And now for something completely different!

Learning-Based 3D

Goal

- I’d like to answer: what is computer vision and where is it headed?
- In the process, I’d like to give you a sense of what computer vision is like.

What is CV?

Get a computer to understand

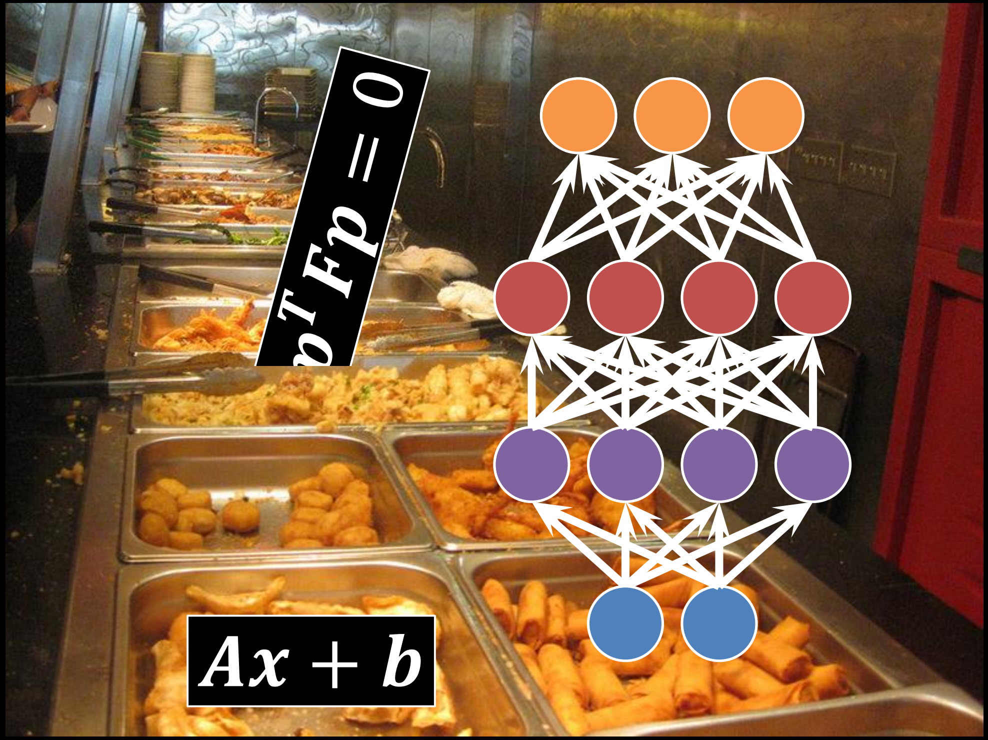
What is CV?

This could incorporate:

- Naming things (recognition)
- Reconstruction (geometry)
- Understanding opportunities for action (didn't cover – call me in 10 years)
- In the process, requires building up tools for processing images and fitting models

Reality of "What is CV?"

Right now: most people would say a buffet of techniques & accumulated knowledge about geometry, pixels, data and learning.



Don’t Be Disappointed With The Buffet!

Get a computer to understand

Understanding an image is incredibly difficult, involves much of your incredible brain, and is perhaps “AI-Complete”

Plus, email me in 15 years and see if we have non-buffet answers.

Where is it Headed?

3 topics that I am betting on or which other people are betting on:

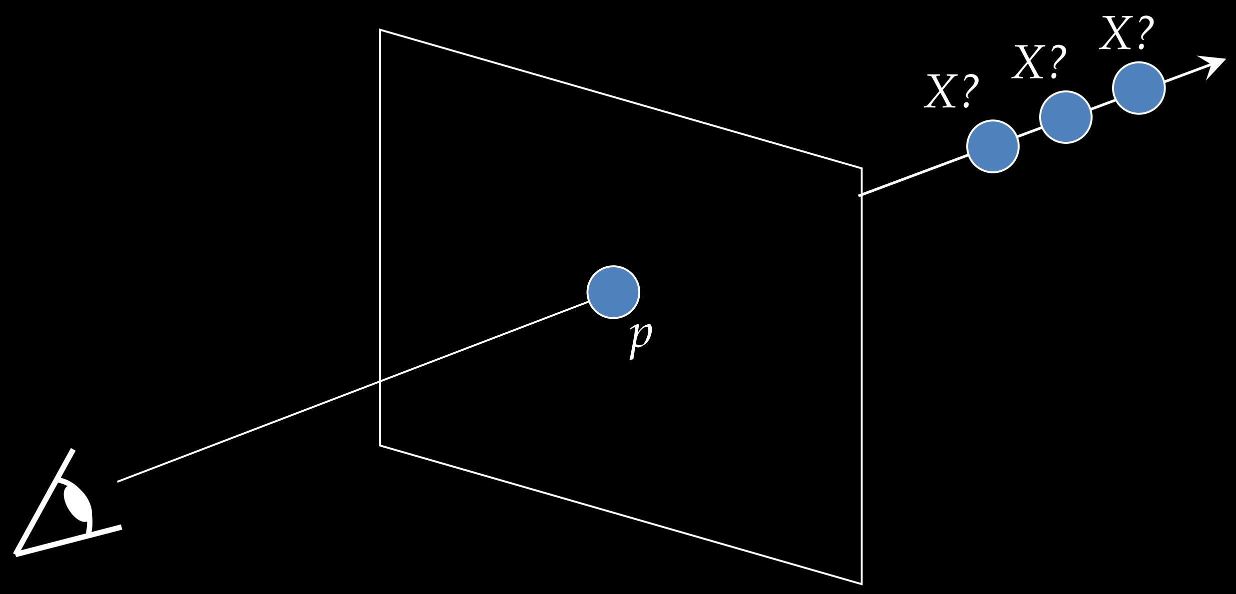
- Learning and geometry (Today)
- Embodied agents (Thursday)
- Vision and language (Next Tuesday)

Some slides won't be posted since I'm borrowing heavily from others' current research slides that they've been generous to share.

In the Process

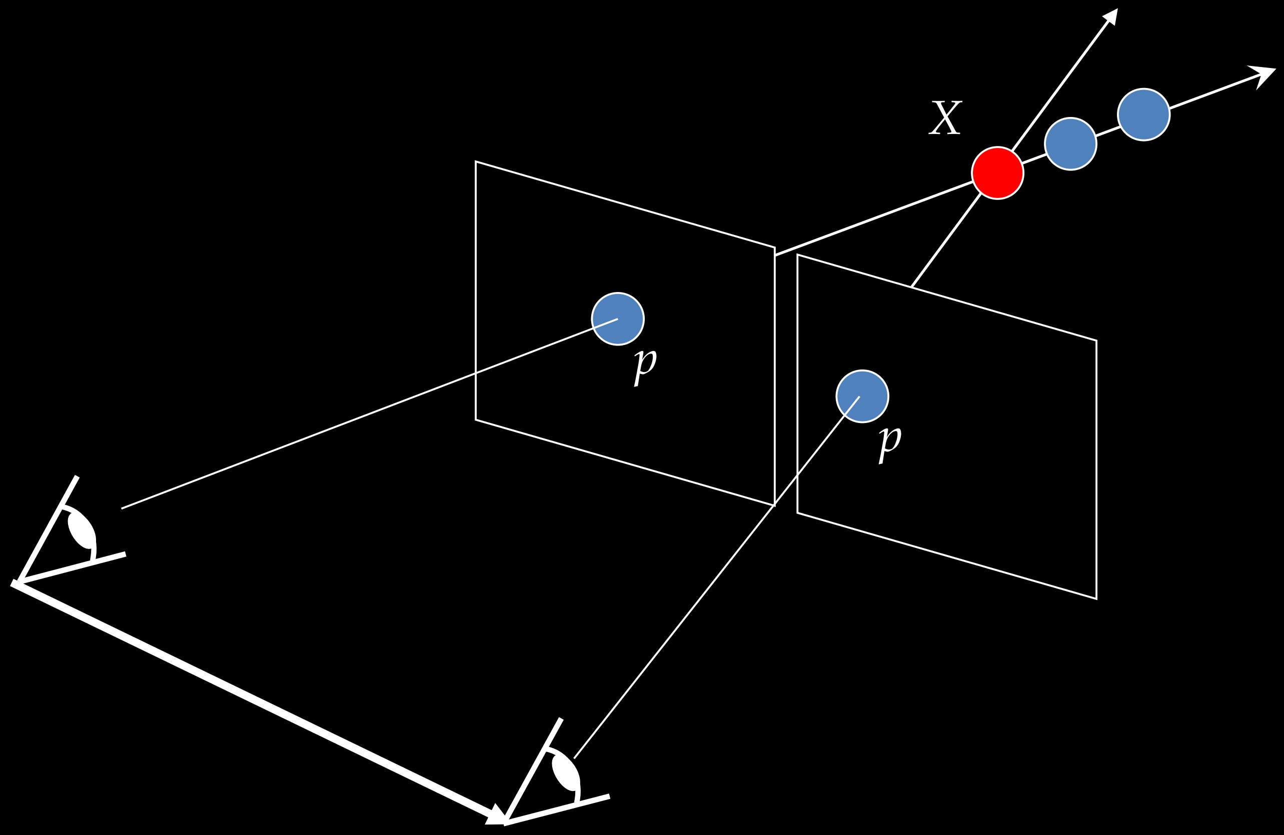
- I hope to give you a sense of how:
- Research in vision is conducted
- We think we know we’ve succeeded
- We think we know we’re not fooling ourselves!

Cues For 3D



- Given a *calibrated camera* and an image, we only know the ray corresponding to each pixel.
- Nowhere near enough constraints for X

Cues For 3D



- Stereo: given 2 calibrated cameras in different views and correspondences, can solve for X

Cues For 3D

Yet when you look at this…

Pictorial Cues for 3D

$f(\cdot)$

"3D"

Learned from Data

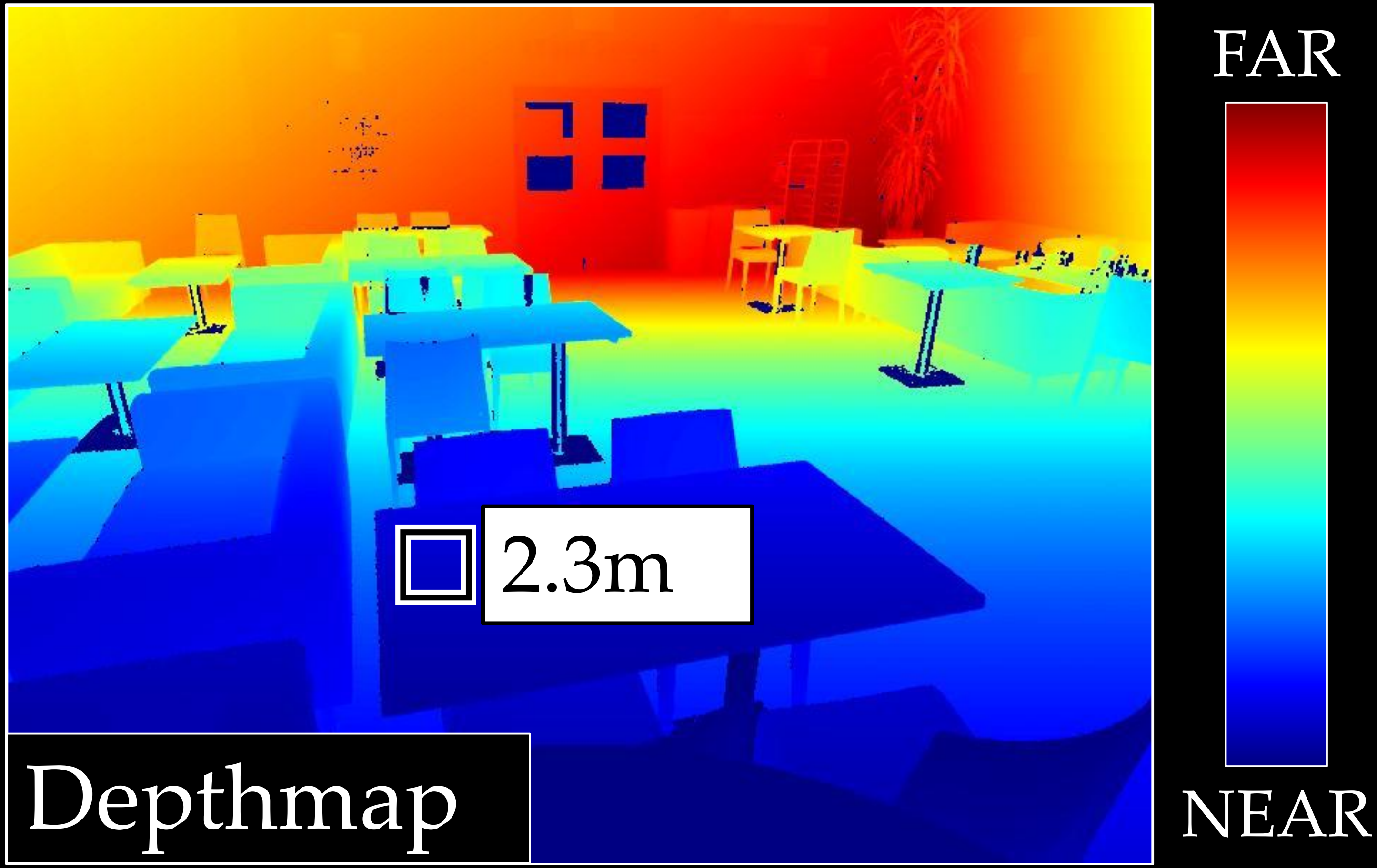
Pictorial Cues for 3D

f(·)

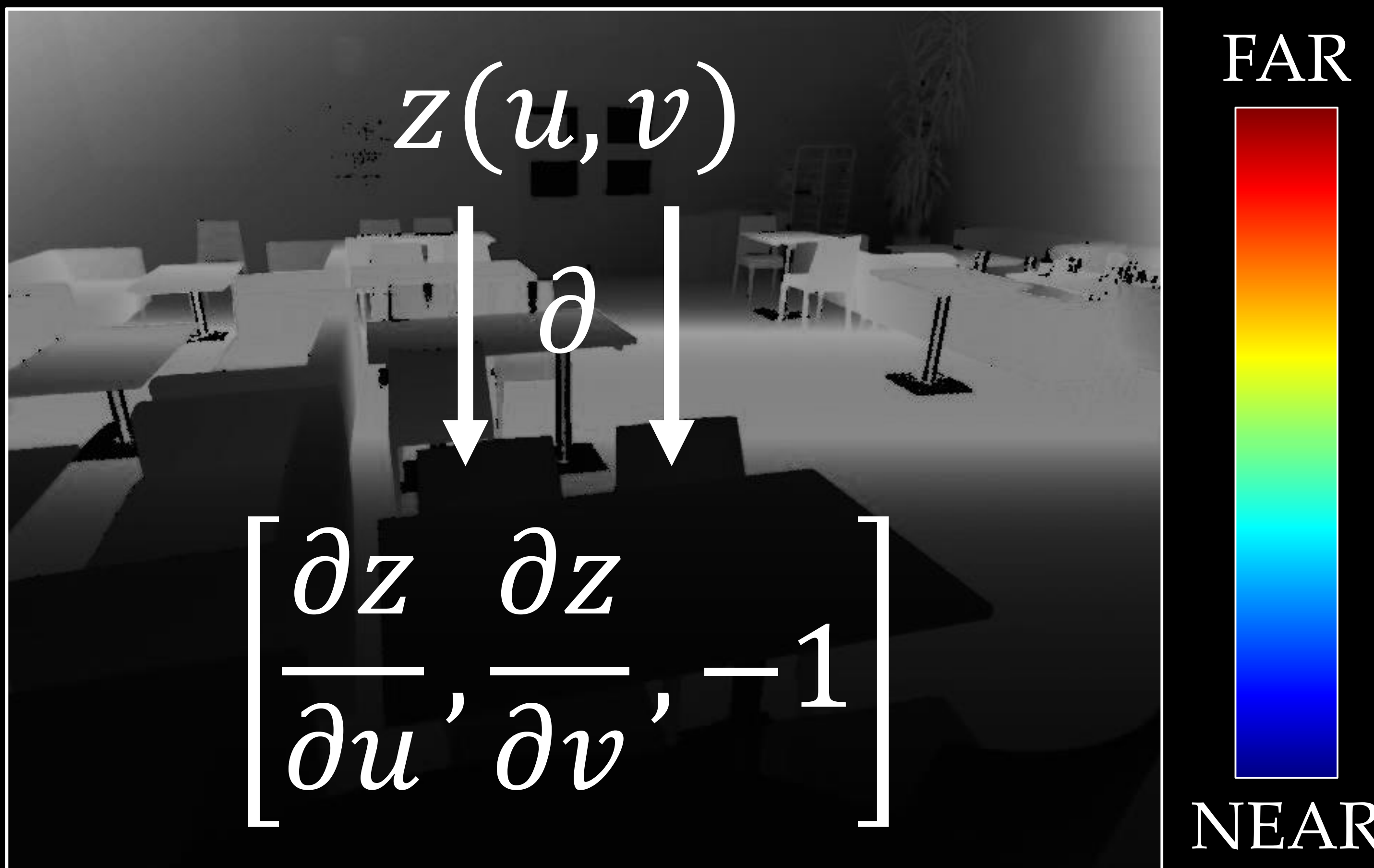
“3D”

Learned from Data

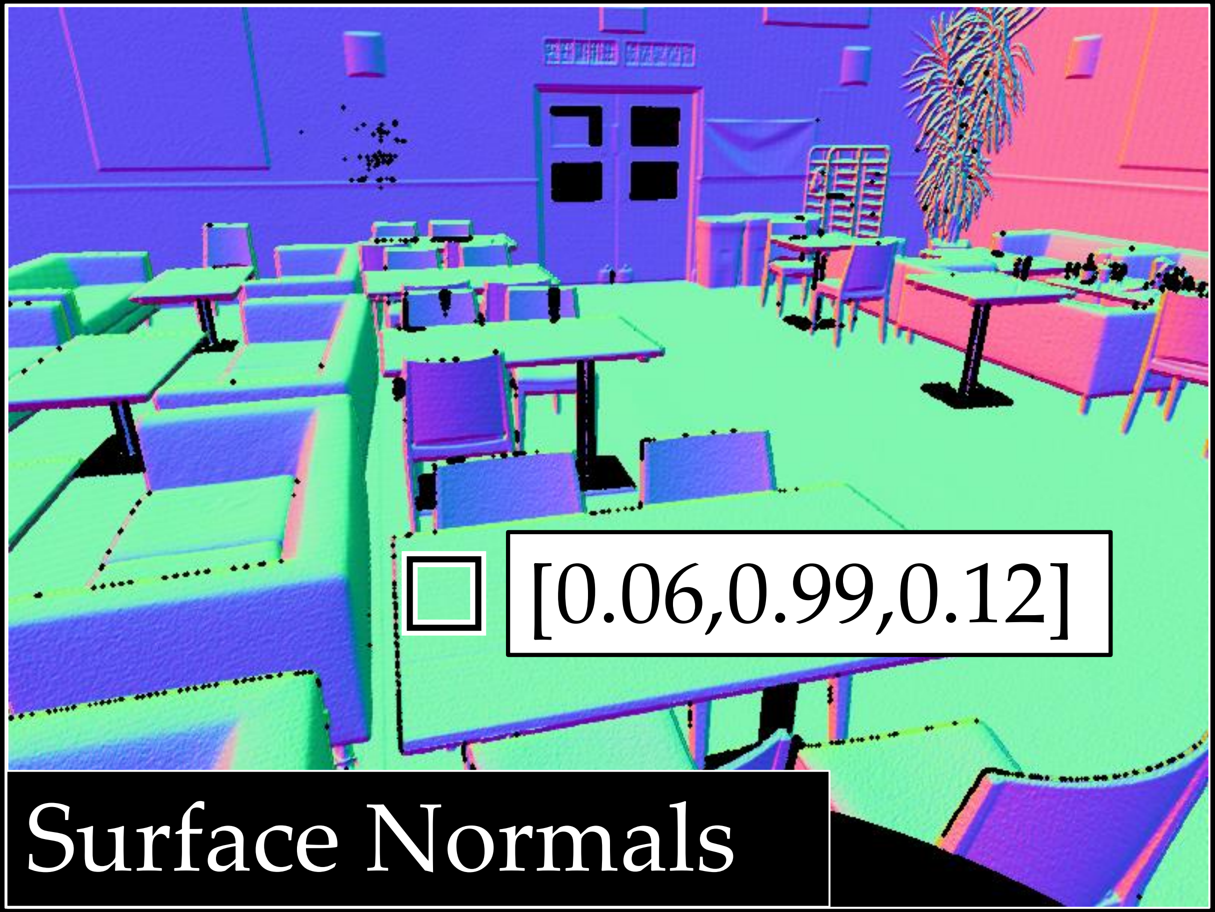
3D Representations



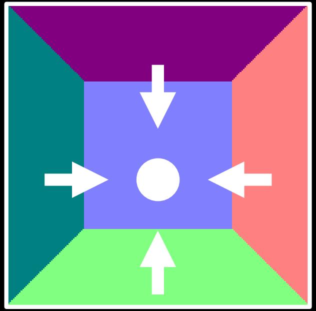
3D Representations



3D Representations



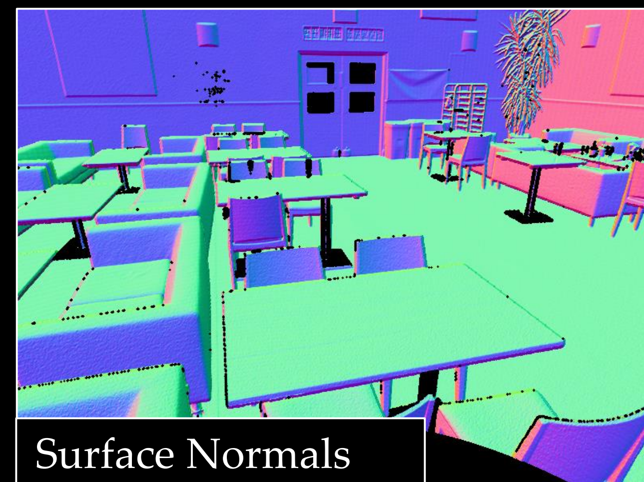
Room



Legend

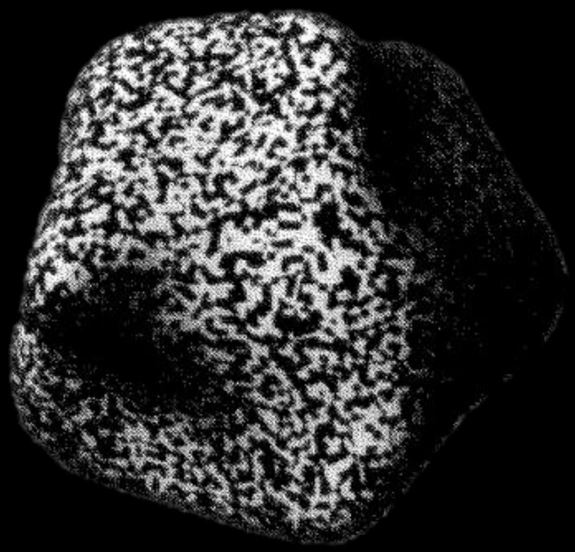
3D Representations

$f(\cdot)$



D.F. Fouhey, A. Gupta, M. Hebert. *Data-Driven 3D Primitives for Single Image Understanding*. ICCV 2013.

