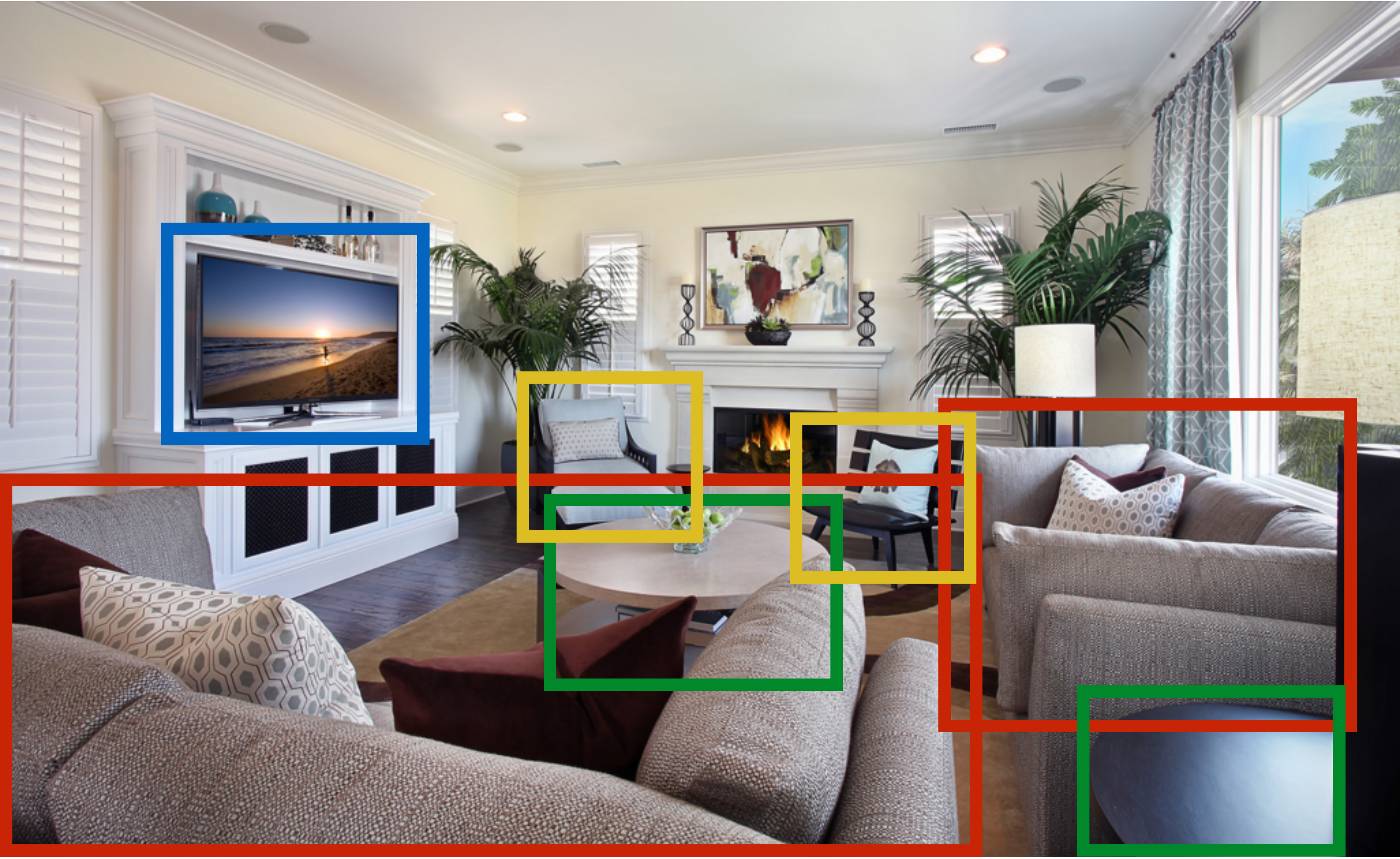
3D Visual Understanding

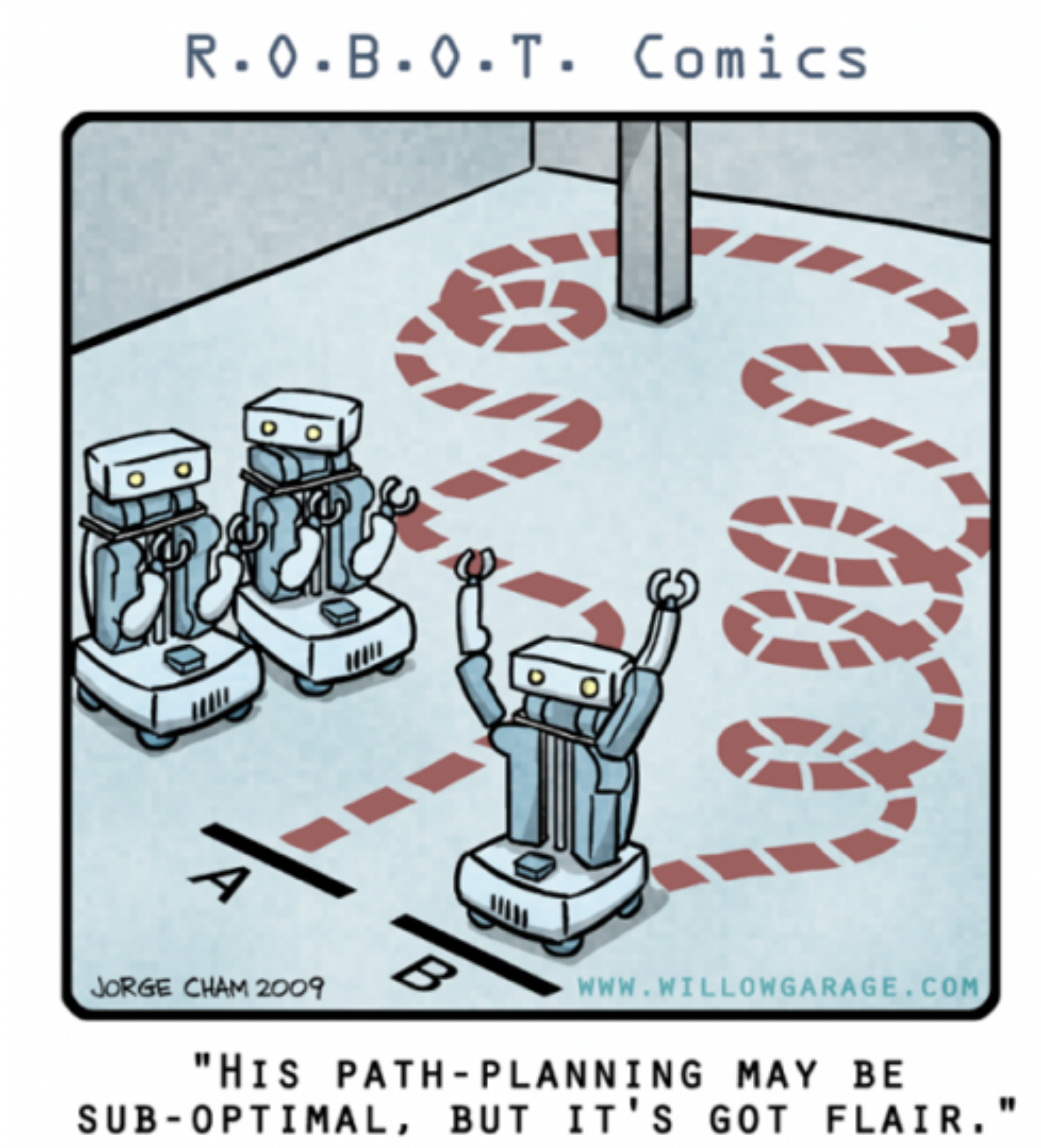
Shubham Tulsiani
University of California, Berkeley

sofa, chair, table, TV…



sofa, chair, table, TV…

Applications

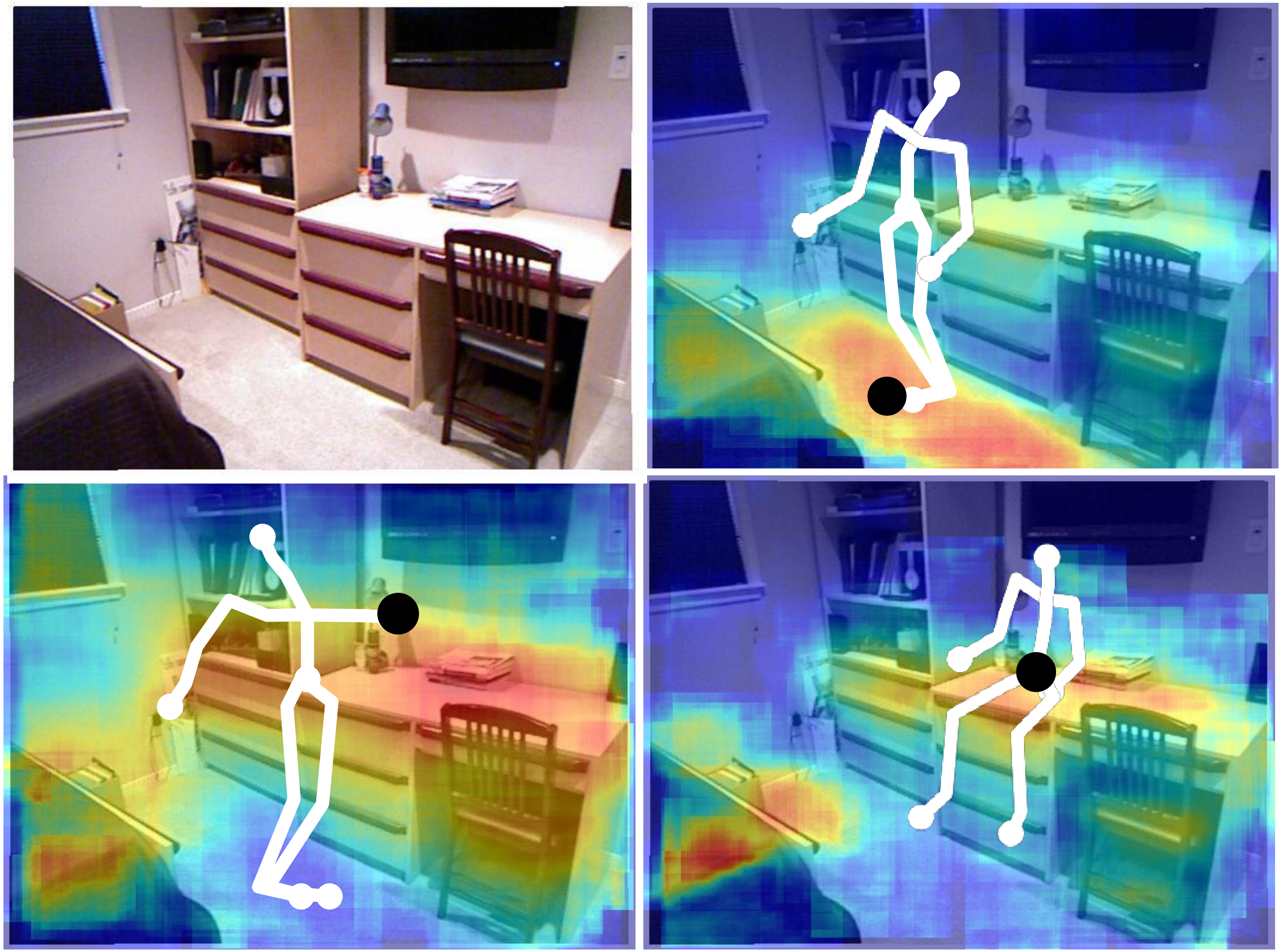


Robot Path Planning

Applications

Object Manipulation

Applications



Affordance Reasoning

Applications

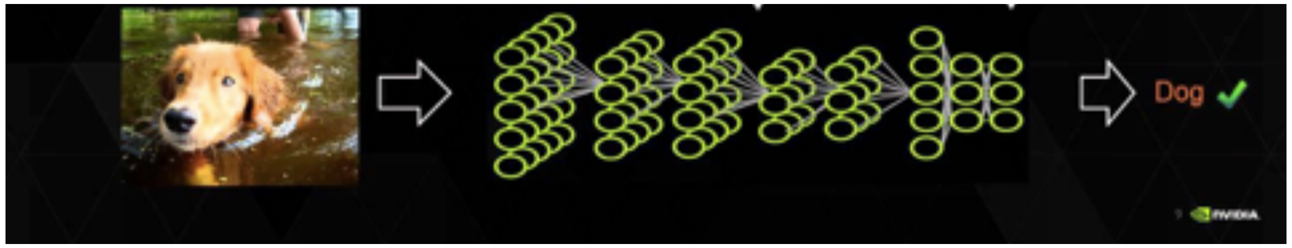
Image Based Graphics

3D Visual Understanding

- Background
- Objects in 3D
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems

3D Visual Understanding

- **Background**
- Objects in 3D
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems



Deep Learning in Computer Vision

Physics of Image Formation

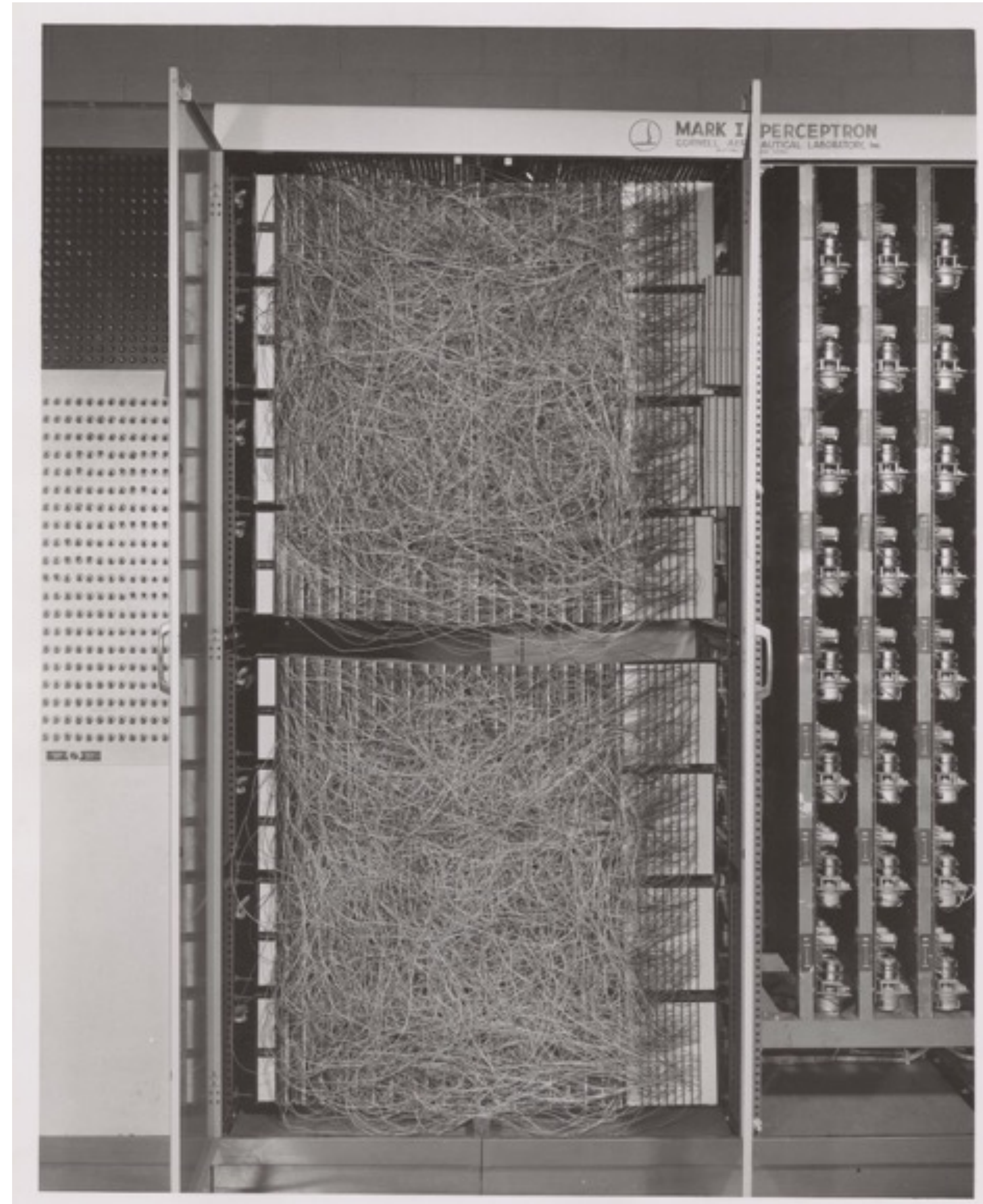
Neural Networks

Neural Networks

Perceptrons. Rosenblatt, 1957

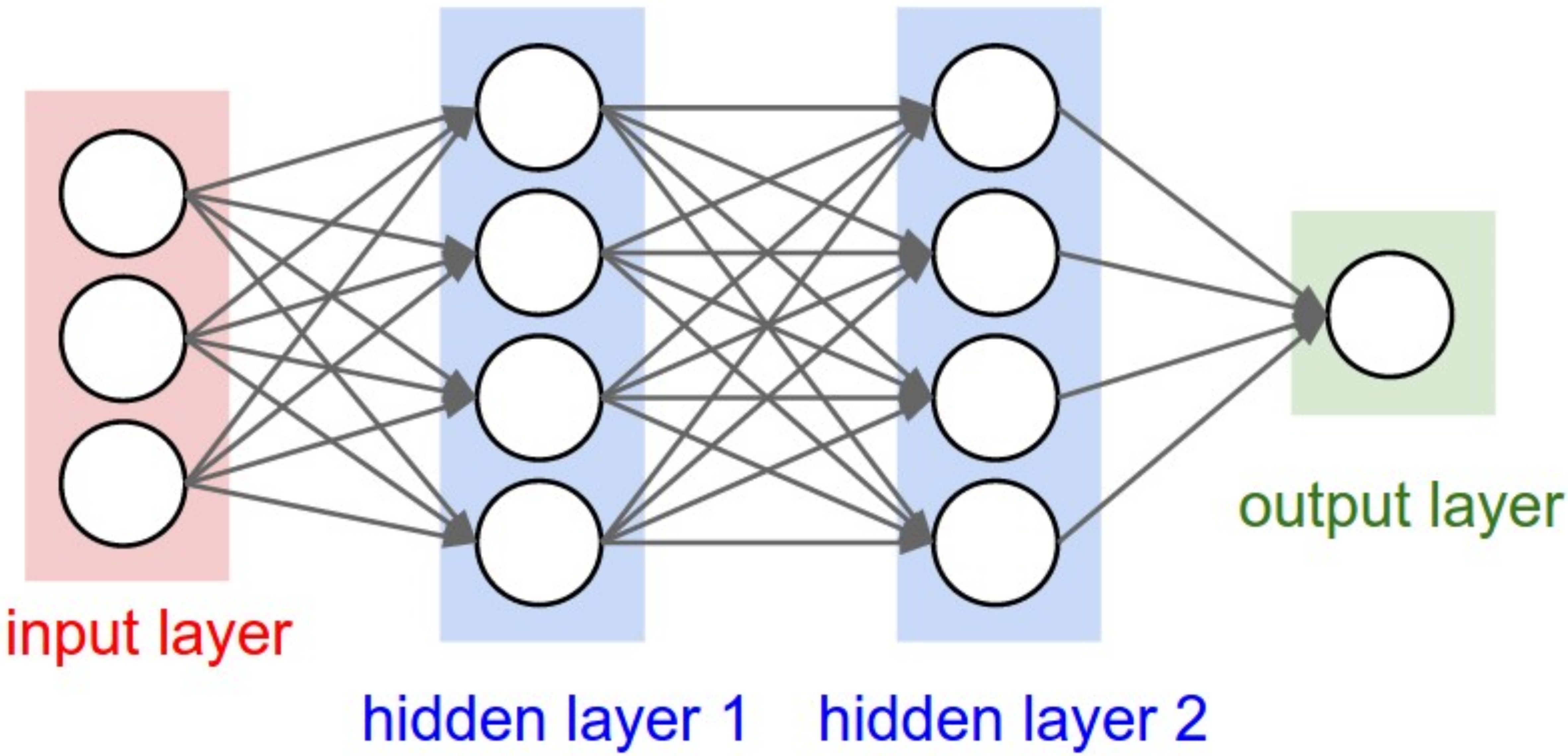
Neural Networks

Perceptrons. Rosenblatt, 1957

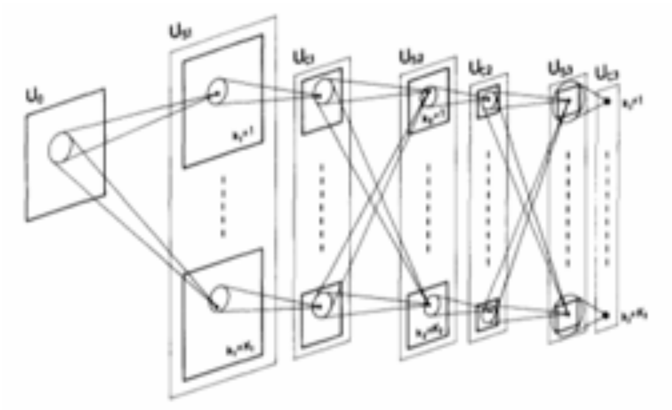


The Mark I Perceptron

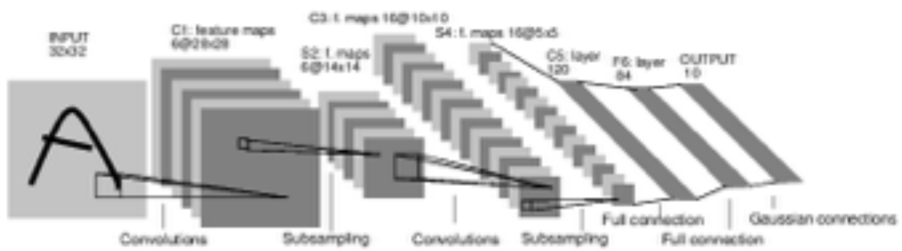
Neural Networks



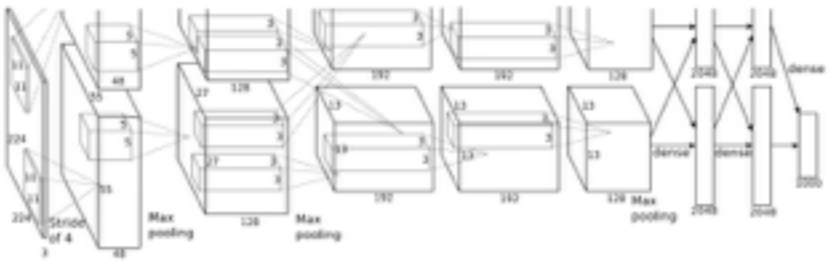
Convolutional Neural Networks



Fukushima, 1980



LeCun et al, 1989



Krizhevsky et al, 2012

- ❑ CNNs are Neural Networks with Convolutional layers – each output unit depends (via a spatially invariant linear function) on a set of neighbouring input units
- ❑ Particularly relevant for input domains with spatial structure (e.g. images)

CNNs in Computer Vision

Object Detection

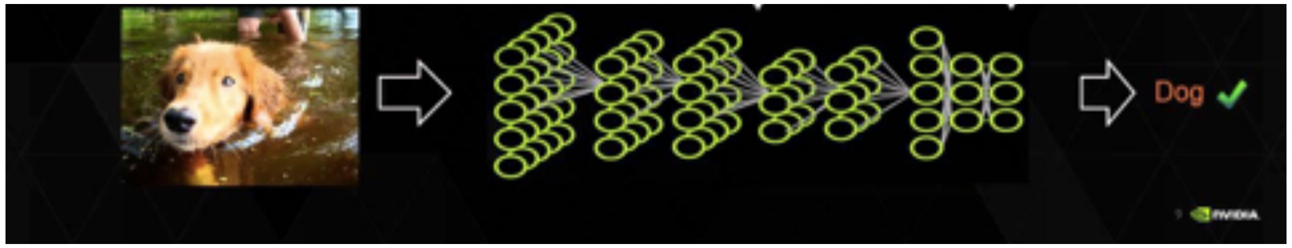
Image Captioning

Semantic Segmentation

Human Pose Estimation

3D Visual Understanding

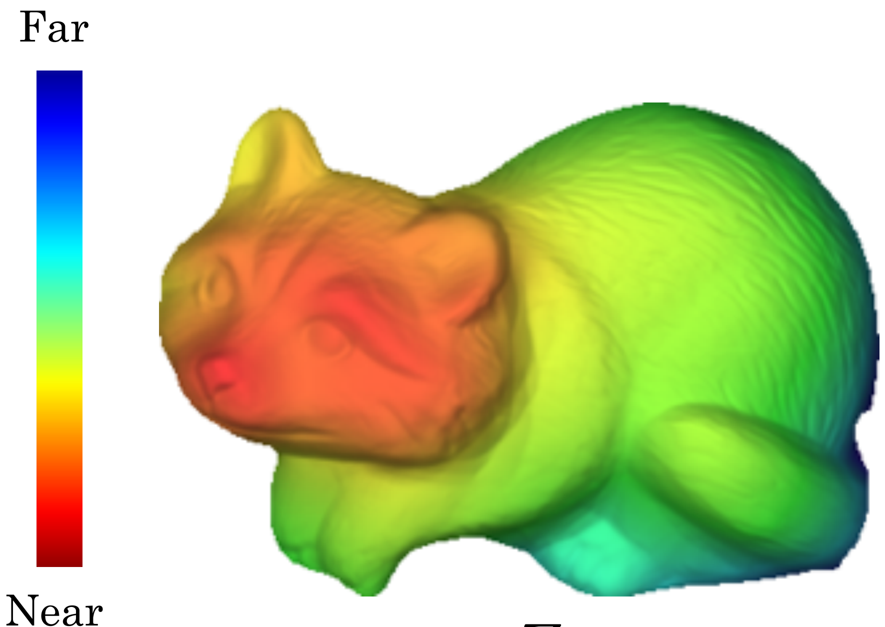
- **Background**
- Objects in 3D
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems



