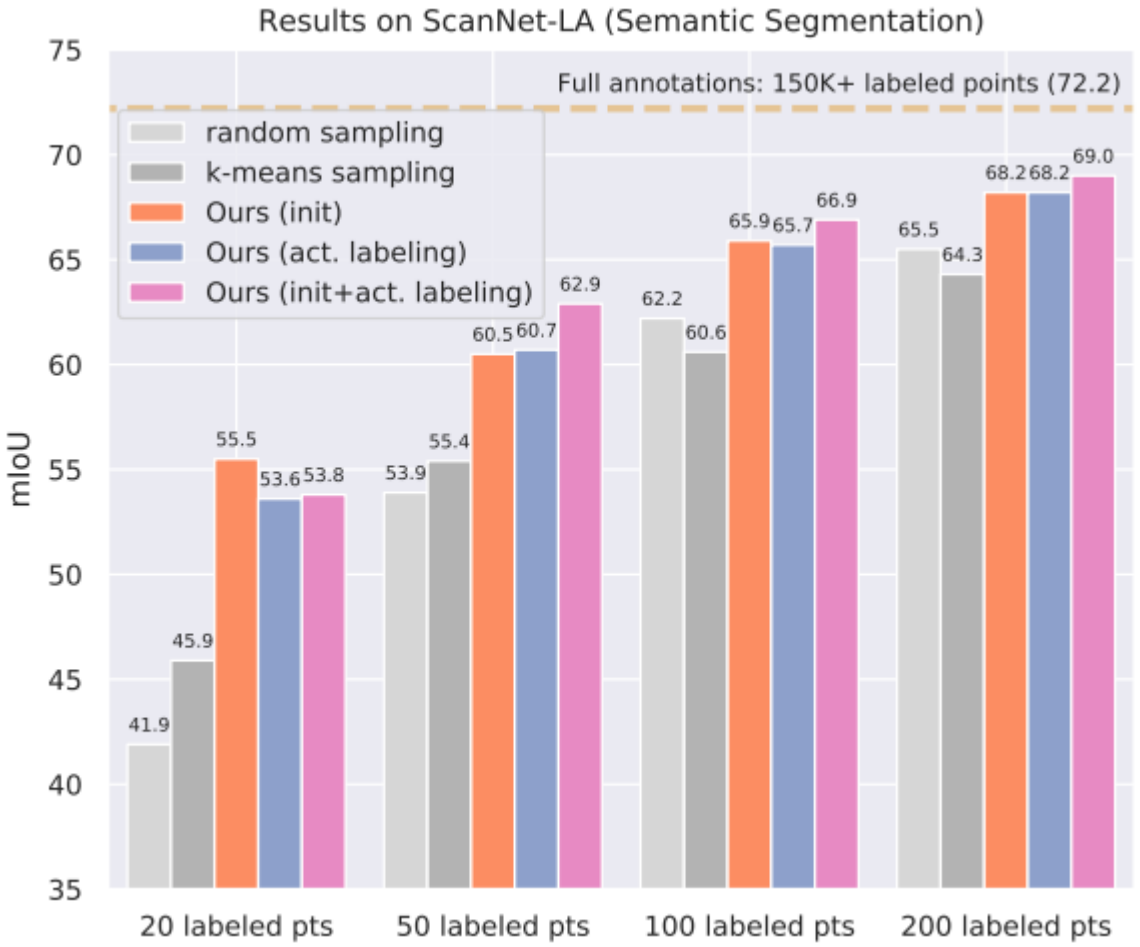
Training from scratch W_{rand}
from random initialization

Pre-training | Efficient Scene Understanding | Baseline

[Hou et al. '21]

Data-Efficient 3D Scene Understanding





[Hou et al. '21]

Can 3D Priors Help 2D Learning?

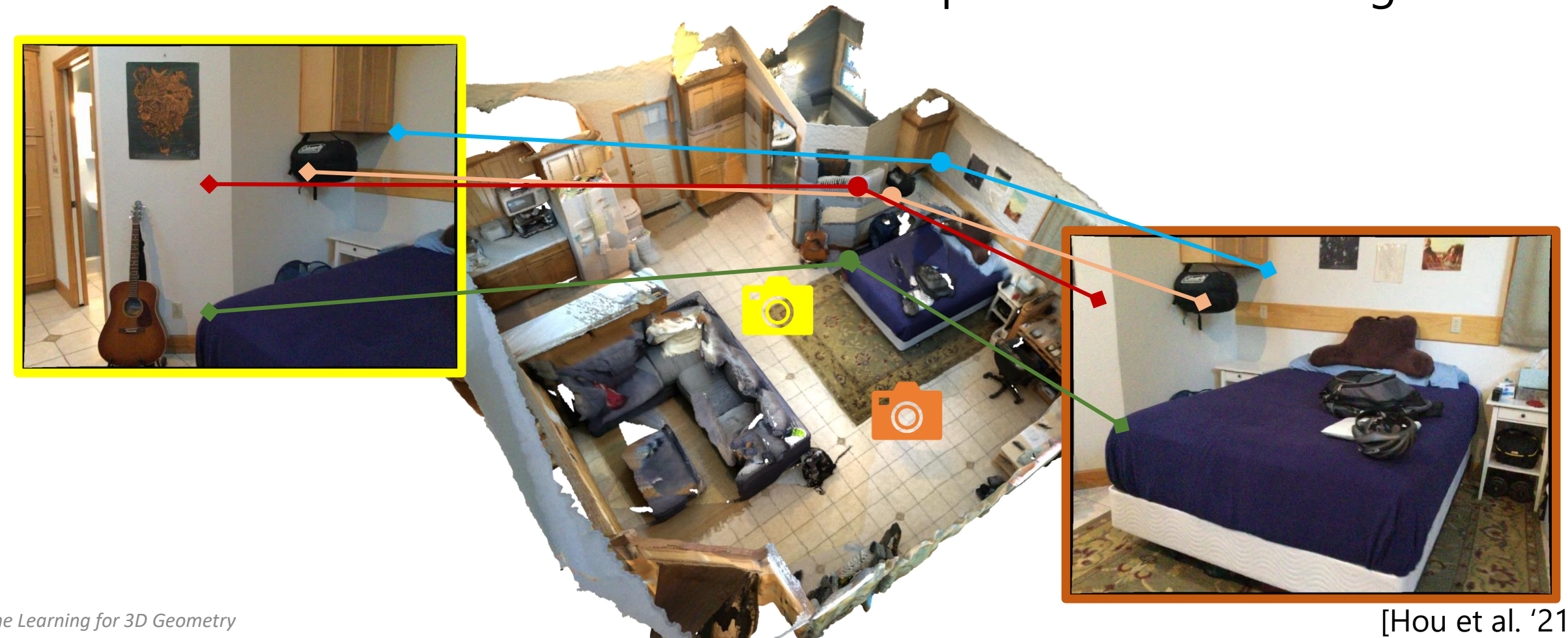
Can 3D Priors Help 2D Learning?

- Use 3D to learn view-invariance in 2D representation learning

[Hou et al. ‘21]

Can 3D Priors Help 2D Learning?

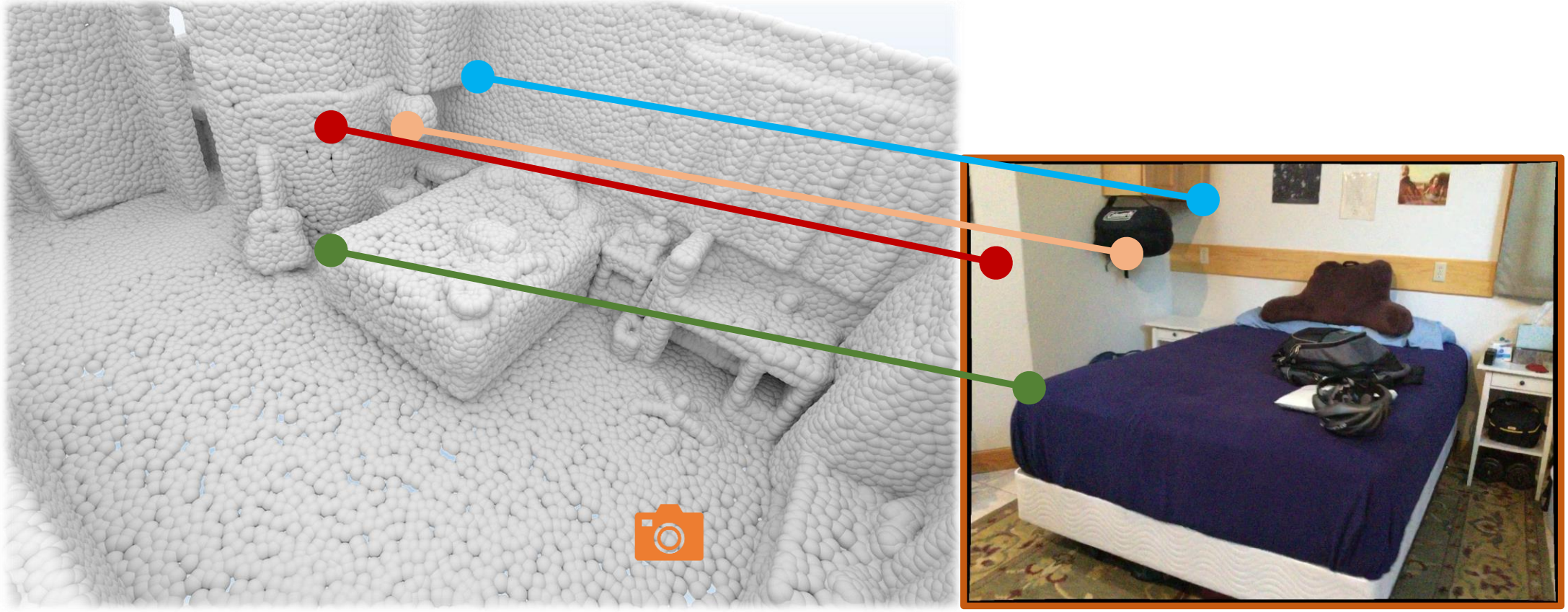
- Use 3D to learn view-invariance in 2D representation learning



[Hou et al. '21]

Can 3D Priors Help 2D Learning?