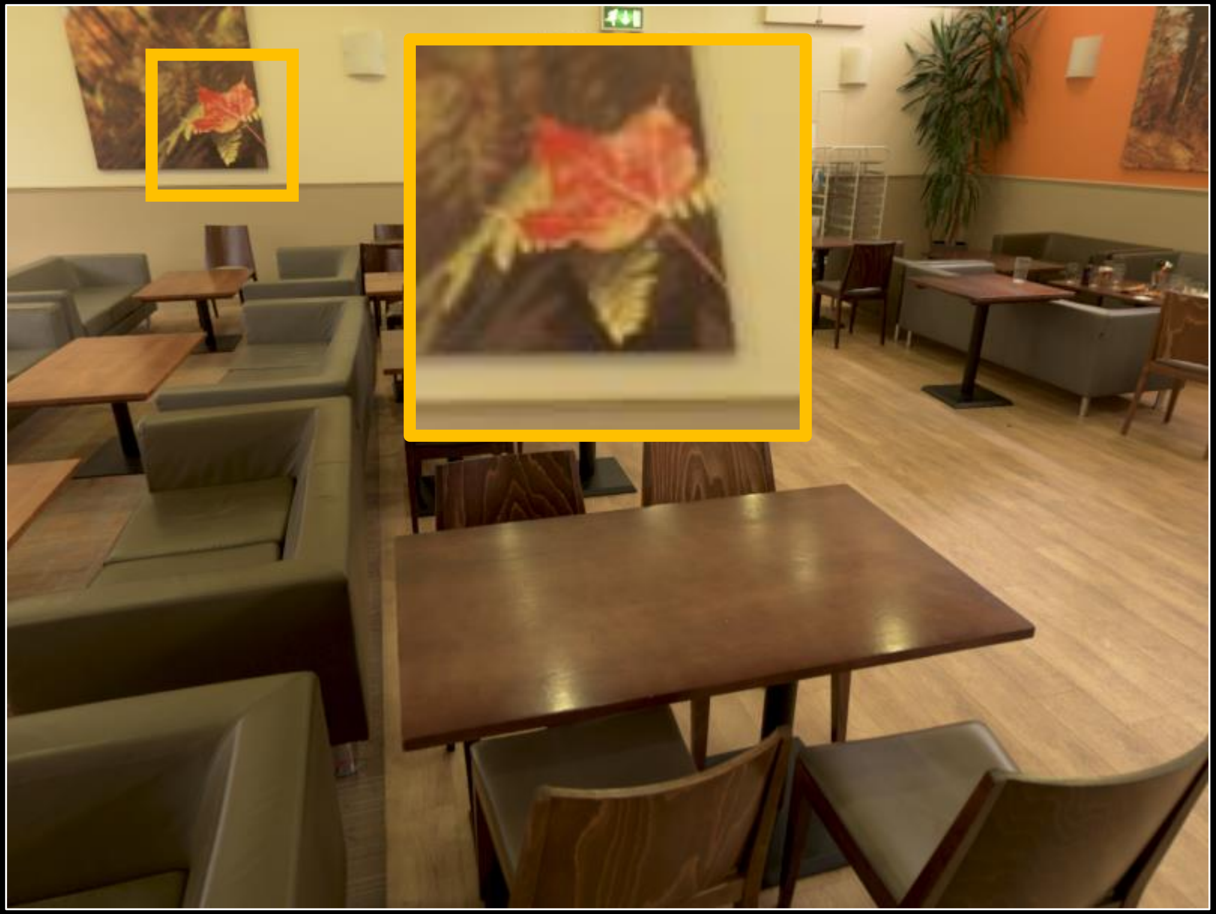
Direct Cues For Normals



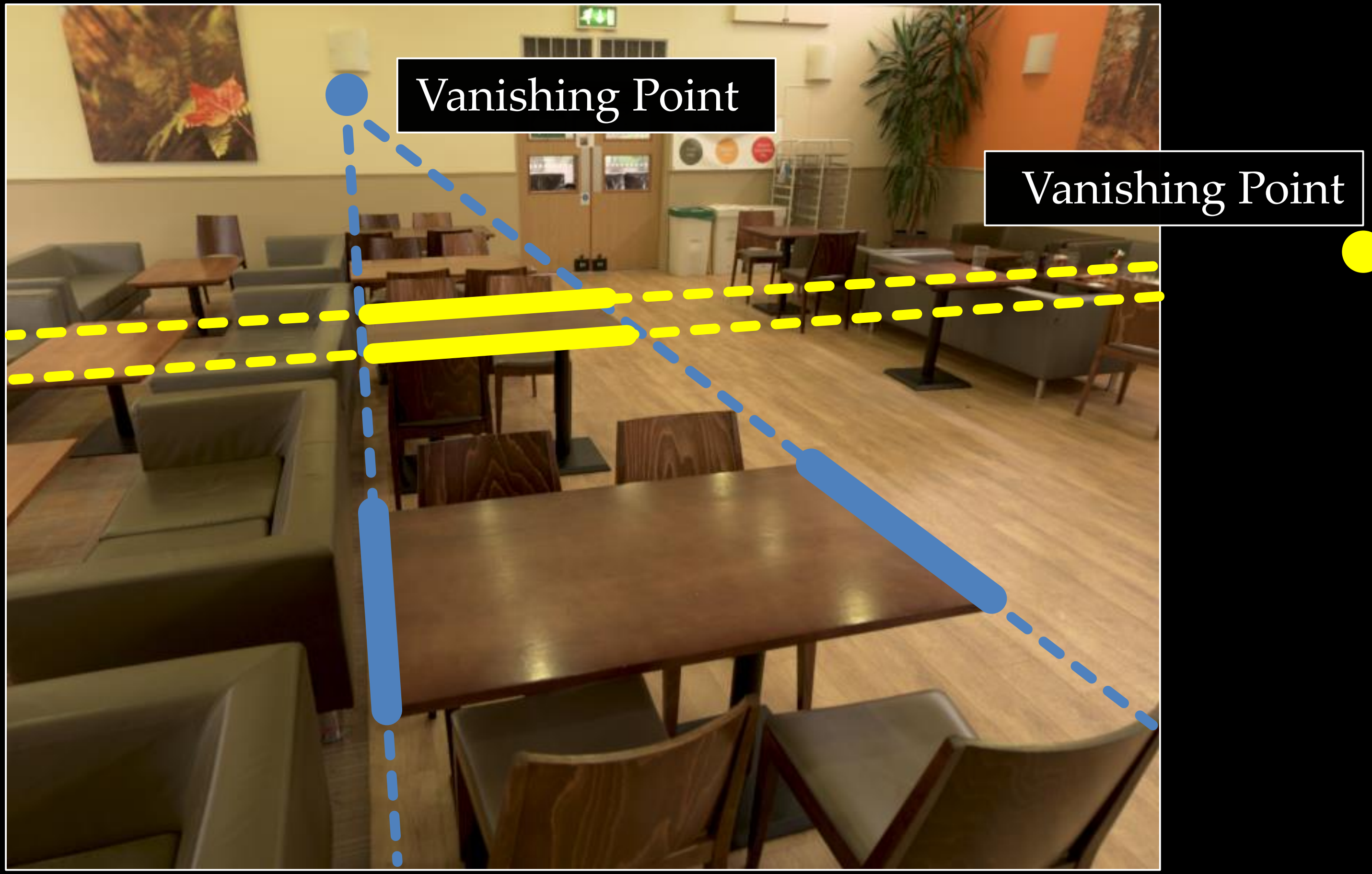
"Depth-difference judgments and attitude settings [surface normals] appear to be independent tasks."
-Koenderink, van Doorn, Kappers '96

Norman and Todd, *The Discriminability of Local Surface Structure.* Perception 1996
Koenderink, Van Doorn, Kappers. *Pictorial Surface attitude and Local Depth Comparisons.* Perception & Psychophysics, 1996
Johnston and Passmore, *Independent Encoding of Surface Orientation and Surface Curvature.* Vision Research, 1994
etc.

Direct Cues For Normals



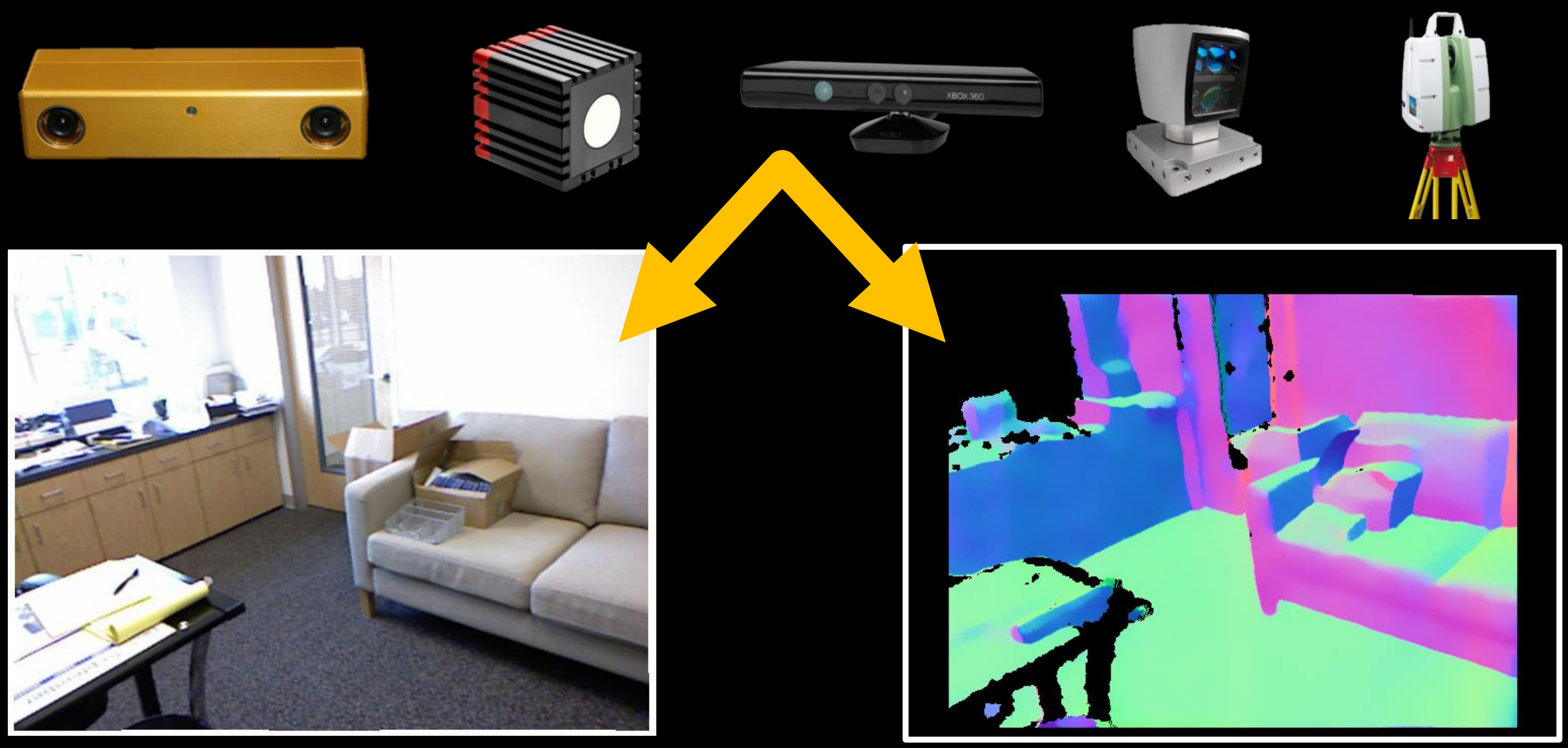
Direct Cues to Normals



Comment – Representations

- For something as simple as whether to predict depth (z(u,v)) or the orientation of the plane ($\left[\frac{\partial z}{\partial u}, \frac{\partial z}{\partial v}, -1\right]$), there are different:
- Metrics (duh)
- Methods (hmm) – these typically aim to take advantage of special structure of the problem
- Applications of techniques

Surface Normals



Color Image

Normals

D. F. Fouhey, A. Gupta, M. Hebert. *Data-Driven 3D Primitives for Single Image Understanding*. ICCV 2013.
D. F. Fouhey, A. Gupta, M. Hebert. *Unfolding an Indoor Origami World*. ECCV 2014.
X. Wang, D.F. Fouhey, A. Gupta. *Designing Deep Networks for Surface Normal Estimation*. CVPR 2015.

Surface Normals

$$f(\cdot)$$

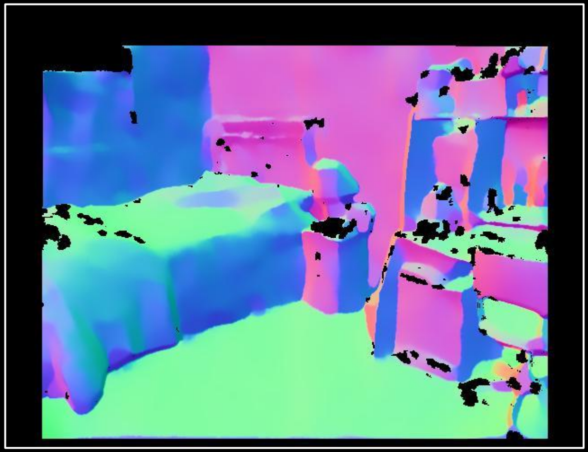
D. F. Fouhey, A. Gupta, M. Hebert. *Data-Driven 3D Primitives for Single Image Understanding*. ICCV 2013.
D. F. Fouhey, A. Gupta, M. Hebert. *Unfolding an Indoor Origami World*. ECCV 2014.
X. Wang, D.F. Fouhey, A. Gupta. *Designing Deep Networks for Surface Normal Estimation*. CVPR 2015.

Applying Deep Learning

Input

CNN

Output



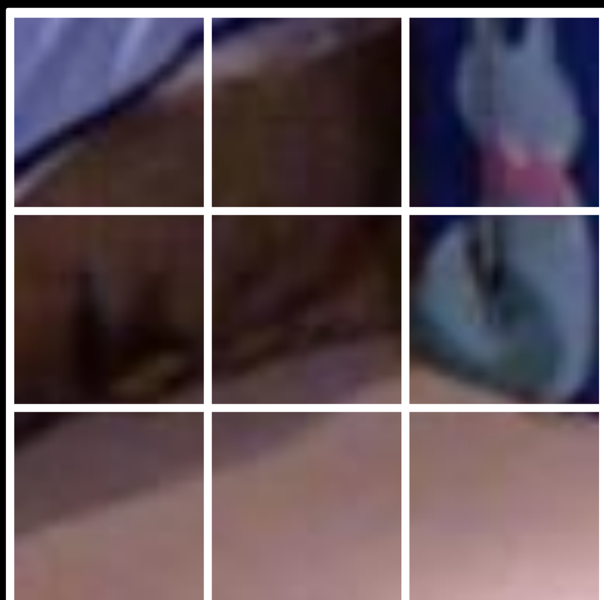
How do we incorporate constraints?

How do we represent the output?

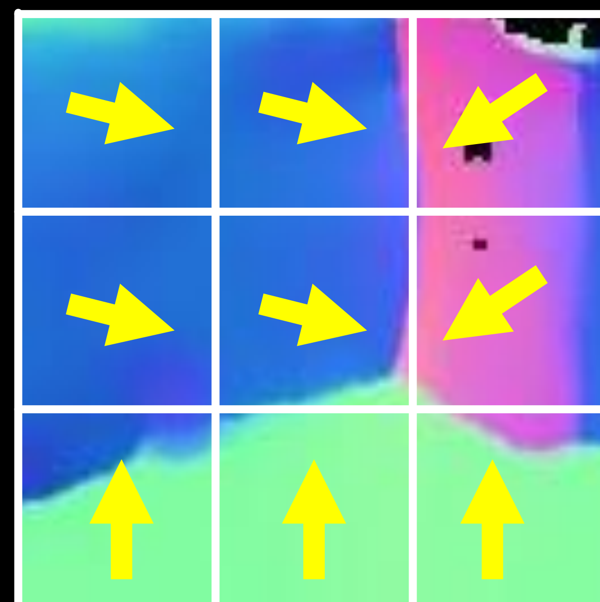
X. Wang, D.F. Fouhey, A. Gupta. *Designing Deep Networks for Surface Normal Estimation.* CVPR 2015

Representation and Objective

Input



Ground Truth



Quantized Normals

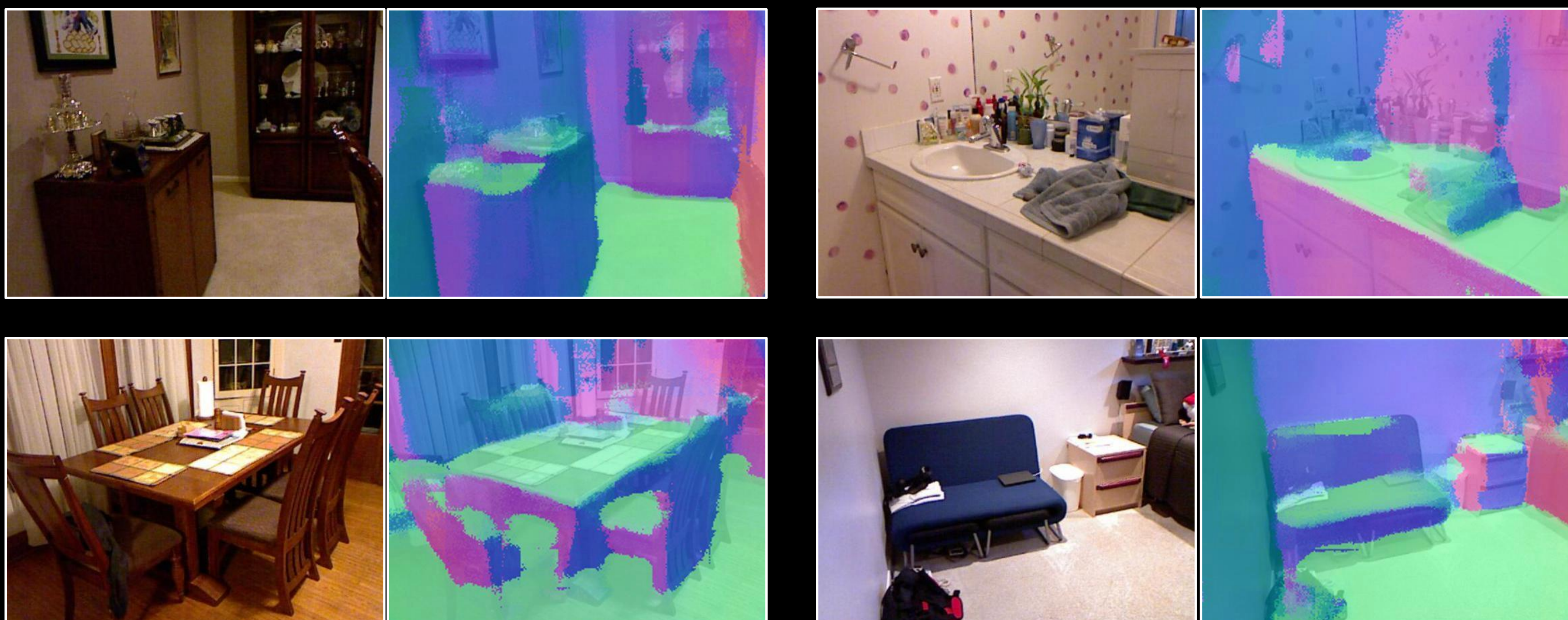
Class 1: Class 2: … Class K:

Normal quantization scheme from Ladicky et al. 2014

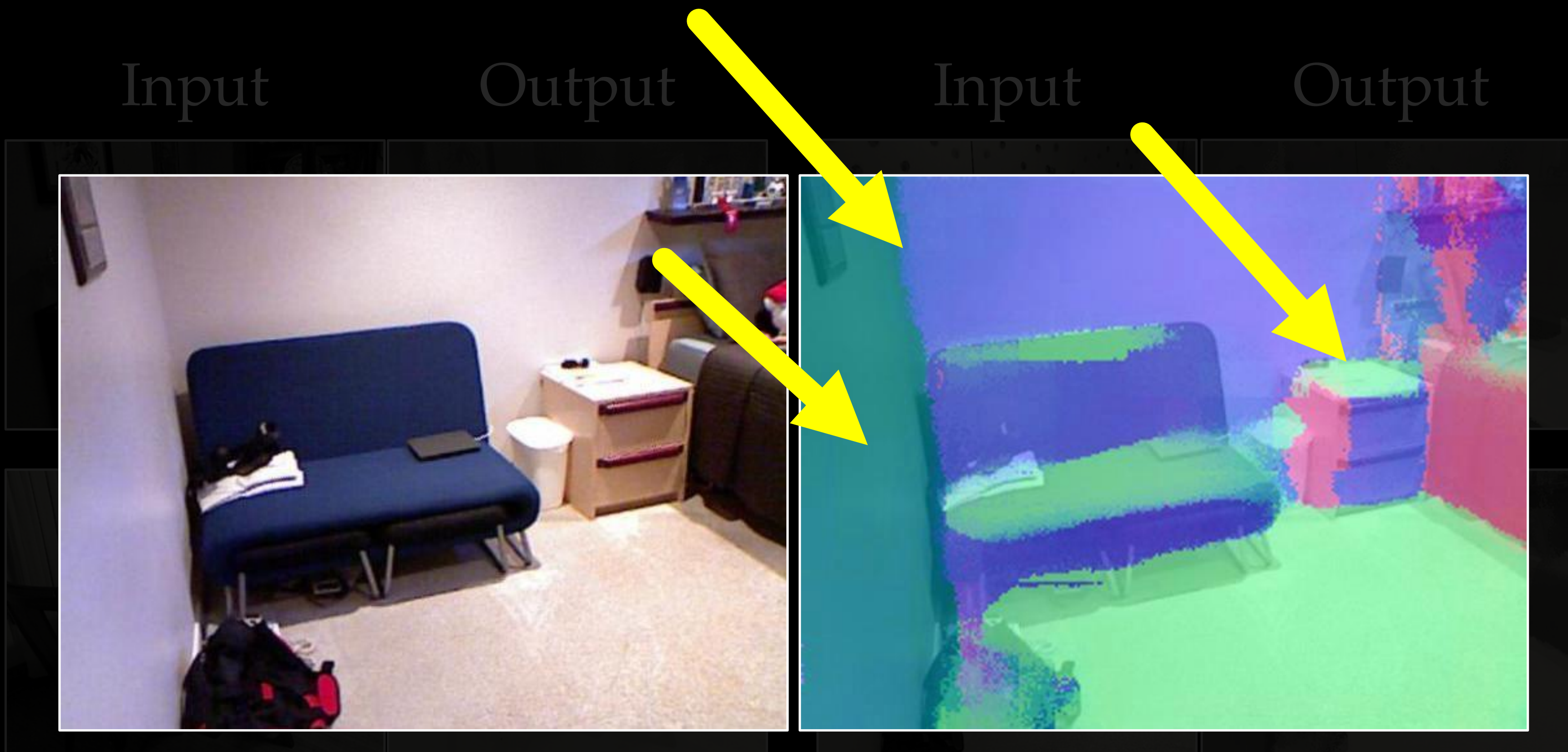
Results

Results

Input | Output | Input | Output



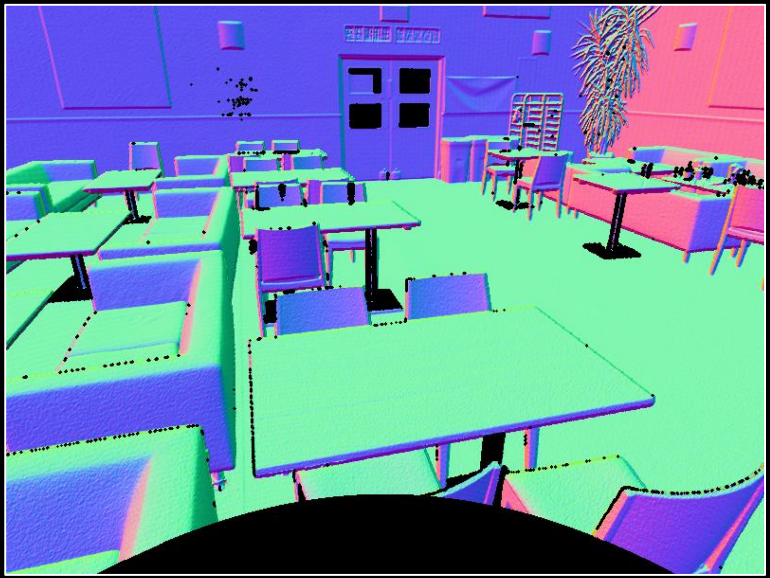
Results



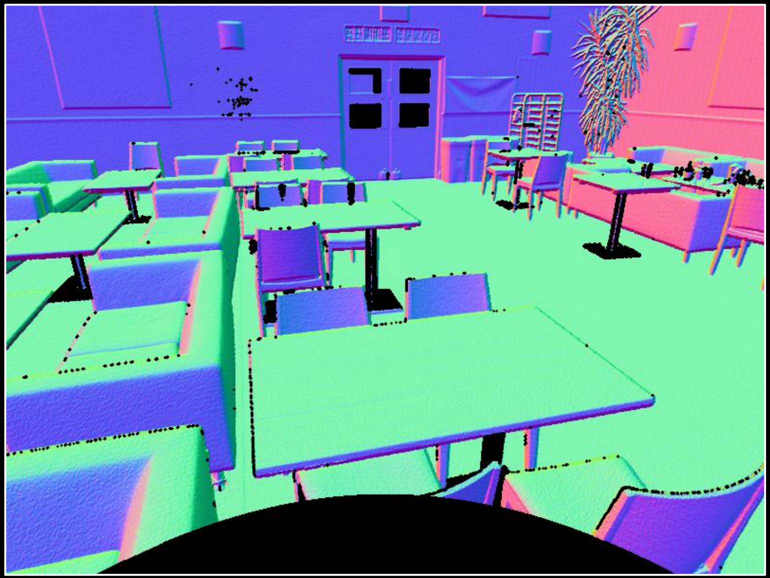
Comment – Picking Problems

- I’d show results, and the response from many people would be “sure, sure, neat but I’ll just buy a Kinect”

3D Representations



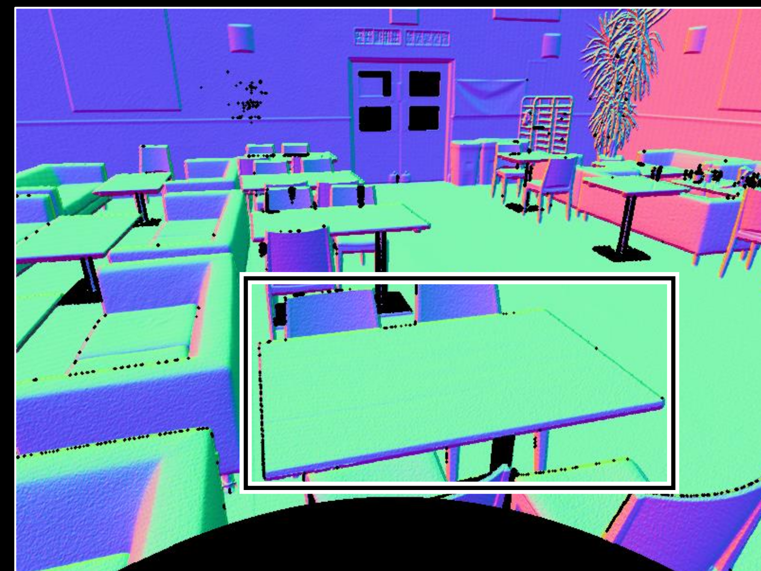
3D Representations



~$50K, 6.5 minutes an image

Adams et al., Scientific Reports, 2016

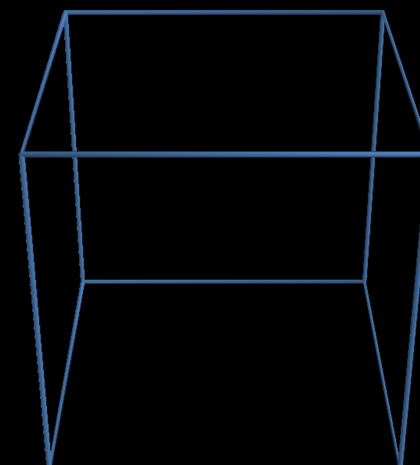
3D Representations



How thick is the table?
What's behind it?
Is the chair attached to the table?

3D Representations

$f(\cdot)$



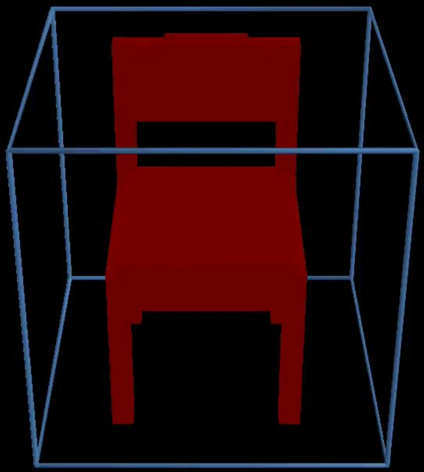
RGB Image

Voxels

R. Girdhar, D. F. Fouhey, M. Rodriguez, A. Gupta.
Learning a predictable and generative vector representation for objects. ECCV 2016
Contemporary work also proposing to predict voxels: C. Choy et al. ECCV 2016.

3D Representations

$f(\cdot)$

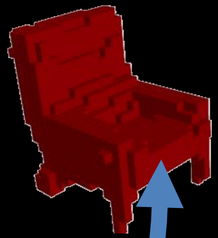


RGB Image

Voxels

R. Girdhar, D. F. Fouhey, M. Rodriguez, A. Gupta.
Learning a predictable and generative vector representation for objects. ECCV 2016
Contemporary work also proposing to predict voxels: C. Choy et al. ECCV 2016.

Approach



20x20x20
Voxel output
S

$S_{i,0}$ P(voxel i empty) 0.1

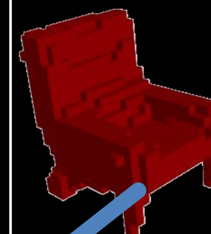
$S_{i,1}$ P(voxel i filled) 0.9

S_i

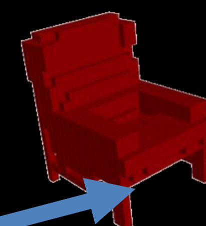
Approach

224x224x3
Image Input
I

$$S = f(I;\theta)$$



20x20x20
Voxel output
S



20x20x20
Voxel Truth
Y

$$-\sum_{i=1}^{20^3}\log(S_{(i,Y_i)})$$

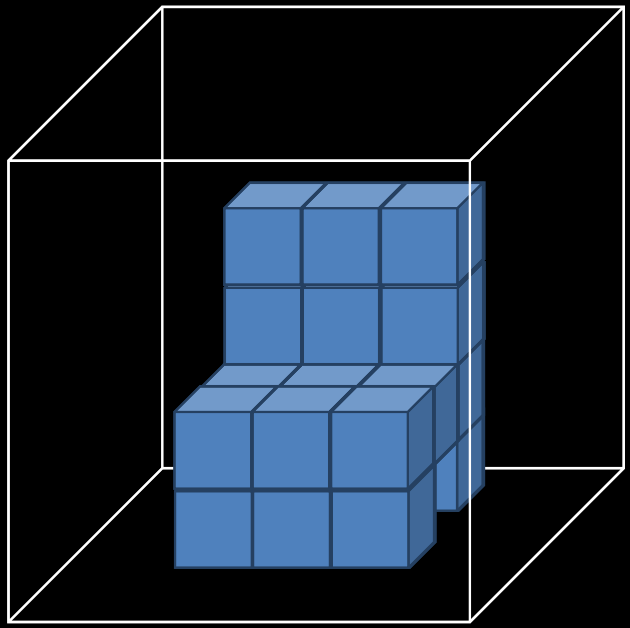
Main Idea

Representation should satisfy

- **Generative in 3D:** should be able to construct objects
- **Predictable from 2D:** should be able to infer from an ordinary image

Output Representation – Voxels

- Binary 3D Pixels
- Size fixed in advance
- Spatially organized



Motivation

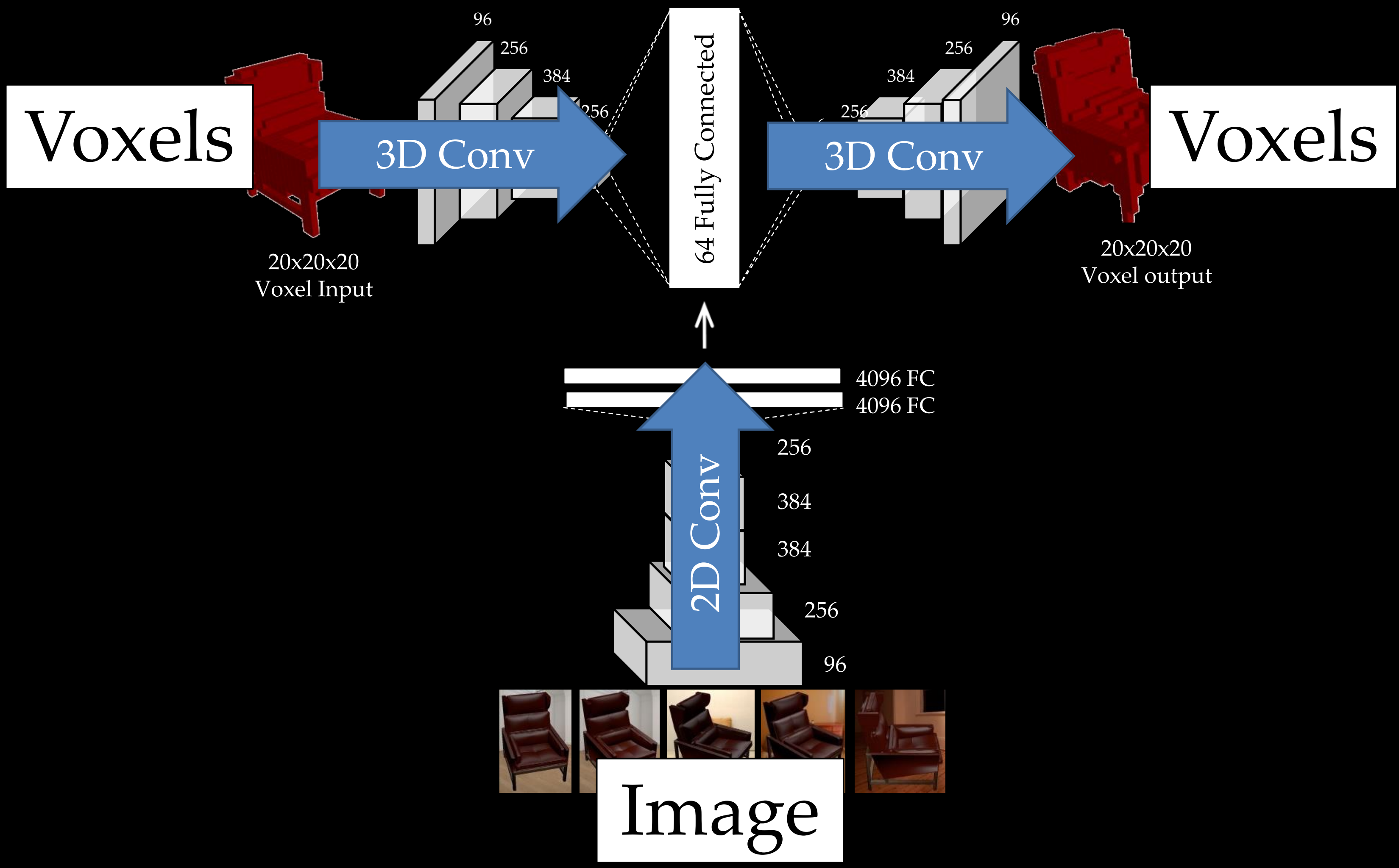
How many couches are there really?

1737662031938094565999824459494356270619397861001172505471732865032623760224580084650943336301208543380003
194362163007597987225472483598640843335685441710193966274131338557192586399006789292714554767500194796127
964596906605976605873665859580600161998556511368530960400907199253450604168622770350228527124626728538626
805418833470107651091641919900725415994689920112219170907023561354484047025713734651608777544579846111001
059482132180956689444108315785401642188044178788629853592228467331730519810763559577944882016286493908631
503101121166109571682295769470379514531105239965209245314082665518579335511291525230373316486697786532335
206274149240813489201828773854353041855598709390675430960381072270432383913542702130202430186637321862331
068861776780211082856984506050024895394320139435868484643843368002496089956046419964019877586845530207748
994394501505588146979082629871366088121763790555364513243984244004147636040219136443410377798011608722717
131323621700159335786445601947601694025107888293017058178562647175461026384343438874861406516767158373279
032321096262126551620255666605185789463207944391905756886829667520553014724372245300878786091700563444079
107099009003380230356461989260377273986023281444076082783406824471703499844642915587790146384758051663547
775336021829171033411043796977042190519657861762804226147480755555085278062866268677842432851421790544407
006581148631979148571299417963950579210719961422405768071335213324842709316205032078384168750091017964584
060285240107161561019930505687950233196051962261970932008838279760834318101044311710769457048672103958655
016388894770892065267451228938951370237422841366052736174160431593023473217066764172949768821843606479073
866252864377064398085101223216558344281956767163876579889759124956035672317578122141070933058555310274598
884089982879647974020264495921703064439532898207943134374576254840272047075633856749514044298135927611328
433323640657533550512376900773273703275329924651465759145114579174356770593439987135755889403613364529029
604049868233807295134382284730745937309910703657676103447124097631074153287120040247837143656624045055614
076111832245239612708339272798262887437416818440064925049838443370805645609424314780108030016683461562597
569371539974003402697903023830108053034645133078208043917492087248958344081026378788915528519967248989338
59202712442391408339177188452446496864505205821815101050847125828590768535580722988074767763478937 6

Motivation

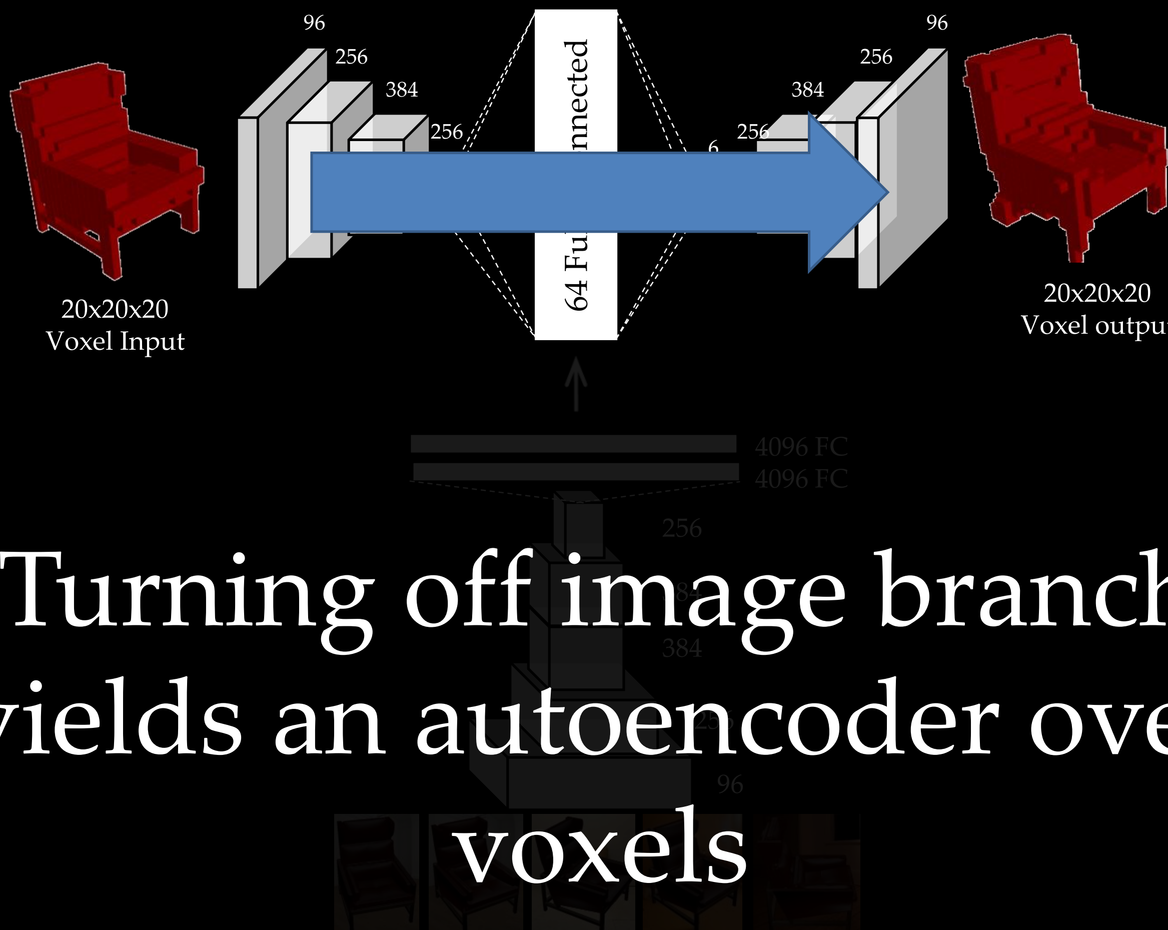
How many couches are there really?