Deep Learning in Computer Vision

Physics of Image Formation

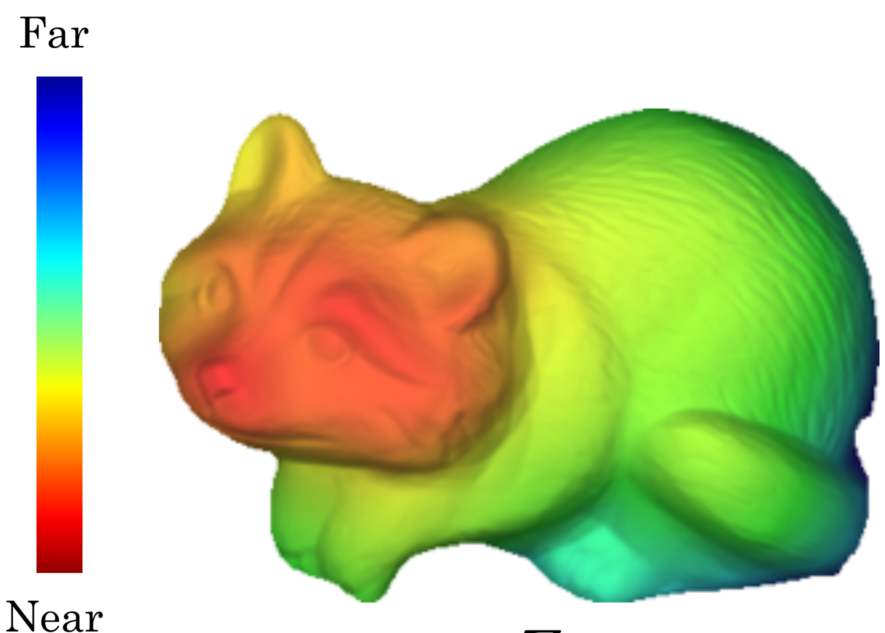
Forward Optics



Z
shape / depth

Slide Courtesy : Jon Barron

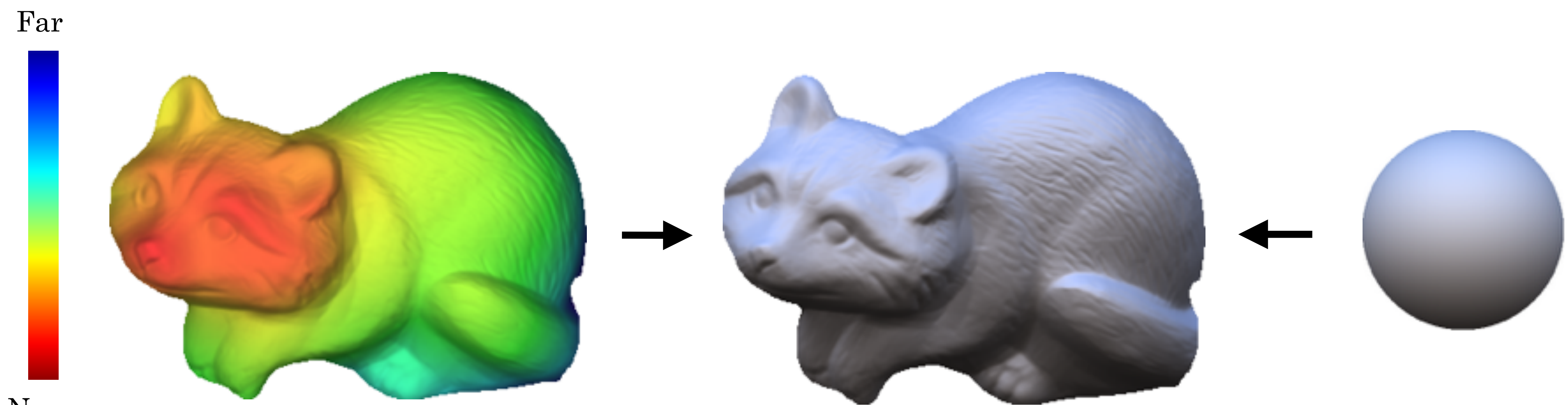
Forward Optics



Z
shape / depth

L
illumination

Forward Optics



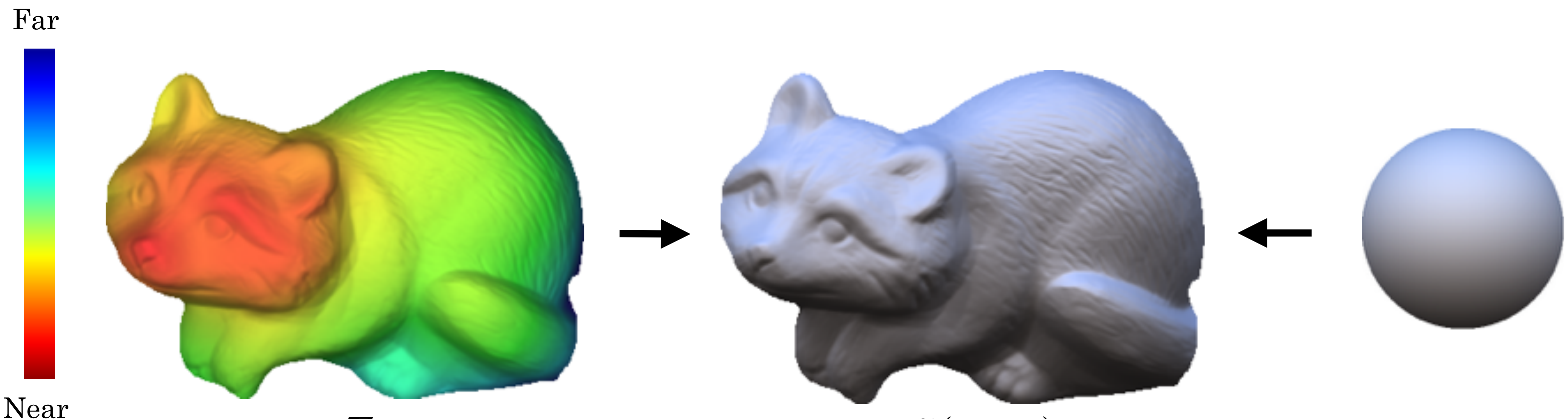
Z
shape / depth

$S(Z, L)$
log-shading image of Z and L

L
illumination

Slide Courtesy : Jon Barron

Forward Optics



Z
shape / depth

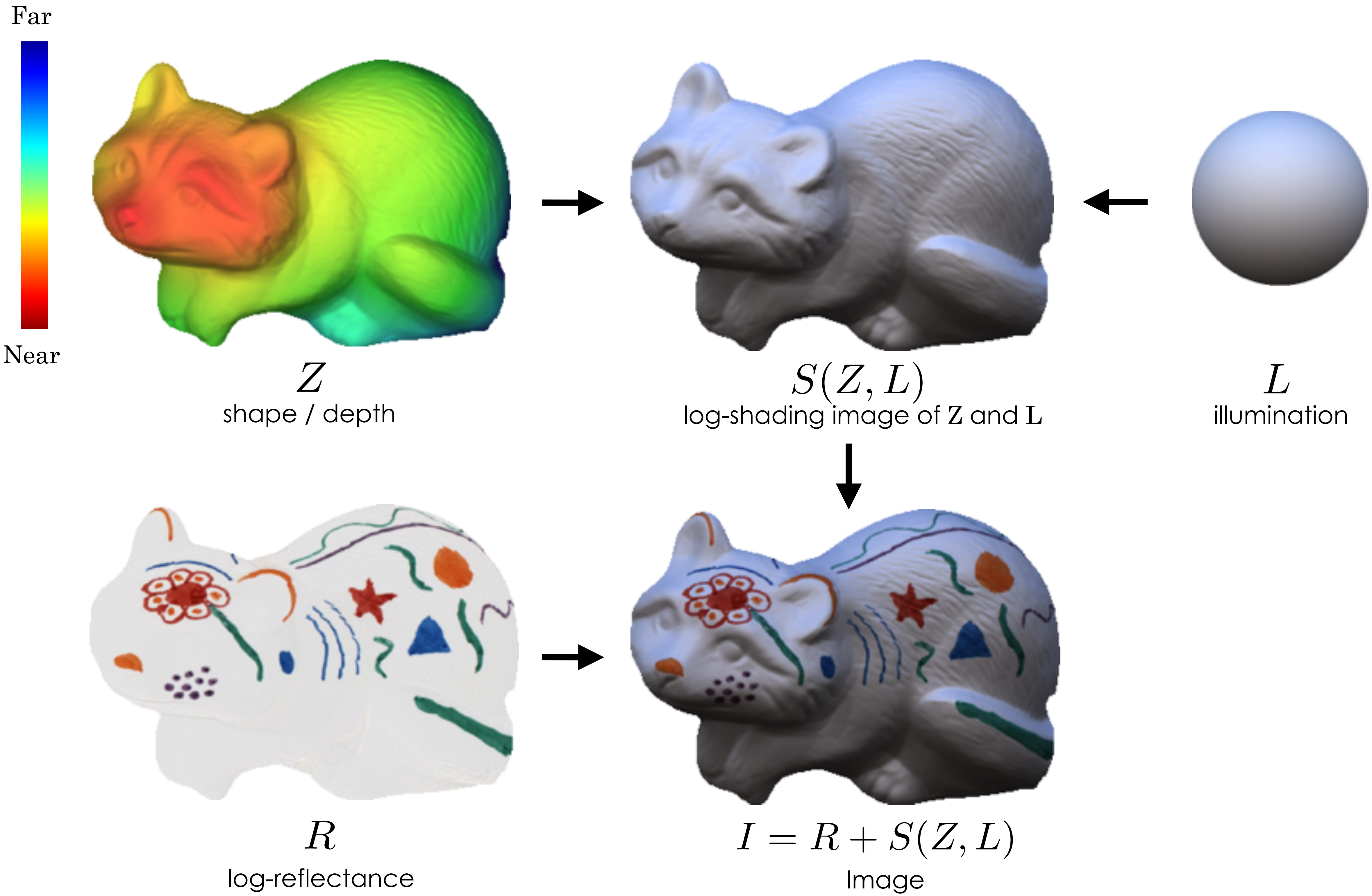
$S(Z, L)$
log-shading image of Z and L

L
illumination

R
log-reflectance

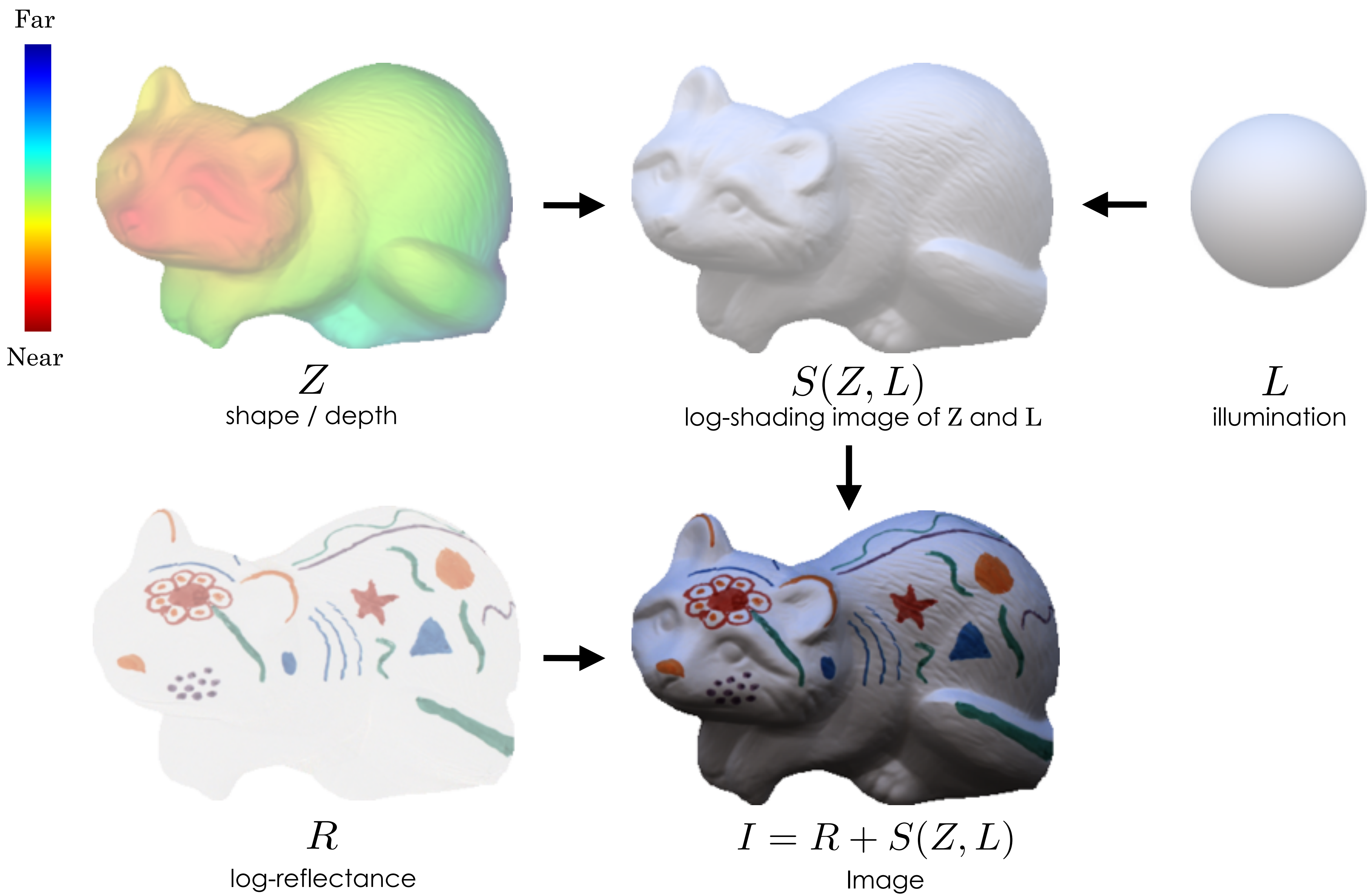
Slide Courtesy : Jon Barron

Forward Optics



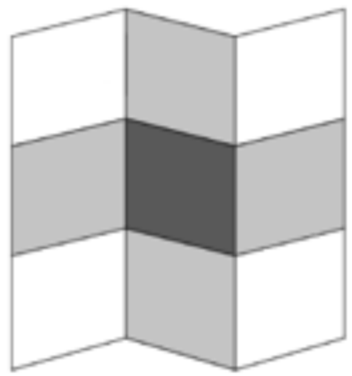
Slide Courtesy : Jon Barron

Physics of Image Formation



Slide Courtesy : Jon Barron

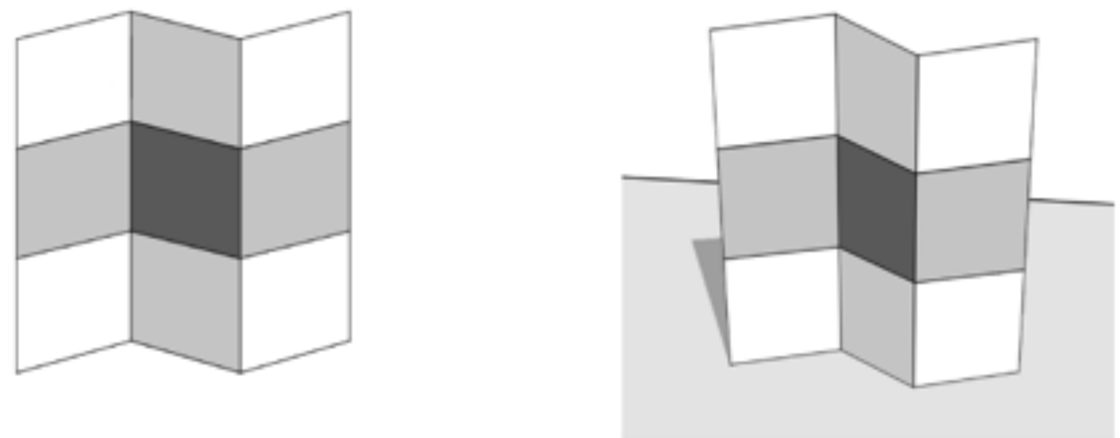
The Workshop Metaphor



(a) an image

E. Adelson and A. Pentland, "The perception of shading and reflectance," *Perception as Bayesian inference*, 1996.

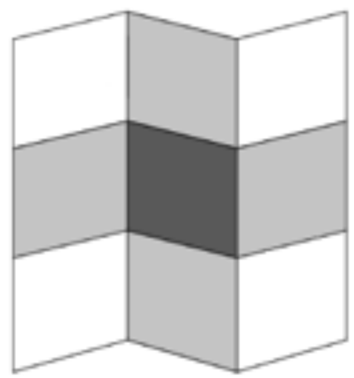
The Workshop Metaphor



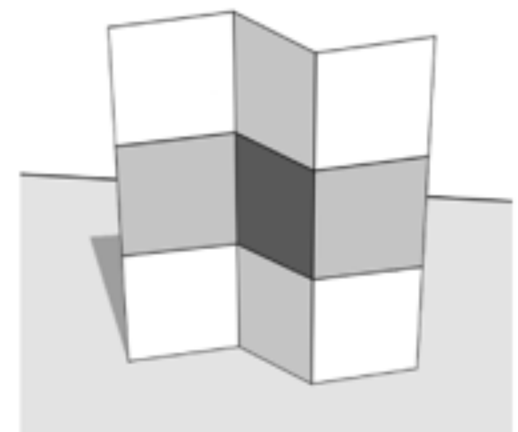
(a) an image (b) a likely explanation

E. Adelson and A. Pentland, "The perception of shading and reflectance," *Perception as Bayesian inference*, 1996.

The Workshop Metaphor



(a) an image

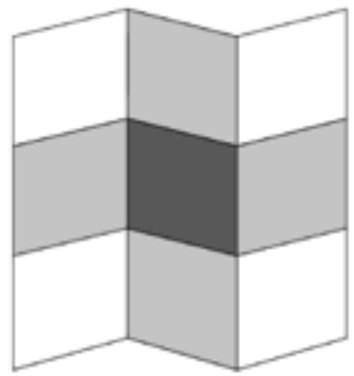


(b) a likely explanation

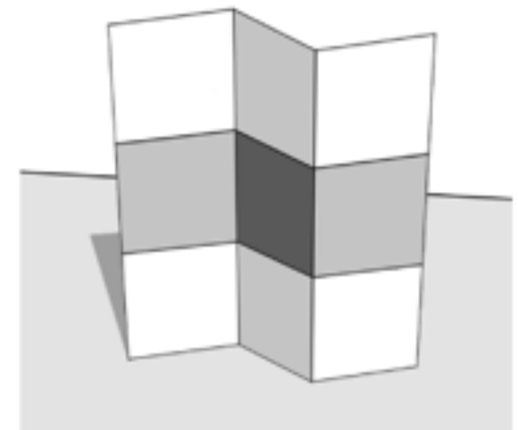
(c) painter's explanation

E. Adelson and A. Pentland, "The perception of shading and reflectance," *Perception as Bayesian inference*, 1996.

The Workshop Metaphor

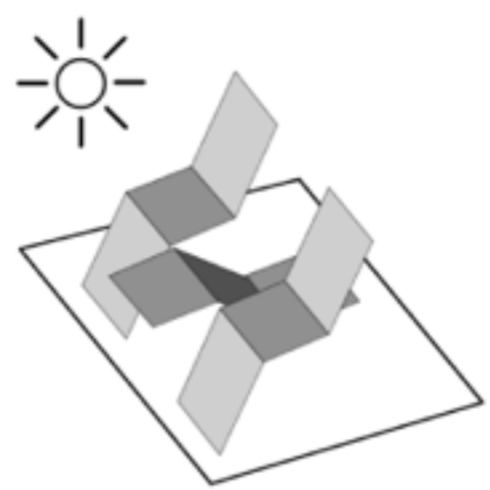


(a) an image



(b) a likely explanation

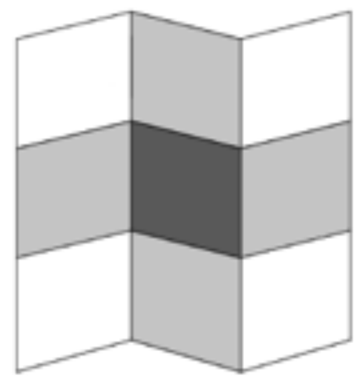
(c) painter's explanation



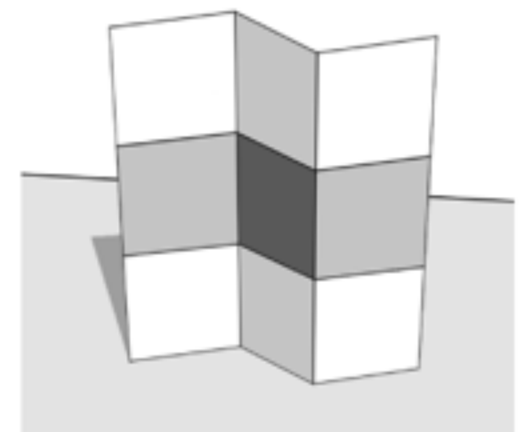
(d) sculptor's explanation

E. Adelson and A. Pentland, "The perception of shading and reflectance," *Perception as Bayesian inference*, 1996.

The Workshop Metaphor

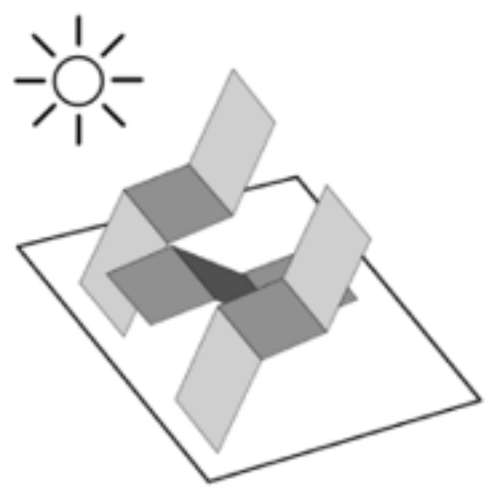


(a) an image

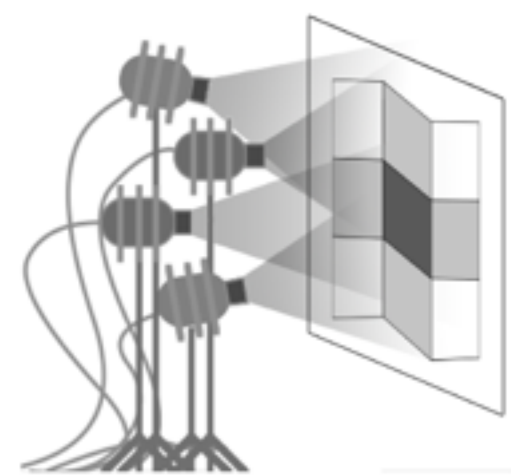


(b) a likely explanation

(c) painter's explanation



(d) sculptor's explanation

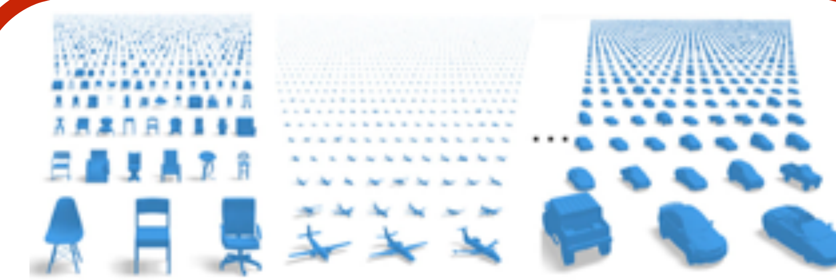


(e) gaffer's explanation

E. Adelson and A. Pentland, "The perception of shading and reflectance," *Perception as Bayesian inference*, 1996.

3D Visual Understanding

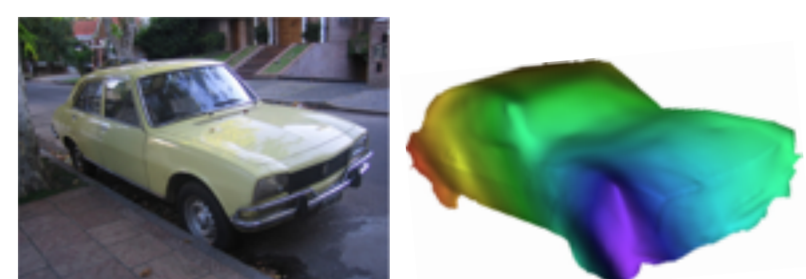
- Background
- **Objects in 3D**
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems