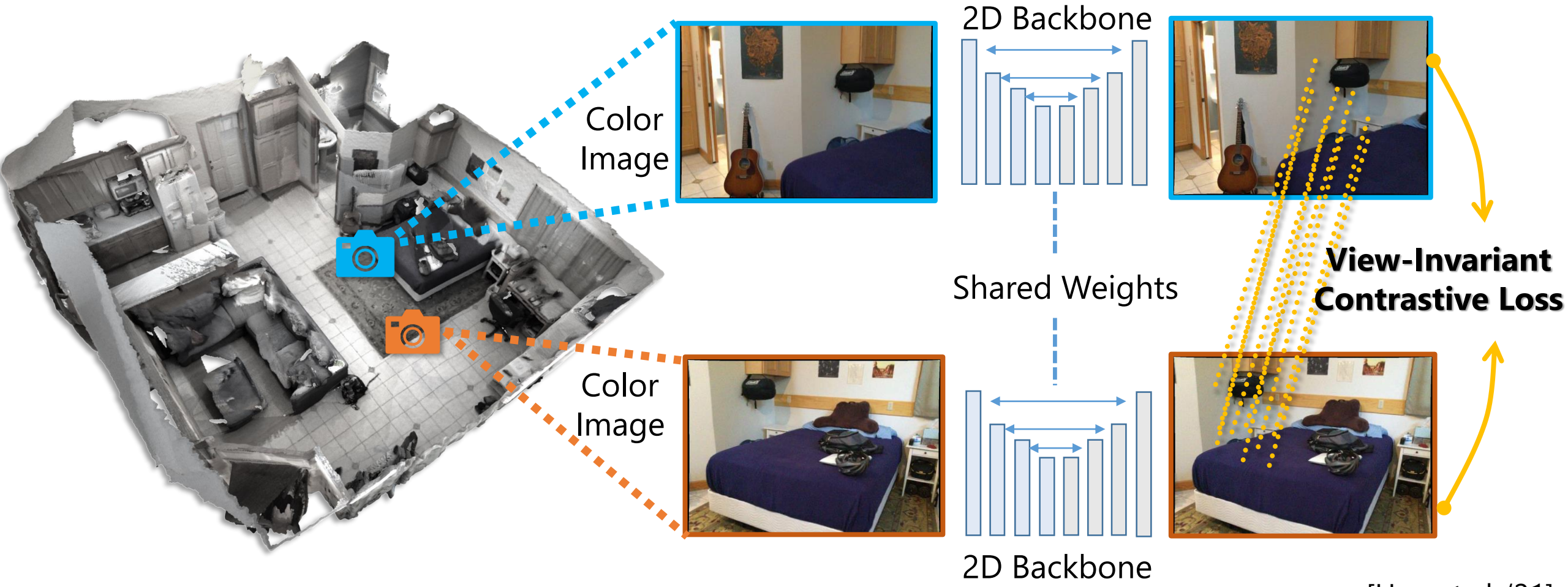
- Encode geometric priors with 2D-3D correspondences



[Hou et al. '21]

Can 3D Priors Help 2D Learning?

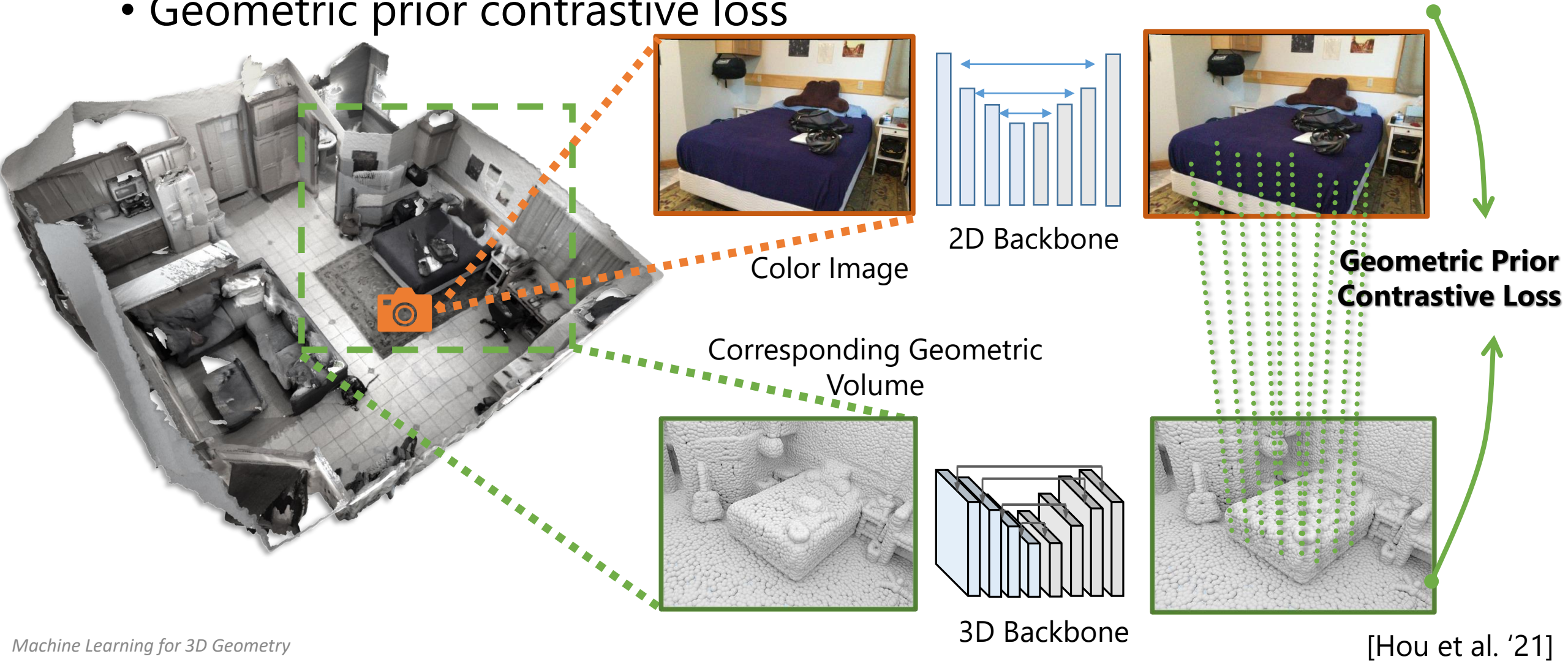
- View-invariant contrastive loss



[Hou et al. '21]

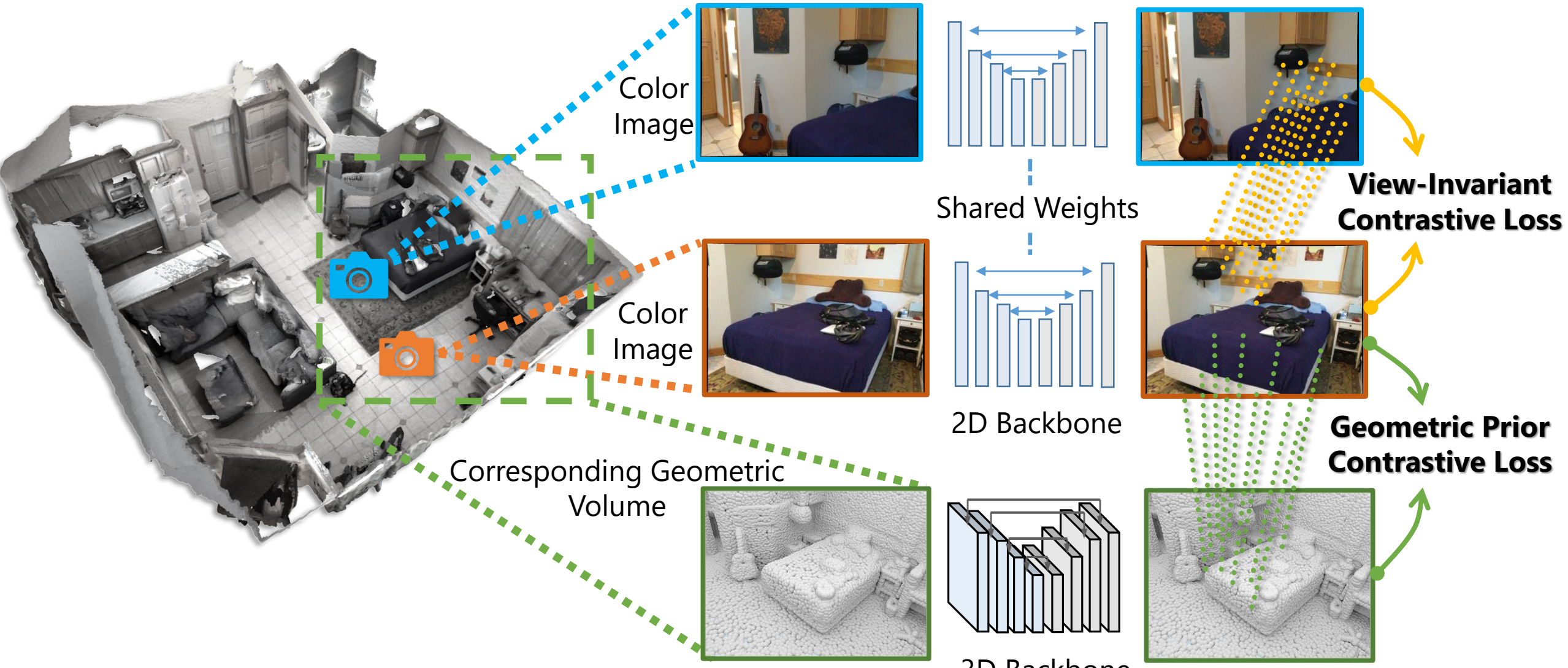
Can 3D Priors Help 2D Learning?

- Geometric prior contrastive loss



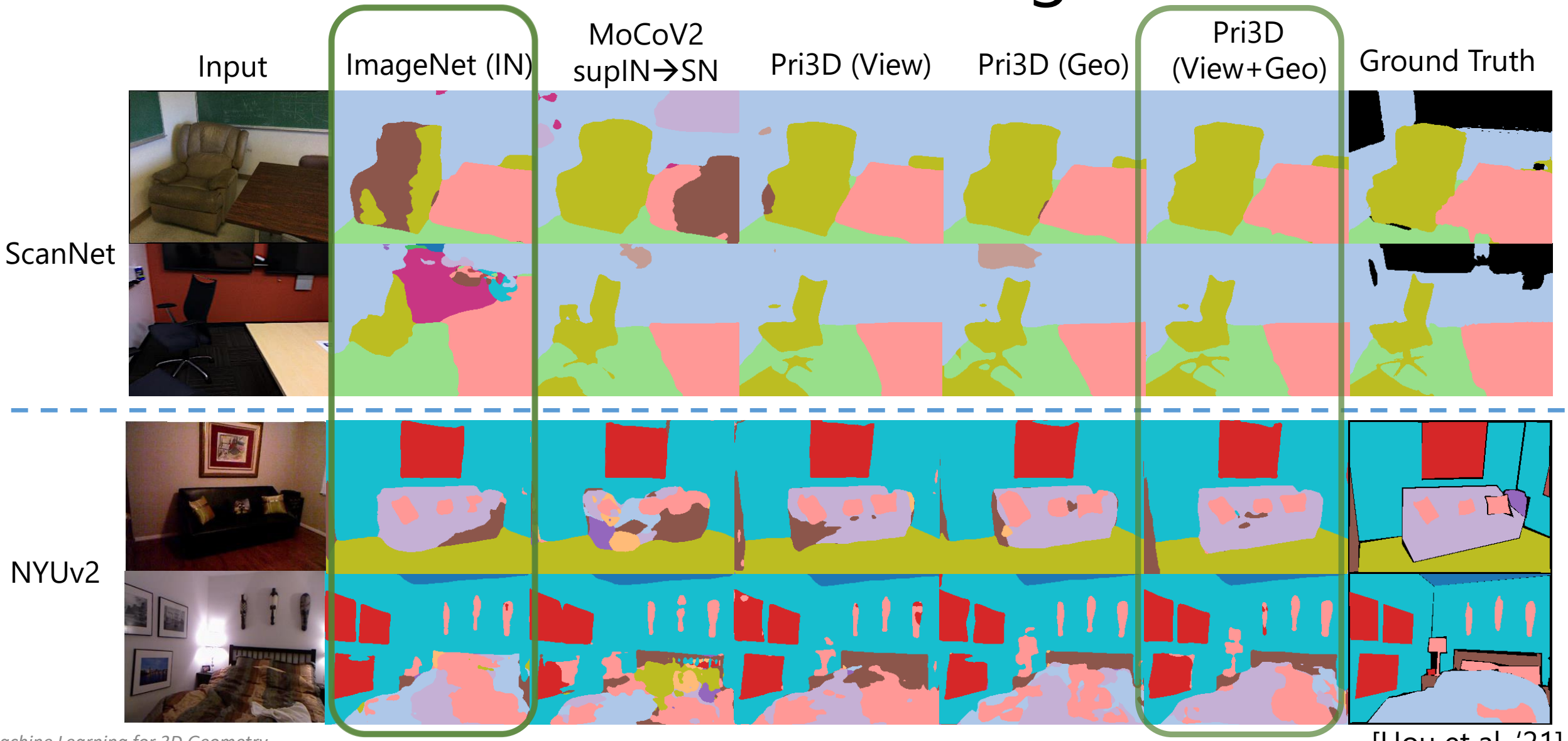
[Hou et al. '21]