Approach



Voxels
20x20x20
Voxel Input
96
256
384
256
3D Conv
64 Fully Connected
96
256
384
256
3D Conv
Voxels
20x20x20
Voxel output
4096 FC
4096 FC
256
384
384
256
96
2D Conv
Image

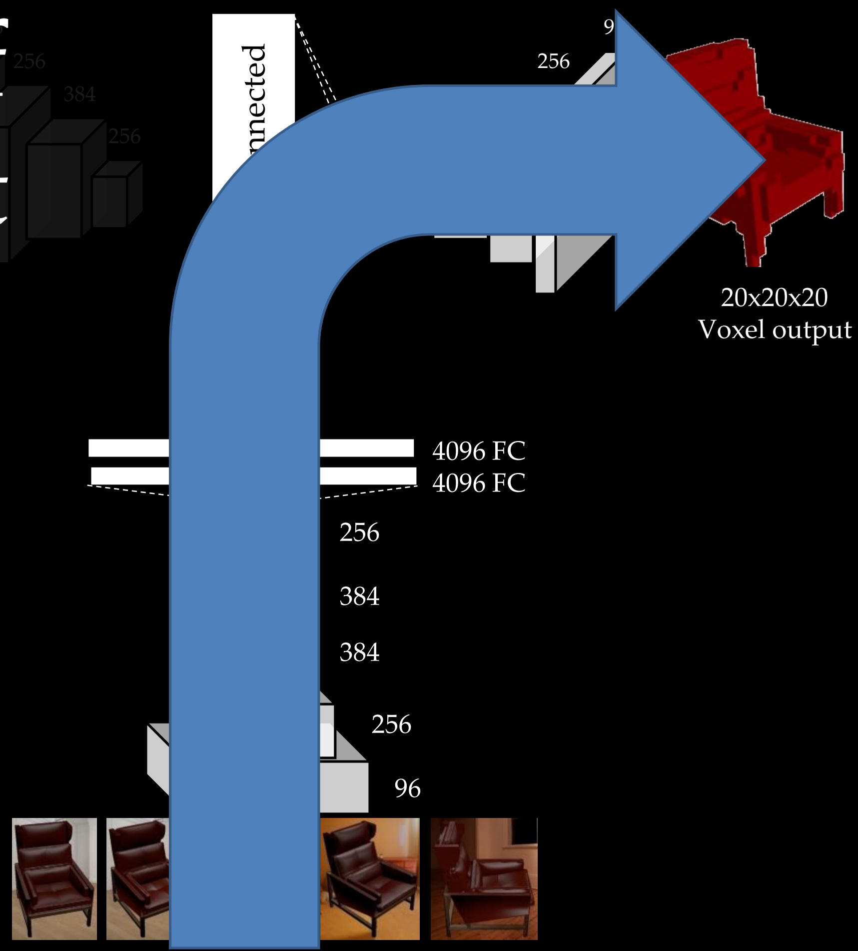
Approach



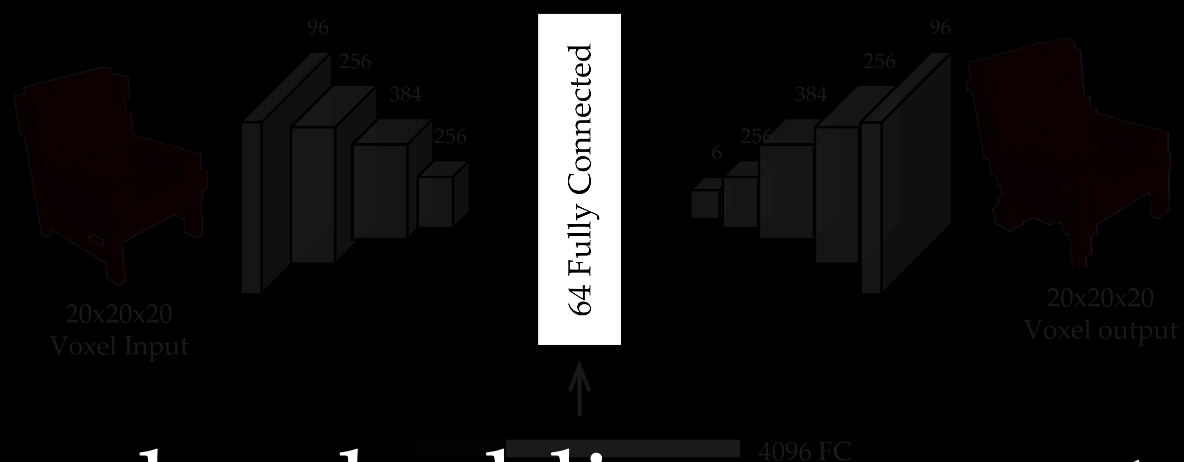
Turning off image branch yields an autoencoder over voxels

Approach

Turning off voxel input yields an image to voxel predictor

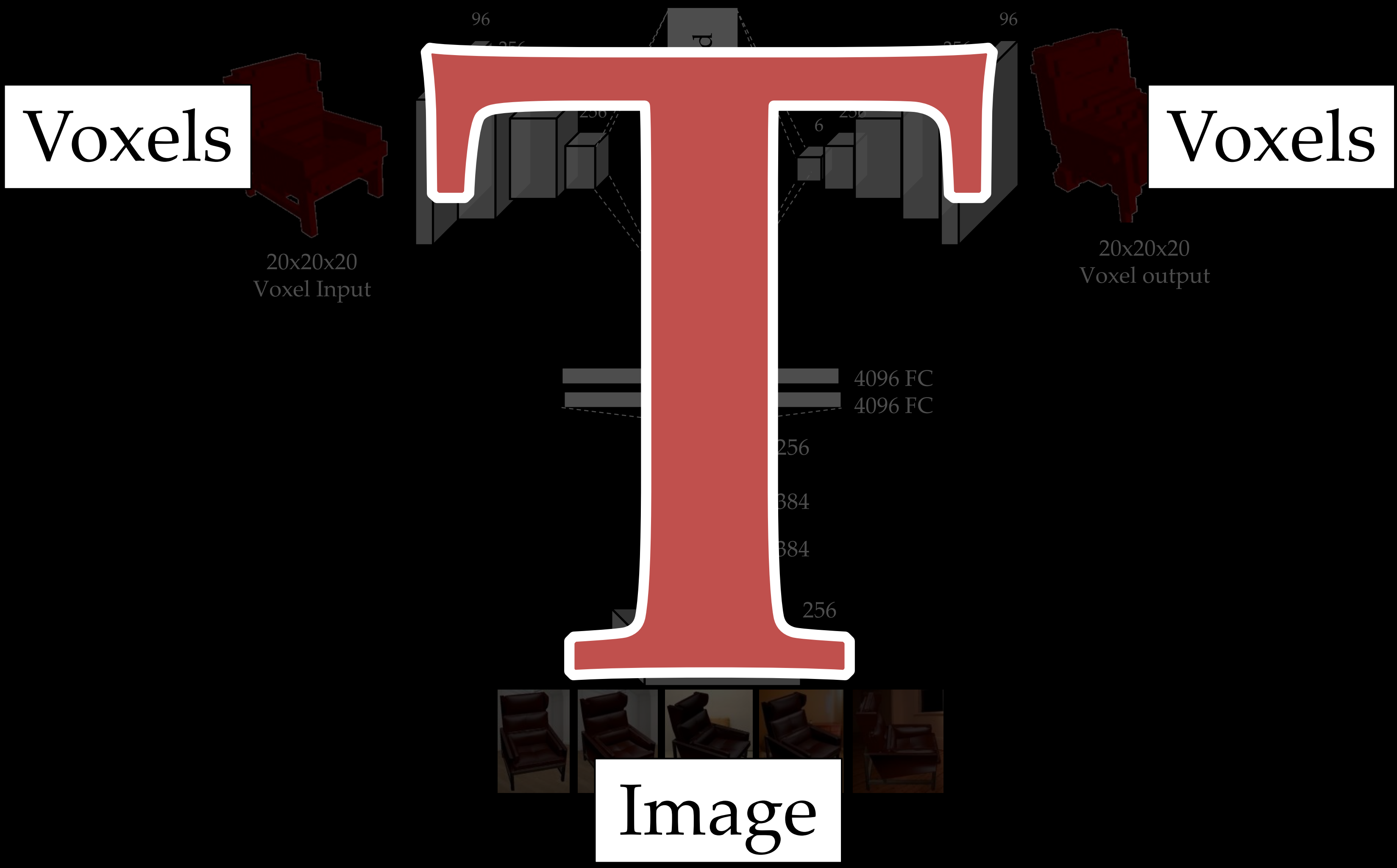


Approach

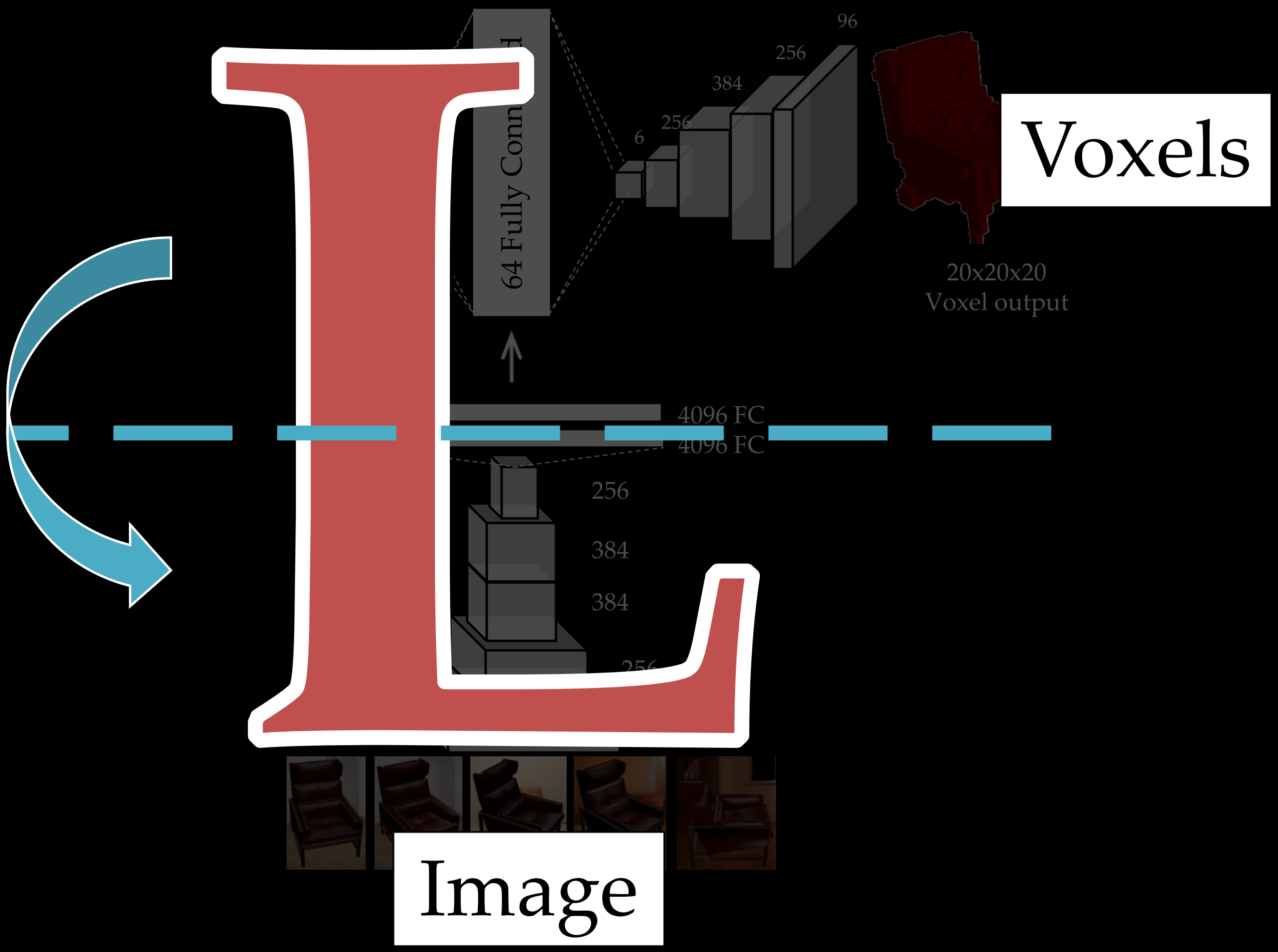


Learned embedding parameterizes shape in a way that is:
(a) generative and predictable
(b) accessible from voxels and images

TL-Network

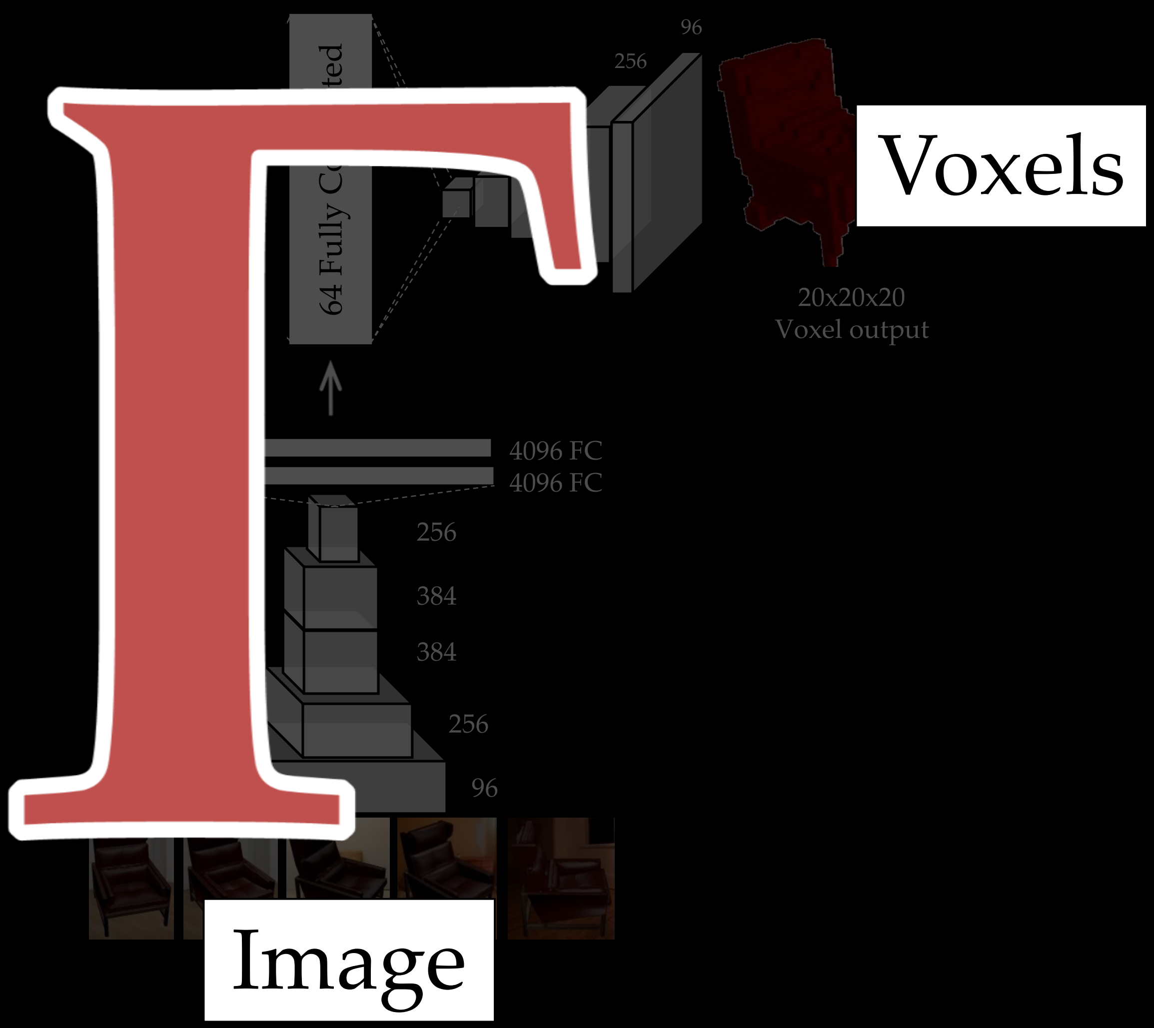

Voxels
Voxels
20x20x20
Voxel Input
20x20x20
Voxel output
96
96
4096 FC
4096 FC
256
384
384
256
Image