Shape Collections

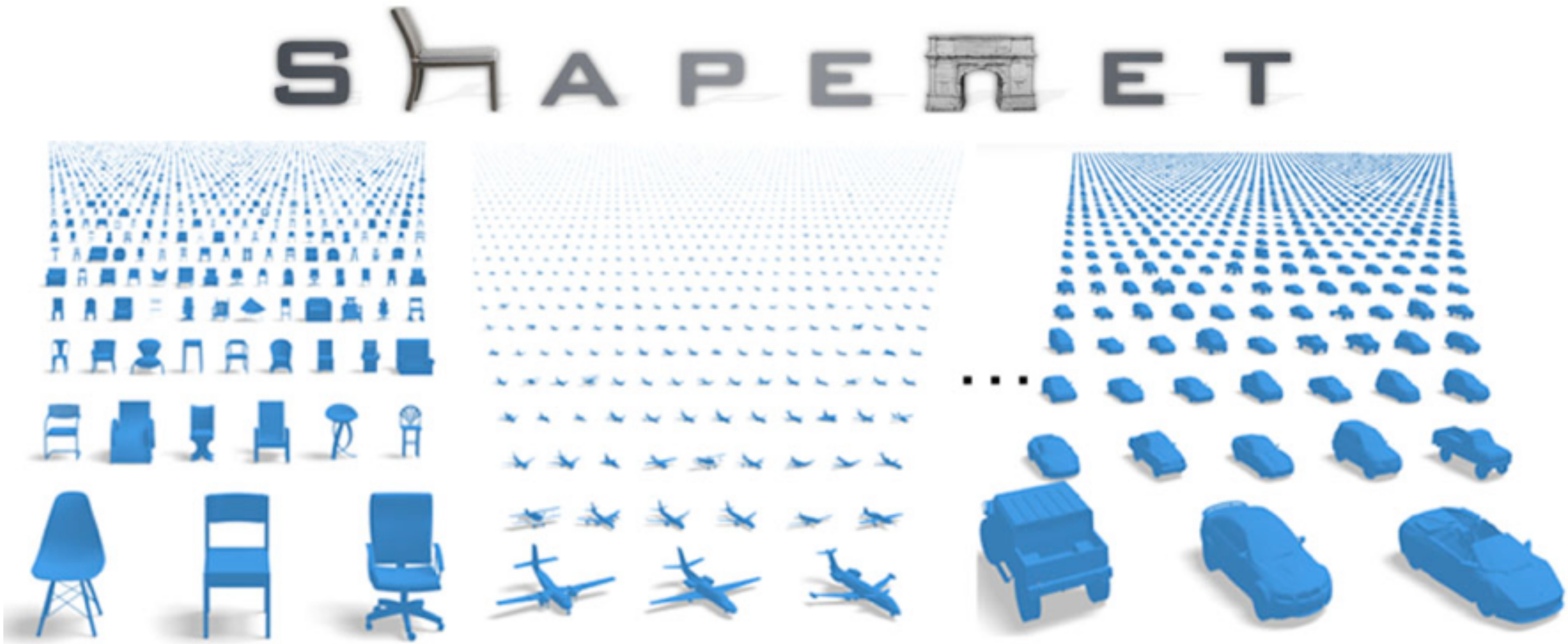
Shape Representation

Object Reconstruction



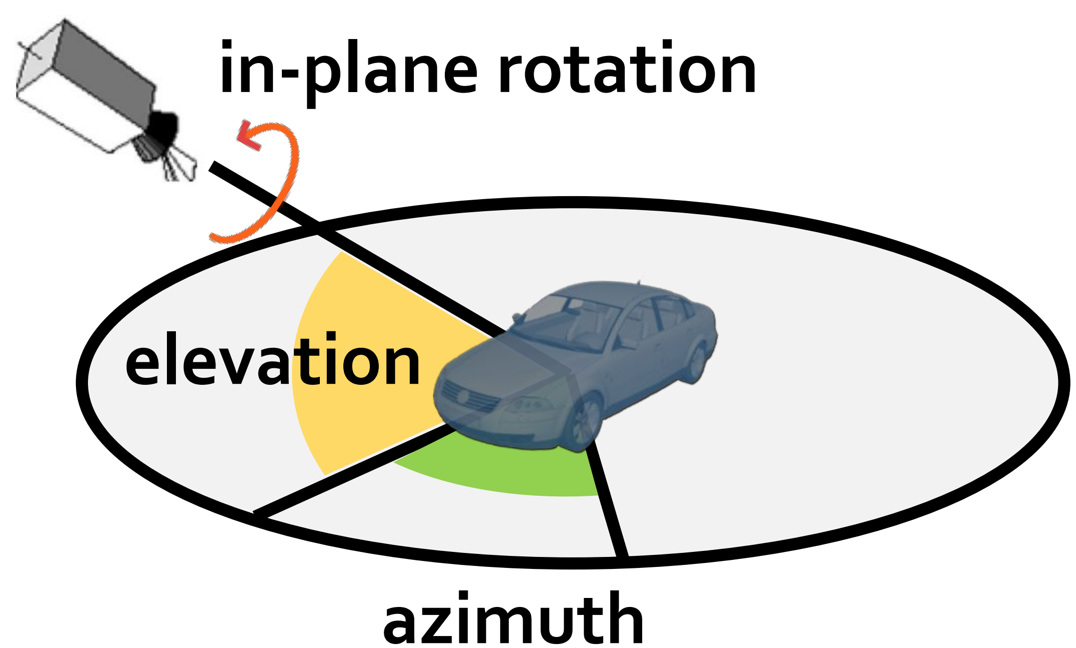
Reconstruction ‘in the wild’

Shape Collections for 3D Understanding



Shape Collections for 3D Understanding

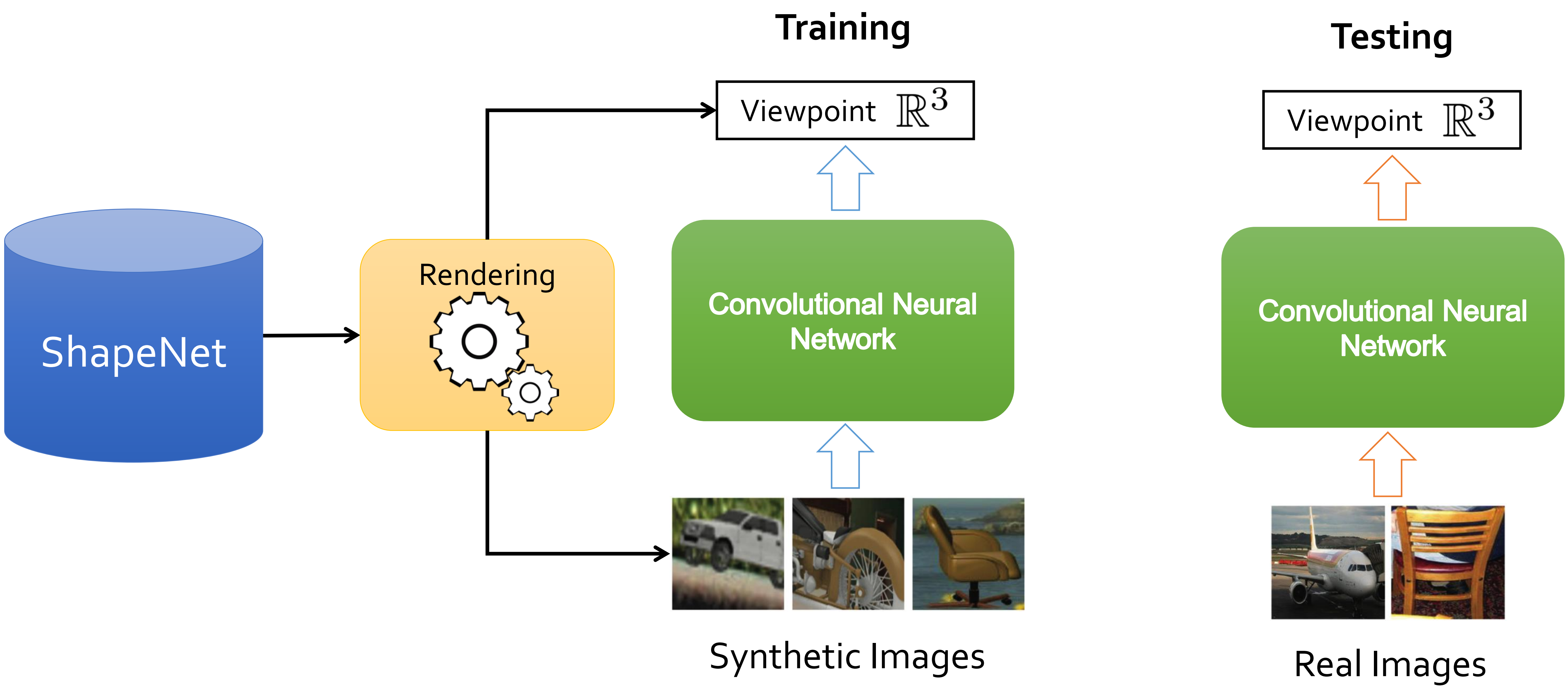
3D Viewpoint Estimation



Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views

Su, Qi, Li, Guibas

Shape Collections for 3D Understanding



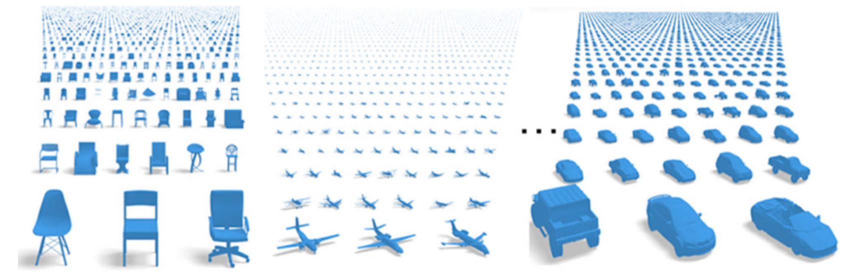
Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views
Su, Qi, Li, Guibas

Shape Collections for 3D Understanding

Render for CNN: Viewpoint Estimation in Images Using CNNs
Trained with Rendered 3D Model Views
Su, Qi, Li, Guibas

3D Visual Understanding

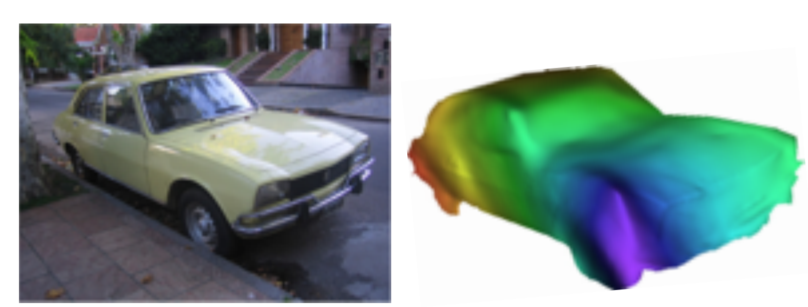
- Background
- **Objects in 3D**
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems



Shape Collections

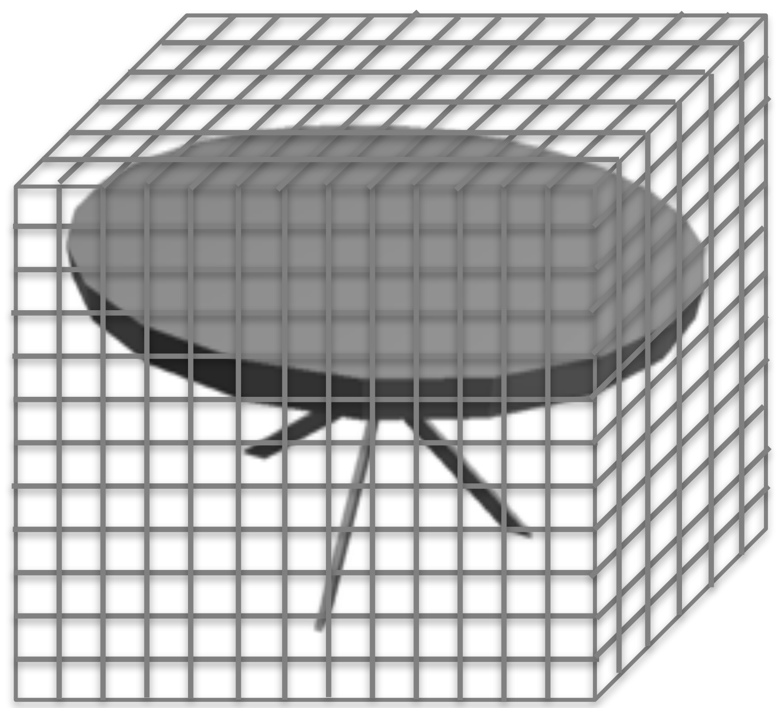
Shape Representation

Object Reconstruction



Reconstruction 'in the wild'

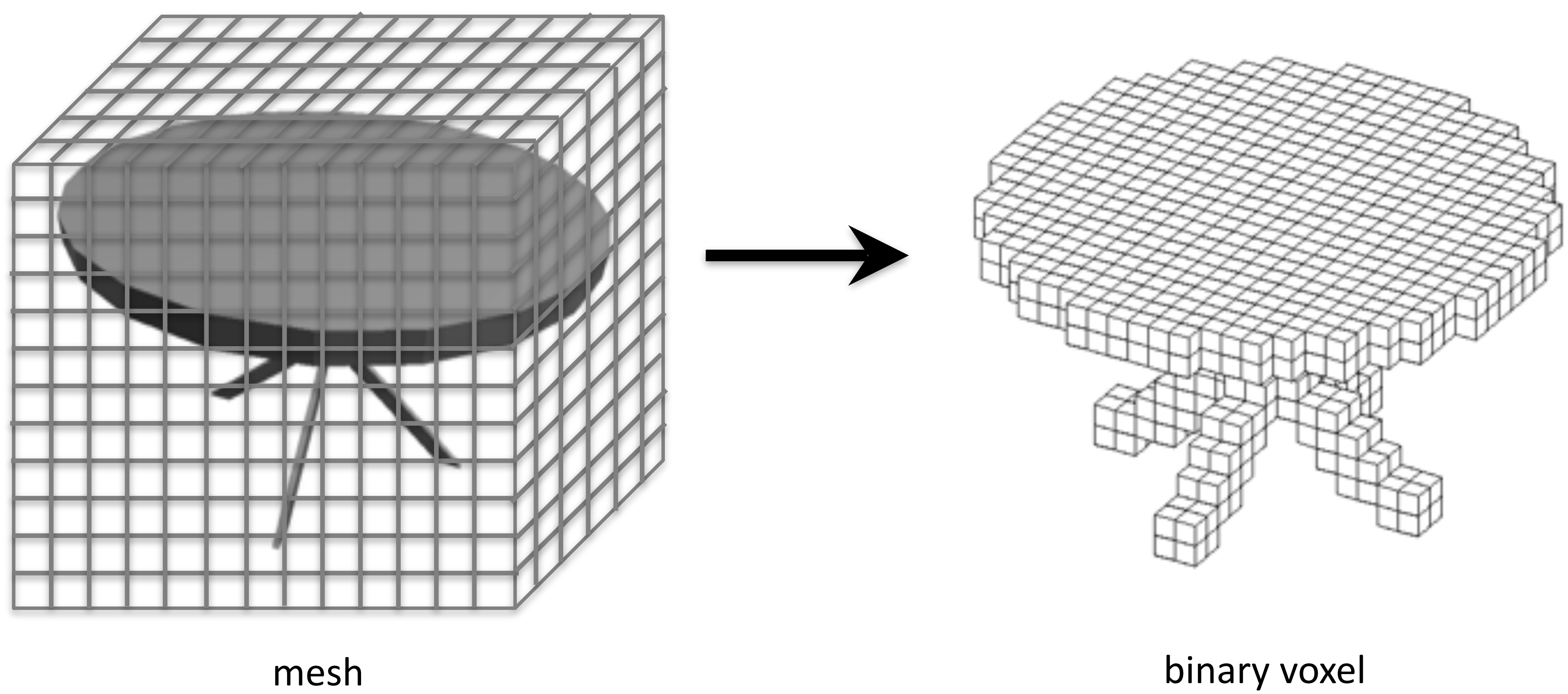
Shape Representations



mesh

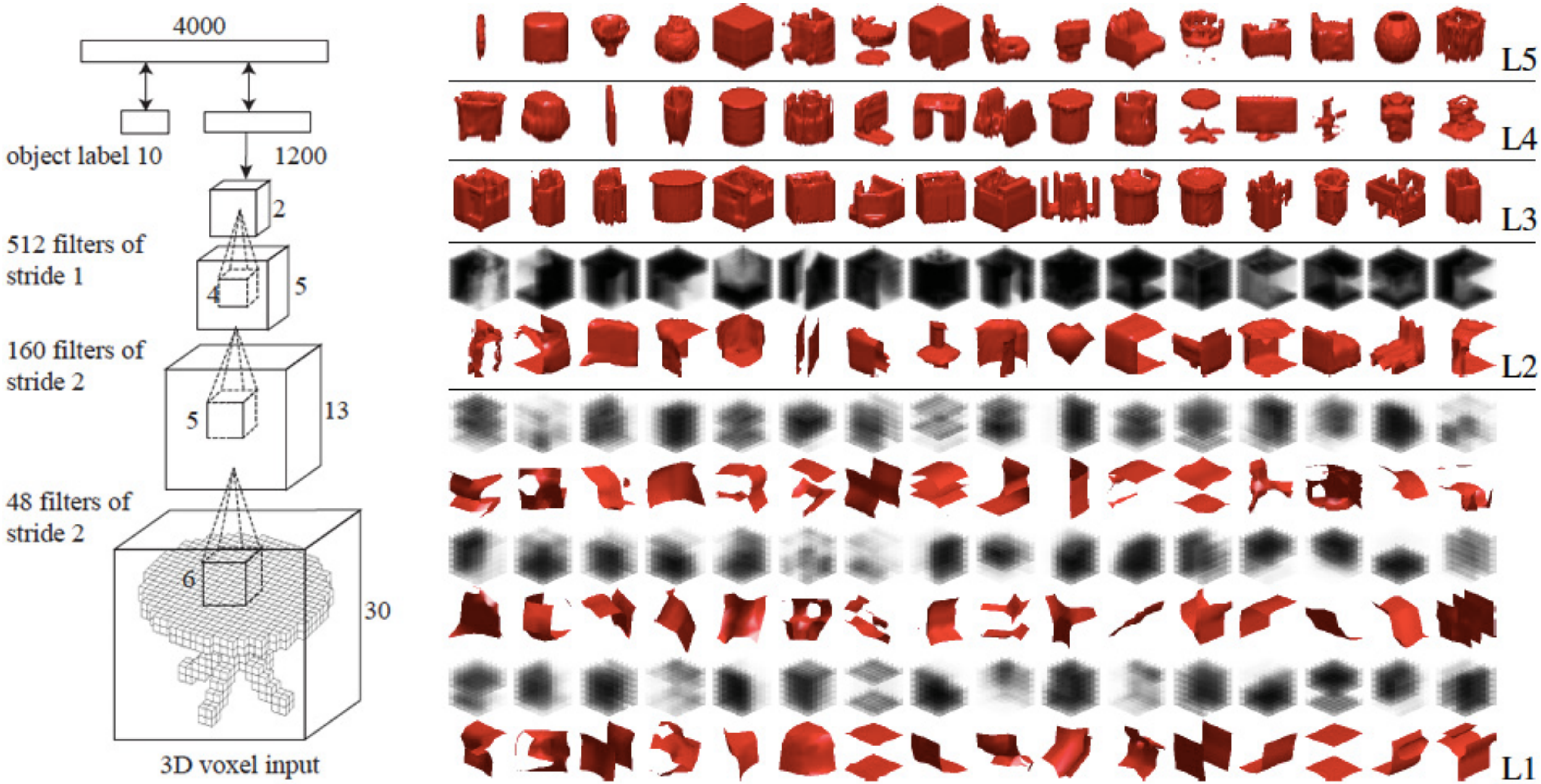
3D ShapeNets: A Deep Representation for Volumetric Shapes
Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao

Shape Representations



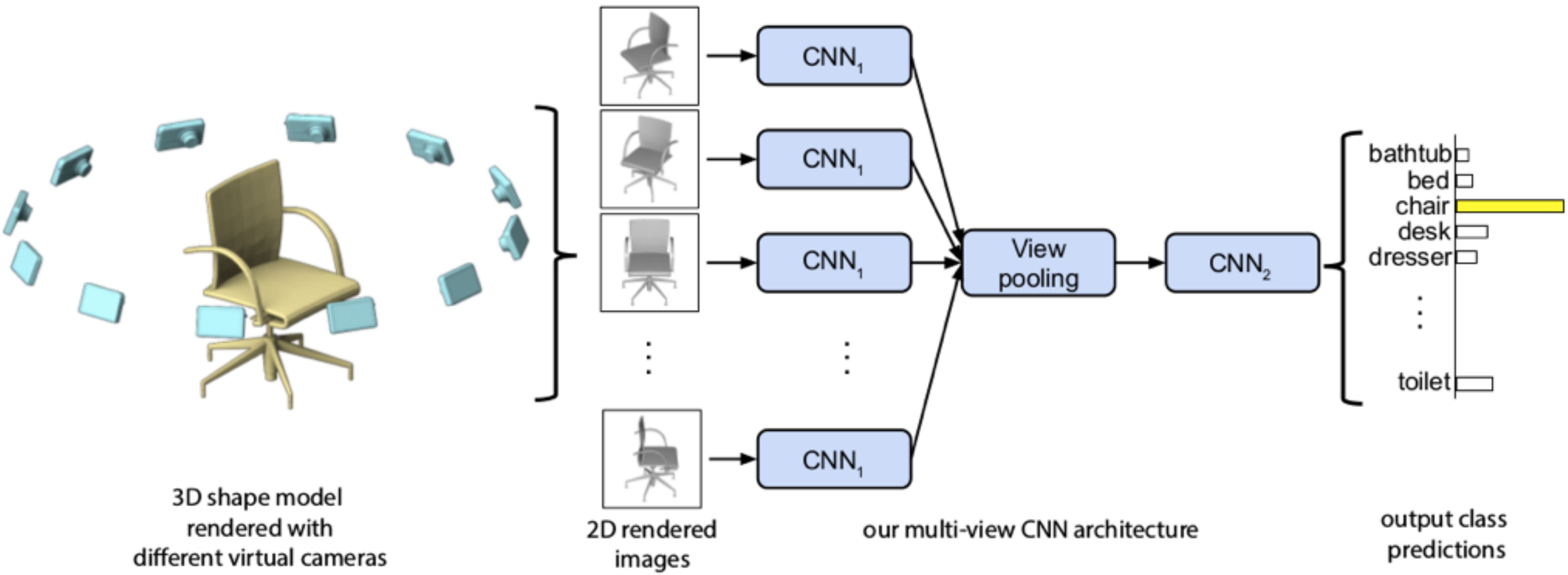
3D ShapeNets: A Deep Representation for Volumetric Shapes
Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao

Shape Representations



3D ShapeNets: A Deep Representation for Volumetric Shapes
Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao

Shape Representations



Multi-view Convolutional Neural Networks for 3D Shape Recognition

Su, Maji, Kalogerakis, Learned-Miller

Shape Representations

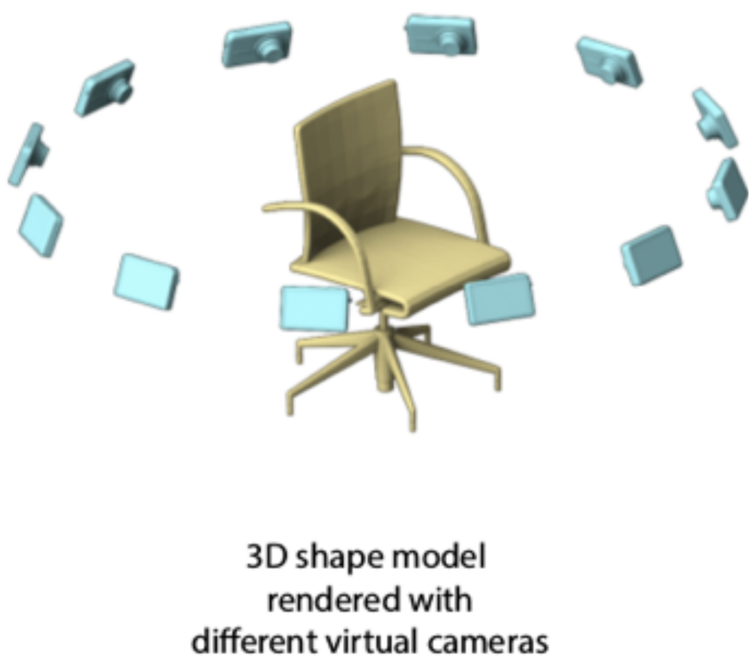
Method	Training Config.			Test Config.	Classification (Accuracy)	Retrieval (mAP)
	Pre-train	Fine-tune	#Views	#Views		
(1) SPH [16]	-	-	-	-	68.2%	33.3%
(2) LFD [5]	-	-	-	-	75.5%	40.9%
(3) 3D ShapeNets [37]	ModelNet40	ModelNet40	-	-	77.3%	49.2%
(4) FV	-	ModelNet40	12	1	78.8%	37.5%
(5) FV, 12×	-	ModelNet40	12	12	84.8%	43.9%
(6) CNN	ImageNet1K	-	-	1	83.0%	44.1%
(7) CNN, f.t.	ImageNet1K	ModelNet40	12	1	85.1%	61.7%
(8) CNN, 12×	ImageNet1K	-	-	12	87.5%	49.6%
(9) CNN, f.t.,12×	ImageNet1K	ModelNet40	12	12	88.6%	62.8%
(10) MVCNN, 12×	ImageNet1K	-	-	12	88.1%	49.4%
(11) MVCNN, f.t., 12×	ImageNet1K	ModelNet40	12	12	89.9%	70.1%
(12) MVCNN, f.t.+metric, 12×	ImageNet1K	ModelNet40	12	12	89.5%	**80.2**%
(13) MVCNN, 80×	ImageNet1K	-	80	80	84.3%	36.8%
(14) MVCNN, f.t., 80×	ImageNet1K	ModelNet40	80	80	**90.1**%	70.4%
(15) MVCNN, f.t.+metric, 80×	ImageNet1K	ModelNet40	80	80	**90.1**%	79.5%

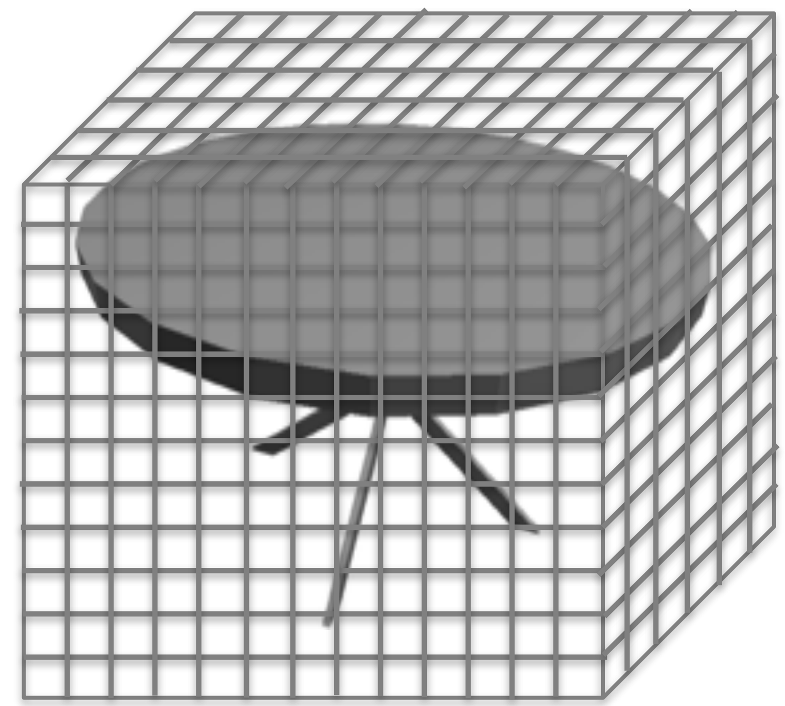
* f.t.=fine-tuning, metric=low-rank Mahalanobis metric learning

Multi-view Convolutional Neural Networks for 3D Shape Recognition

Su, Maji, Kalogerakis, Learned-Miller

Shape Representations



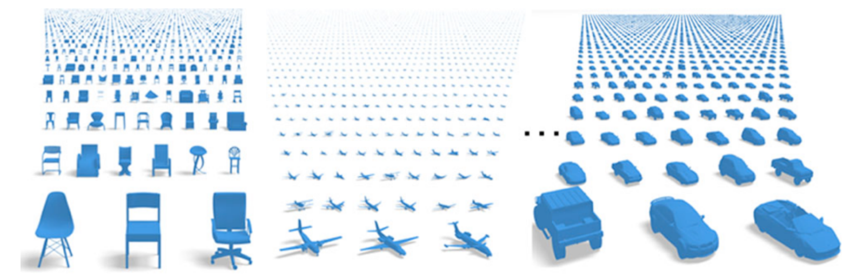


mesh

- Do we represent shapes as features in image space or mesh space ?