Joint View-Invariant & Geometric Priors



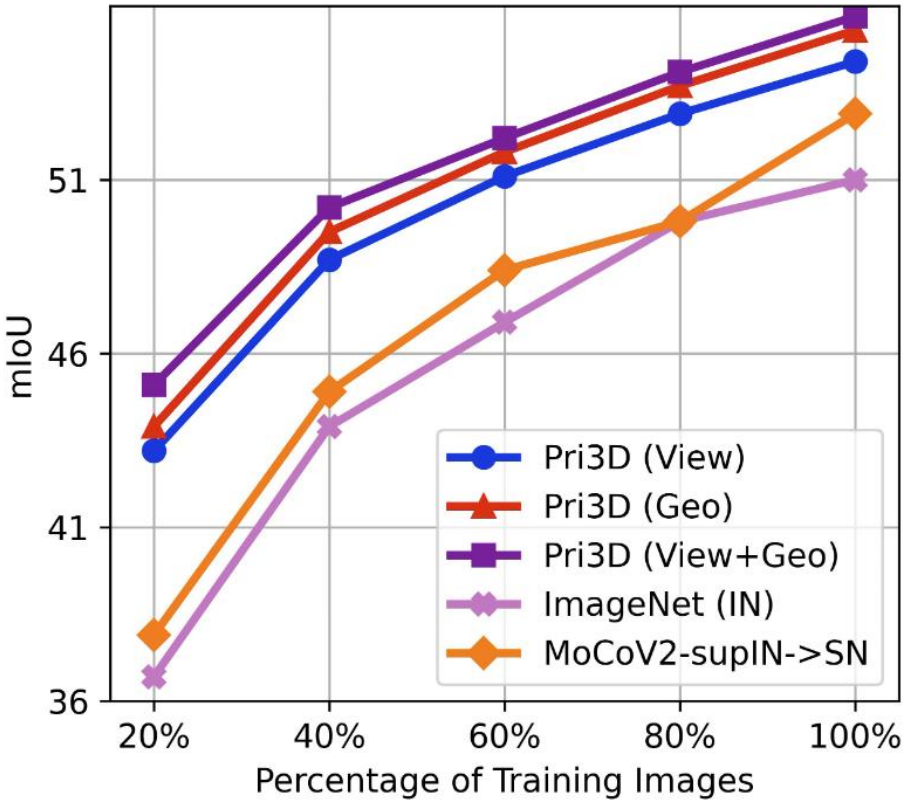
[Hou et al. '21]

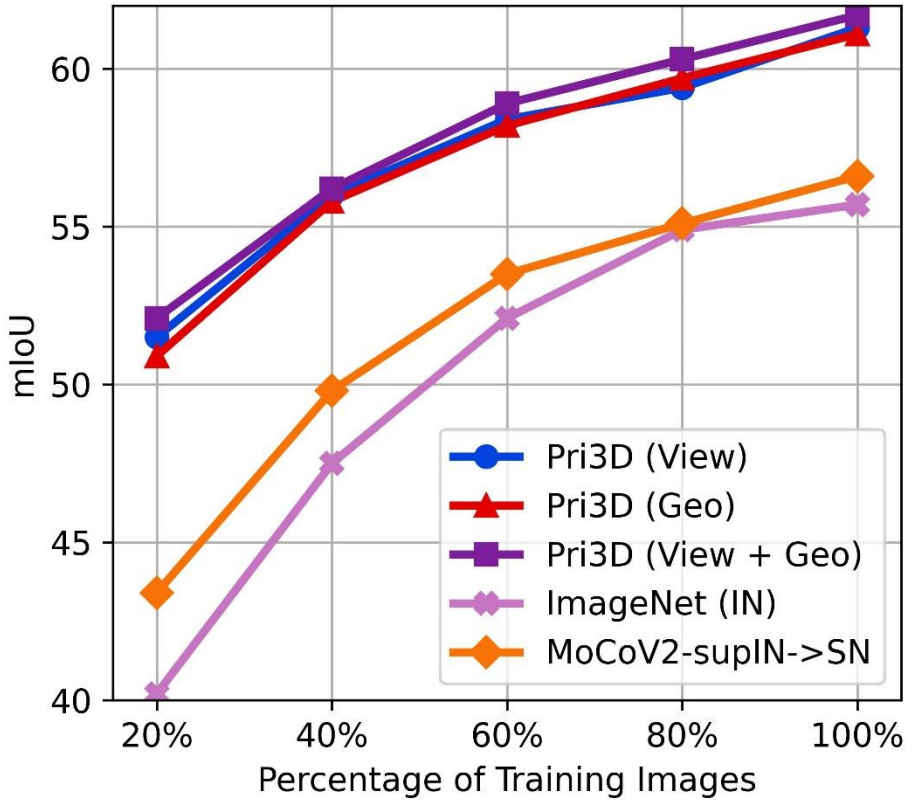
3D Priors for 2D Semantic Segmentation



[Hou et al. '21]

2D Semantic Segmentation with Limited Data





Yes, 3D priors can help learned 2D representations

[Hou et al. '21]

Domain Adaptation

- Two data domains:
- Source: trained model performs well, larger amount of labeled data available
- Target: related data characteristics, less labeled data available

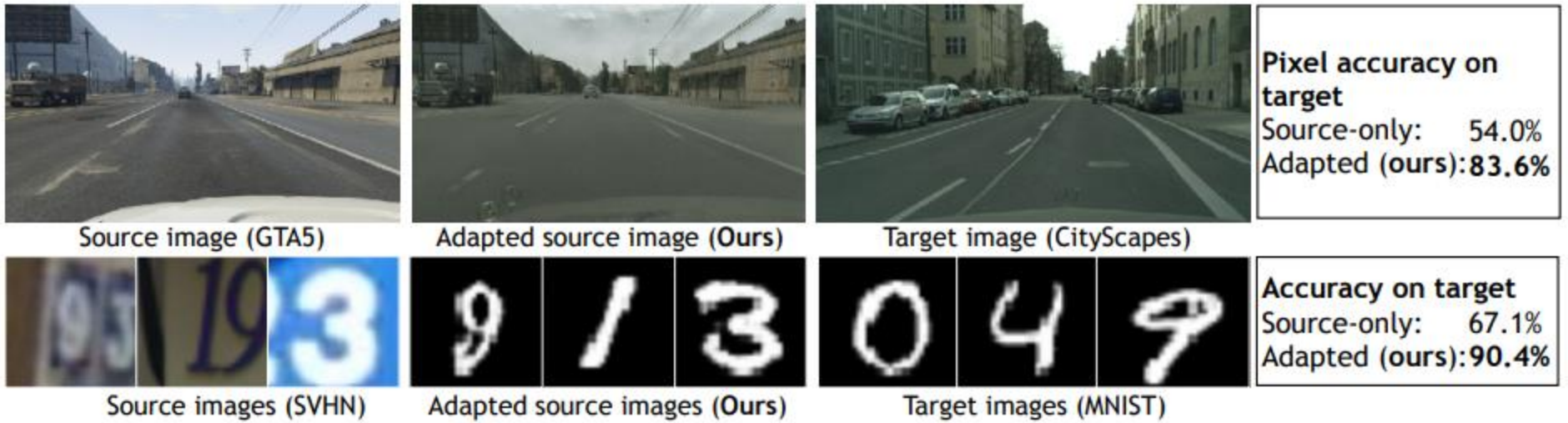
- E.g., synthetic vs. real

Domain Adaptation

- Often unsupervised: typically don’t have 1:1 correspondences between the domains
- Common domains: synthetic/real (e.g., GTA/Cityscapes), weather changes, day/night, different sensors
- Common tasks: semantic segmentation, some object detection

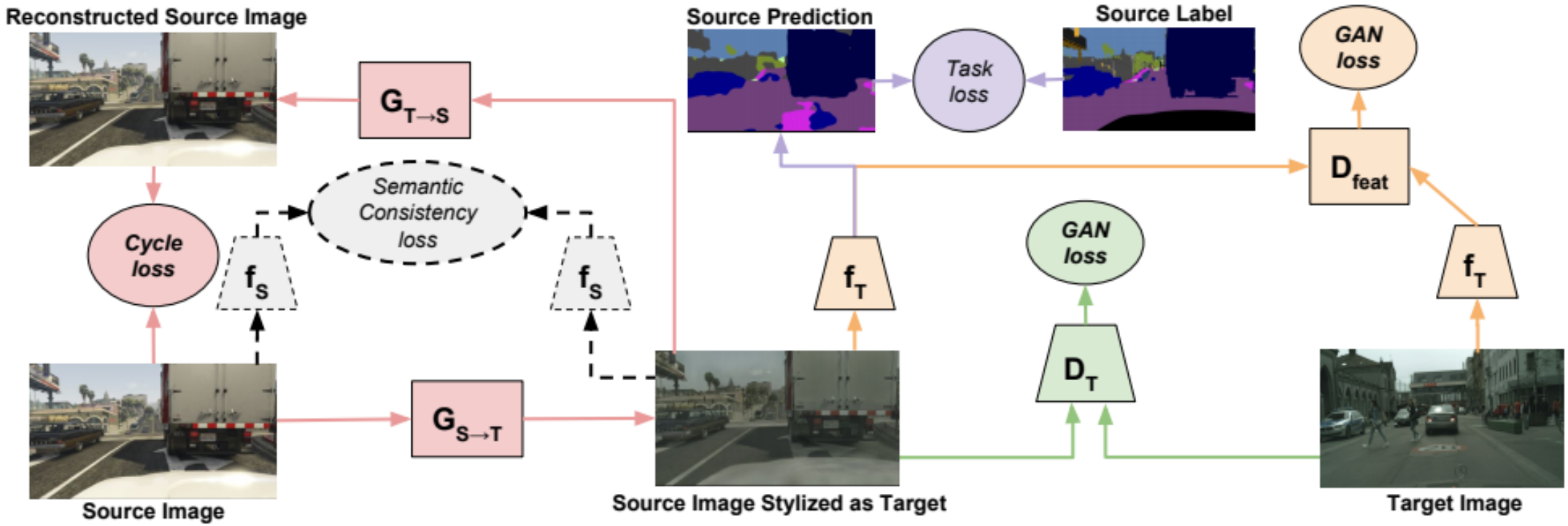
CyCADA: Cycle-Consistent Adversarial Domain Adaptation

- Adapt between image domains
- Constrain latent (feature space) representations to match
- Constrain adapted images to match



[Hoffman et al. '18]

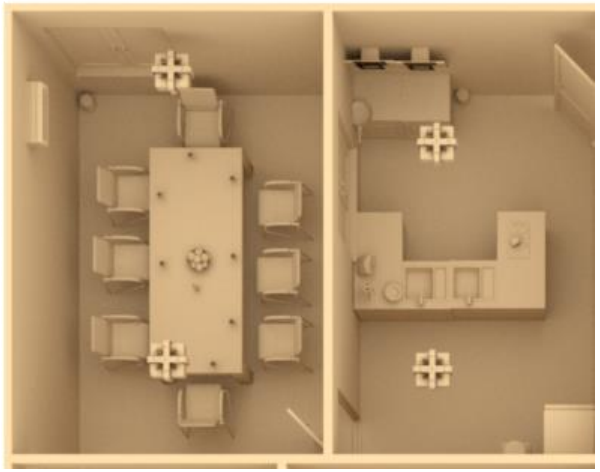
CyCADA: Cycle-Consistent Adversarial Domain Adaptation

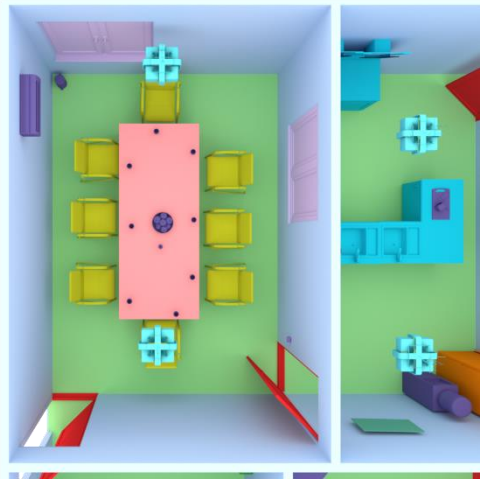


[Hoffman et al. '18]

3D Domain Adaptation?