TL-Network

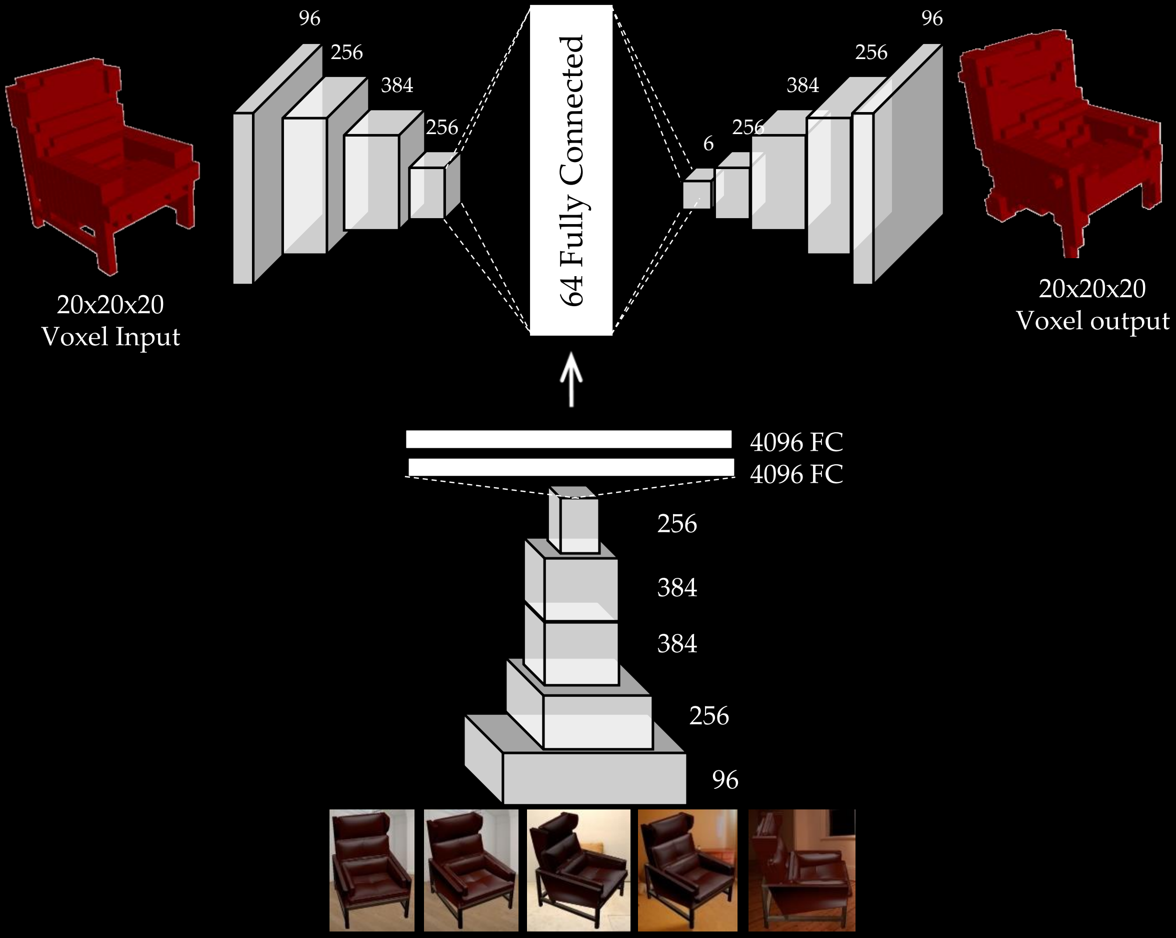


64 Fully Conn
6
256
384
256
96
Voxels
20x20x20
Voxel output
4096 FC
4096 FC
256
384
384
Image

TL-Network

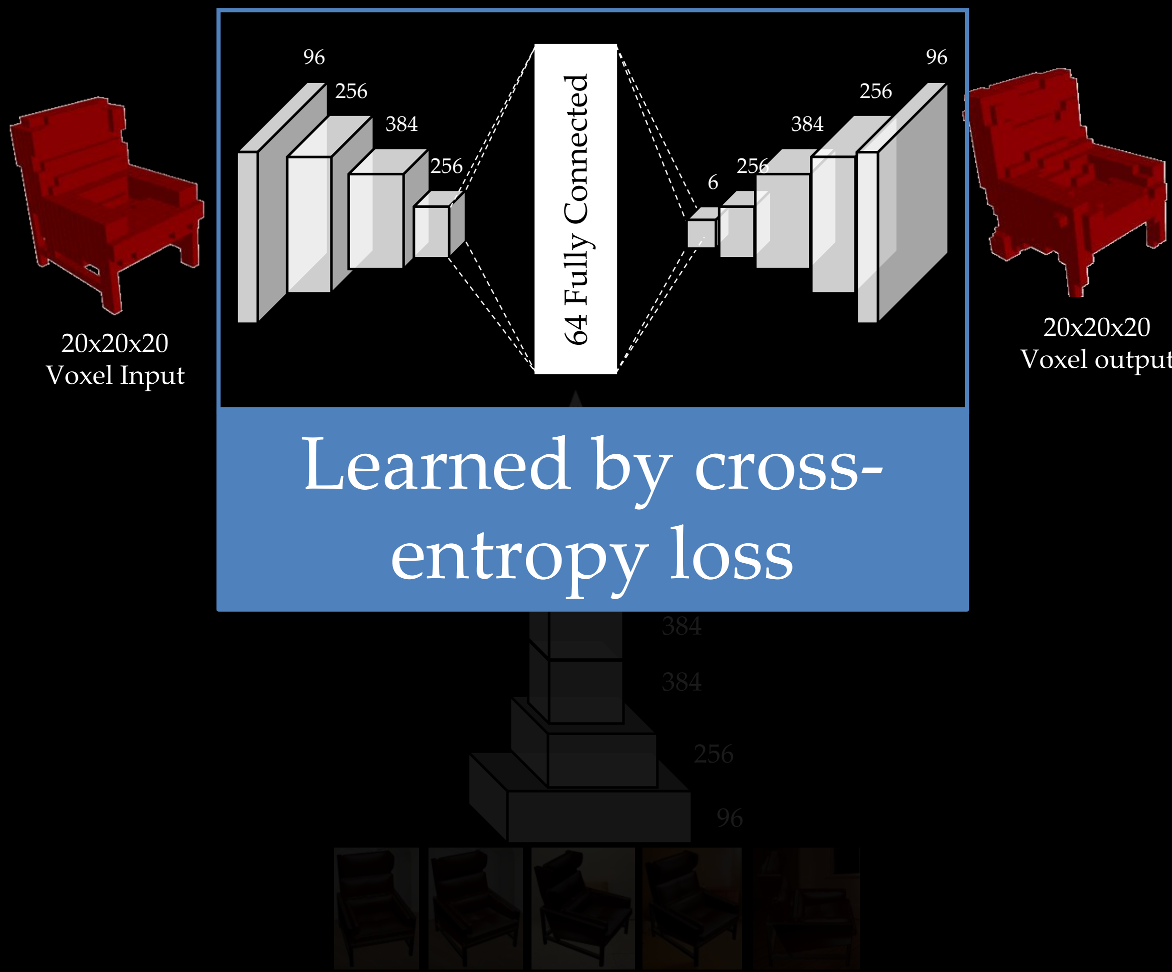

Voxels
96
256
20x20x20
Voxel output
4096 FC
4096 FC
256
384
384
256
96
Image

Training



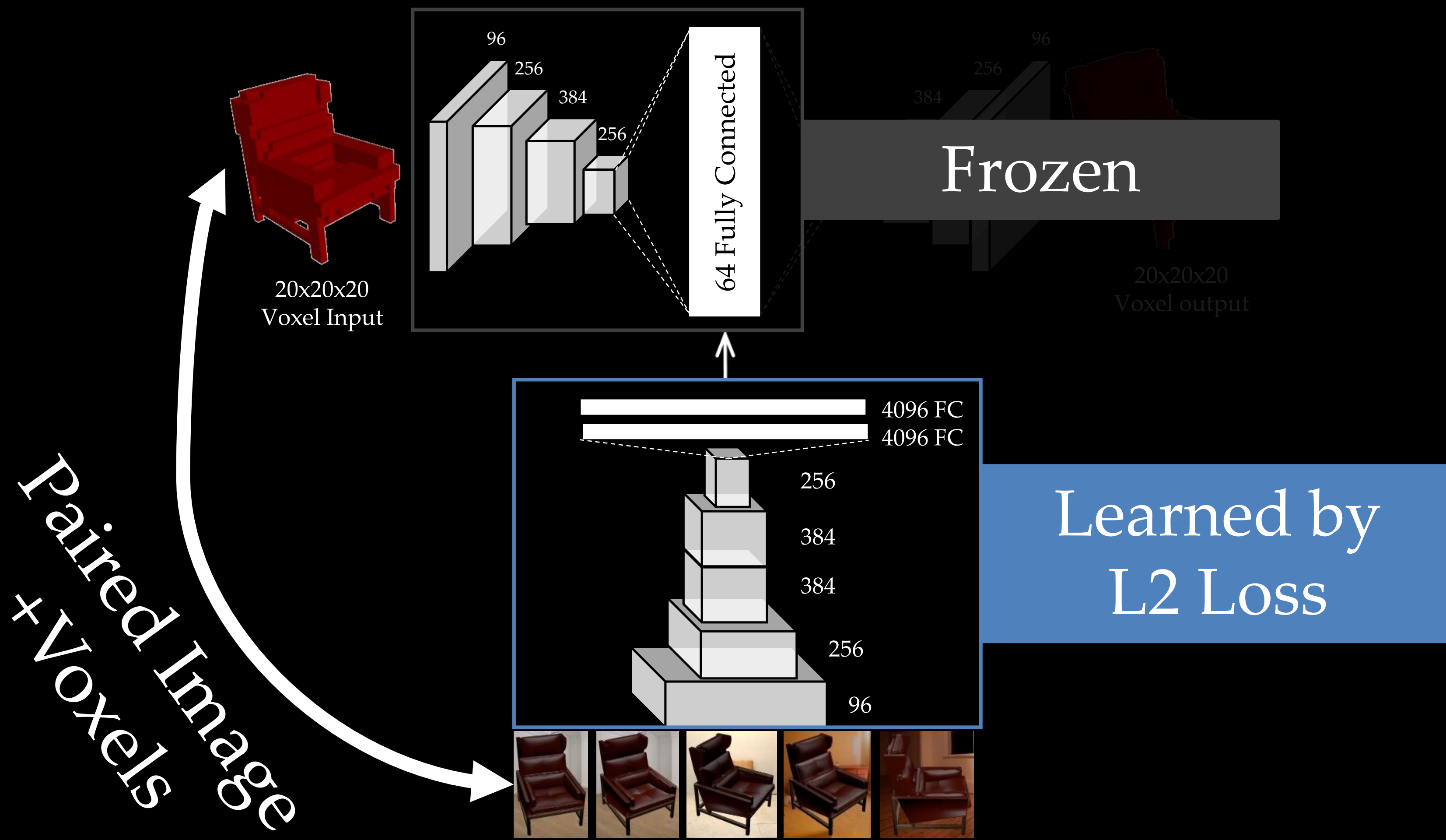
96
256
384
256
64 Fully Connected
6
256
384
256
96
20x20x20
Voxel Input
20x20x20
Voxel output
4096 FC
4096 FC
256
384
384
256
96

Training – Stage 1



Learned by cross-entropy loss

Training – Stage 2

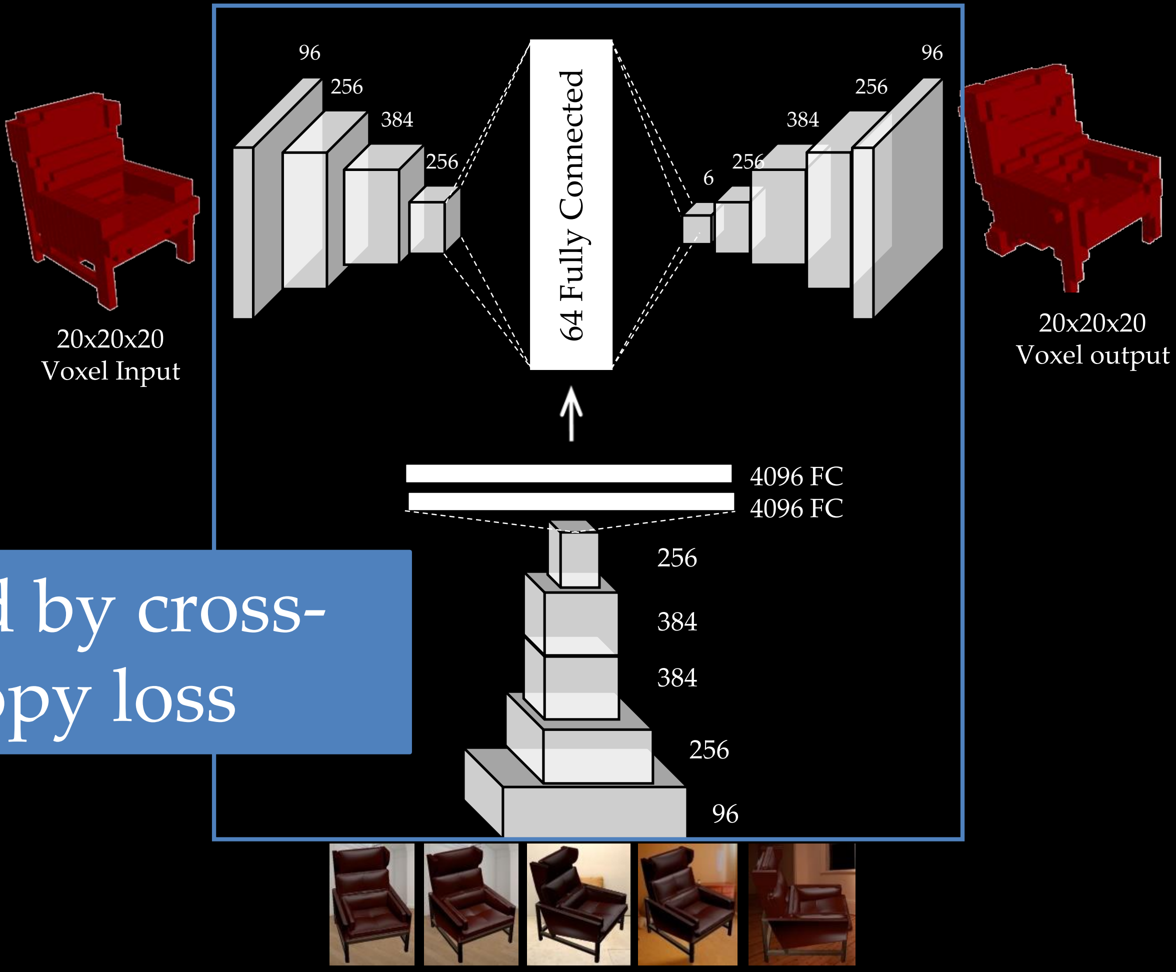


Training Data

- 5 Categories from ShapeNet (**no category labels used in learning**)
- Standard rendering techniques

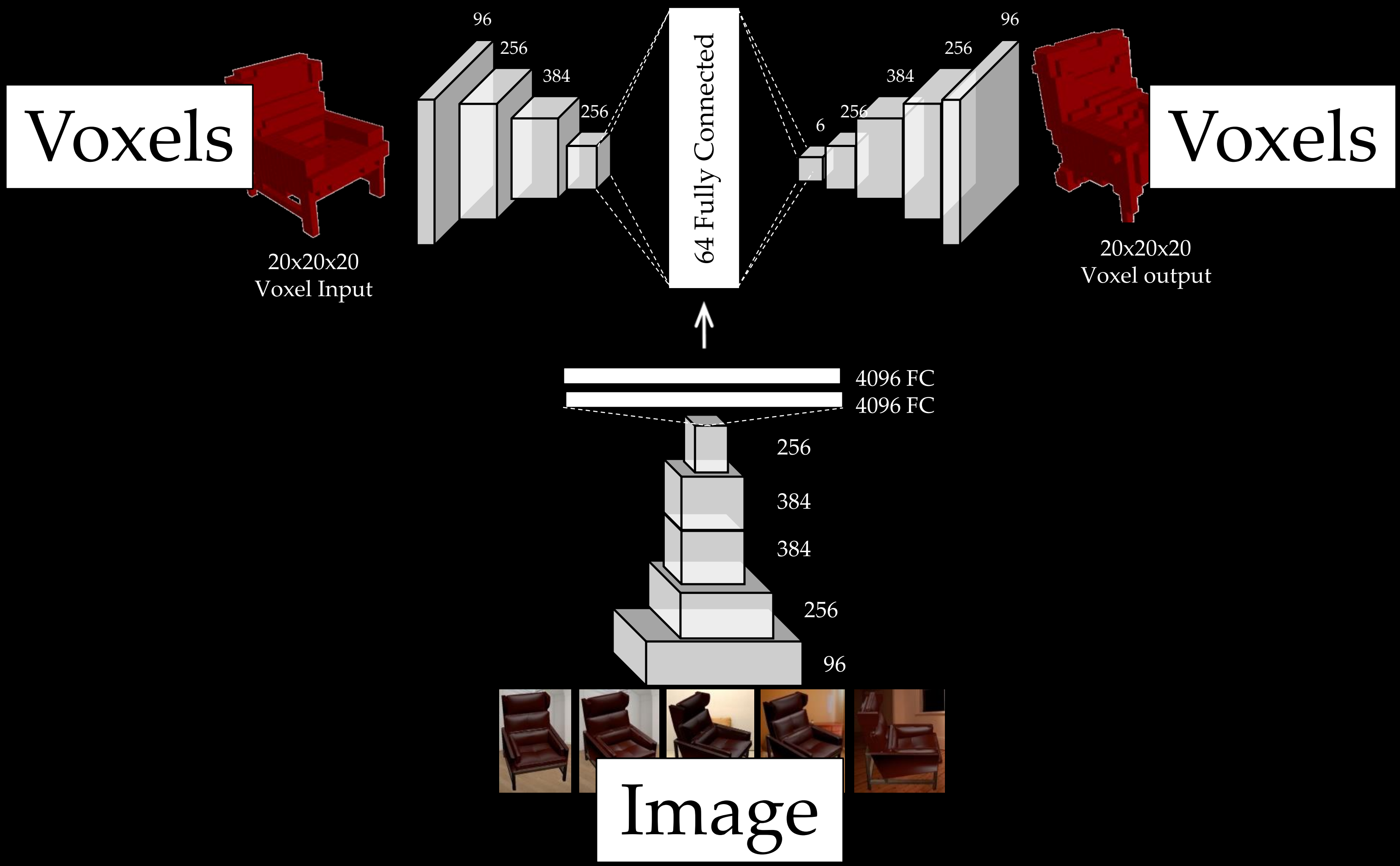
Rendering techniques from Su et al. ICCV15

Training – Stage 3



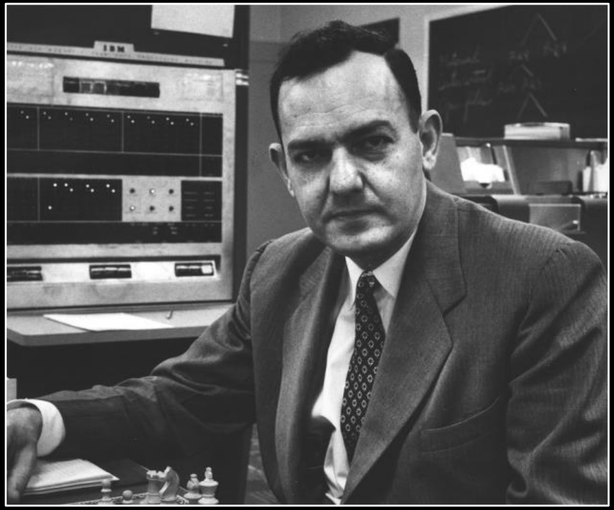
Learned by cross-entropy loss

Experiments

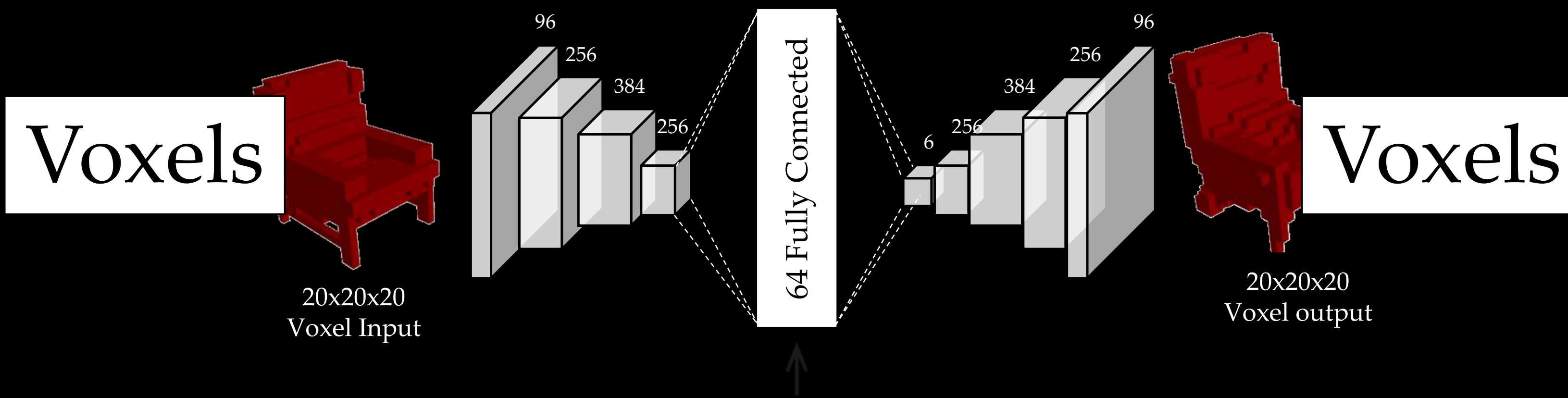


Commentary – Experiments

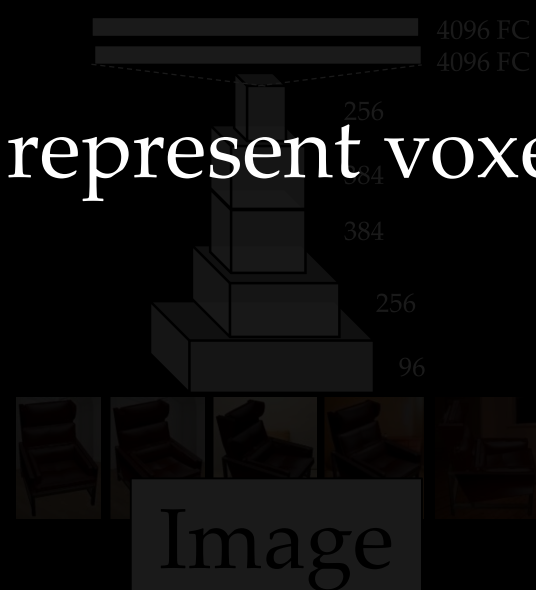
- Main goals of any experiments: Empirically verify that we achieved what we said we would achieve.
- "In the computer field, the moment of truth is a running program; all else is prophecy" – Herbert Simon