3D Visual Understanding

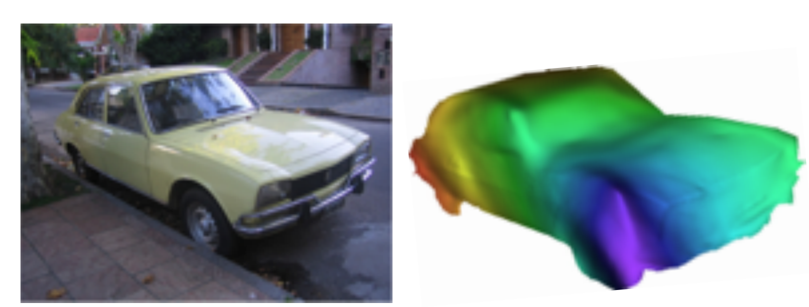
- Background
- **Objects in 3D**
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems



Shape Collections

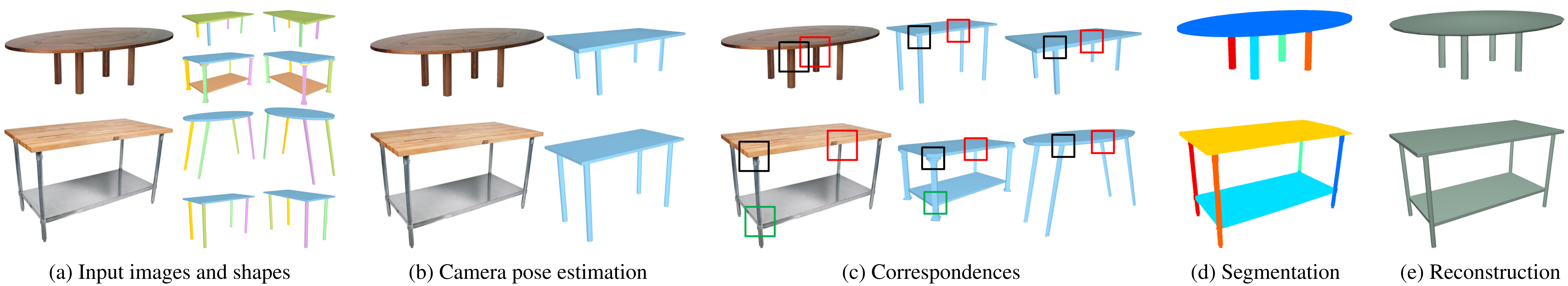
Shape Representation

Object Reconstruction



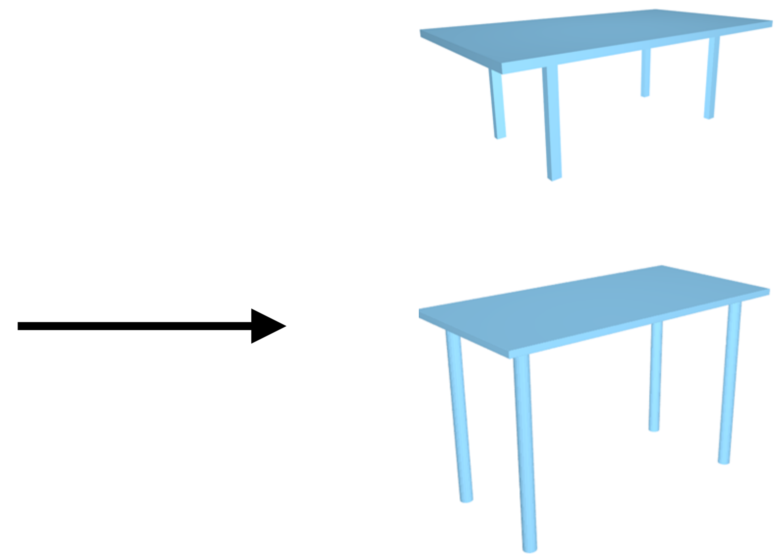
Reconstruction ‘in the wild’

Object Reconstruction



Single-View Reconstruction via Joint Analysis of Image and Shape Collections
Huang, Wang, Koltun

Object Reconstruction

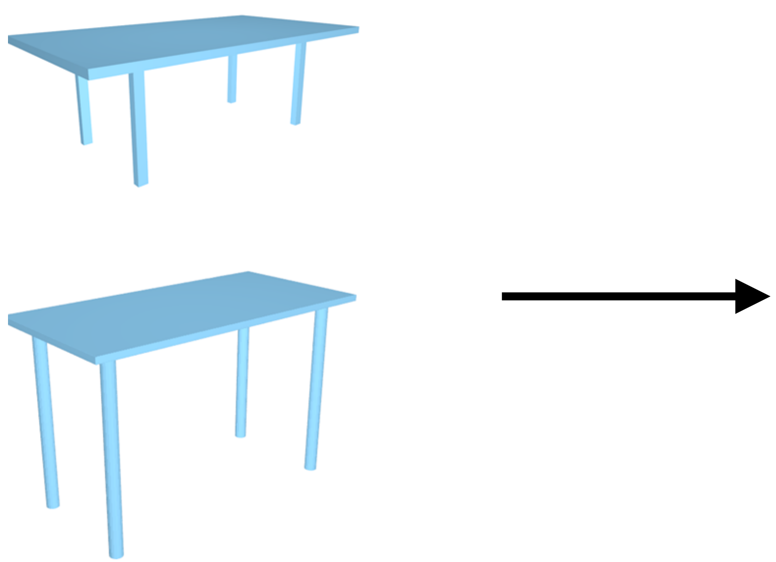


Input Image

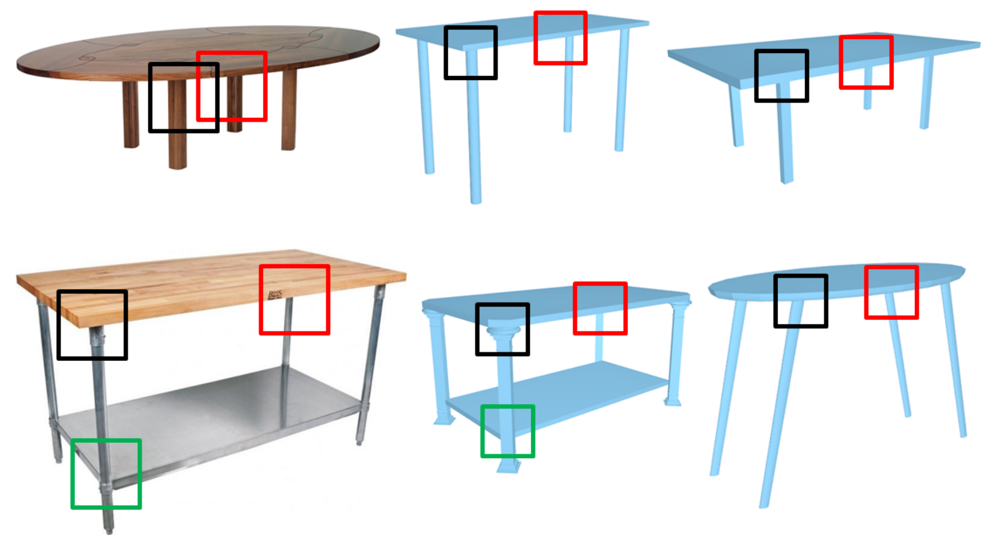
Pose Estimation

Single-View Reconstruction via Joint Analysis of Image and Shape Collections
Huang, Wang, Koltun

Object Reconstruction



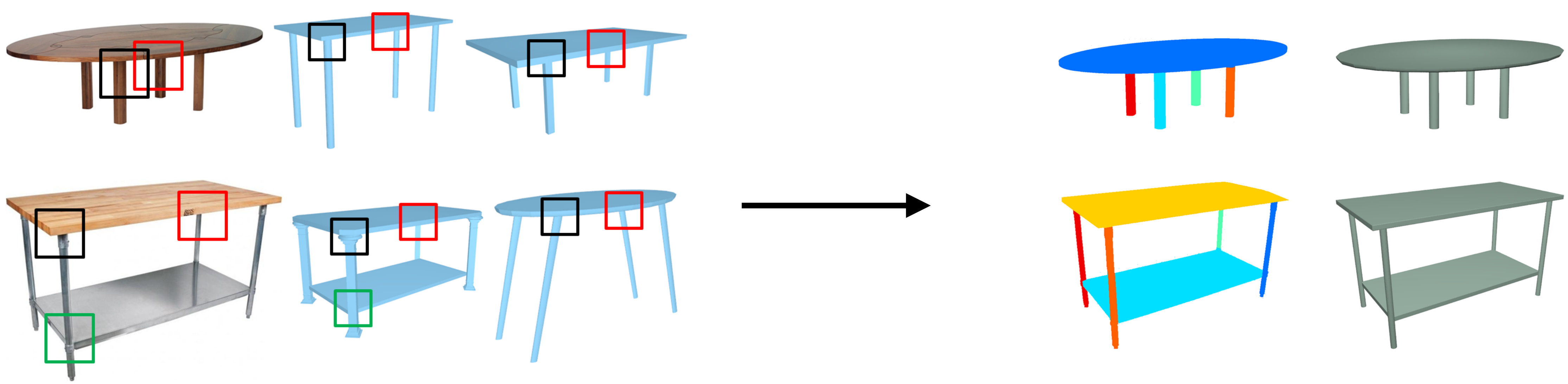
Pose Estimation



Similar Models and Correspondences

Single-View Reconstruction via Joint Analysis of Image and Shape Collections
Huang, Wang, Koltun

Object Reconstruction



Similar Models and Correspondences

Part Segmentation and Reconstruction

Single-View Reconstruction via Joint Analysis of Image and Shape Collections
Huang, Wang, Koltun

Object Reconstruction

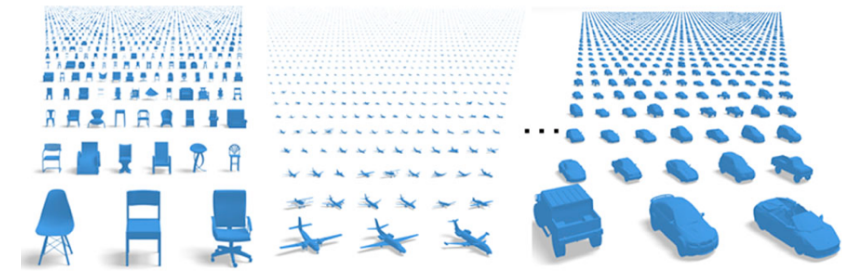
Figure 8: *Results on four datasets. From left to right in each column: Web image, computed segmentation, 3D model reconstructed by our approach (two views, green), and closest pre-existing model, shown for reference (blue).*

Results

Single-View Reconstruction via Joint Analysis of Image and Shape Collections
Huang, Wang, Koltun

3D Visual Understanding

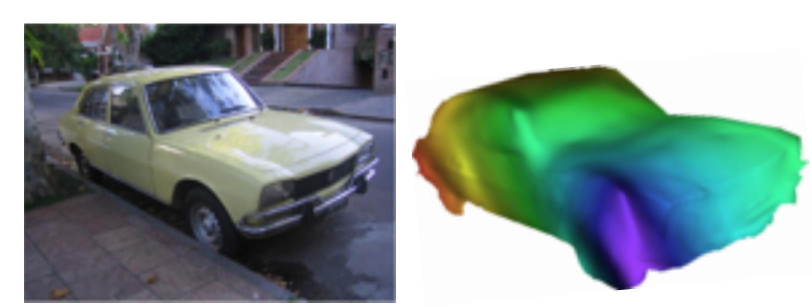
- Background
- **Objects in 3D**
- Scenes in 3D
- 3D Understanding without Understanding 3D
- Open Problems



Shape Collections

Shape Representation

Object Reconstruction



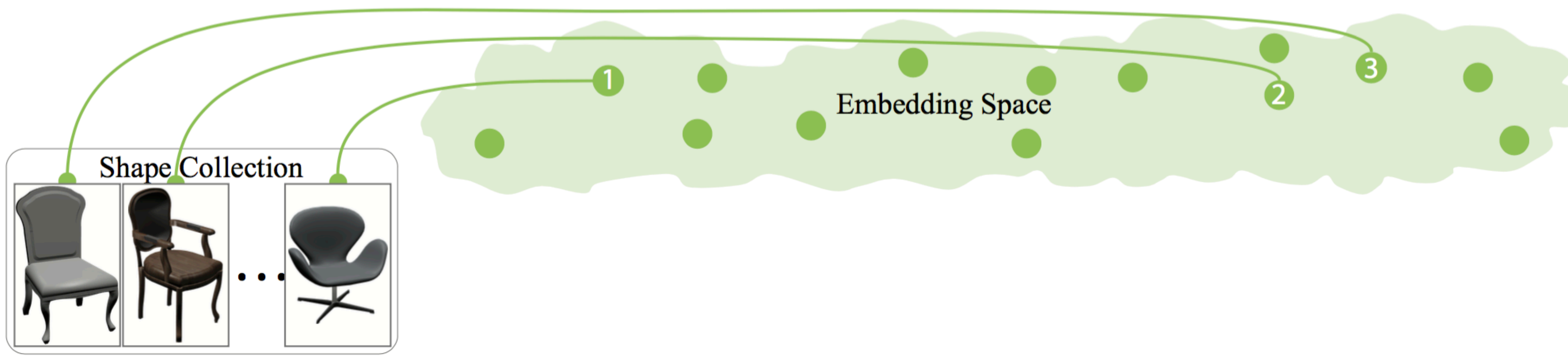
Reconstruction ‘in the wild’

Reconstruction in the ‘Wild’

Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the 'Wild'

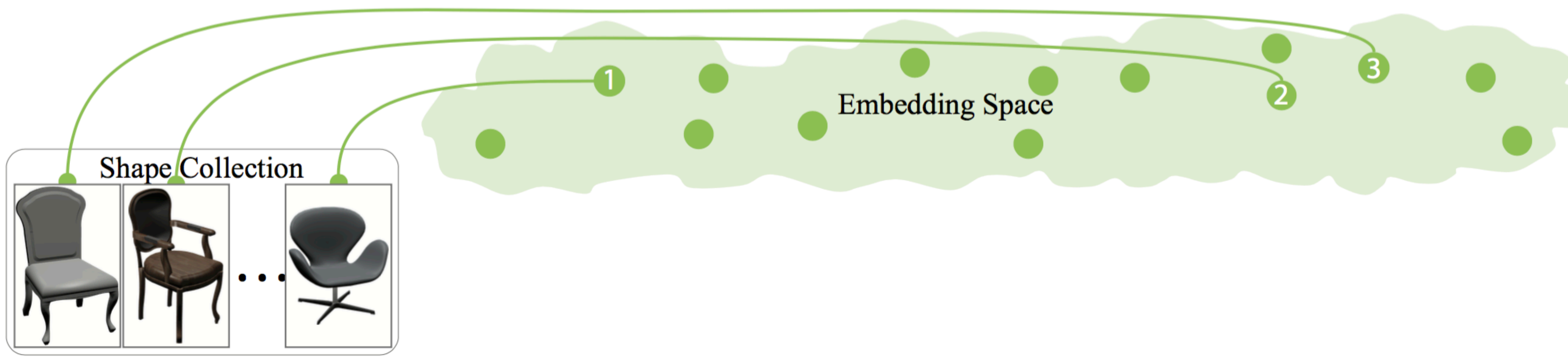
- Learn an embedding of shapes



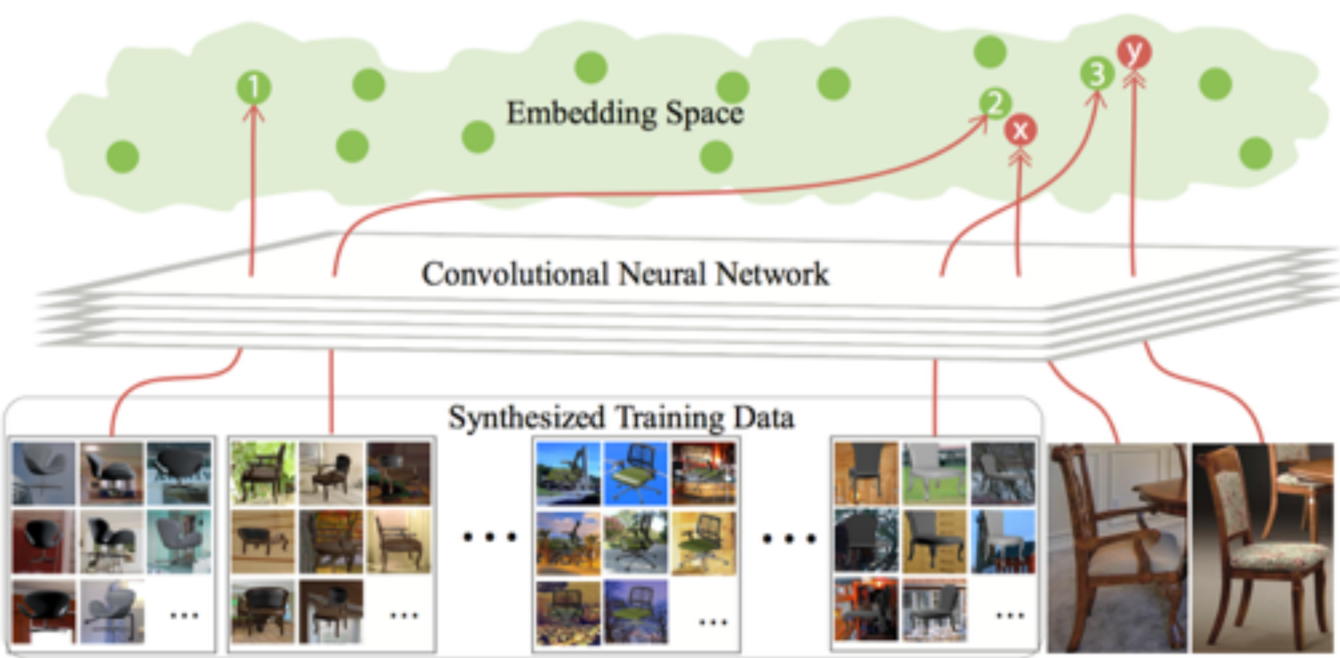
Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the 'Wild'

- Learn an embedding of shapes

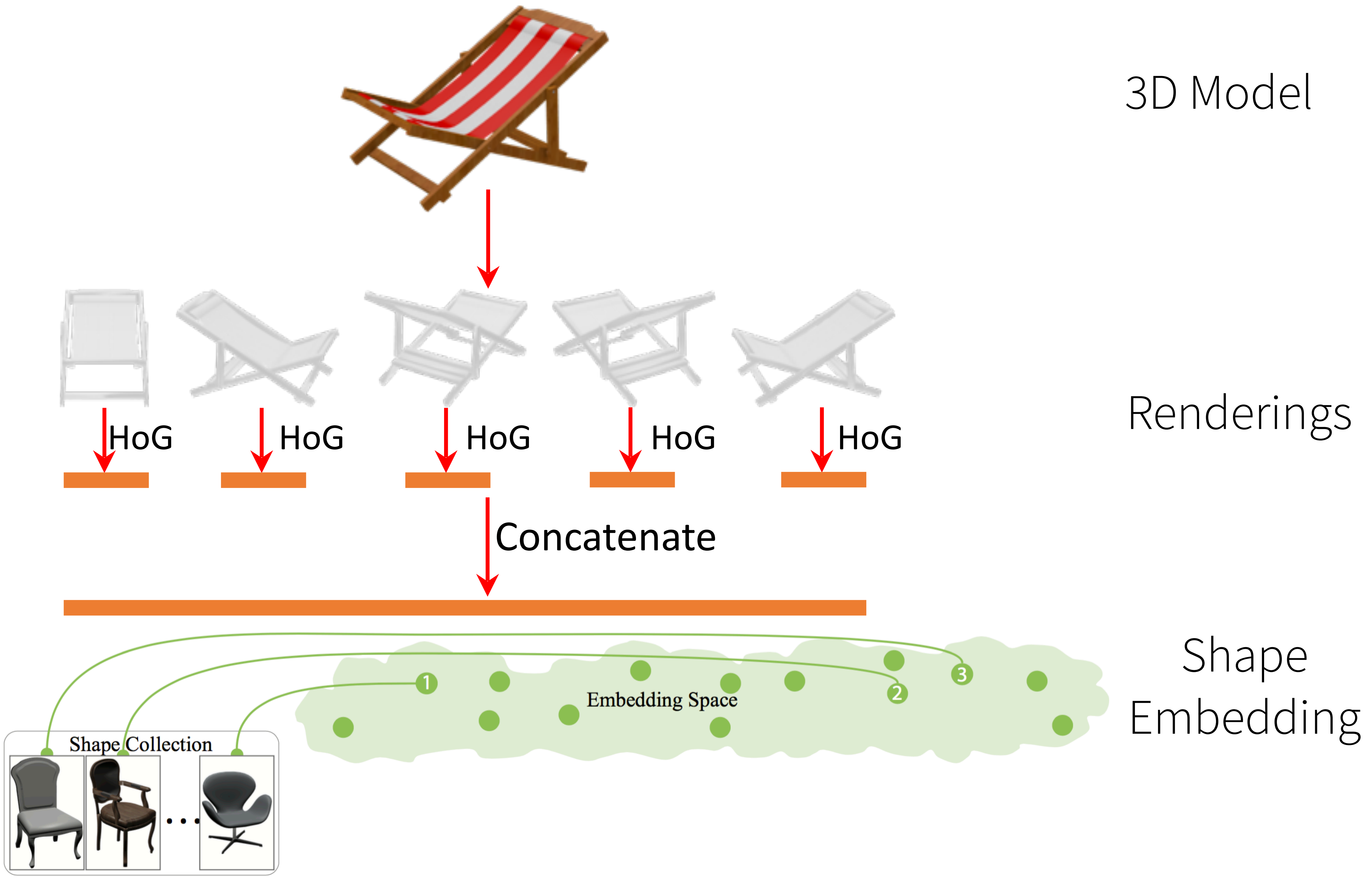


- Populate images in embedding space



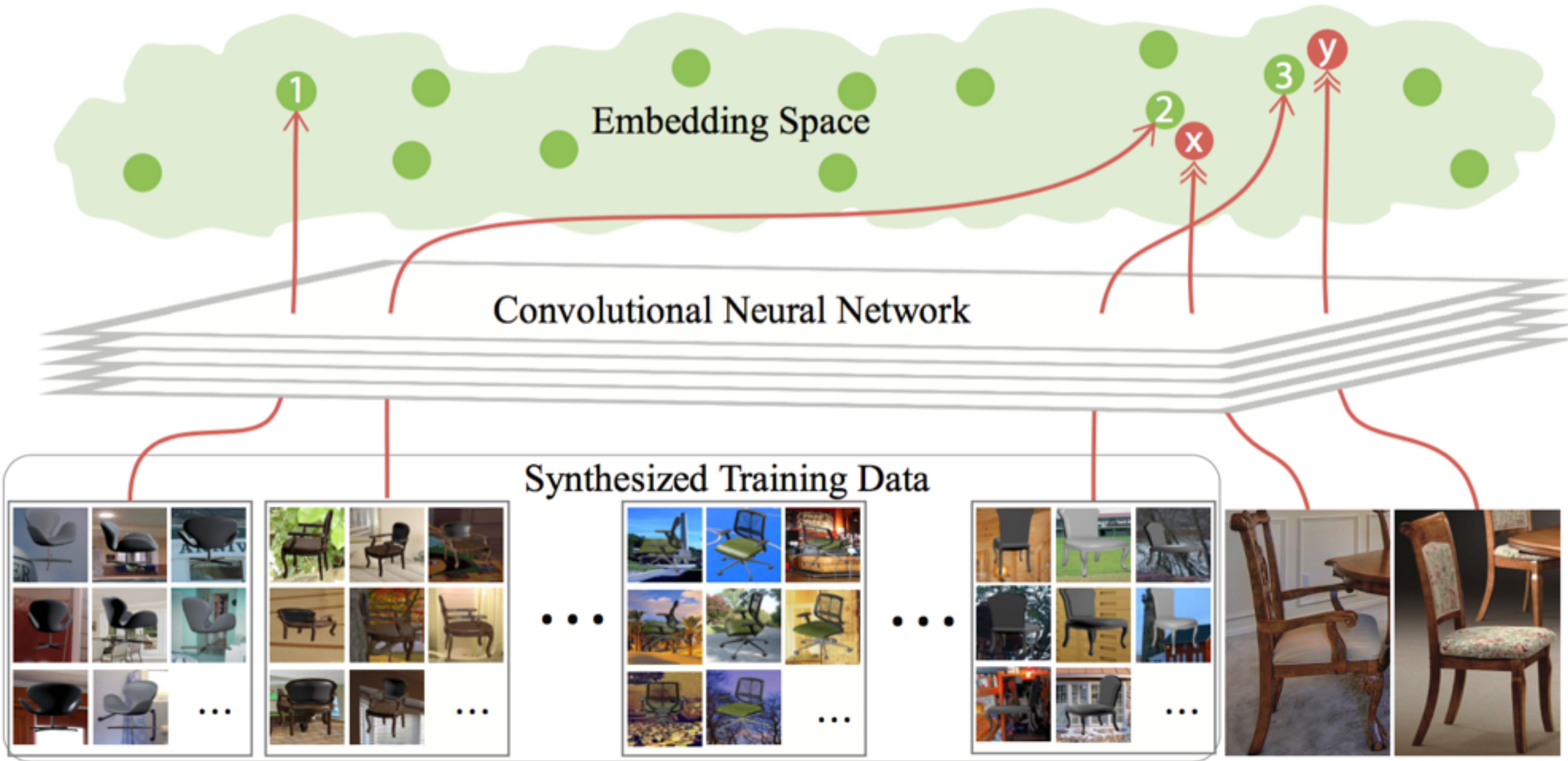
Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the 'Wild'