- Rendering synthetic scenes -> generate lots of labeled data





- CAD retrieval to real-world scans or images

You may have seen…

Can we distill 3D information from text/image models?

Guide 3D features with text encodings

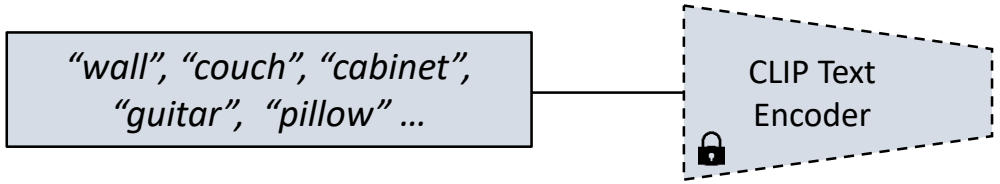
- 3D: data is limited and heavily imbalanced
  - Long-tail class features suffer
- Text: tons of data, less imbalance
- Guide 3D feature learning with pre-trained text embeddings

Language-Grounded 3D Learning

“wall”, “couch”, “cabinet”, “guitar”, “pillow” …

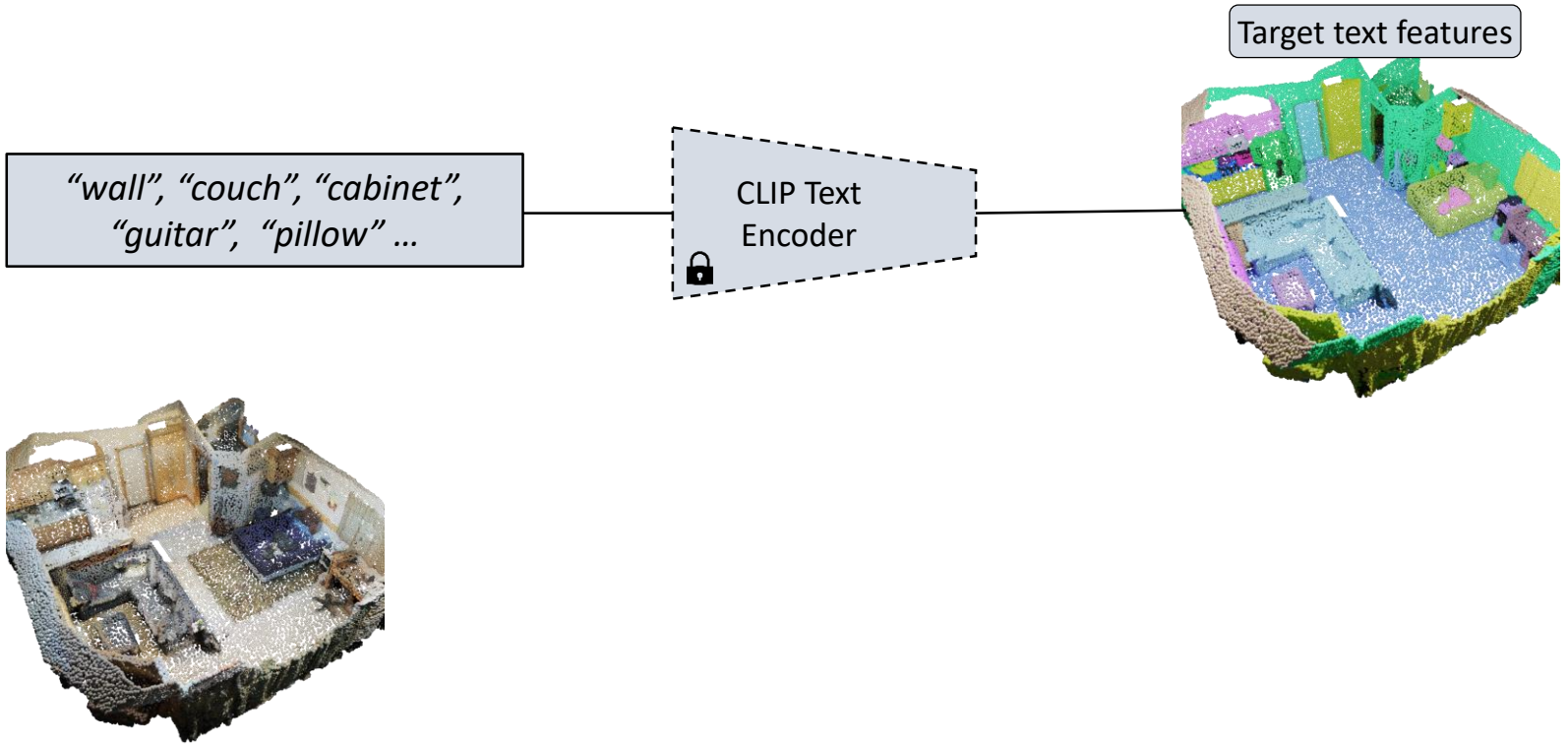
[Rozenberszki et al. ‘22]

Language-Grounded 3D Learning



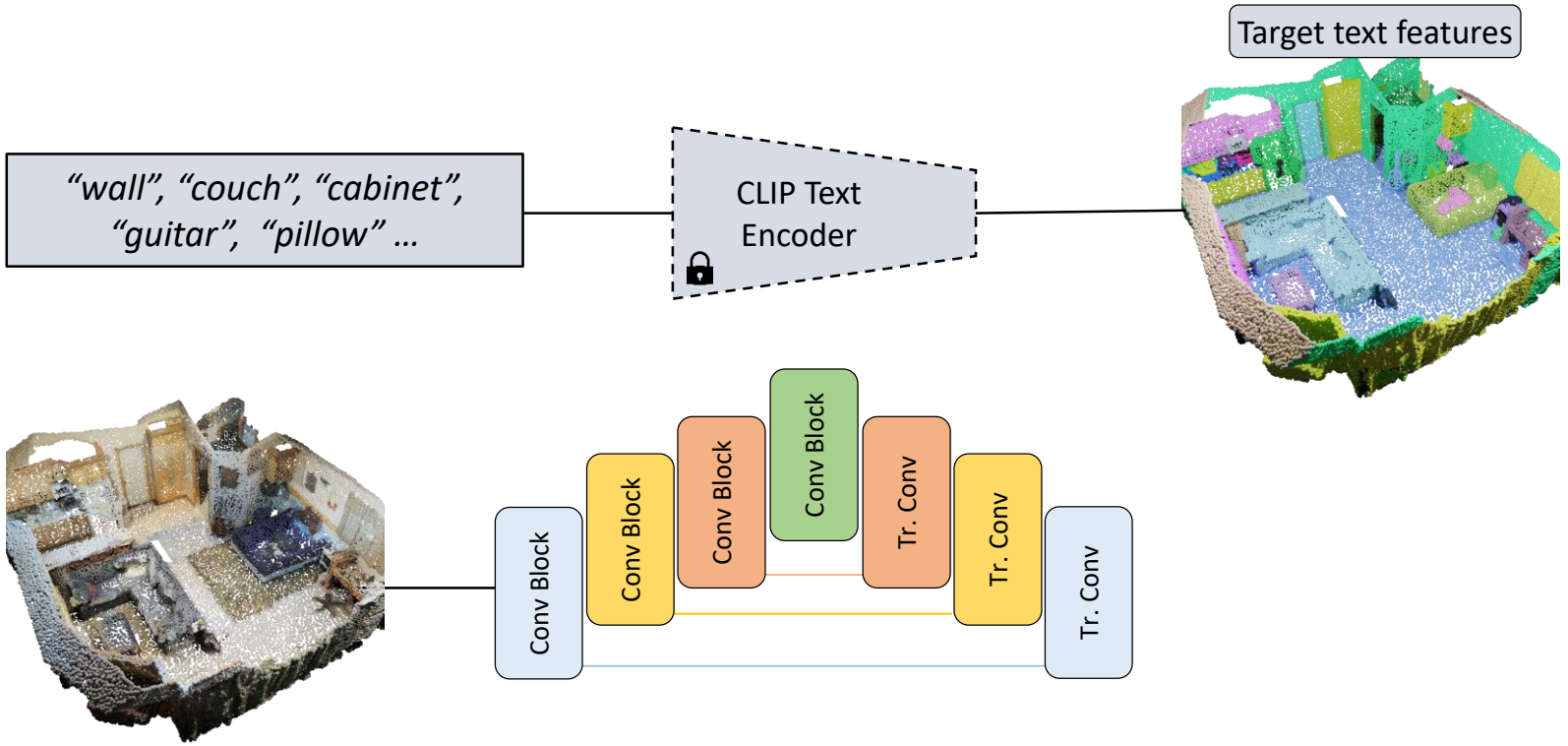
[Rozenberszki et al. ‘22]