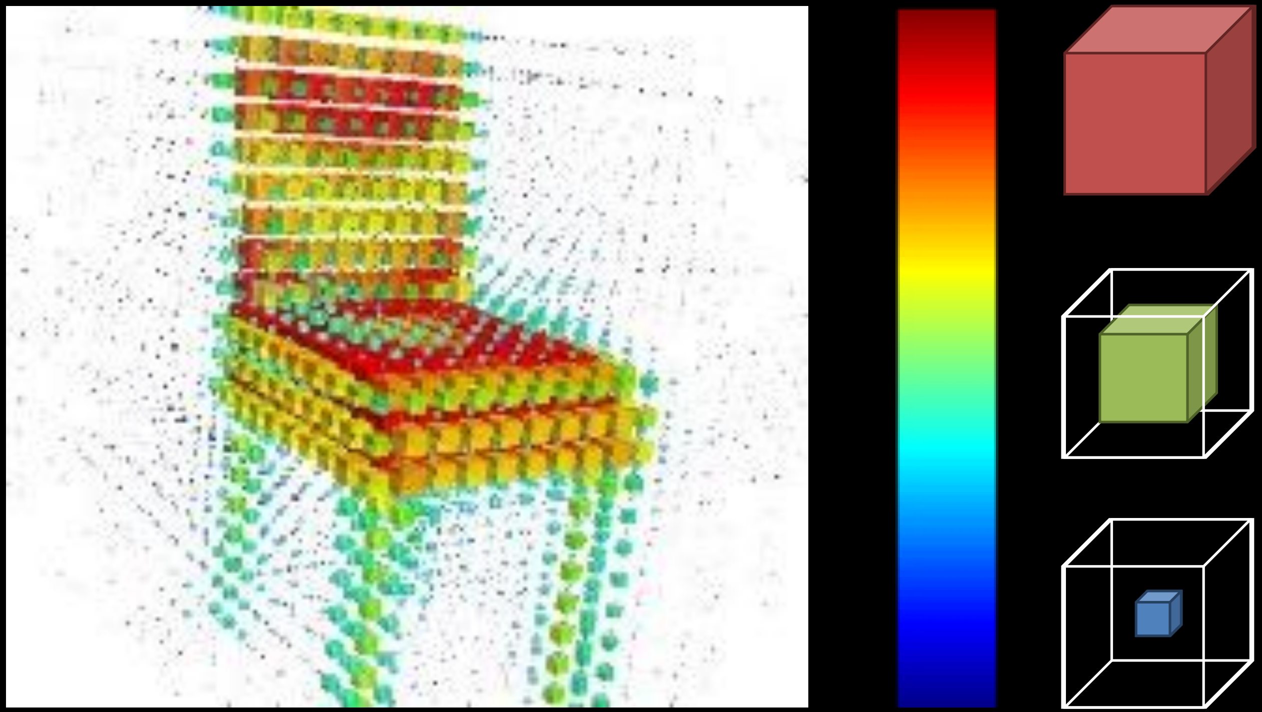
Experiments



Does it represent voxels well?



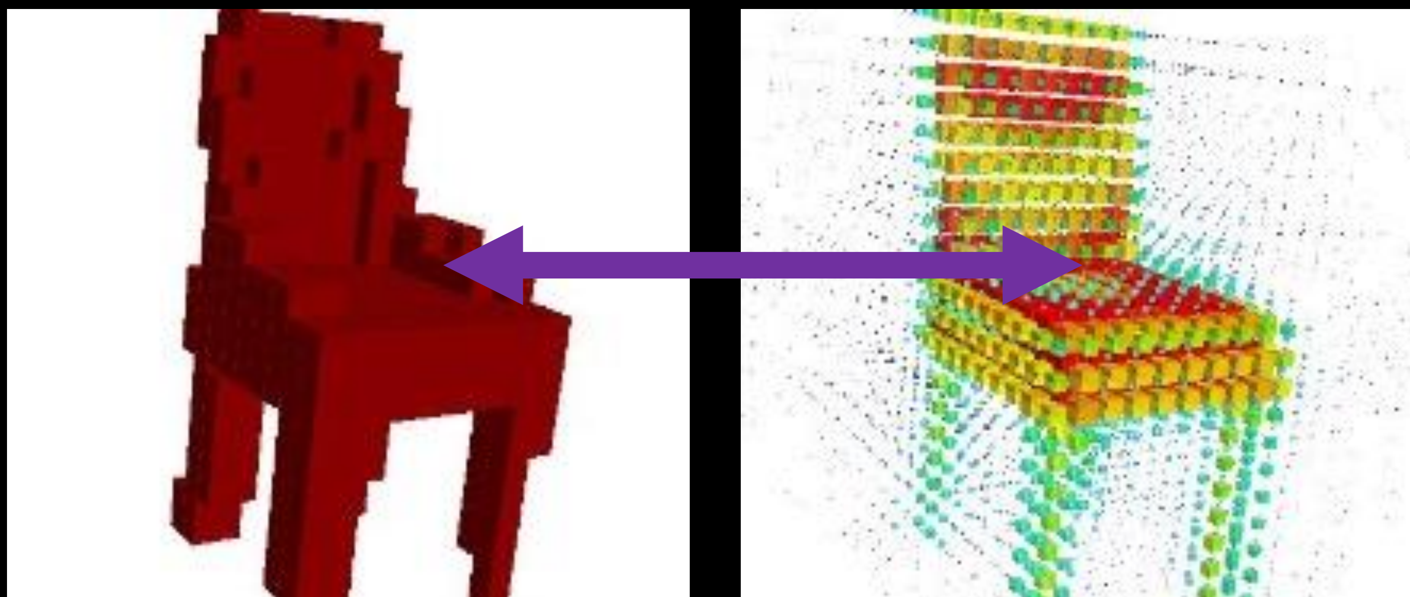
Visualization



Quantifying Performance

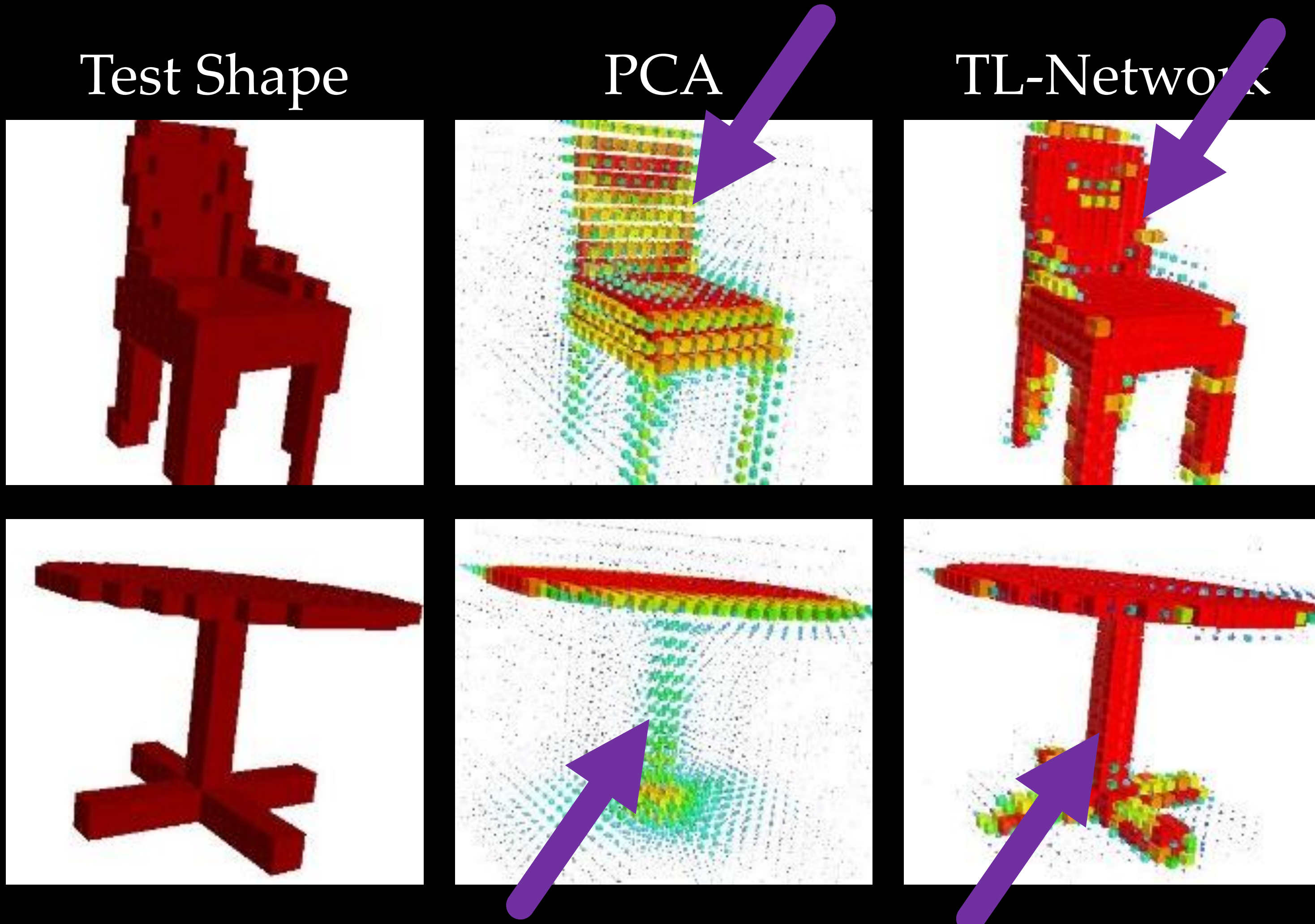
Ground-Truth Voxels

Predicted P(Occupied)



Voxel Representation

Test Shape | PCA | TL-Network



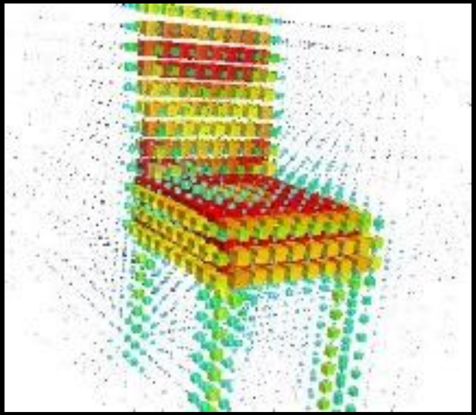
Commentary – Baselines

- **Is 95% good or bad?**
- It depends! You might want to know: how well does something simple do? How well does a known method do?
- Typically comparison points: past methods, linear models, nearest neighbors.
- Considered embarrassing if someone later finds something simple that beats your complex method!

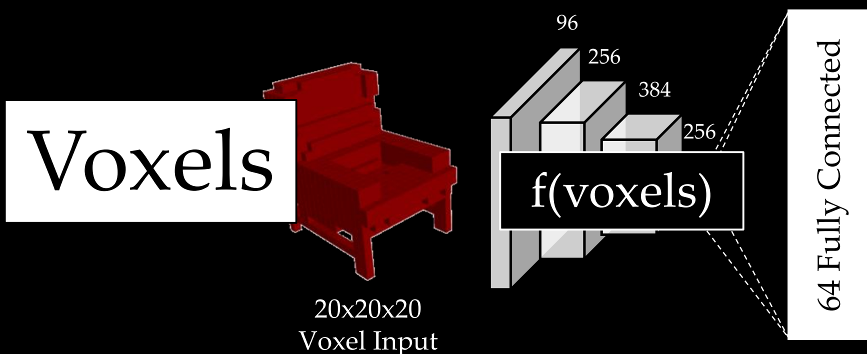
Reconstruction Accuracy

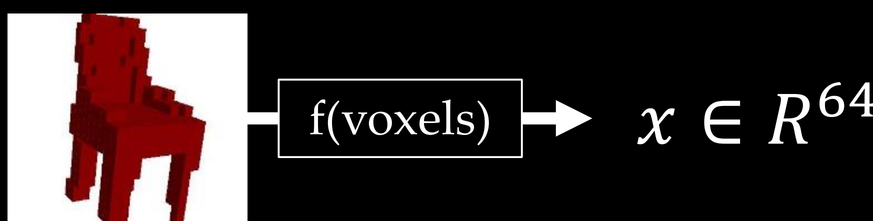
Average precision	
PCA	TL
96.8	97.6

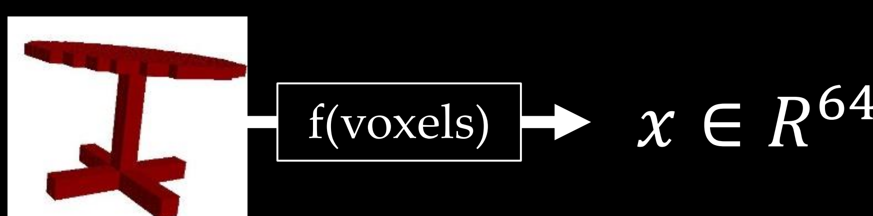
Qualitatively a pretty big gap,
but quantitatively not so.
Because metric isn't quite right.



Voxel Representation







Does this feature contain useful information for distinguishing these categories?

Commentary – Alternate Tasks

- Convnets can have hundreds of million of degrees of freedom
- Biggest fear of many researchers: are we actually learning the thing we set out to learn or something entirely different?
- This can have profound issues, especially if you deploy this
- One solution: test the features on an entirely different task

Voxel Representation

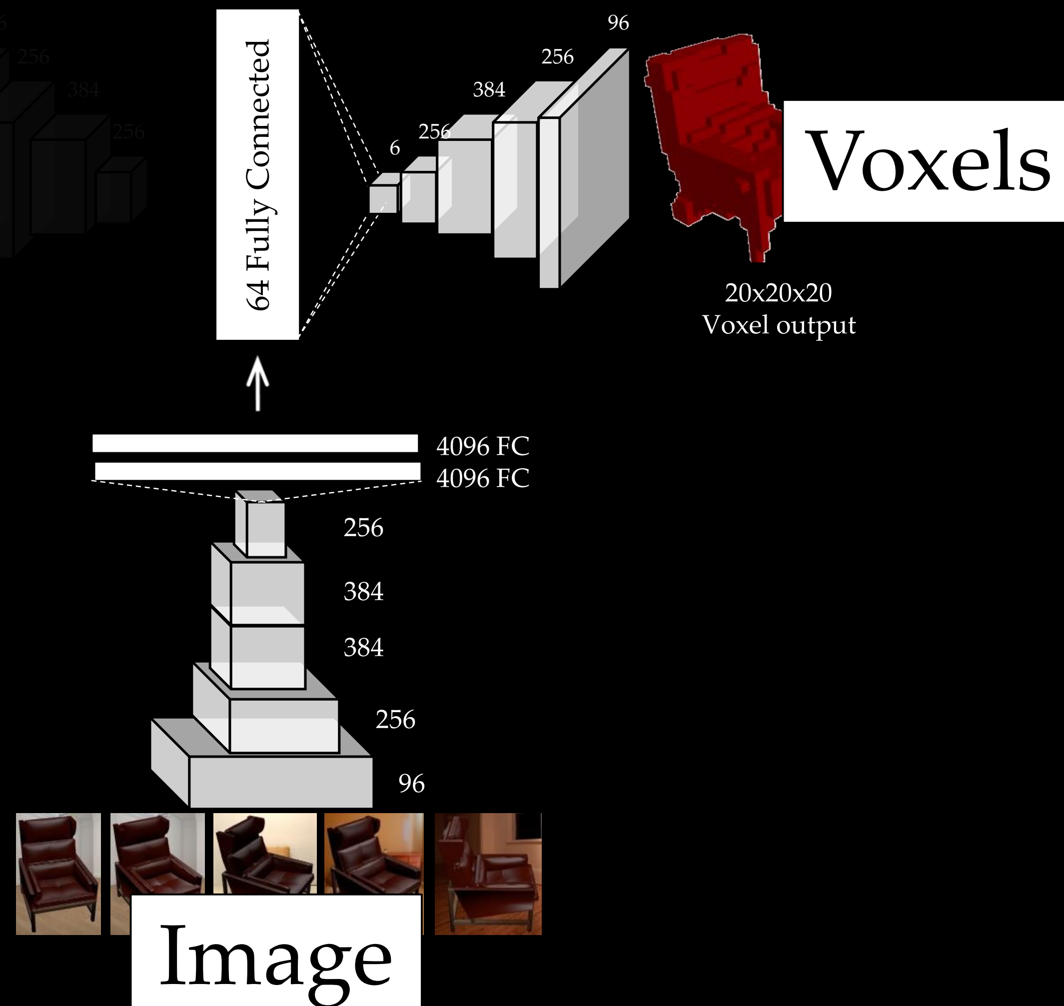
- Classification of 3D shape categories (e.g., toilet) on ModelNet40
- TL was <u>not</u> trained for this task; support vector fit on features

No Class Info.		Class Info. Used
PCA	TL	3D ShapeNets
68.4	74.4	**<u>77.3</u>**

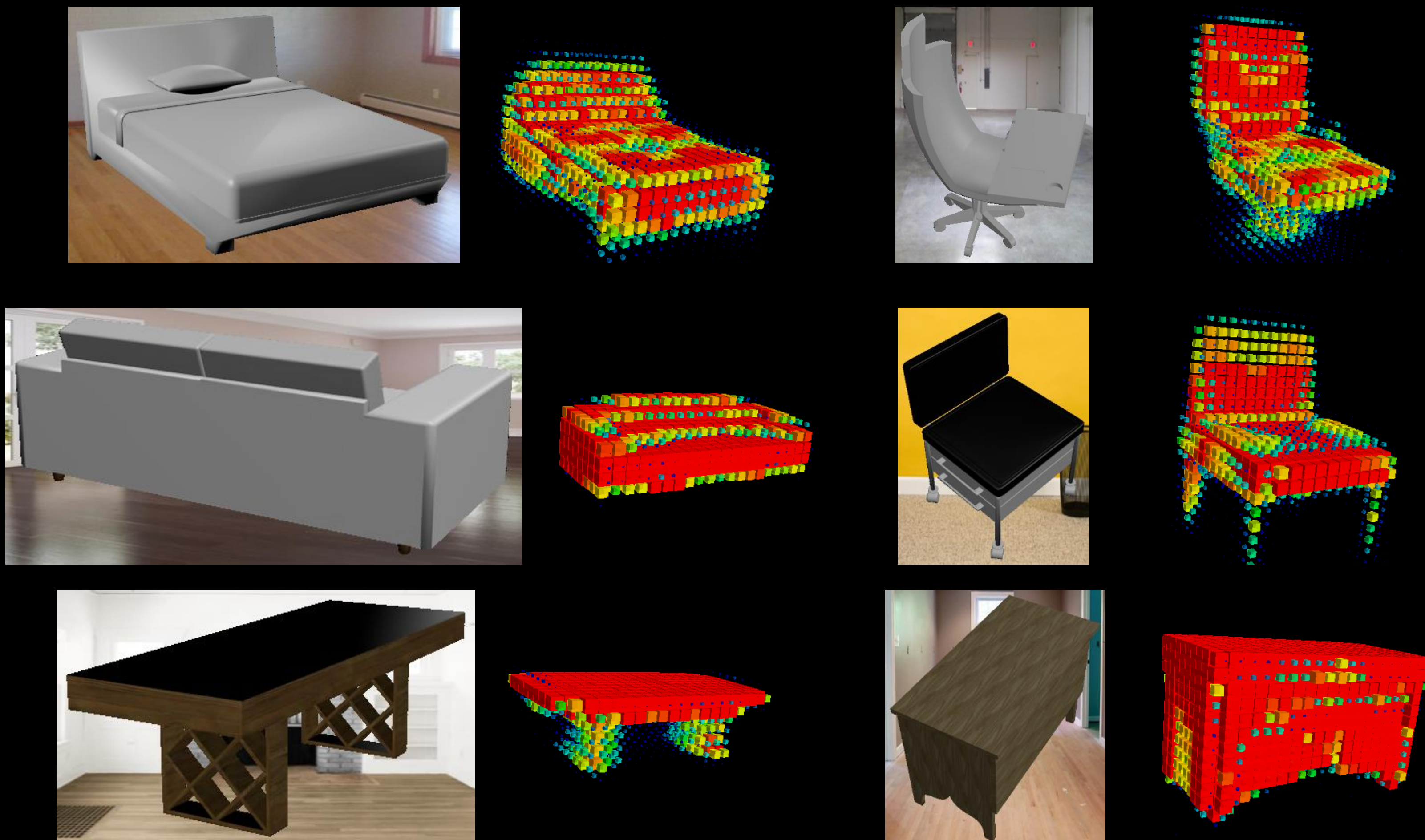
Wu et al., 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling, CVPR '15

Experiments

Can you predict voxels from images?