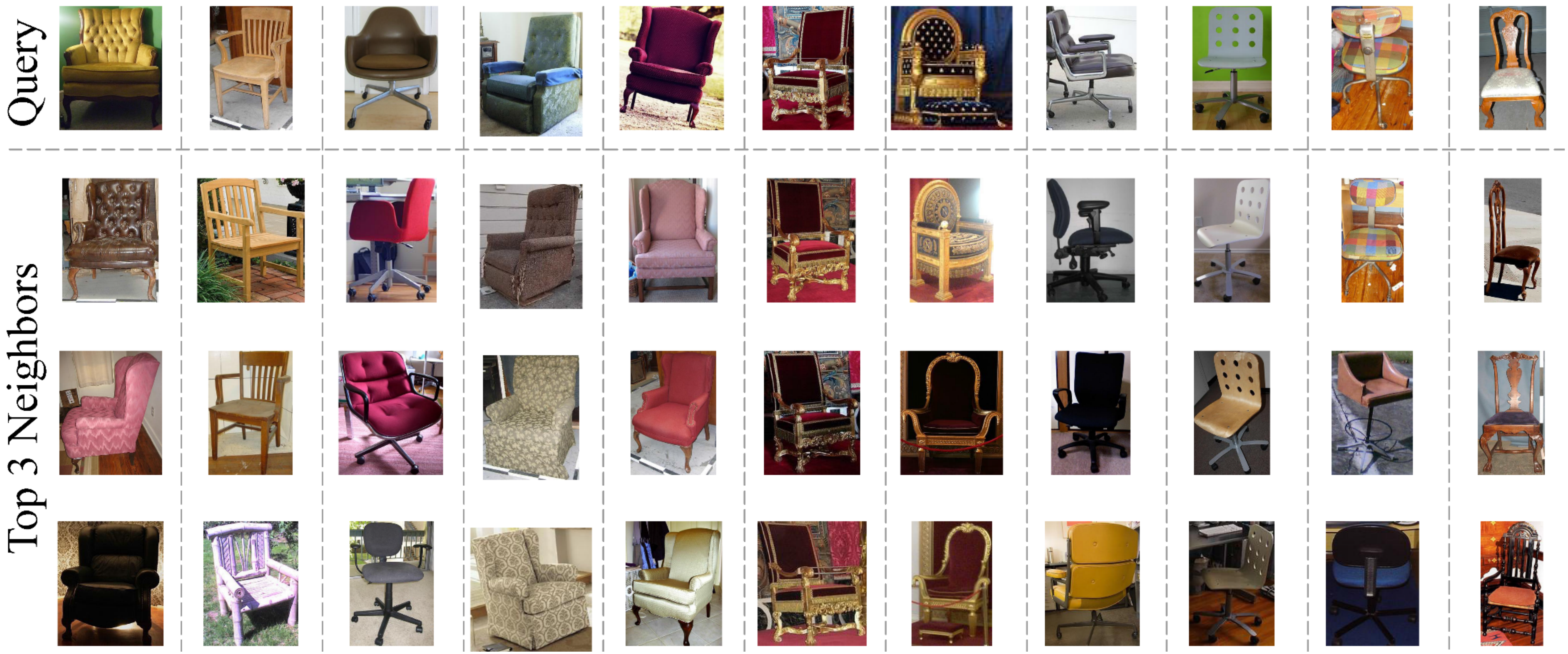
Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the 'Wild'



Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the 'Wild'



Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the 'Wild'

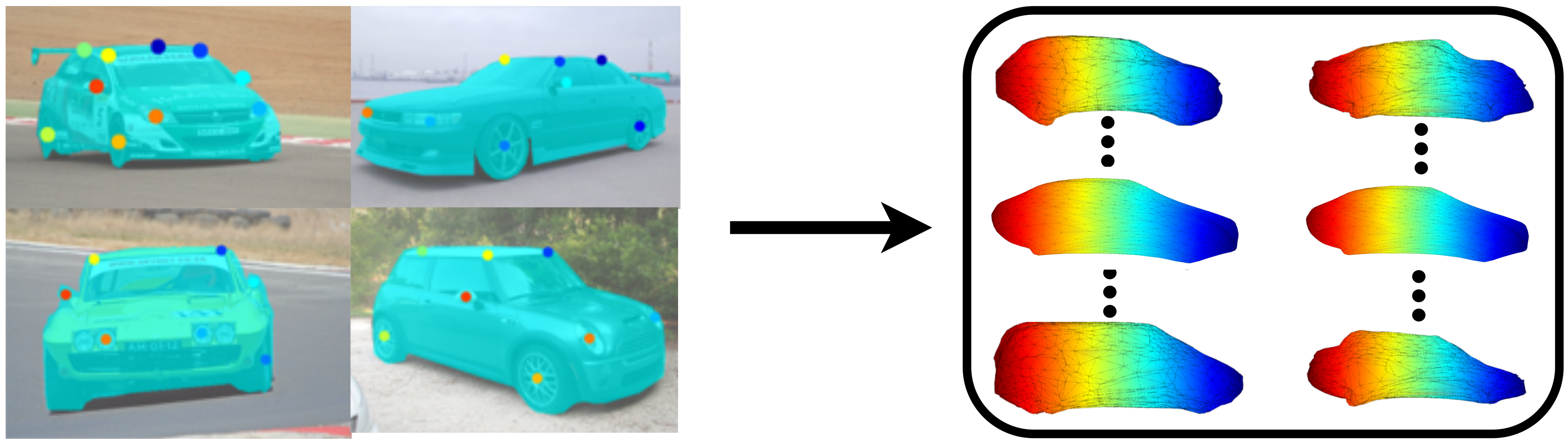
Joint Embeddings of Shapes and Images via CNN Image Purification
Li, Su, Qi, Fish, Cohen-Or, Guibas

Reconstruction in the ‘Wild’

Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the 'Wild'

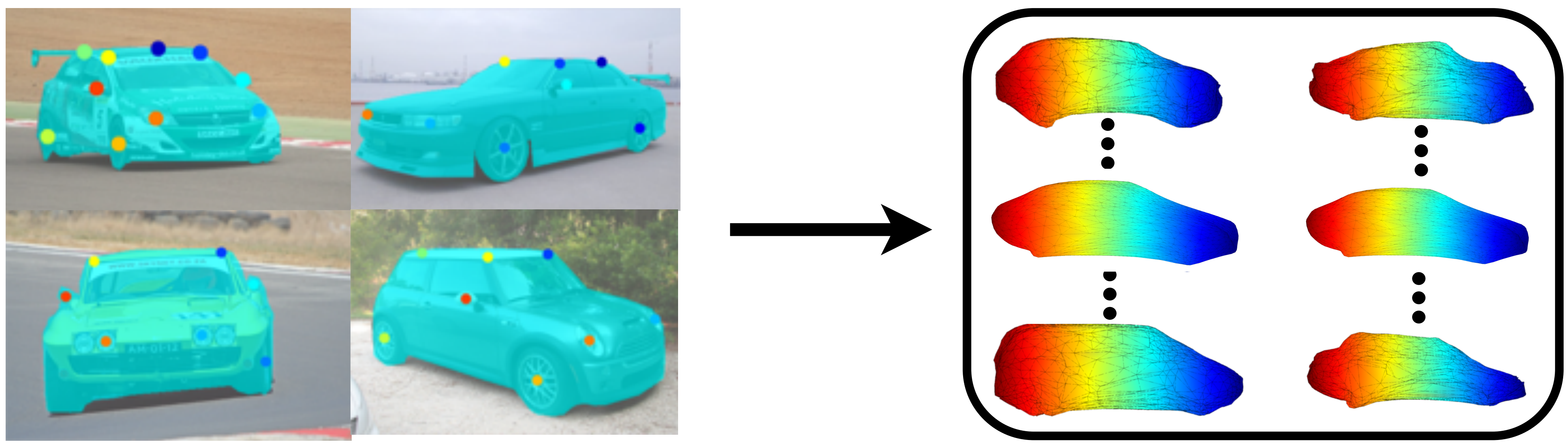
- Learn category specific 3D models from 2D images of objects



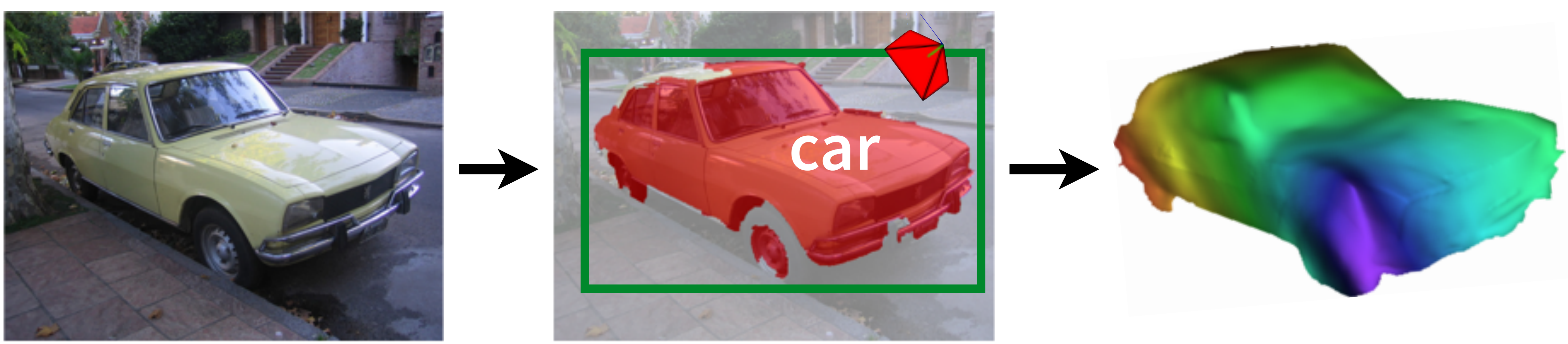
Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the 'Wild'

- Learn category specific 3D models from 2D images of objects



- 3D reconstruction of objects from a single image of a scene



Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the ‘Wild’

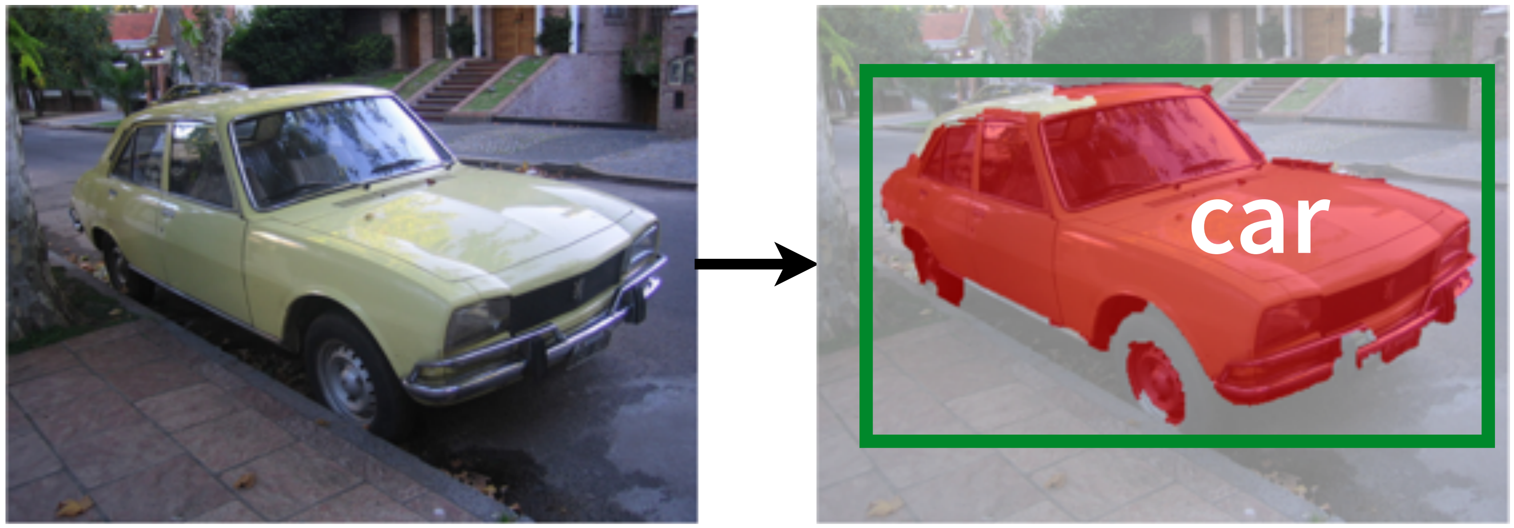
Image

Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the 'Wild'

Image

Object Detection and
Instance Segmentation



Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the ‘Wild’

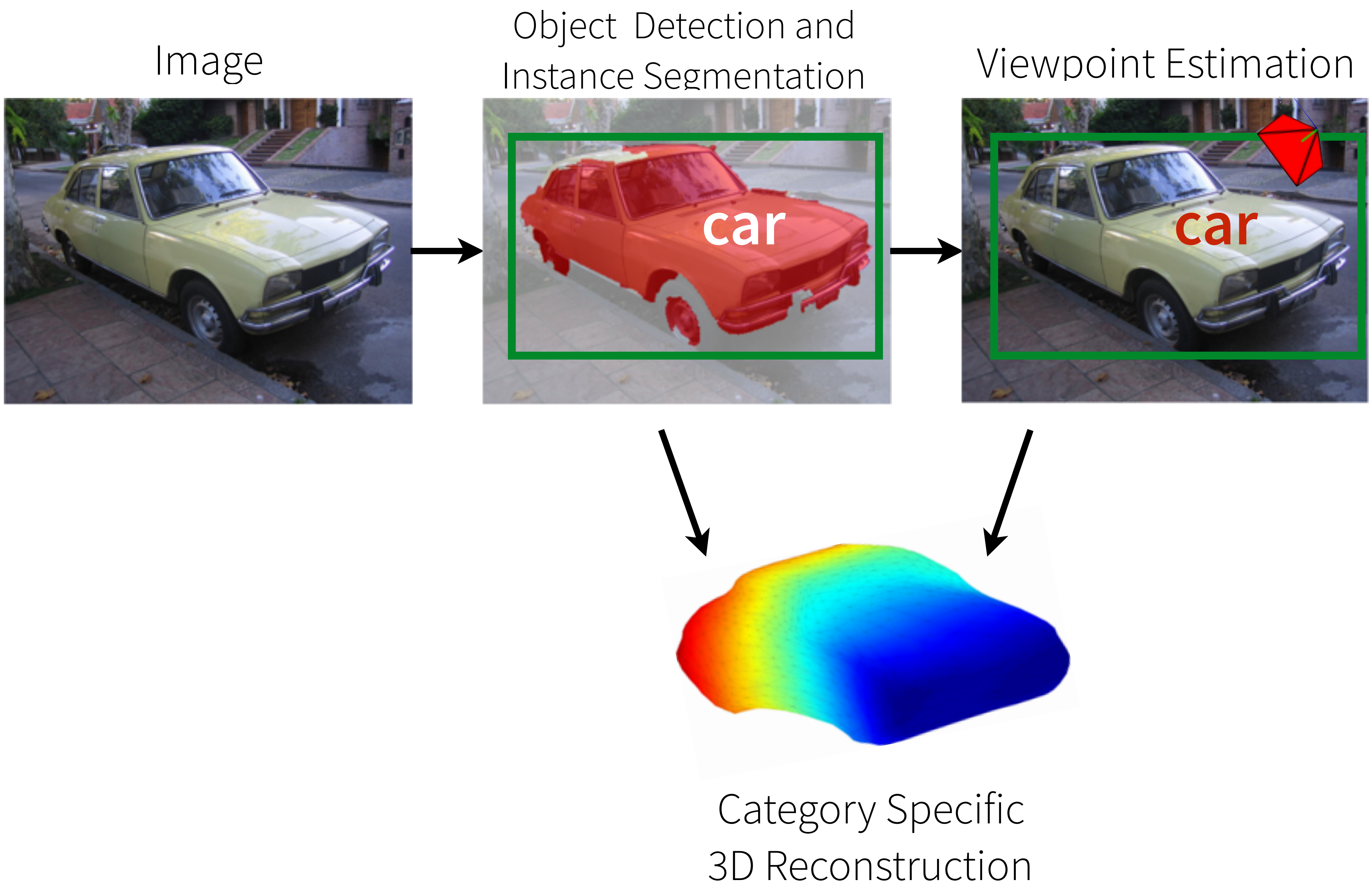
Image

Object Detection and Instance Segmentation

Viewpoint Estimation

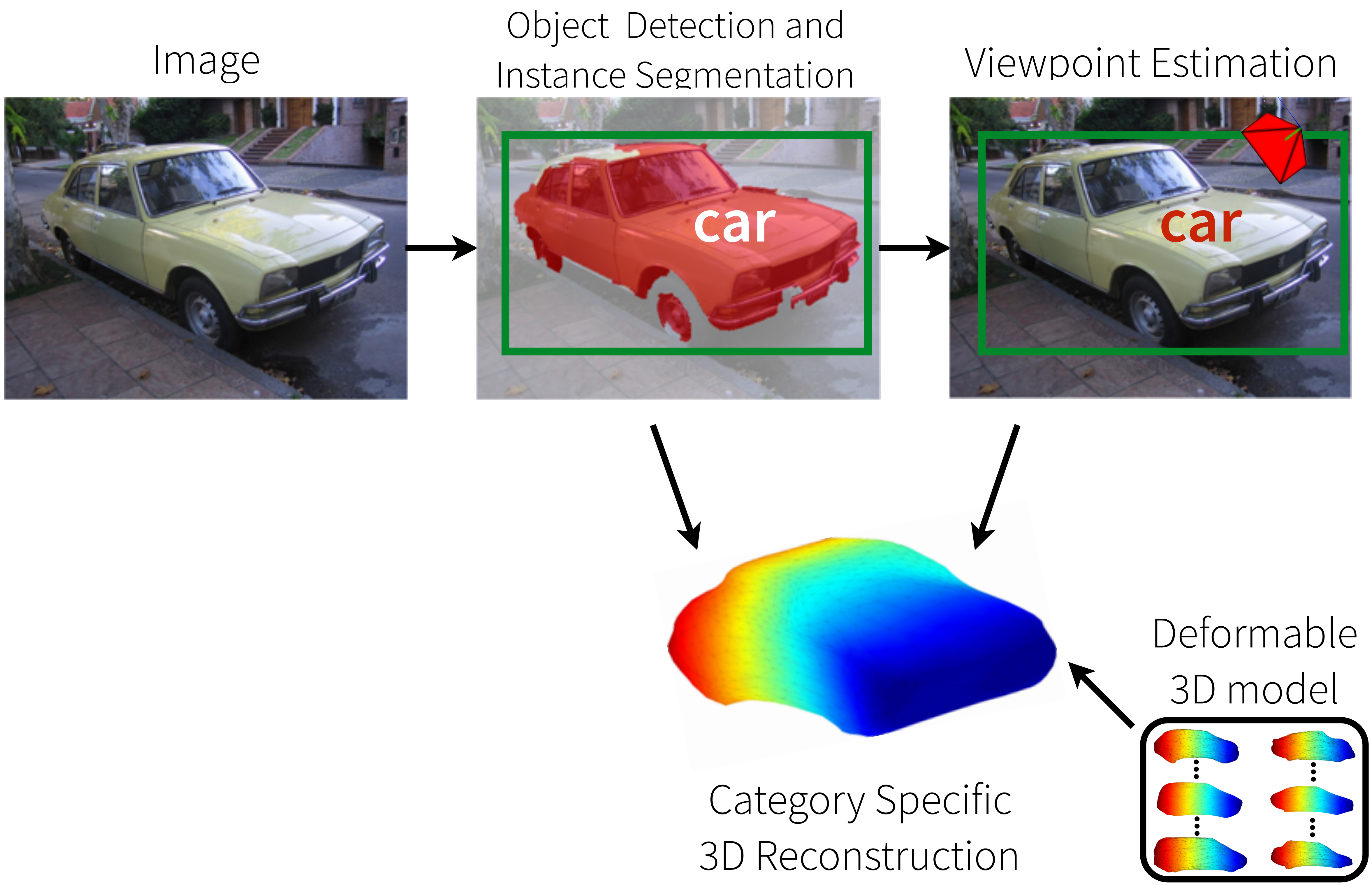
Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the 'Wild'



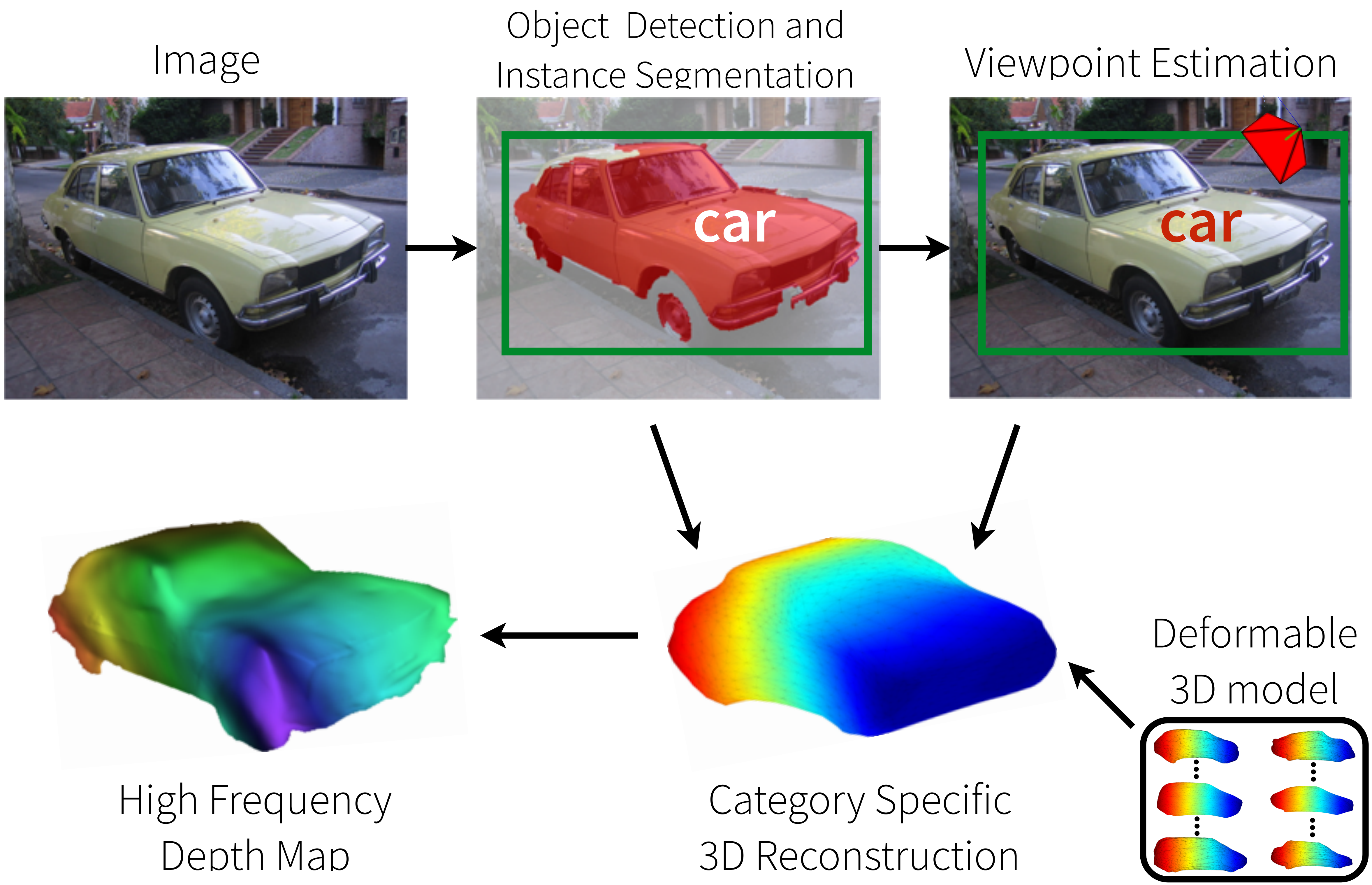
Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the 'Wild'