Language-Grounded 3D Learning



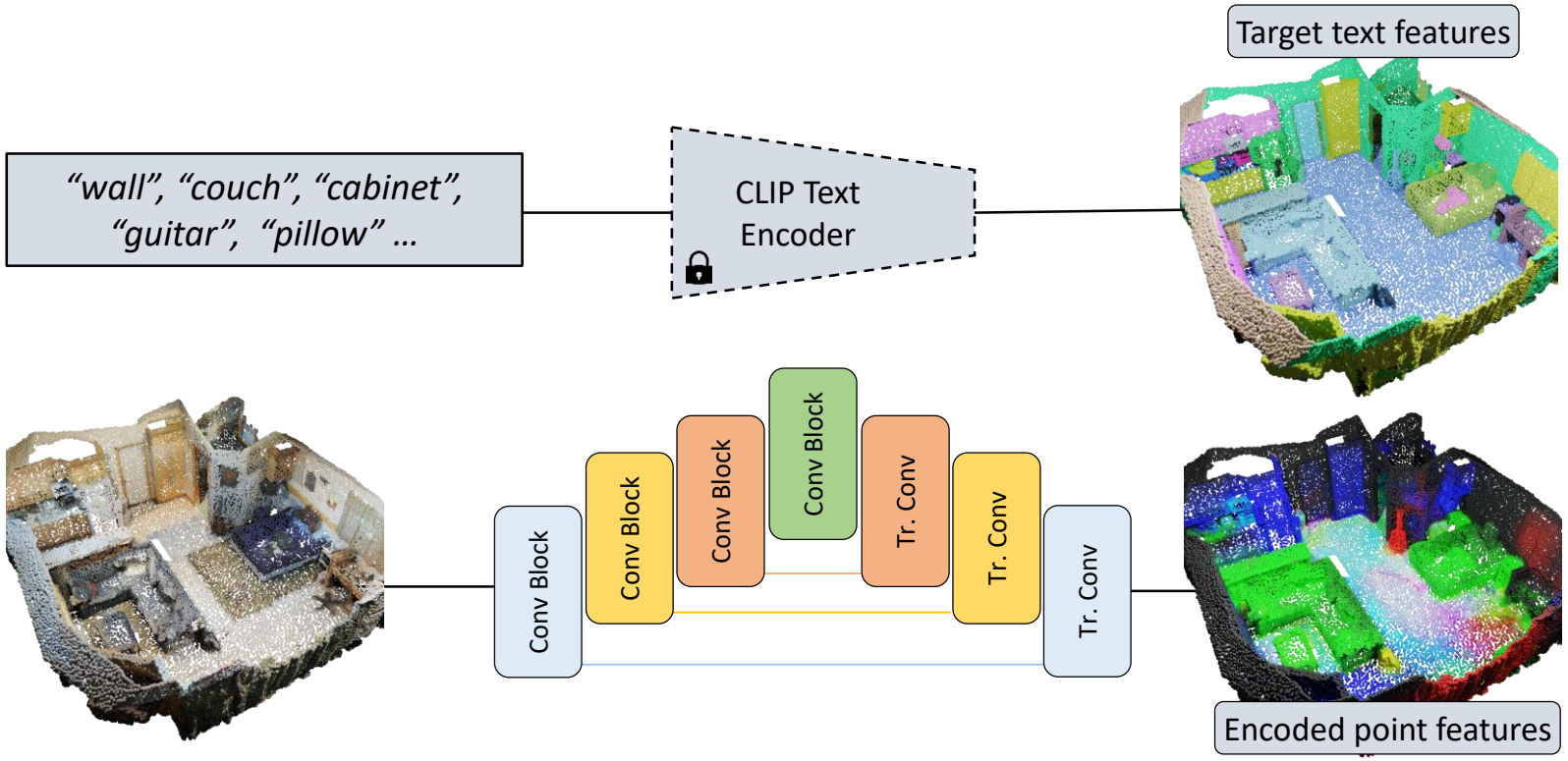
[Rozenberszki et al. ‘22]

Language-Grounded 3D Learning



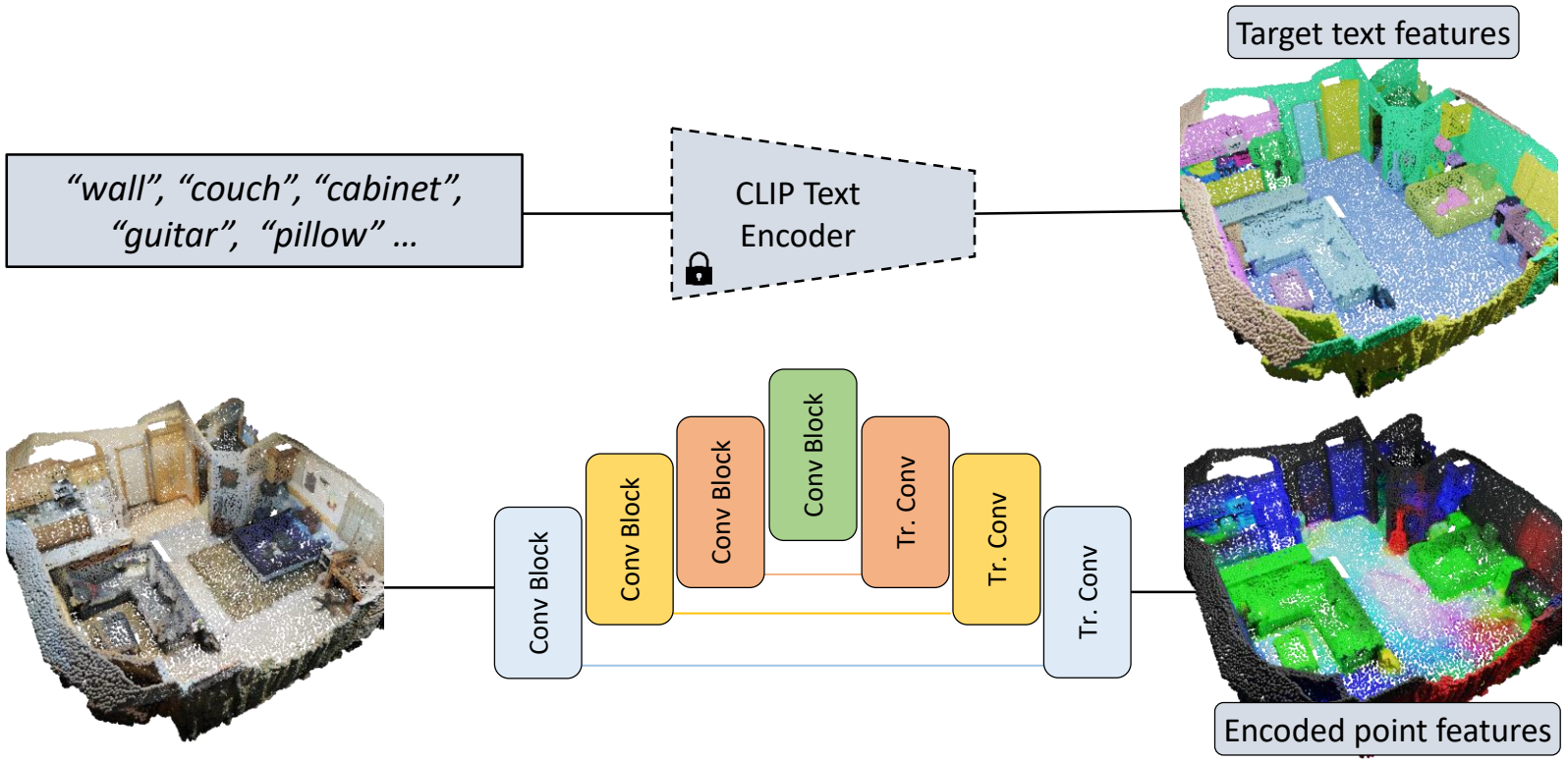
[Rozenberszki et al. '22]

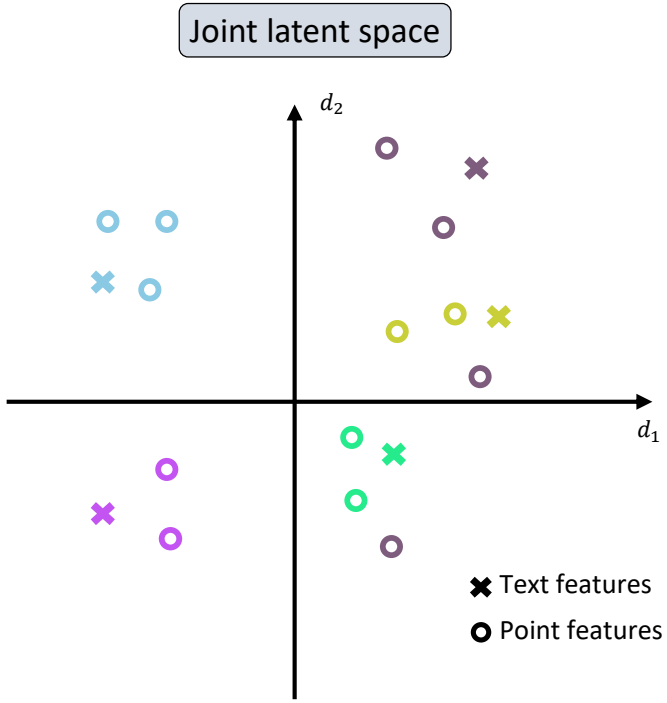
Language-Grounded 3D Learning



[Rozenberszki et al. '22]