Reconstructing Test Models



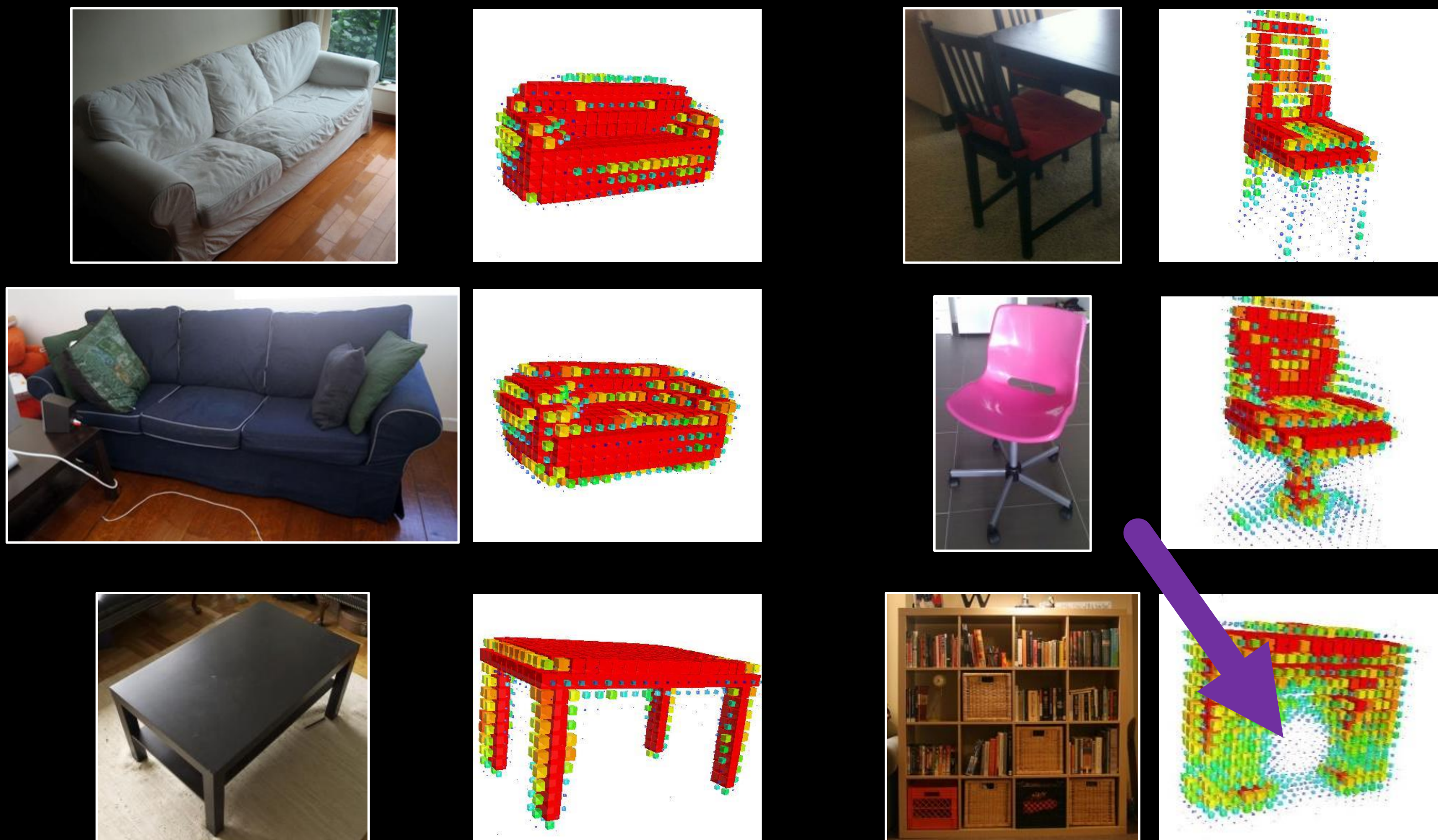
Comments

- Train on synthetic images, test on synthetic images. **Any issues?**

- Is the network actually learning something or cheating by using cues left in by the renderer

- One solution: test on new, non-rendered data

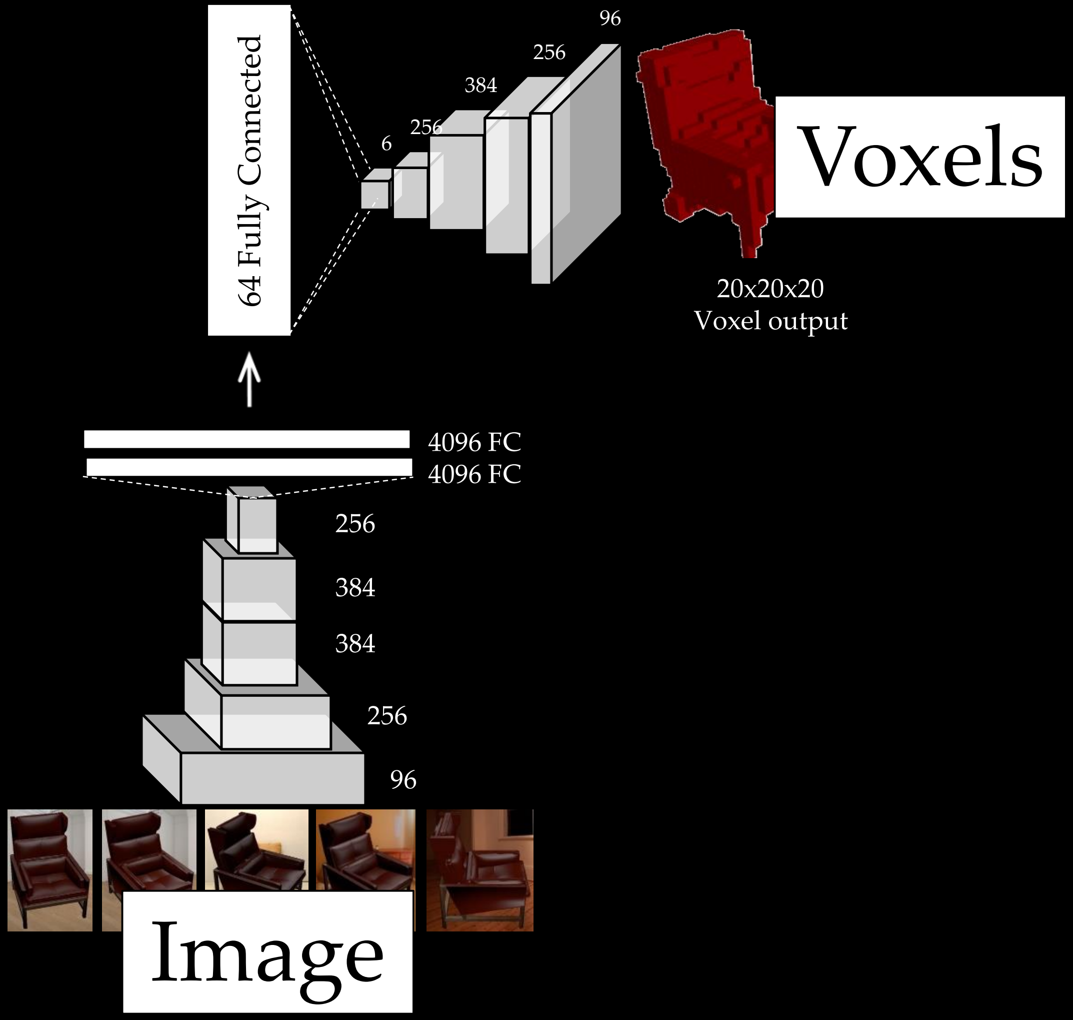
Reconstructing IKEA



Data from Lim et al., Parsing IKEA Objects: Fine Pose Estimation, ICCV 2013

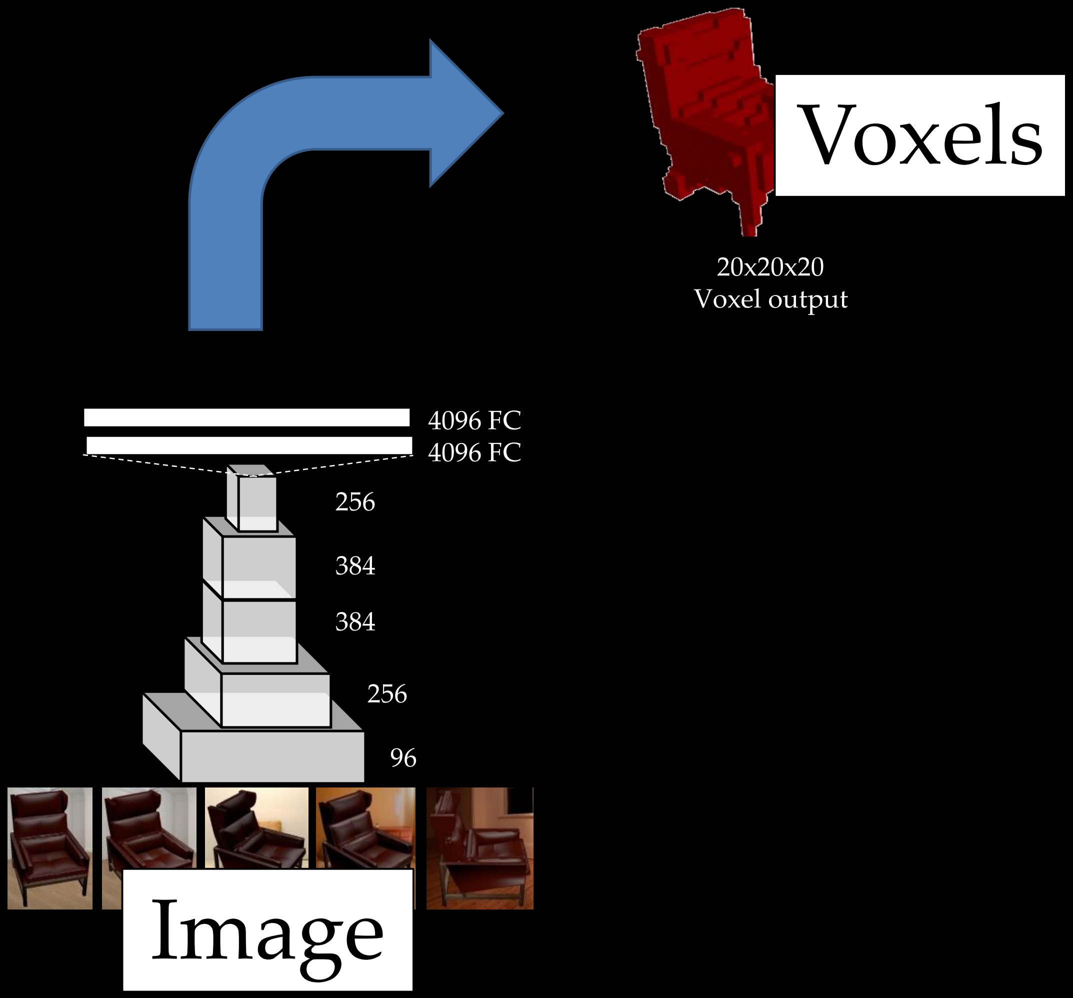
Baseline

Current
Setup



Baseline

Go directly
for voxels



Comments – Baselines

- Not quite the right baseline: just tests whether the 3D convolutional structure is necessary.
- But you don’t always get things right the first time around!
- Research is a process, and the real knowledge comes from multiple papers in a whole series, typically from different authors, not from just one paper

Quantitatively

	Direct to Voxels		
	Conv4	FC8	TL Networks
CAD	38.0	24.8	**65.4**
IKEA	31.1	19.8	**38.3**

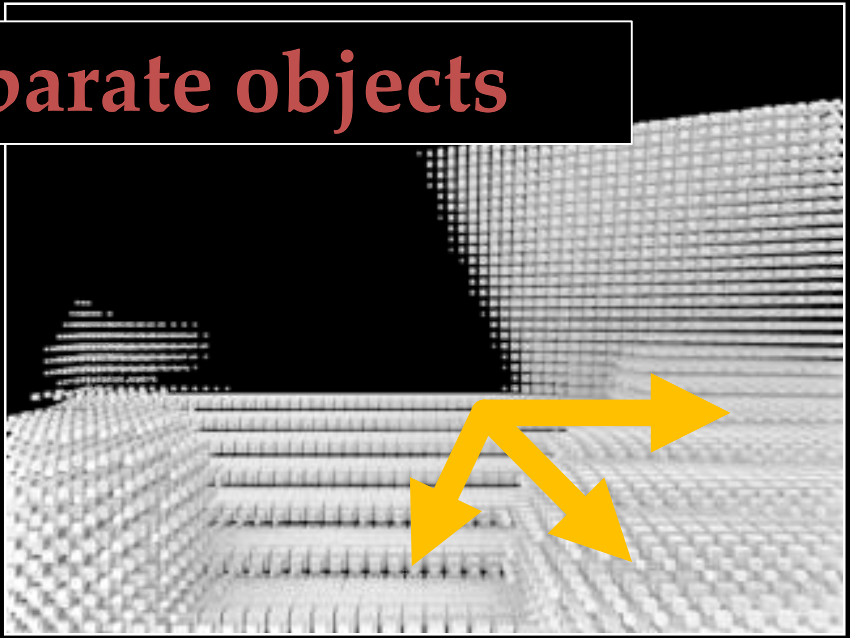
CAD

IKEA

Applying to Scenes

Input: RGB Image

Output: Voxels



1. Doesn't separate objects

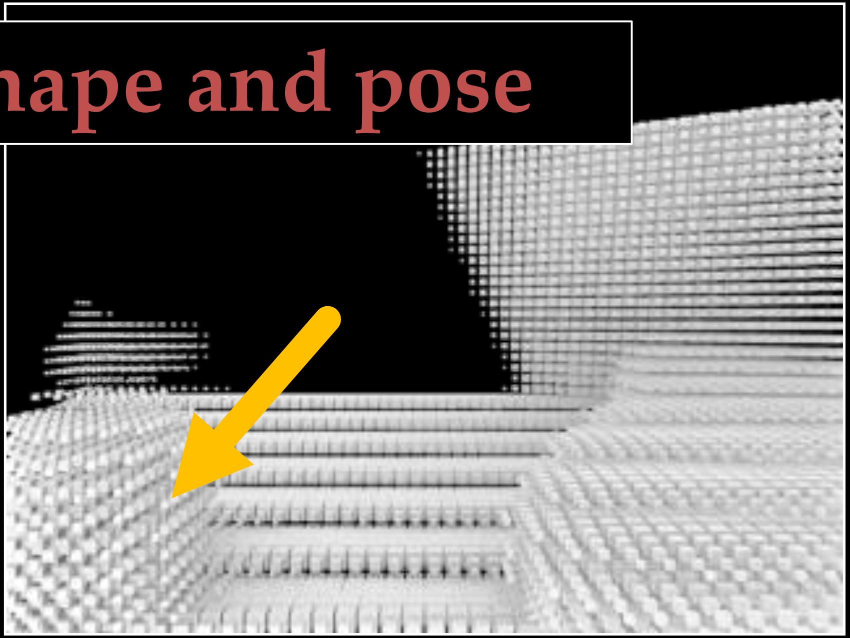
S. Tulsiani, S. Gupta, D.F. **Fouhey**, A.A. Efros, J. Malik.
Factoring Shape, Pose, and Layout from the 2D Image of a 3D Scene. To appear at CVPR 2018.