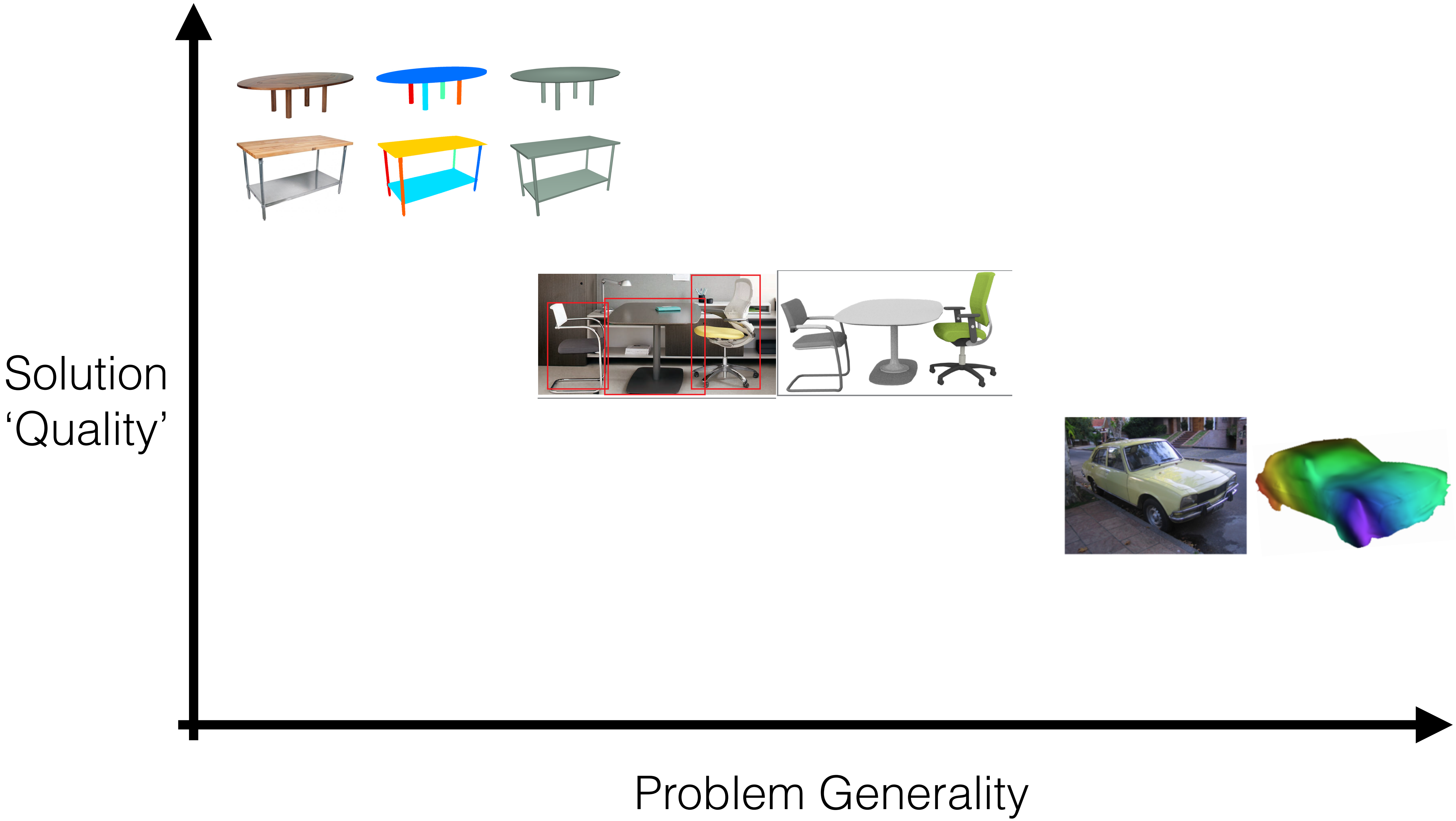
Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Reconstruction in the 'Wild'



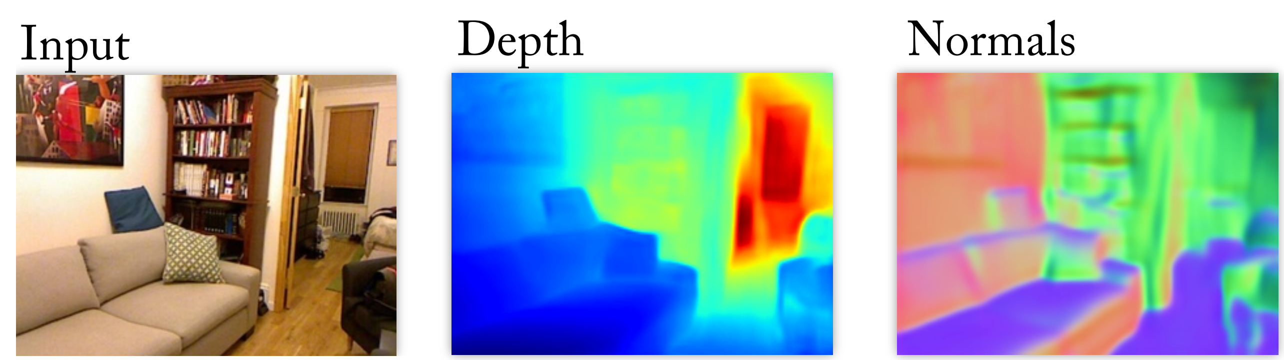
Category-Specific Object Reconstruction from a Single Image
Kar, Tulsiani, Carreira, Malik

Objects in 3D



3D Visual Understanding

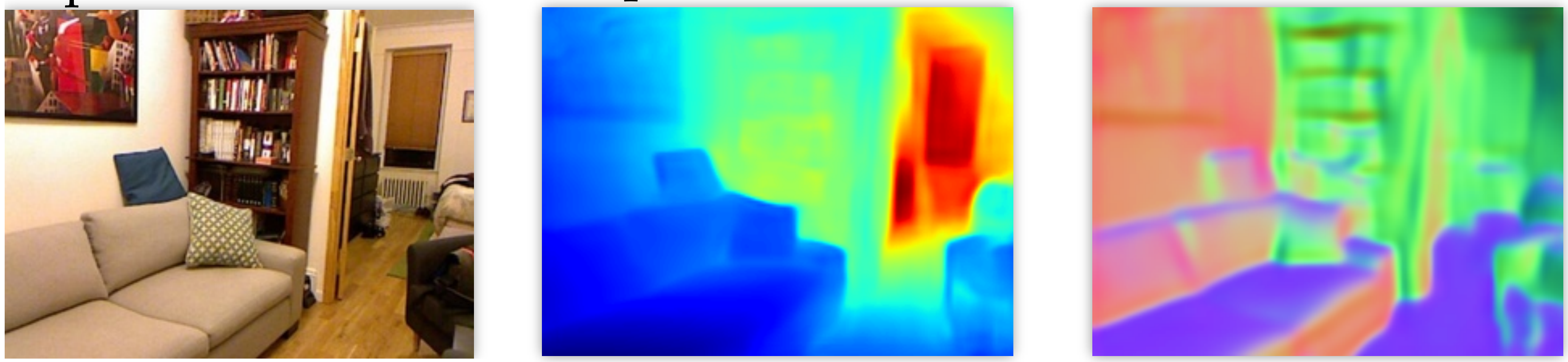
- Background
- Objects in 3D
- **Scenes in 3D**
- 3D Understanding without Understanding 3D
- Open Problems



Depth and Normal Estimation

Scenes in 3D

Input Depth Normals

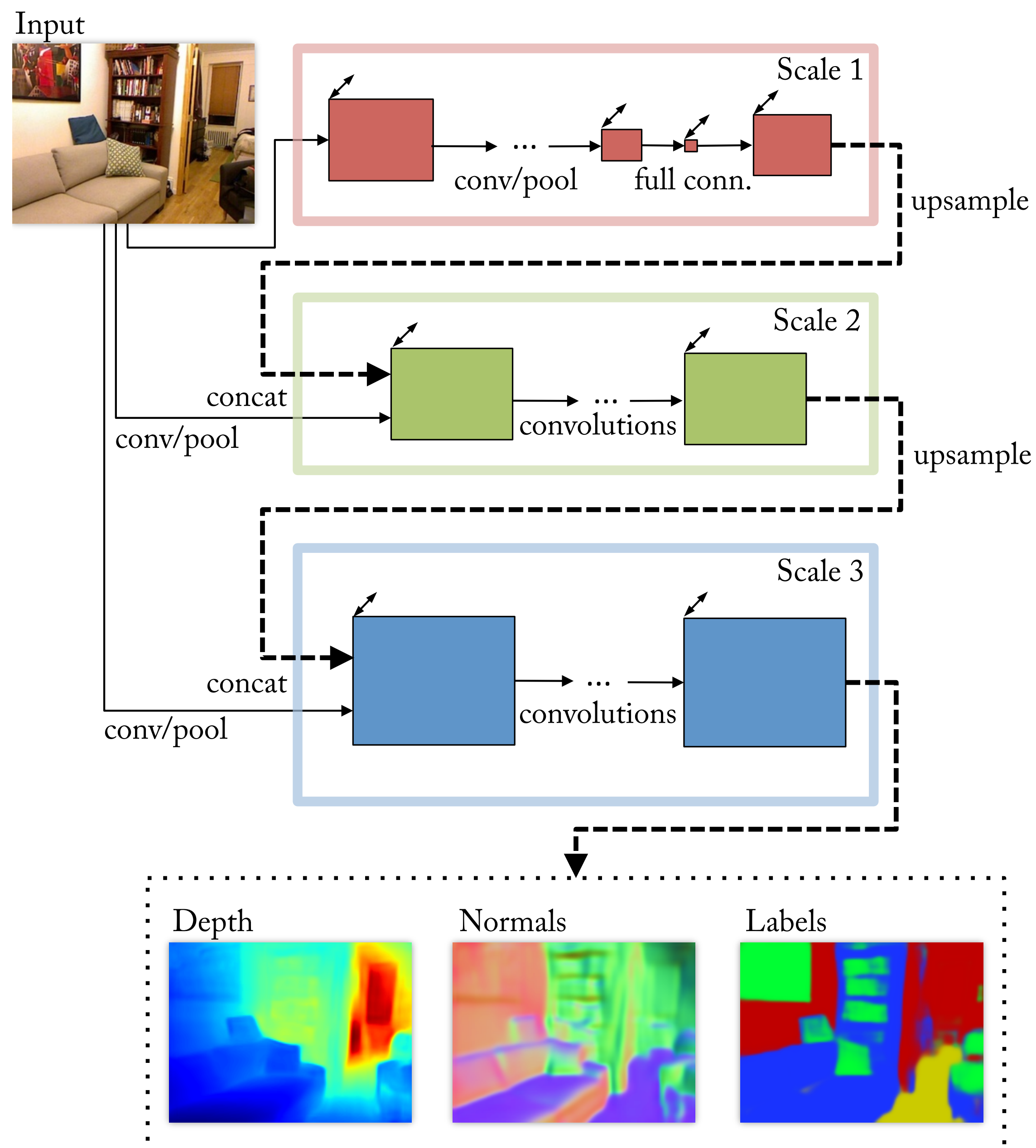


Dataset
(NYU Depth Dataset)

Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture

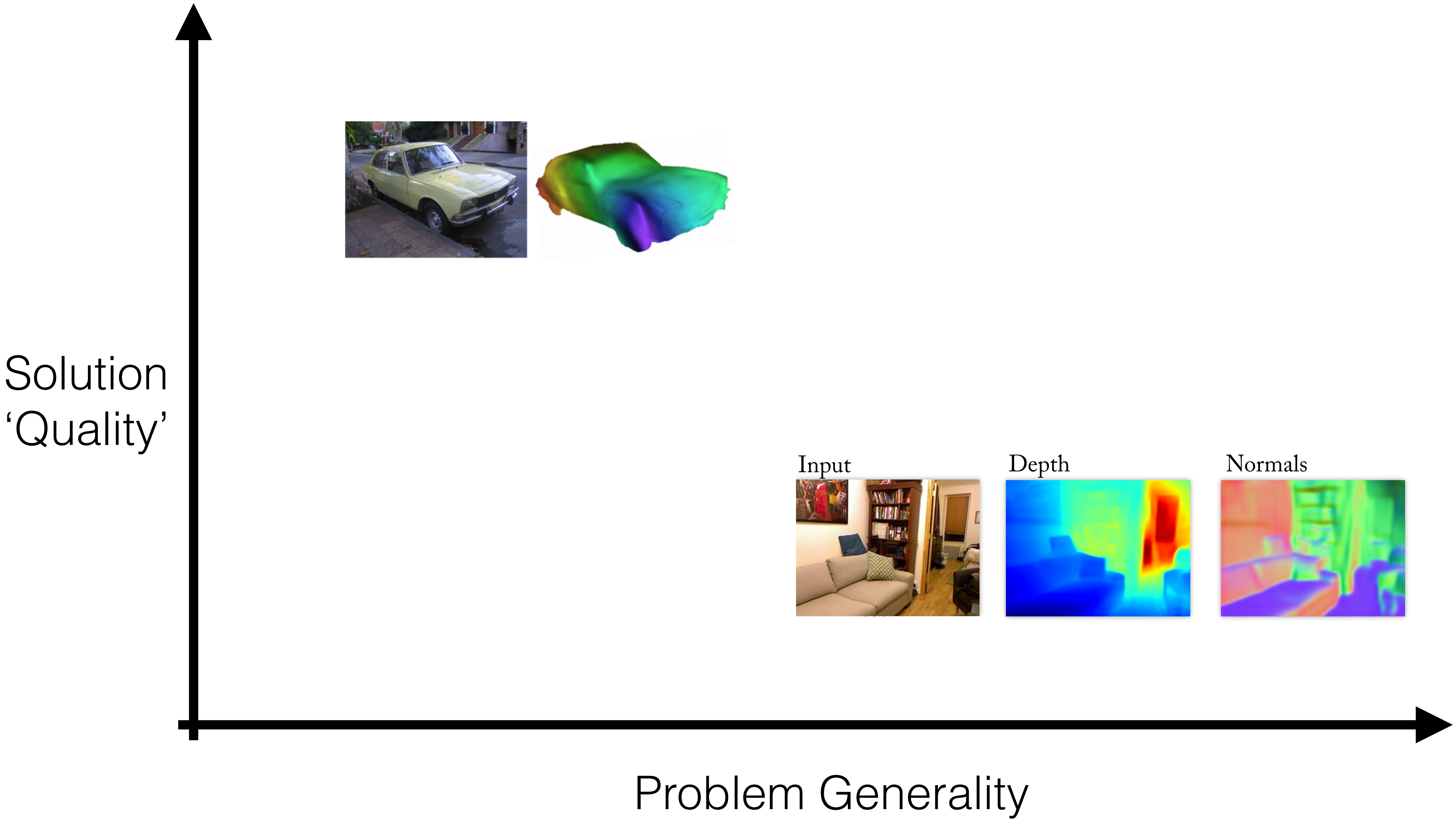
Eigen, Fergus

Scene Reconstruction



Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture

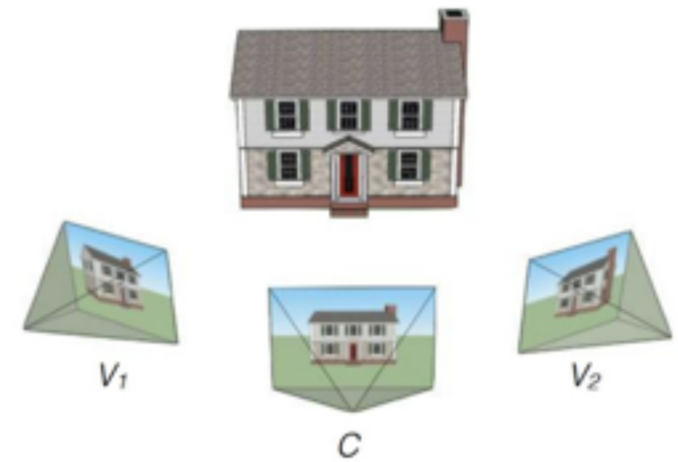
Eigen, Fergus



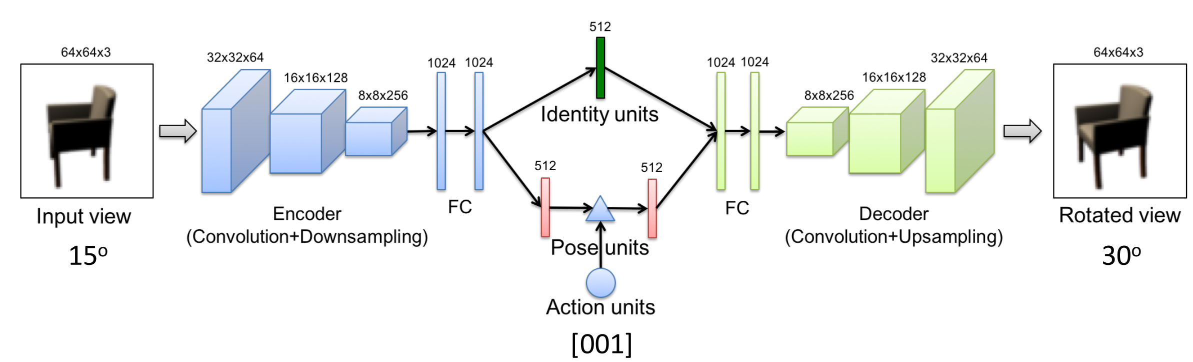
Solution
'Quality'
Input
Depth
Normals
Problem Generality

3D Visual Understanding

- Background
- Objects in 3D
- Scenes in 3D
- **3D Understanding without Understanding 3D**
- Open Problems



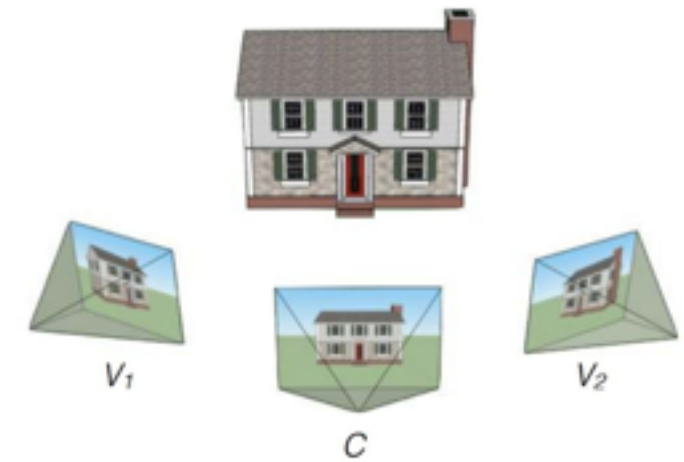
View Interpolation



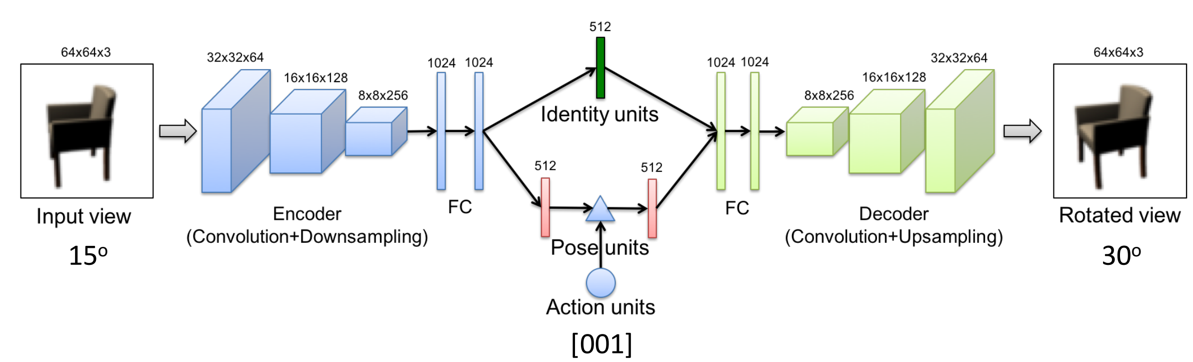
View Synthesis

3D Visual Understanding

- Background
- Objects in 3D
- Scenes in 3D
- **3D Understanding without Understanding 3D**
- Open Problems



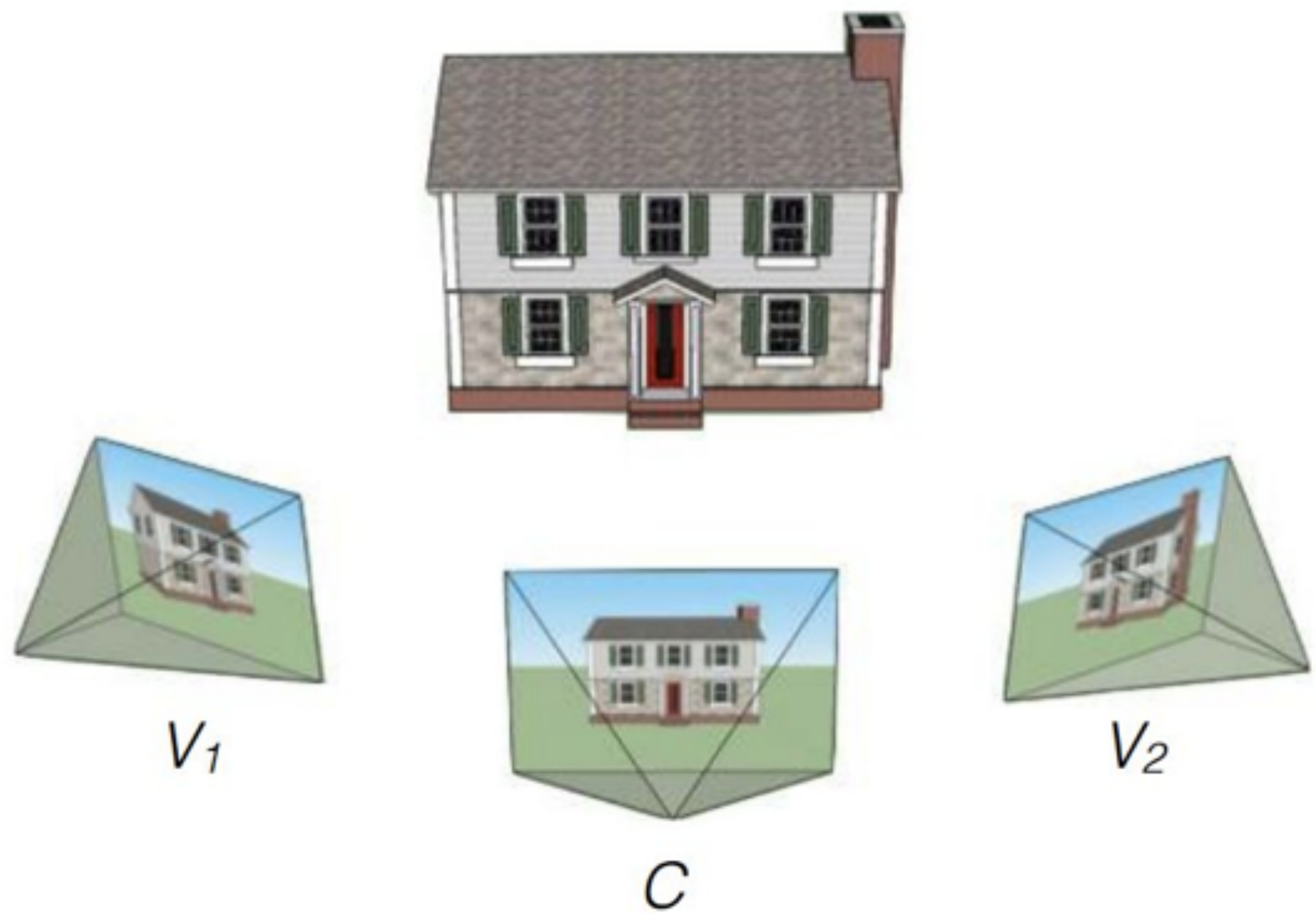
View Interpolation



View Synthesis

View Synthesis

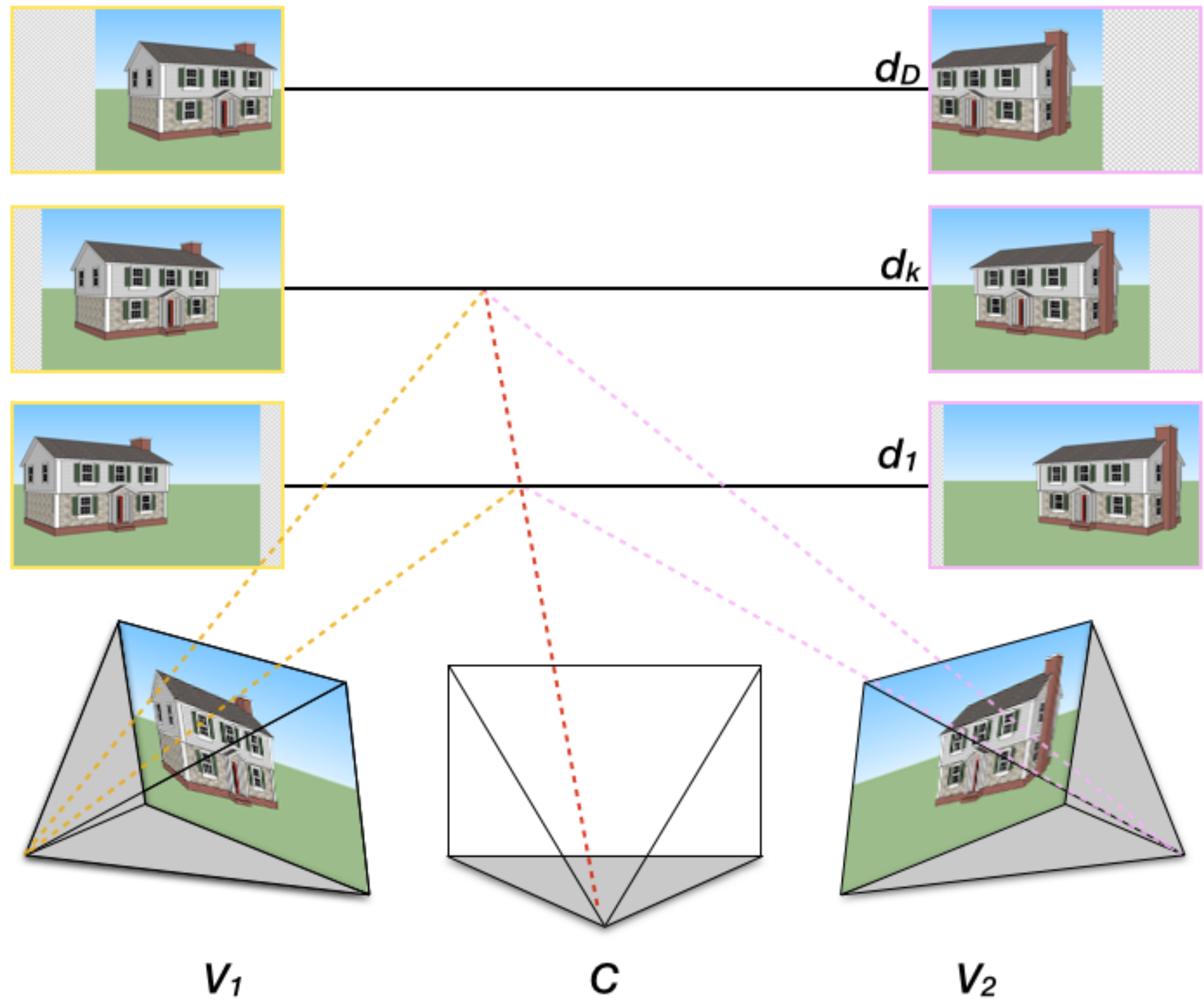
Given two views, we want the image corresponding to the middle view



DeepStereo: Learning to Predict New Views from the World's Imagery
Flynn, Neulander, Philbin, Snavely