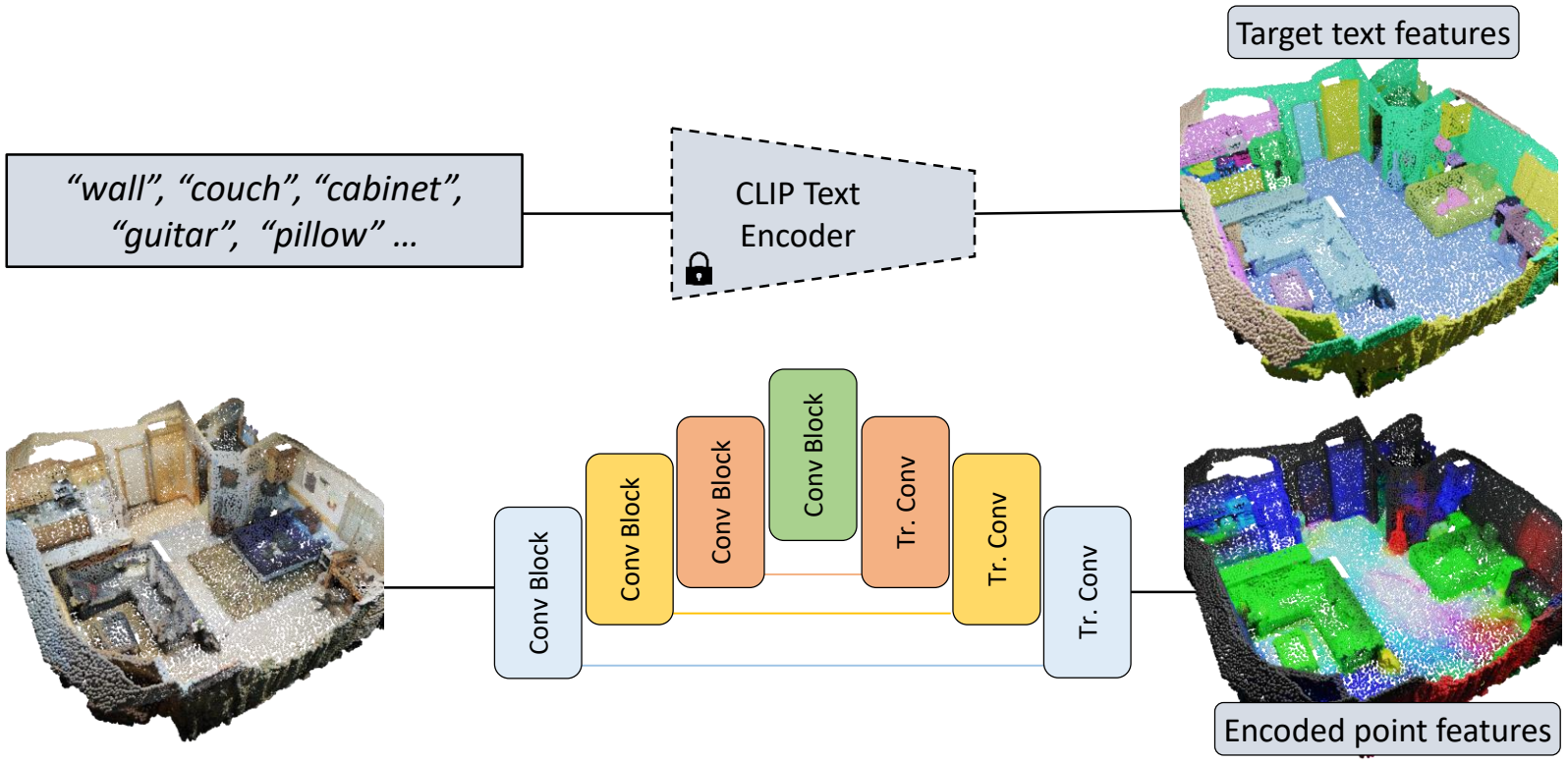
Language-Grounded 3D Learning

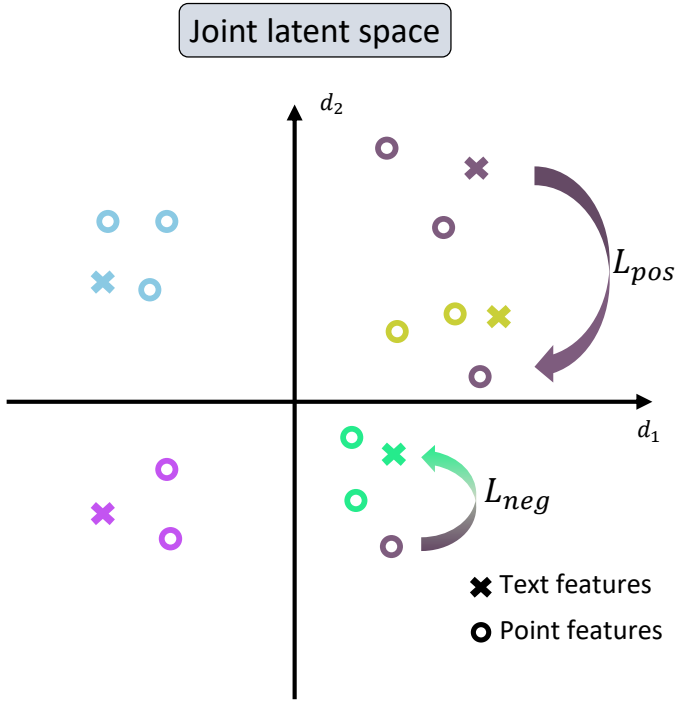




[Rozenberszki et al. '22]

Language-Grounded 3D Learning

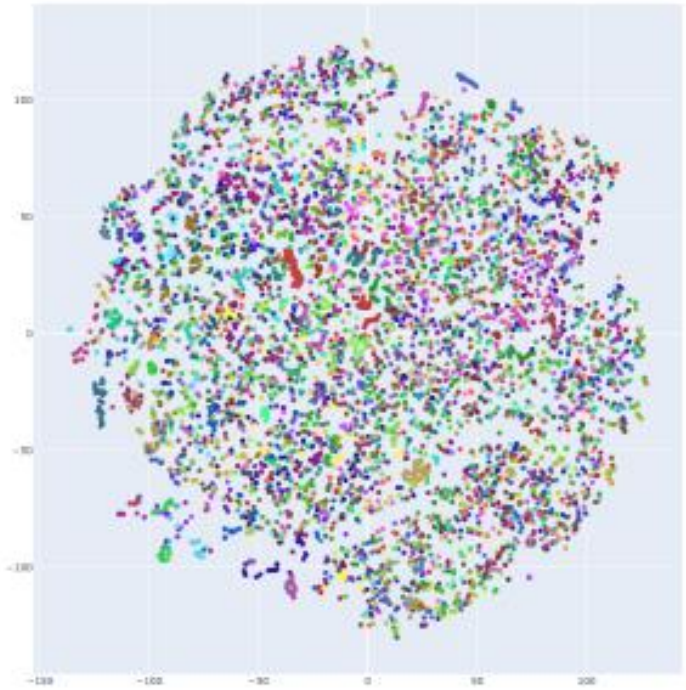




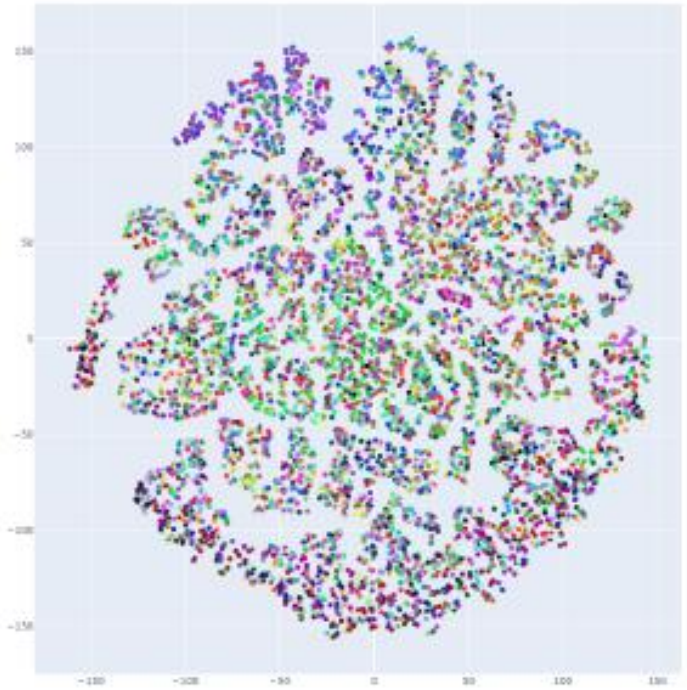
[Rozenberszki et al. '22]

Language-Grounded 3D Learning

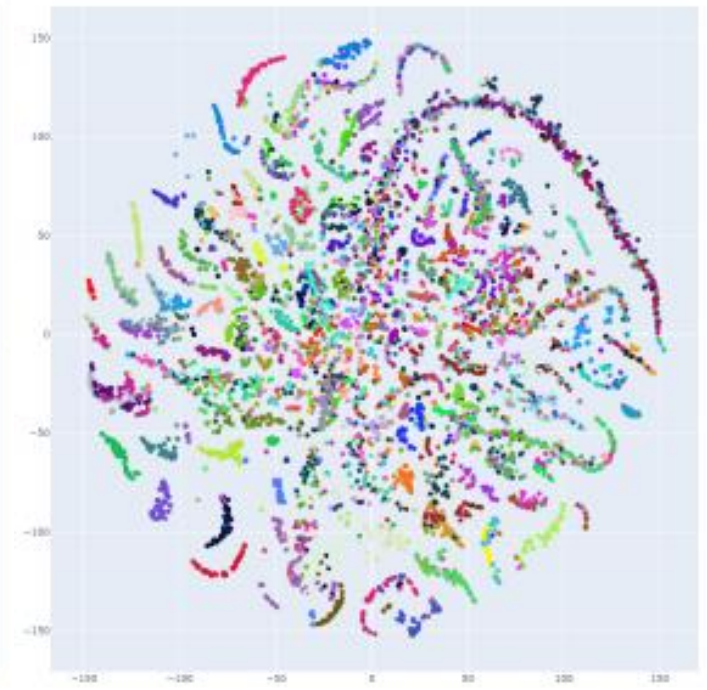
- Learned feature space



CSC [Hou et al. ‘21]



SupCon [Khosla et al. ‘20]

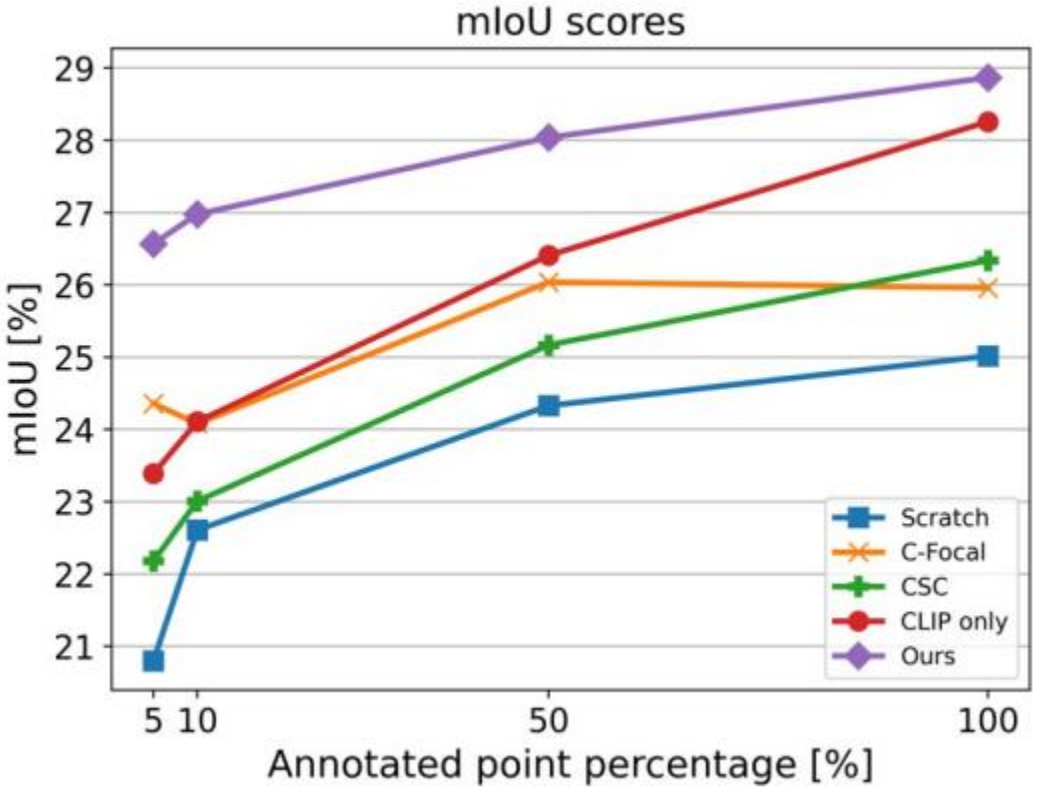


Ours

[Rozenberszki et al. ‘22]