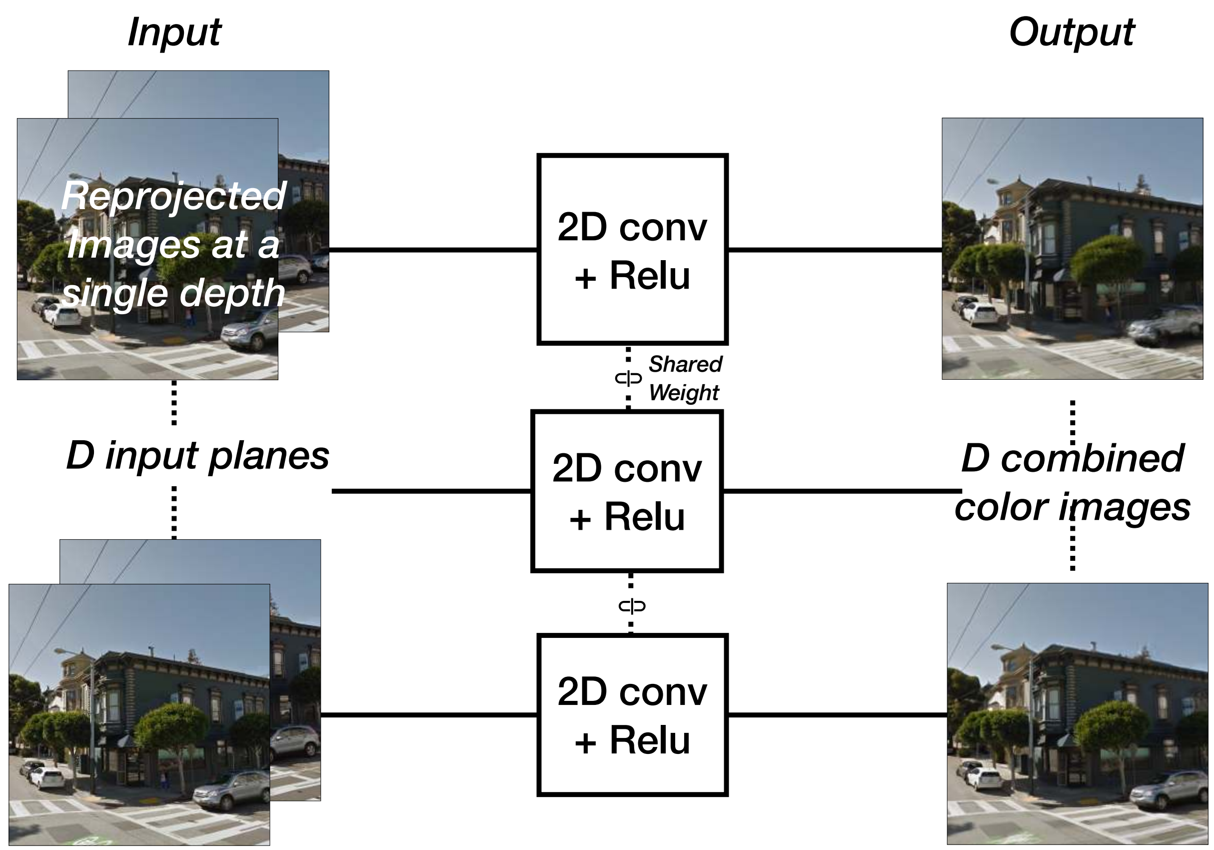
View Synthesis

Images if all pixels were at same depth



DeepStereo: Learning to Predict New Views from the World's Imagery
Flynn, Neulander, Philbin, Snavely

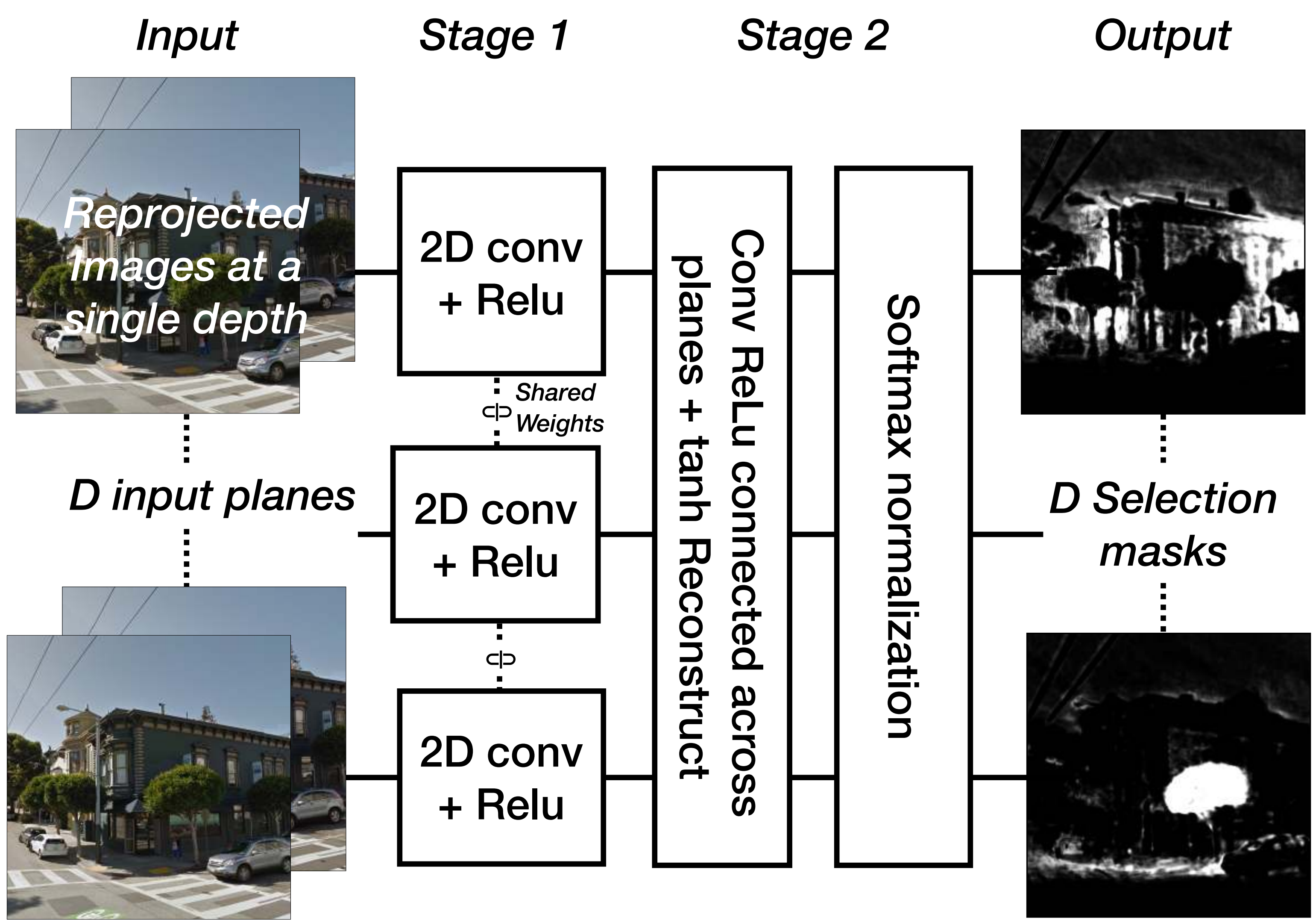
View Synthesis



‘Color CNN’ combines images from left and right cameras at each level

DeepStereo: Learning to Predict New Views from the World's Imagery
Flynn, Neulander, Philbin, Snavely

View Synthesis



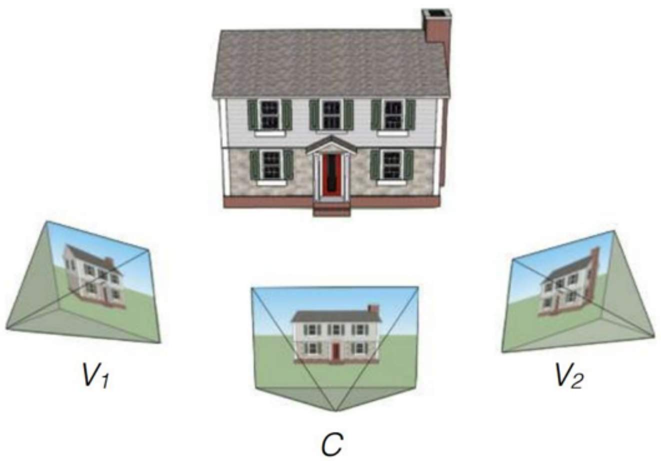
'Selection CNN' chooses which depth level each pixel belongs to.

DeepStereo: Learning to Predict New Views from the World's Imagery
Flynn, Neulander, Philbin, Snavely

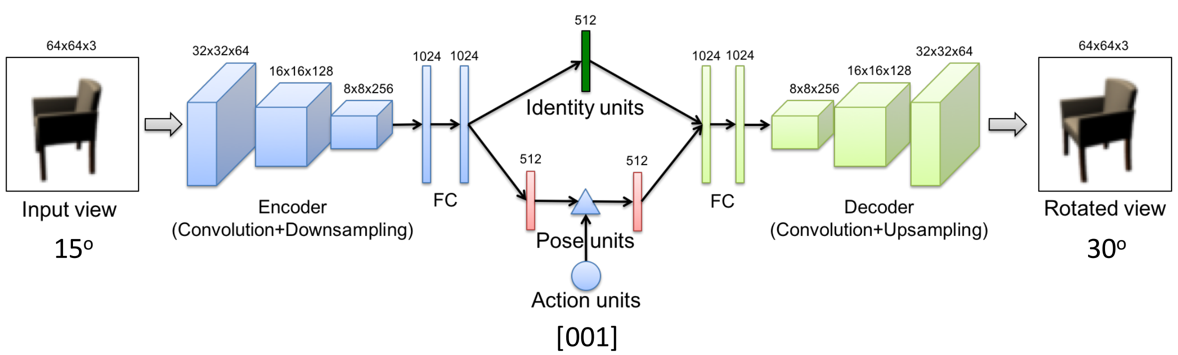
View Synthesis

3D Visual Understanding

- Background
- Objects in 3D
- Scenes in 3D
- **3D Understanding without Understanding 3D**
- Open Problems



View Interpolation



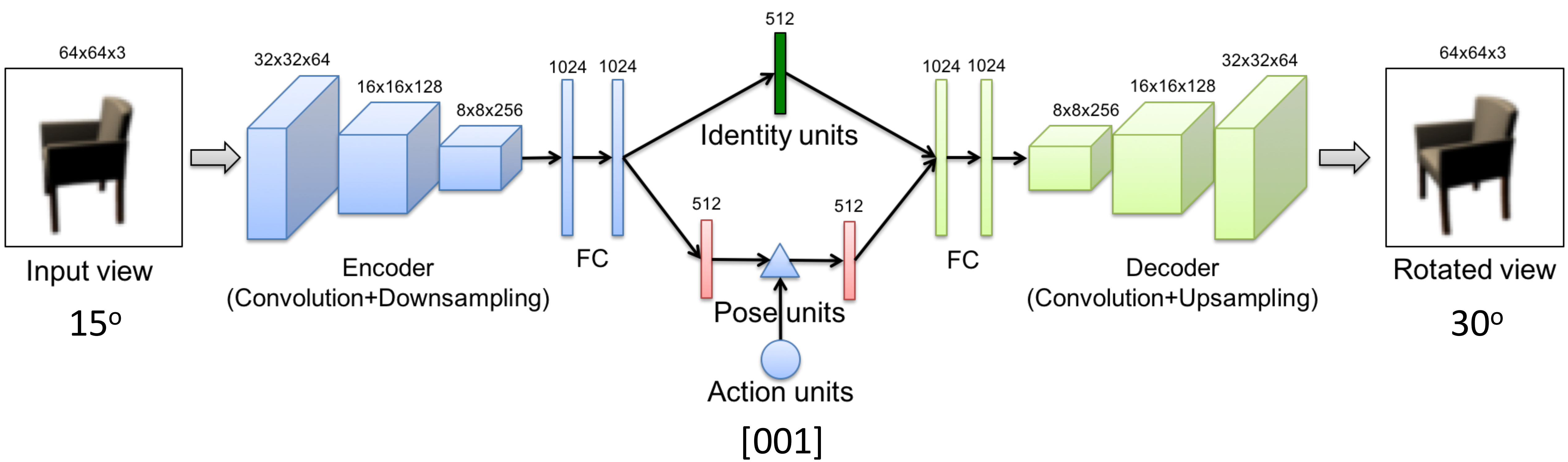
View Extrapolation

View Synthesis

How does the object look from a different view ?

Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis
Yang, Reed, Yang, Lee

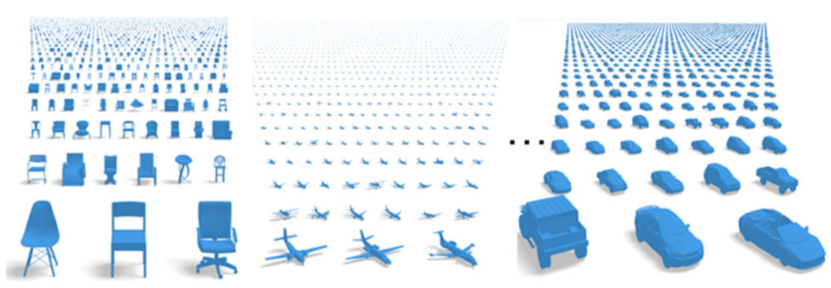
View Synthesis



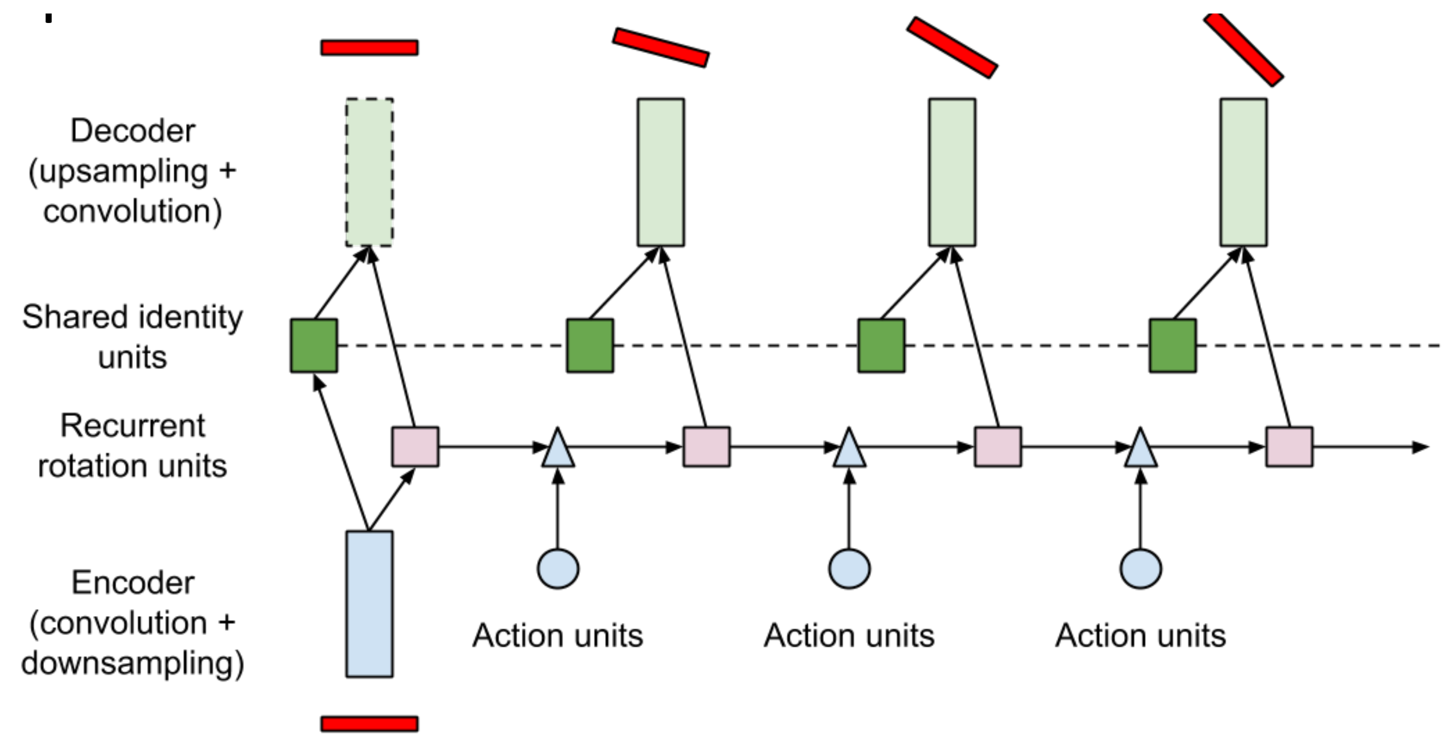
Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis
Yang, Reed, Yang, Lee

View Synthesis

Training Data



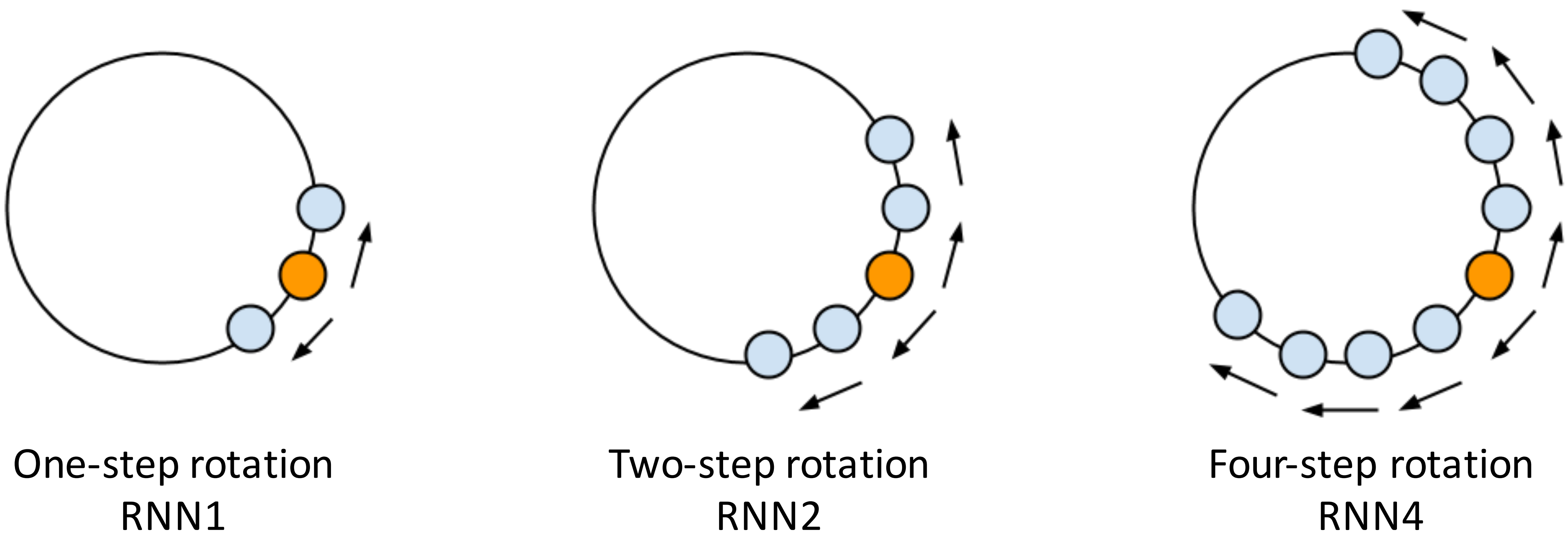
Training



Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis
Yang, Reed, Yang, Lee

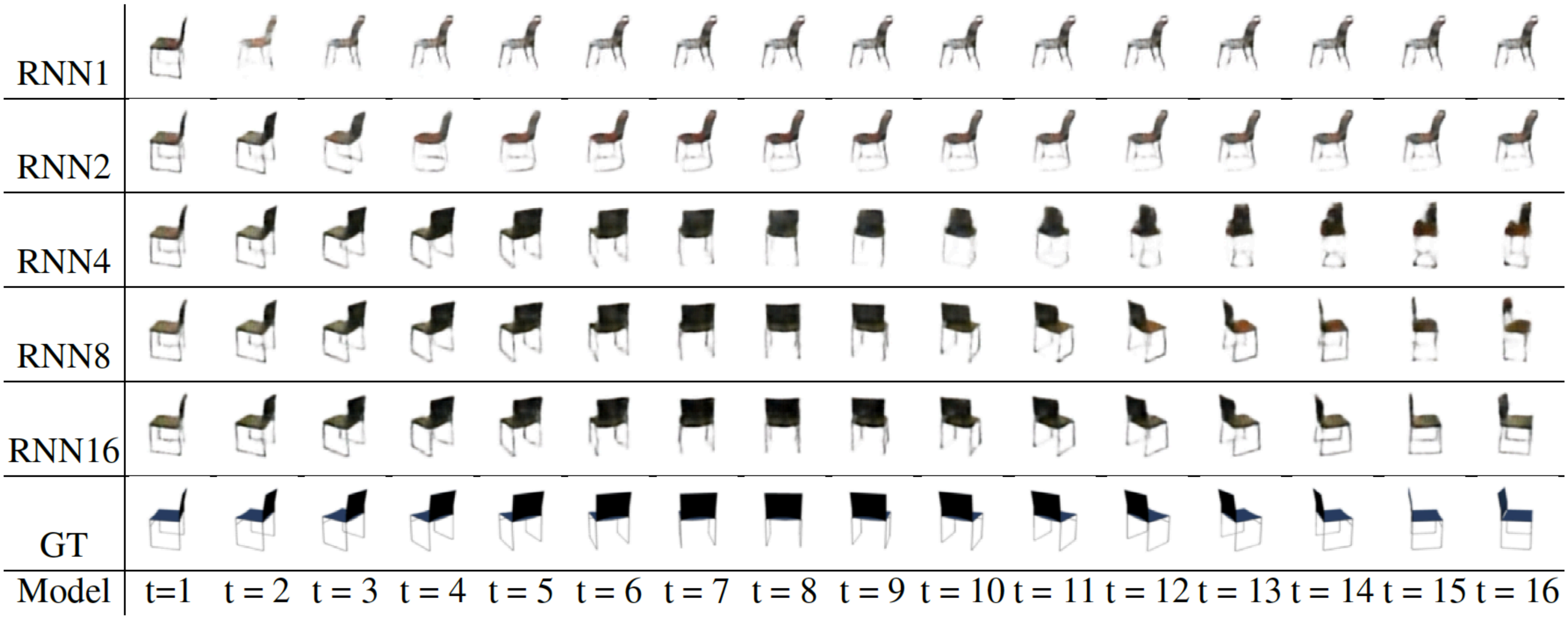
View Synthesis

Curriculum Learning



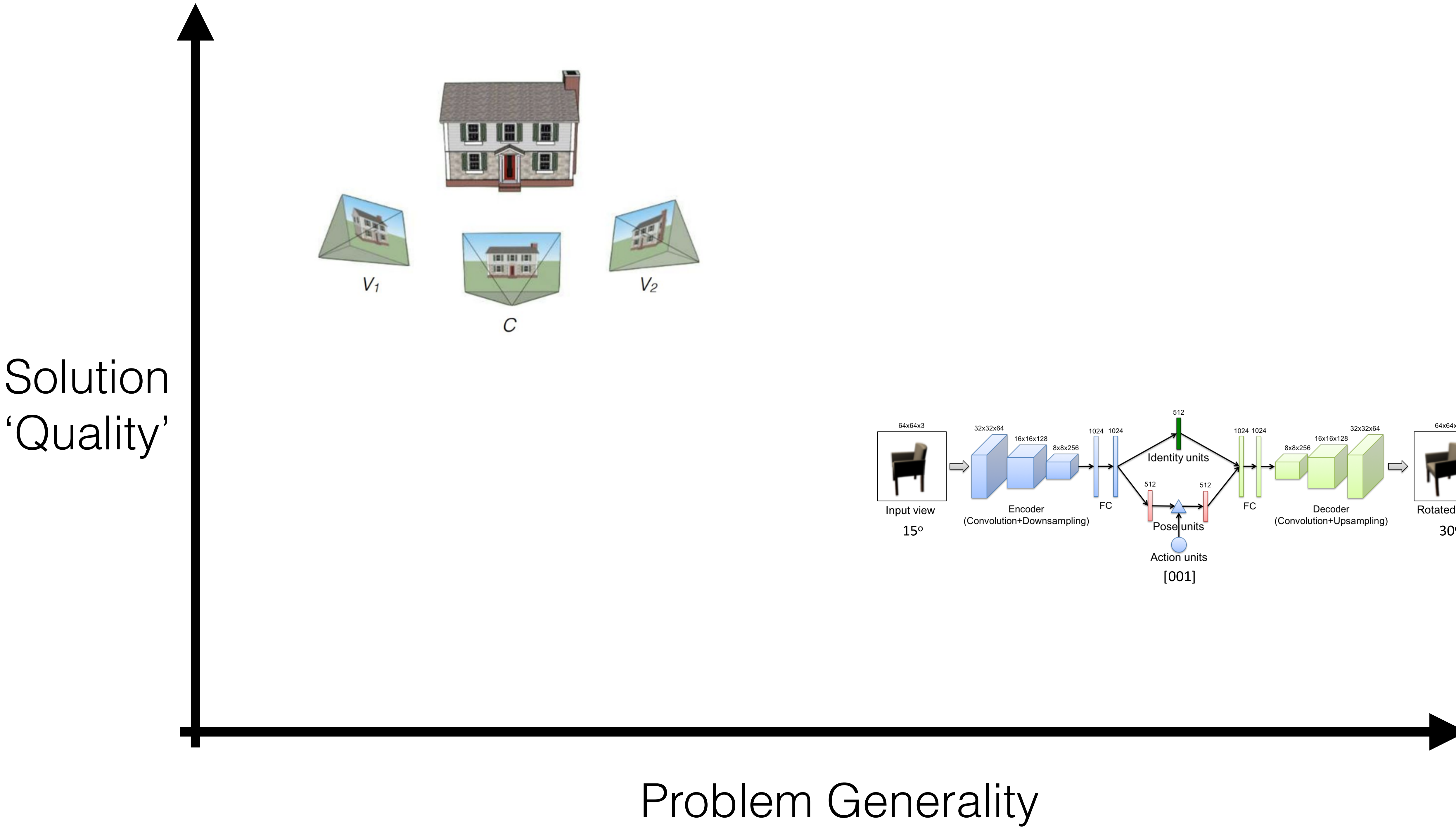
Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis
Yang, Reed, Yang, Lee

View Synthesis



Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis
Yang, Reed, Yang, Lee

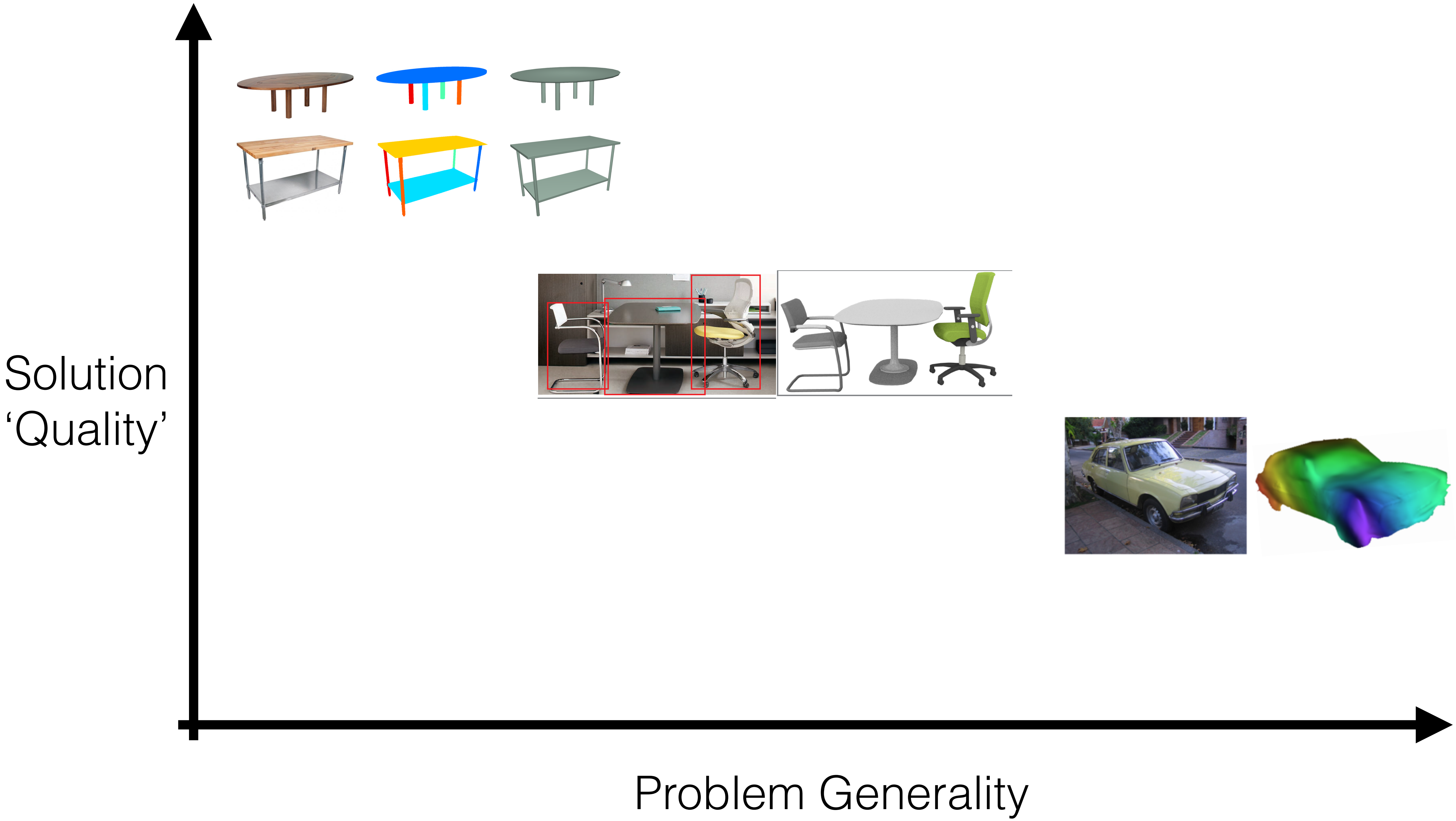
View Synthesis



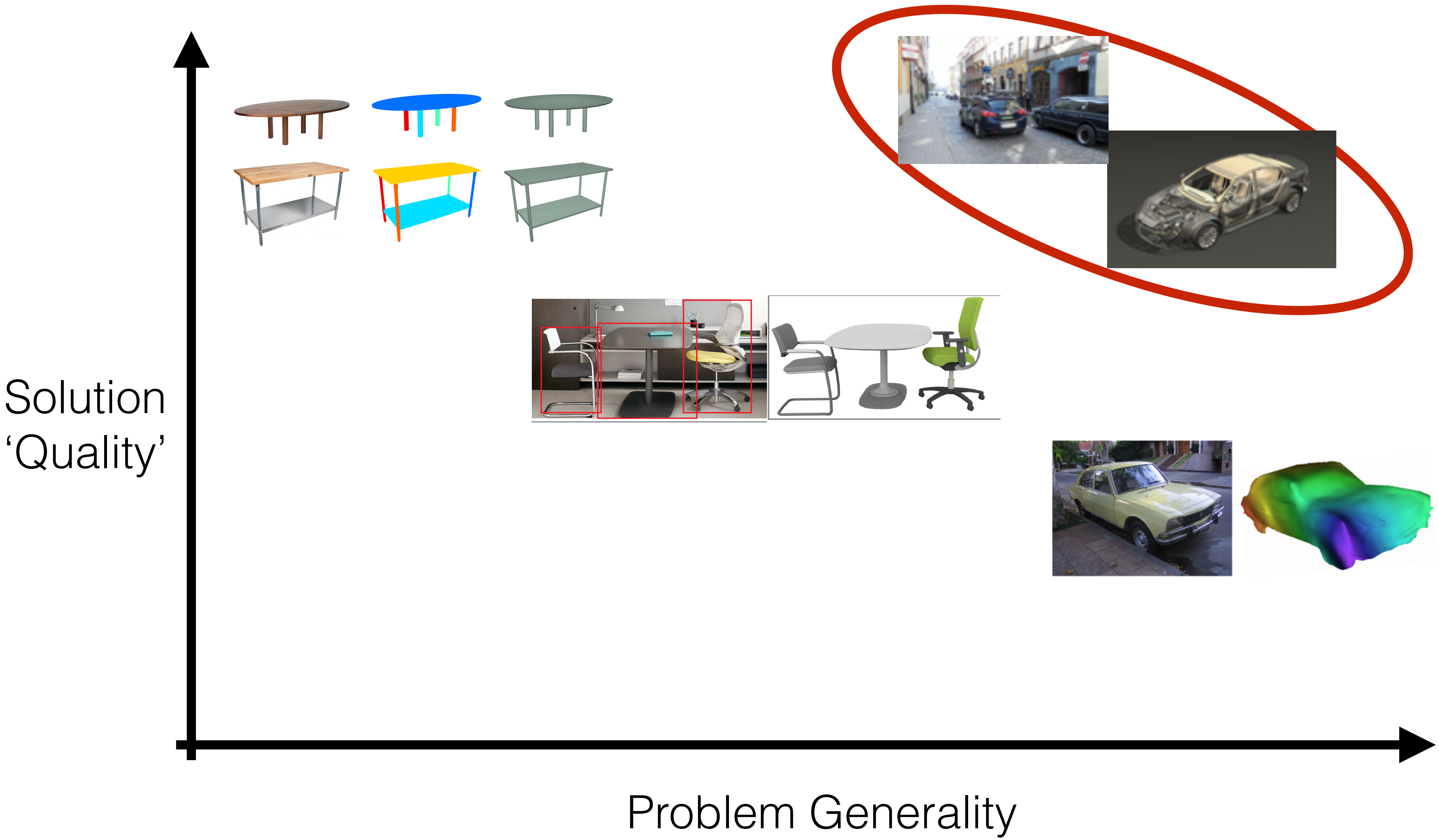
3D Visual Understanding

- Background
- Objects in 3D
- Scenes in 3D
- 3D Understanding without Understanding 3D
- **Open Problems**

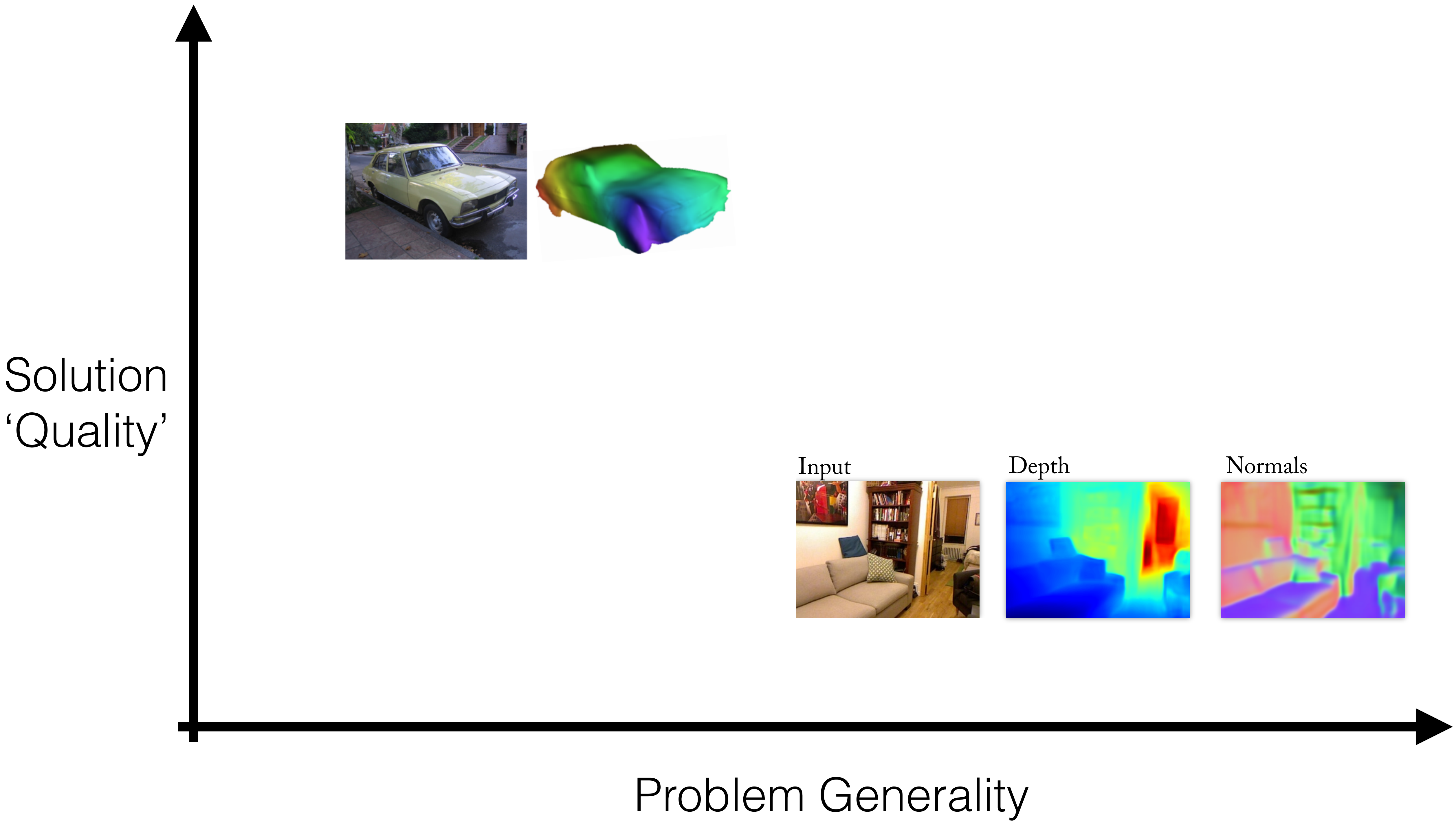
Open Problems



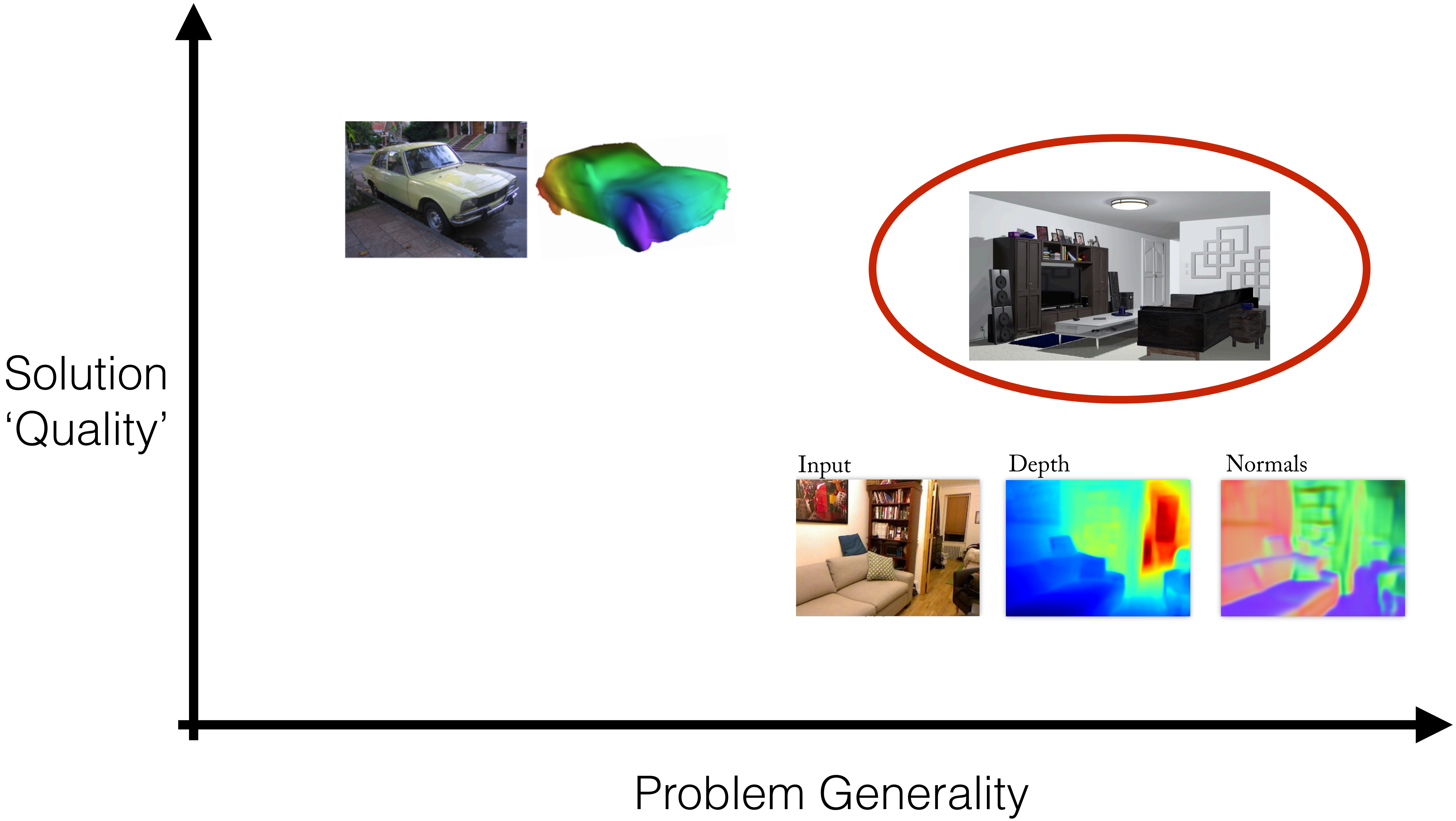
Open Problems



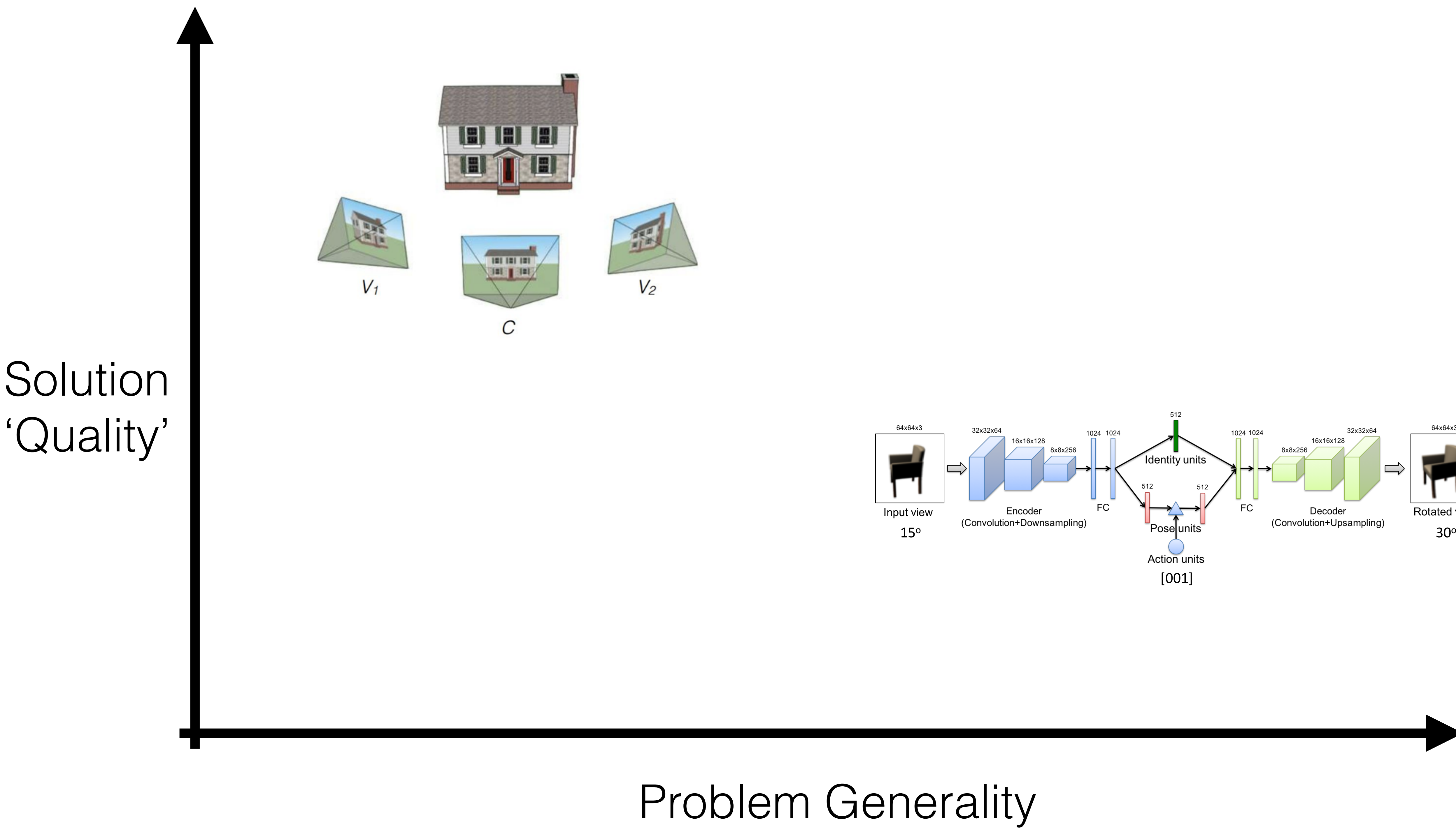
Open Problems



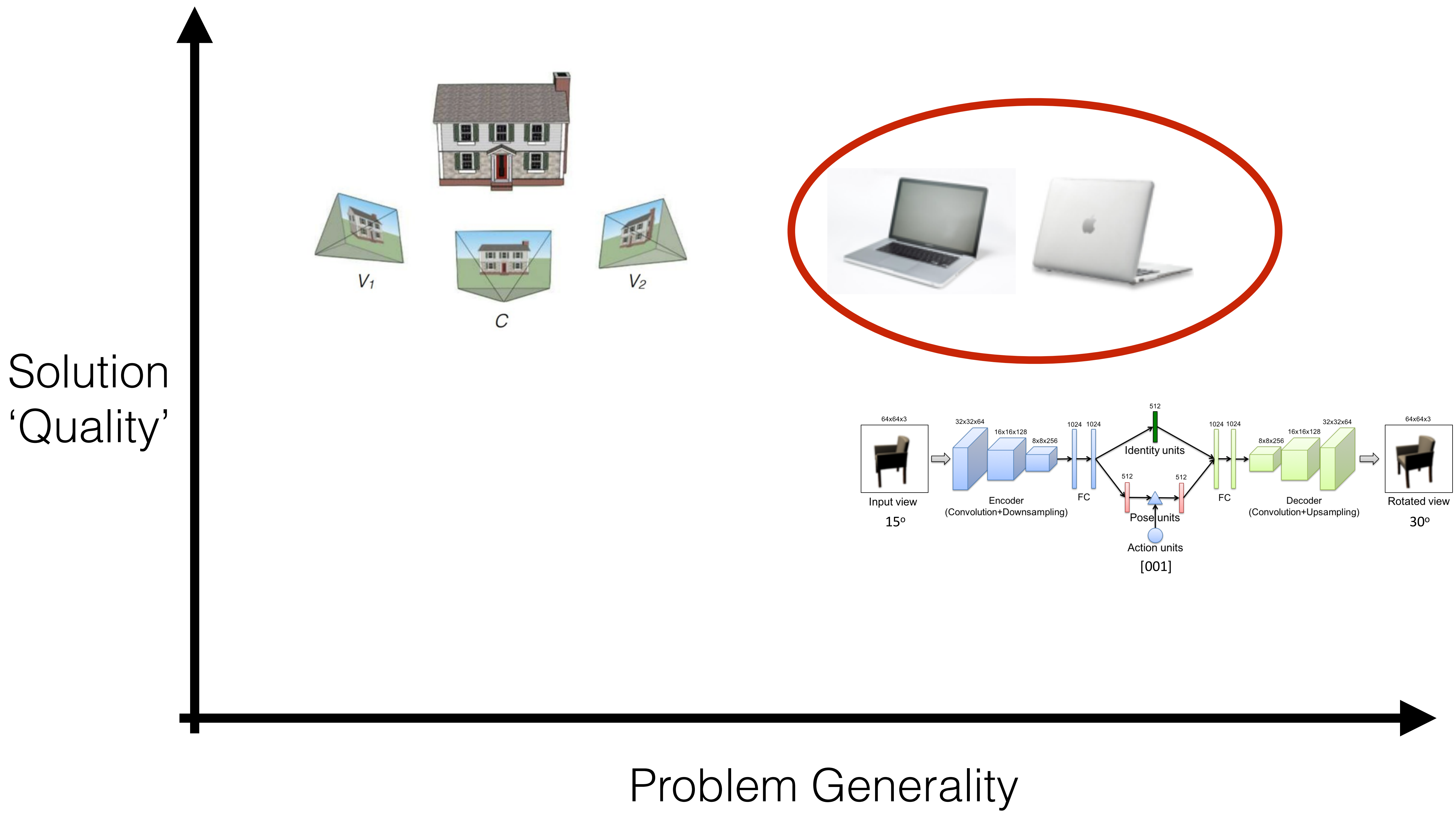
Open Problems



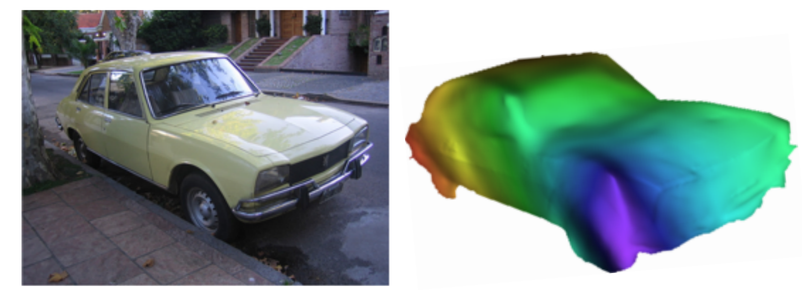
Open Problems



Open Problems



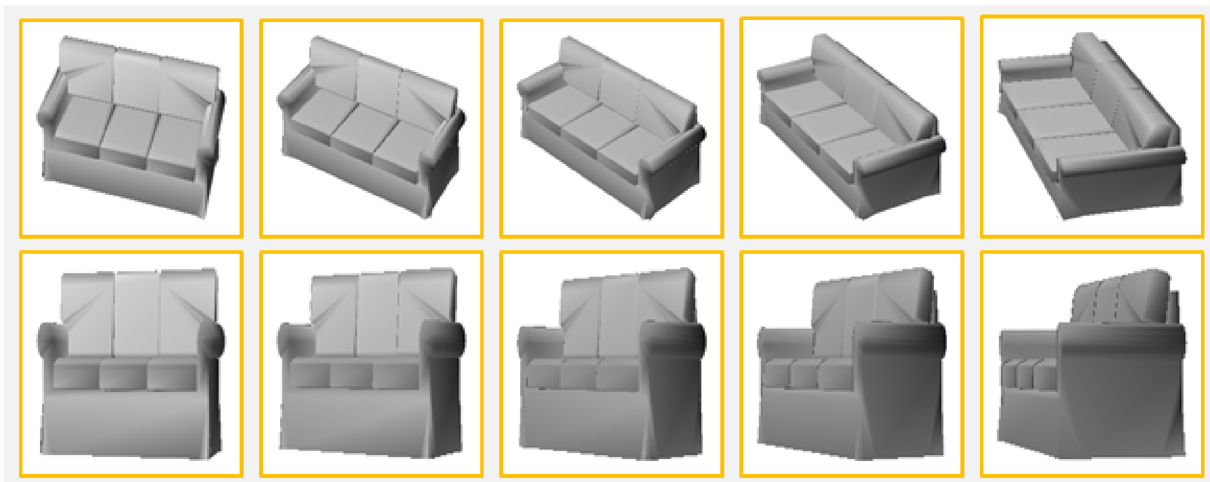
Open Problems : End-to-End Reconstruction



- All methods presented have hand-coded intermediate representations

- The lessons from recent successes of deep learning indicate we might want to instead learn these

Open Problems : Domain Gap



- We don't have real-image annotations for everything (symmetries, part-labels) but we have 3D models
- How can we ensure CNNs trained on synthetic data work on real images ?

Open Problems : Novel Objects

- There are more than 10,000 object categories. How can we learn to make meaningful predictions even on new objects ?

Thank You