Language-Grounded 3D Learning

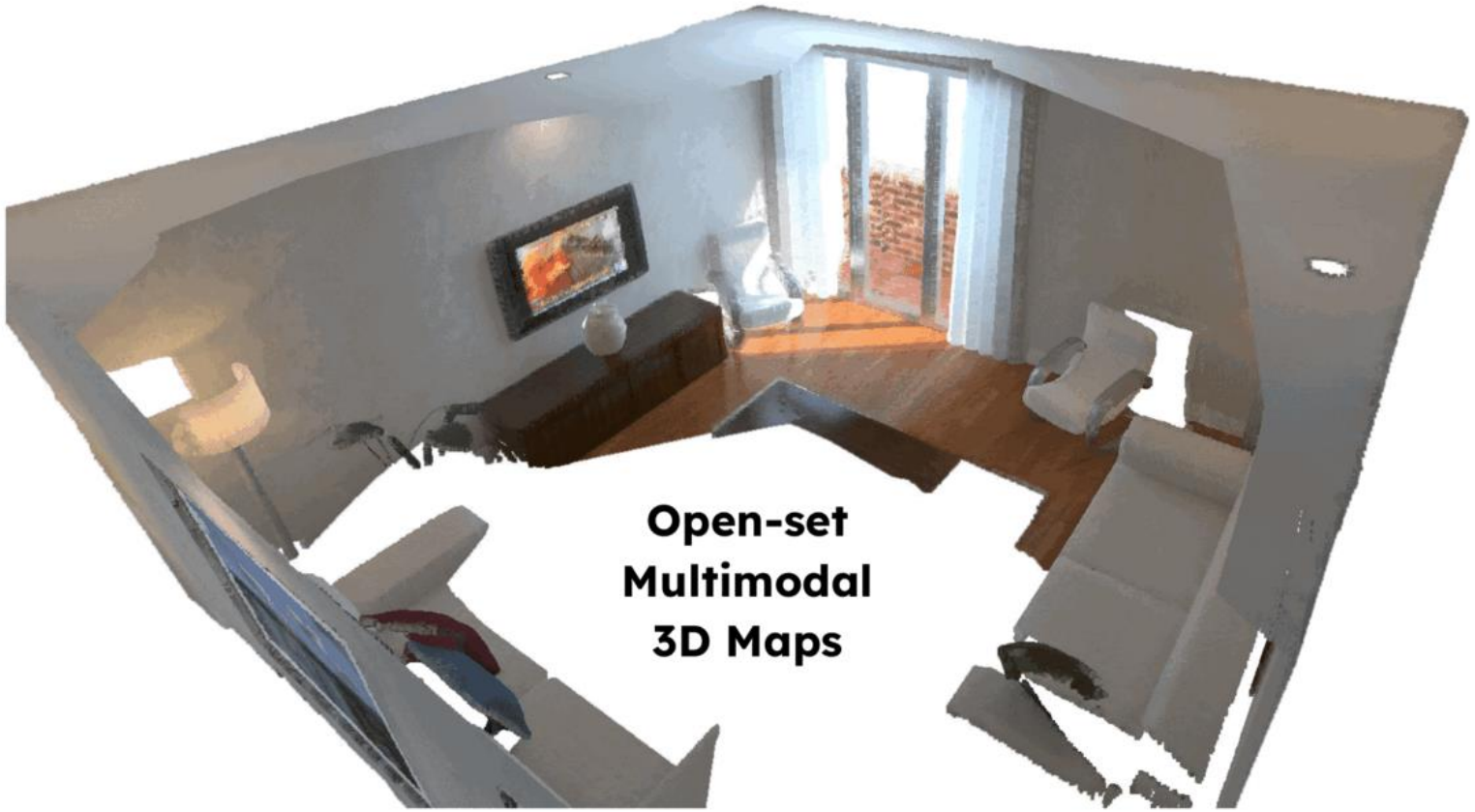
- Semantic segmentation on ScanNet200



[Rozenberszki et al. '22]

Open-set 3D Understanding

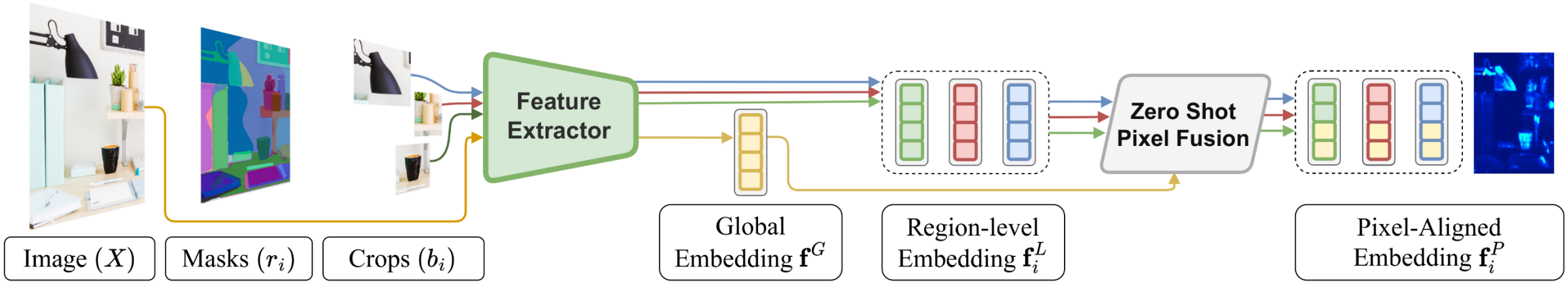
- Multi-modal mapping on 3D scenes



[Jatavallabhula et al. ‘23]

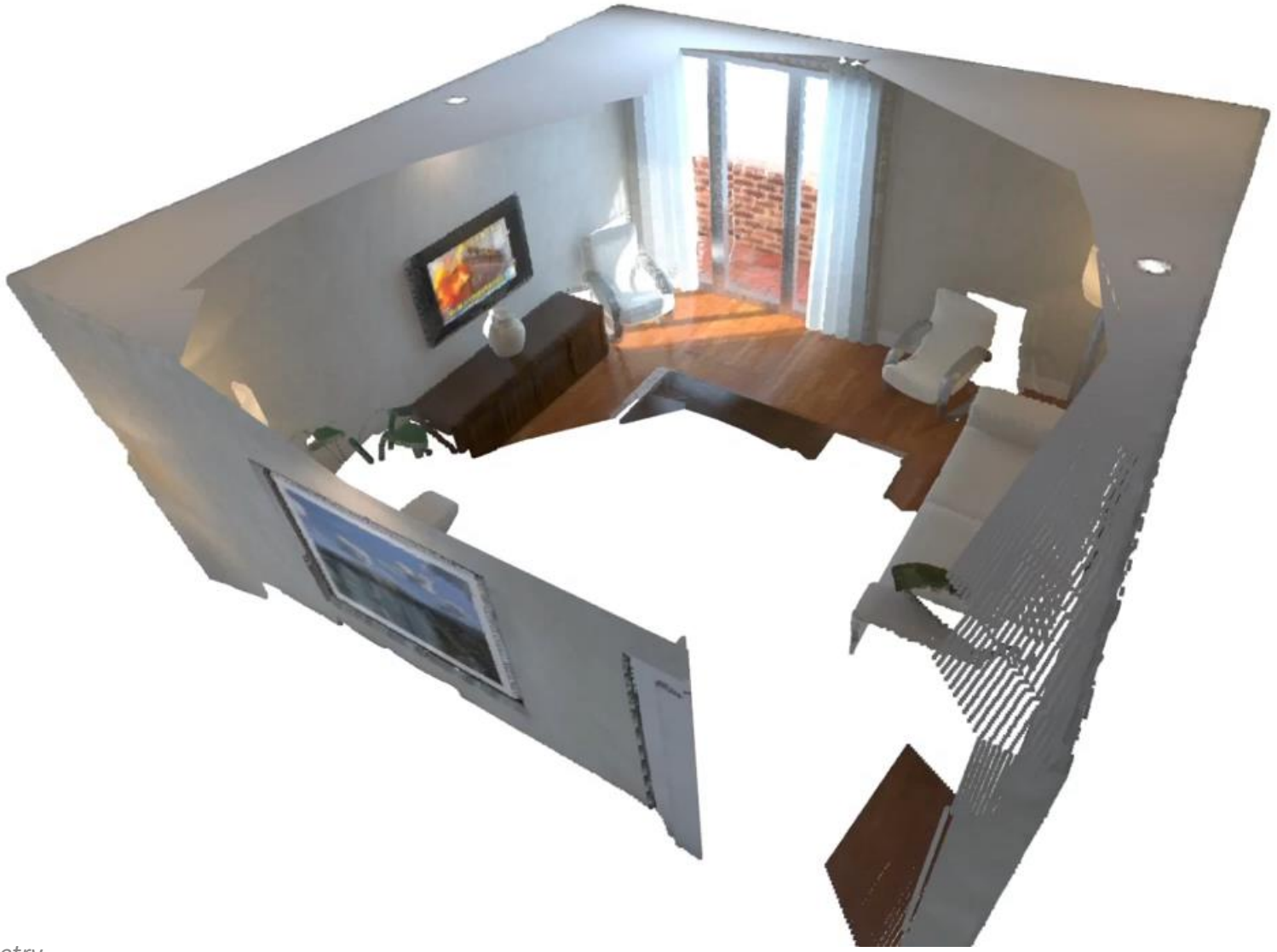
Open-set 3D Understanding

- Multi-modal mapping on 3D scenes
- Get pixel-level information from foundation models (often global per-image)



[Jatavallabhula et al. '23]

Open-set 3D Understanding



[Jatavallabhula et al. ‘23]

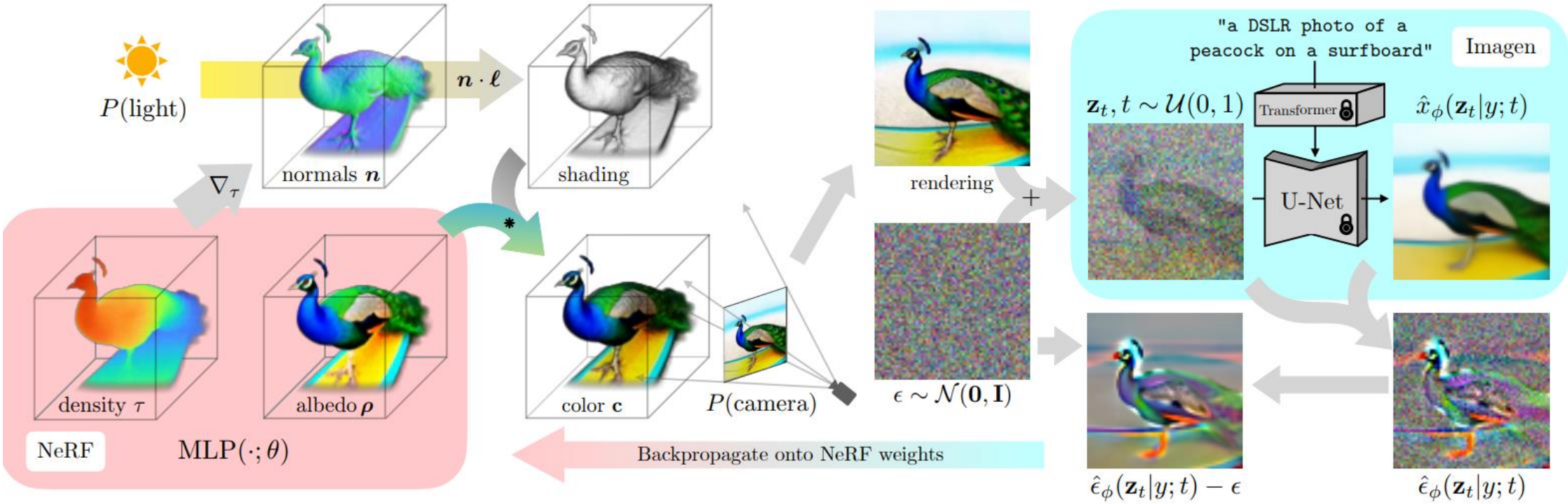
Obtaining 3D from 2D Generative Models

- Recall: DreamFusion can generate multi-view consistent NeRFs from only a 2D image generative model!

[Poole et al. '22]

Obtaining 3D from 2D Generative Models

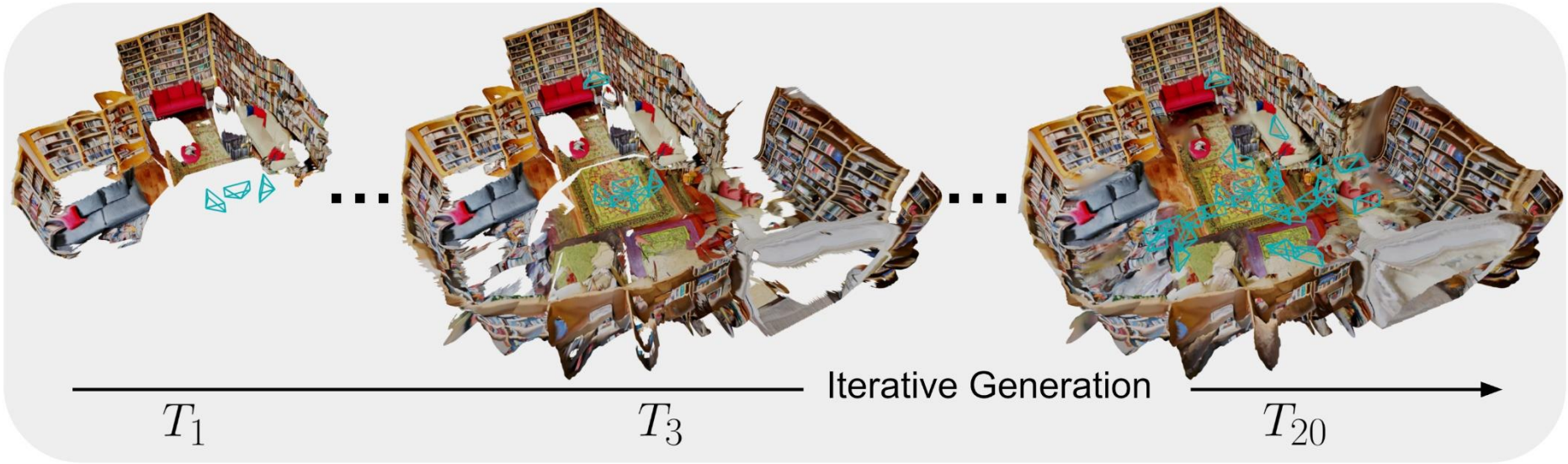
- Recall: DreamFusion can generate multi-view consistent NeRFs from only a 2D image generative model!

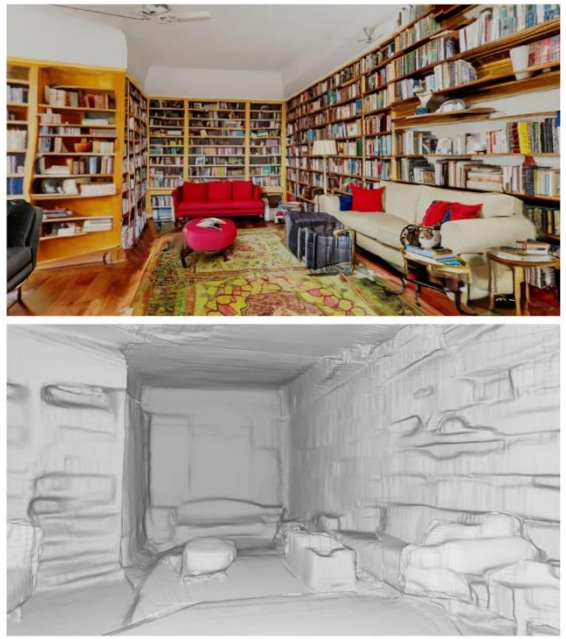


[Poole et al. '22]

Obtaining 3D from 2D Generative Models

- and monocular depth estimation

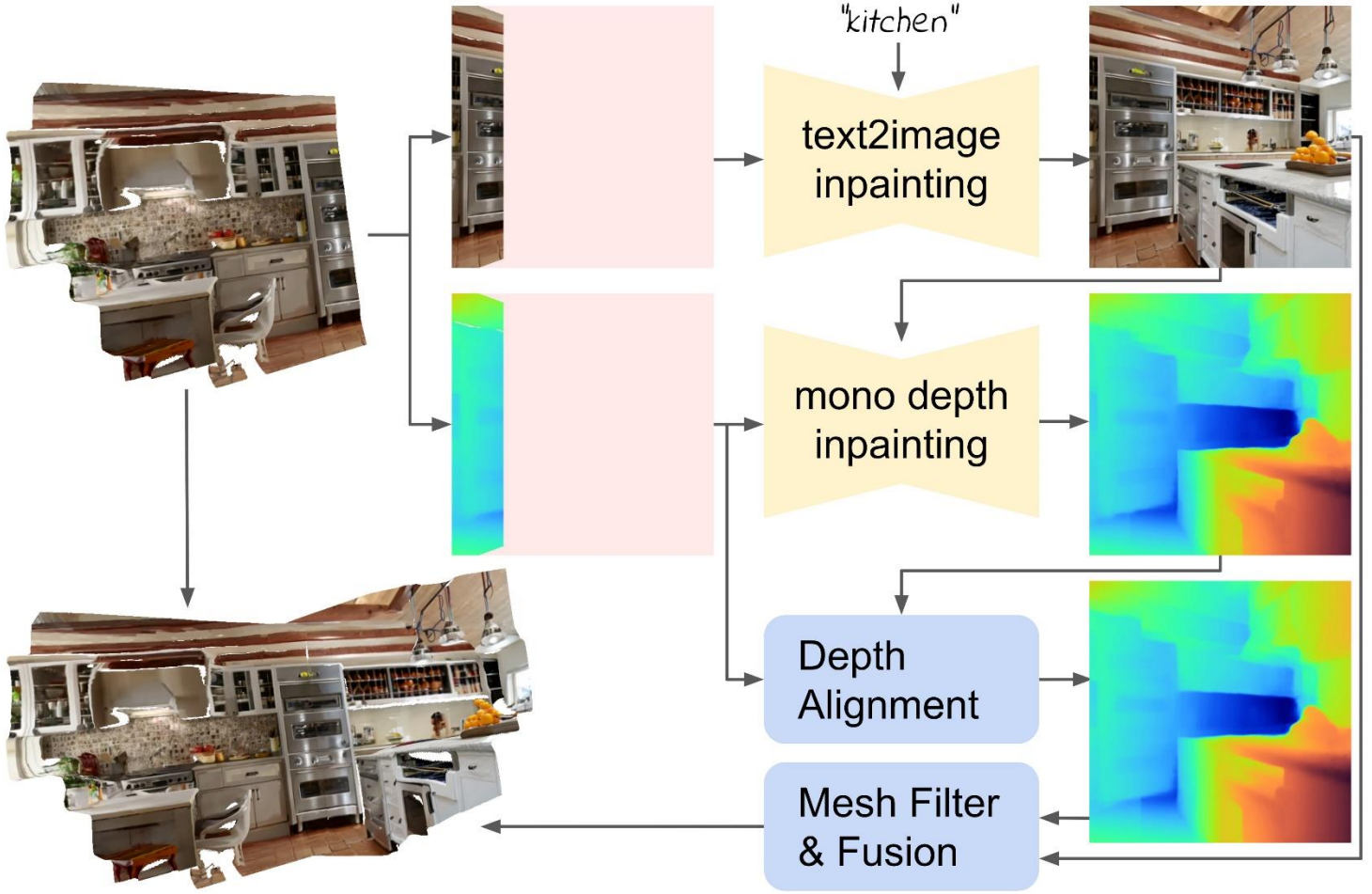




[Höllein and Cao et al. '22]

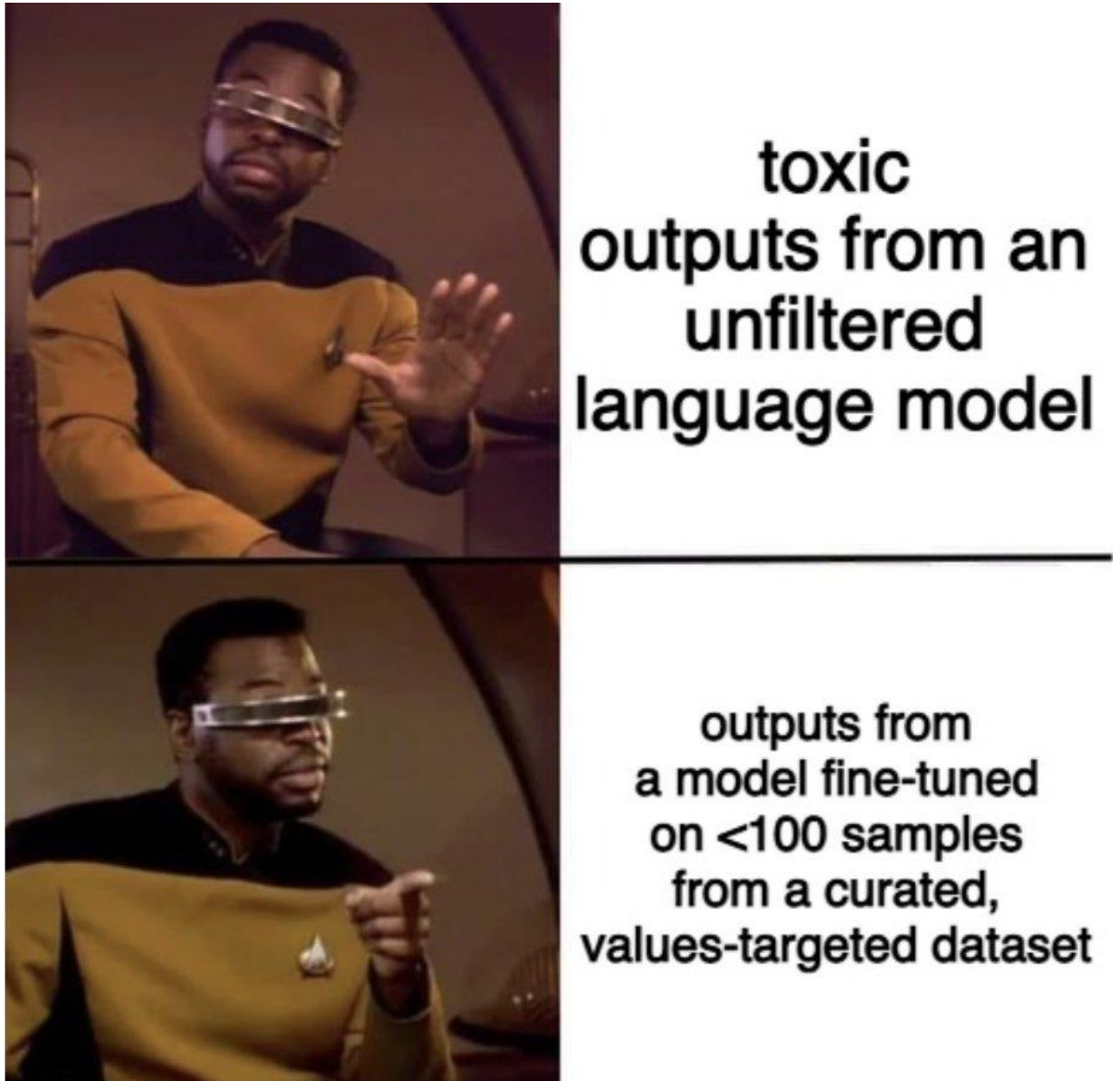
Obtaining 3D from 2D Generative Models

- and monocular depth estimation



[Höllein and Cao et al. '22]

Data?


toxic
outputs from an
unfiltered
language model
outputs from
a model fine-tuned
on <100 samples
from a curated,
values-targeted dataset

Interesting Challenges

- Data-efficient learning
- Combining weak signal from large data sources (e.g., text, images) with strong signal from smaller data sources (e.g., 3D/4D)
- 3D/4D representations and operators
  - Hierarchical/compositional representation?
- Higher-level semantic understanding, e.g., human interactions, functionality