Applying to Scenes

Input: RGB Image

Output: Voxels

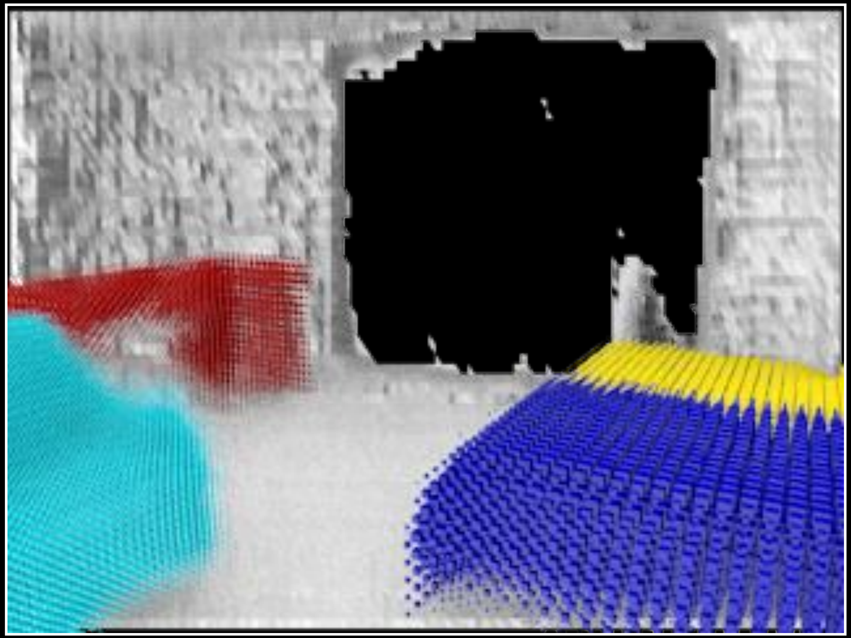
2. Conflates shape and pose



Applying to Scenes

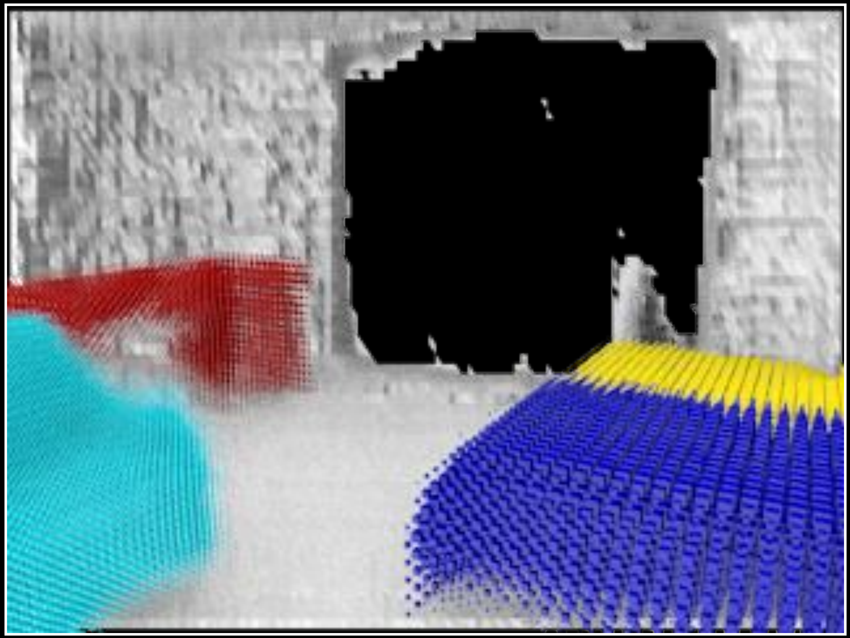
Input: RGB Image

Output: Factored

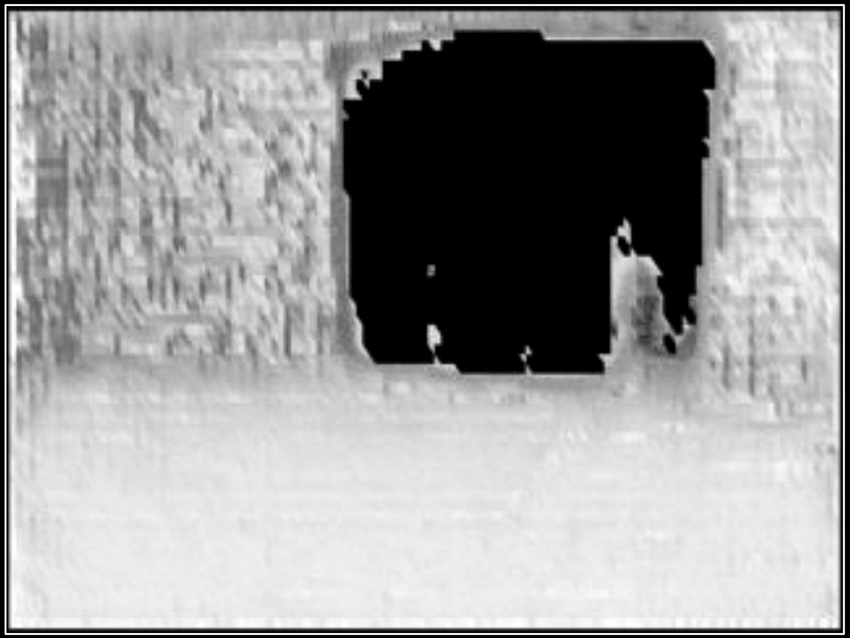


Applying to Scenes

Output: Factored

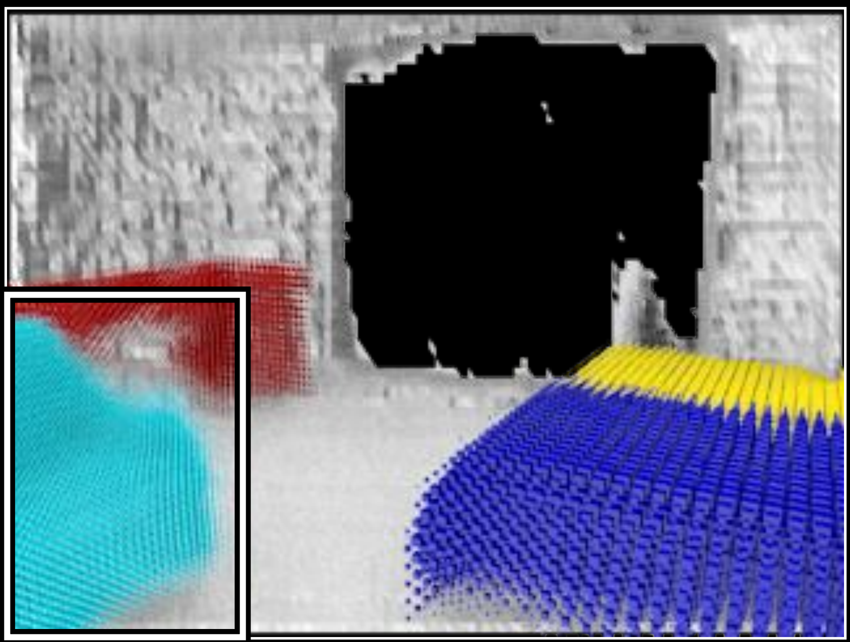


Part 1: Layout

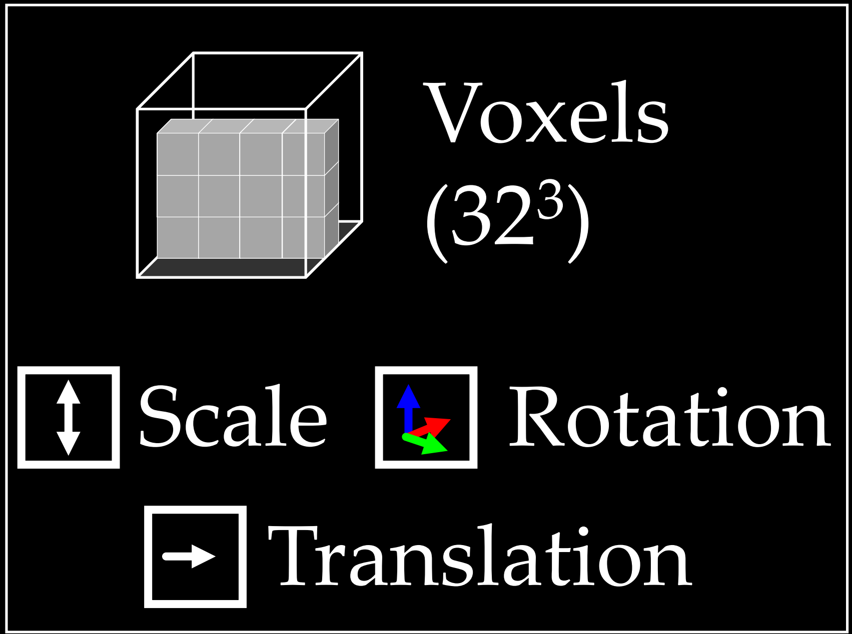


Applying to Scenes

Output: Factored

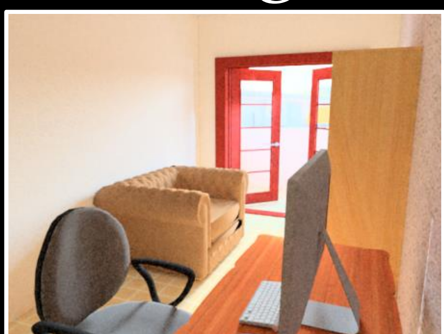


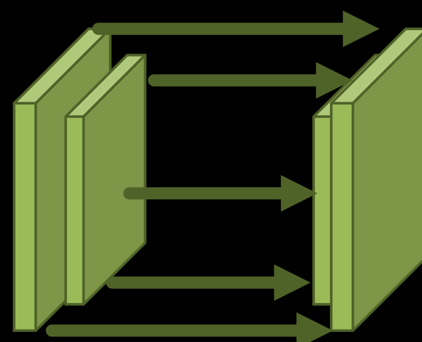
Part 2: Per-Object



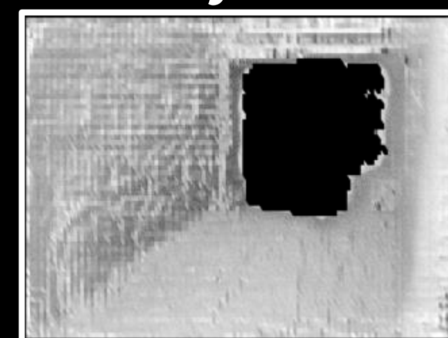
Approach

Image



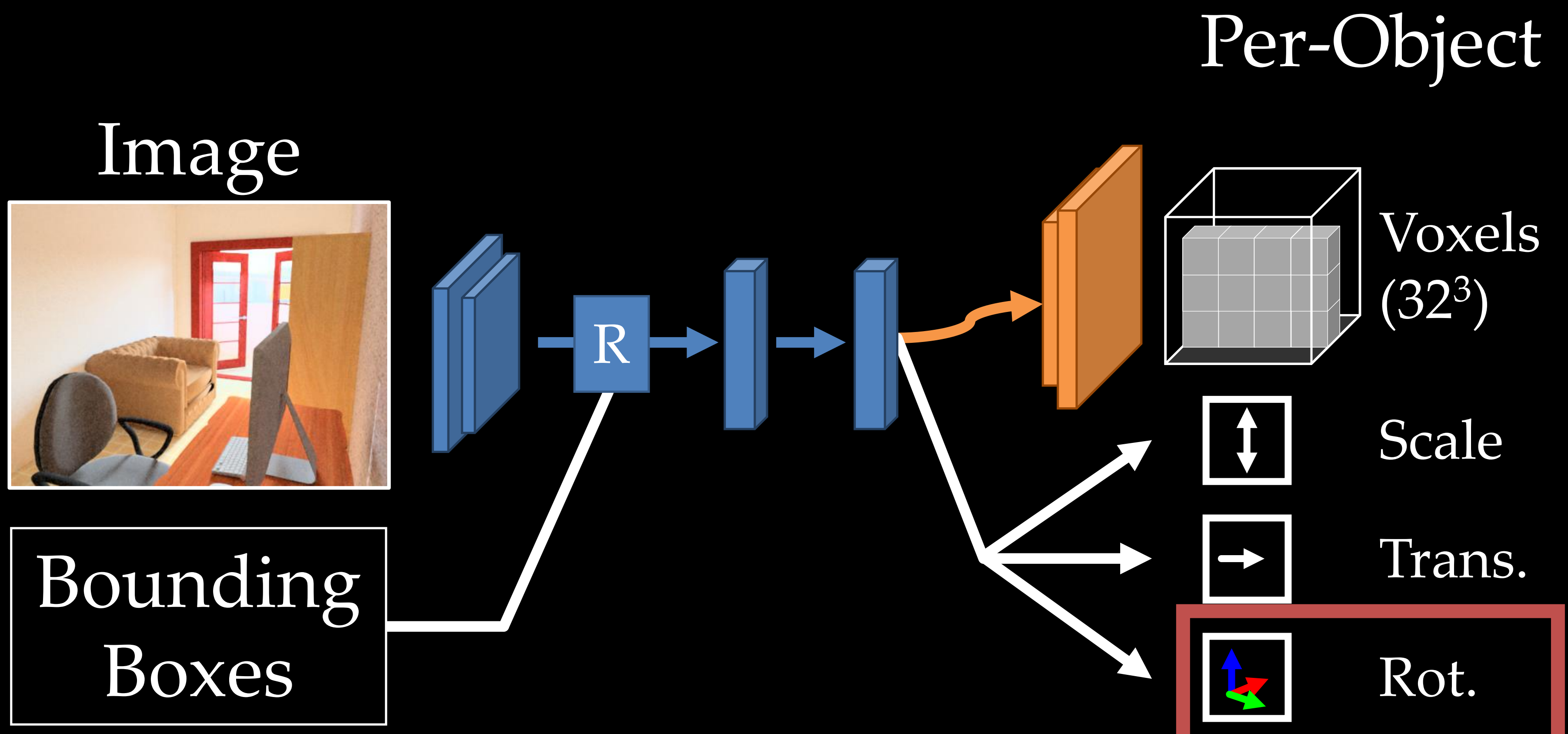


Layout



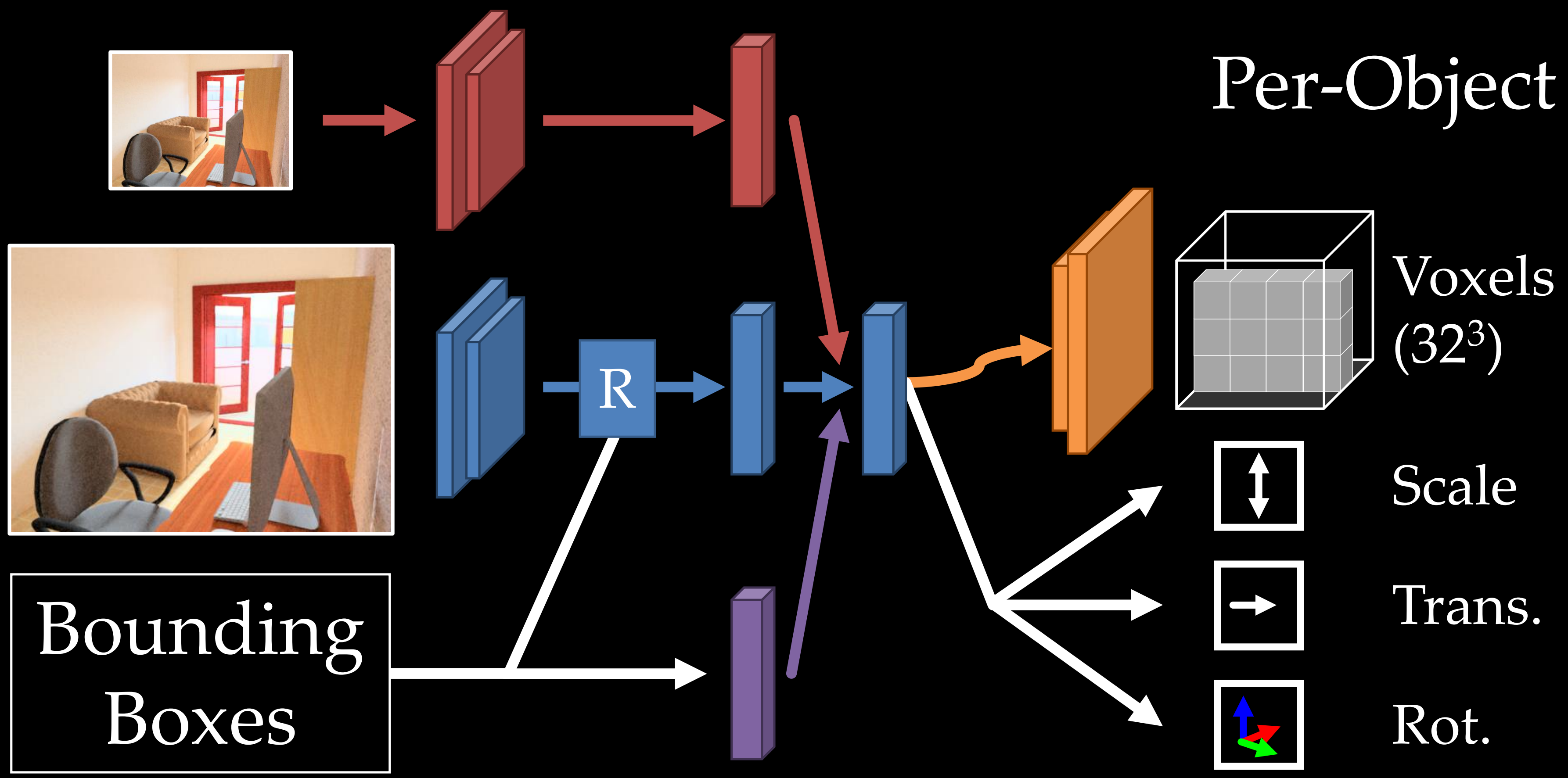
Standard encoder/decoder
with skip connections

Approach



Per-Object
Image
R
Voxels
(32^3)
Scale
Trans.
Rot.
Bounding Boxes

Approach

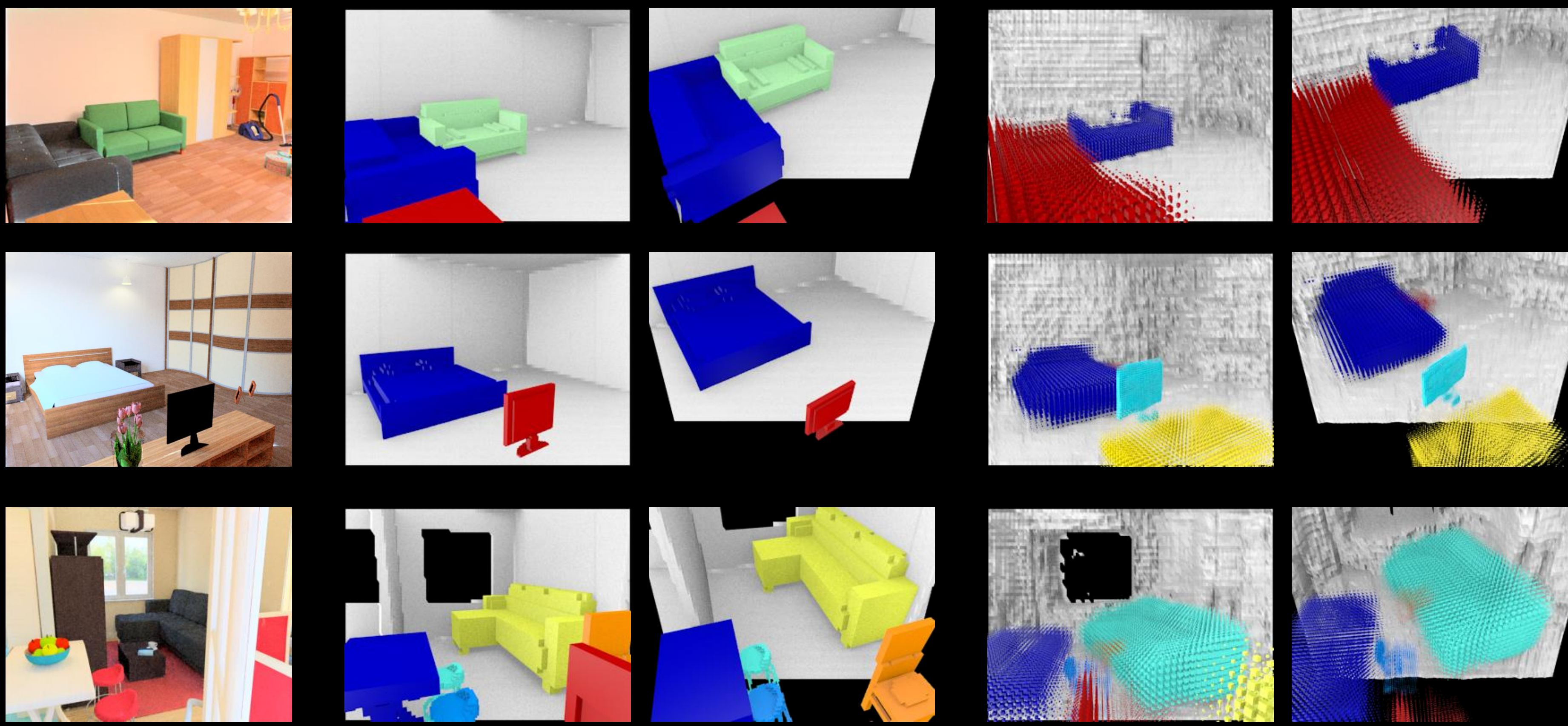


Quantitative Results

	% Rot. < 30°	% Trans. < 1m
Base	**75.2**	**90.7**
No Context	69.3	85.4
Regress Rotation	48.1	--

Results (SUNCG)

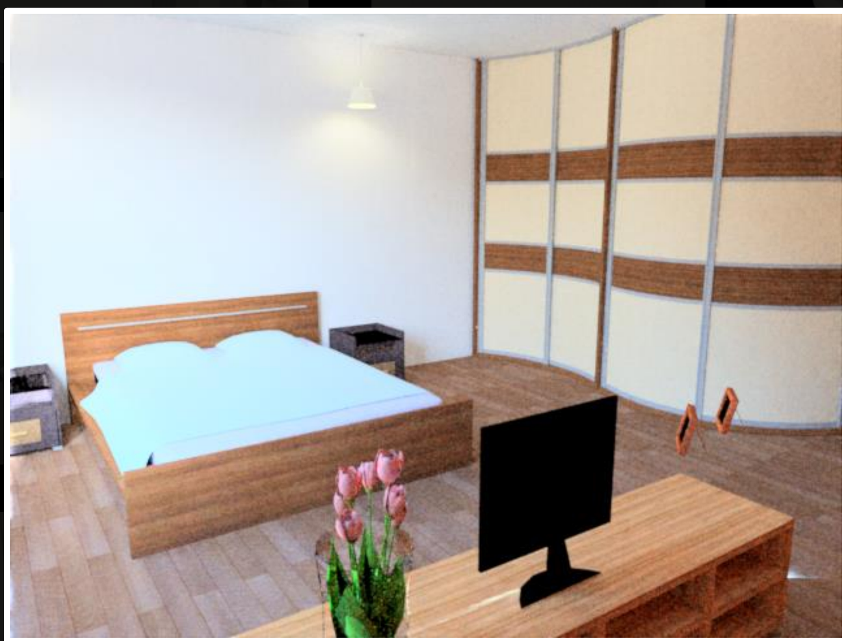
Input Ground-Truth Prediction

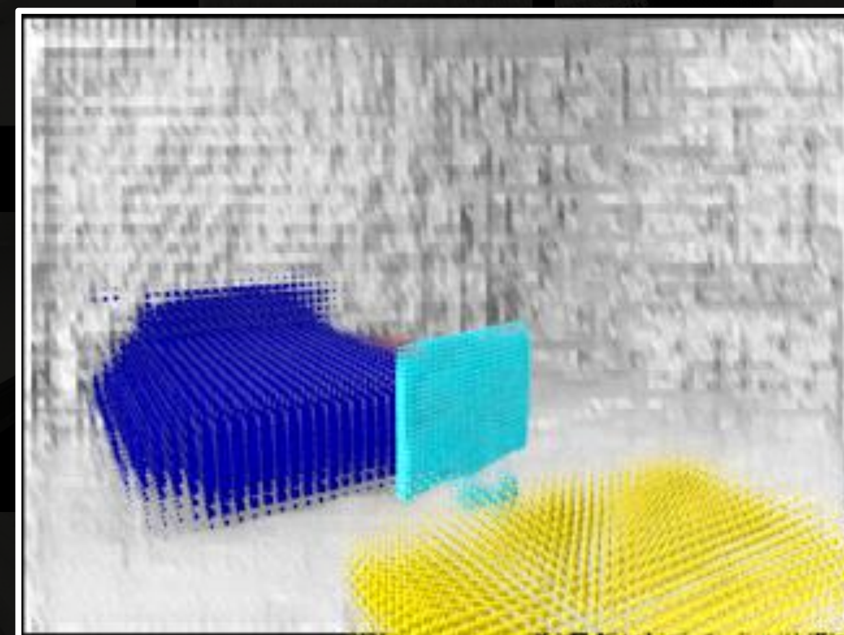


SunCG: Song et al. CVPR 2017, rendered by Zhang et al. CVPR 2017

Results (SUNCG)

Input Ground-Truth Prediction

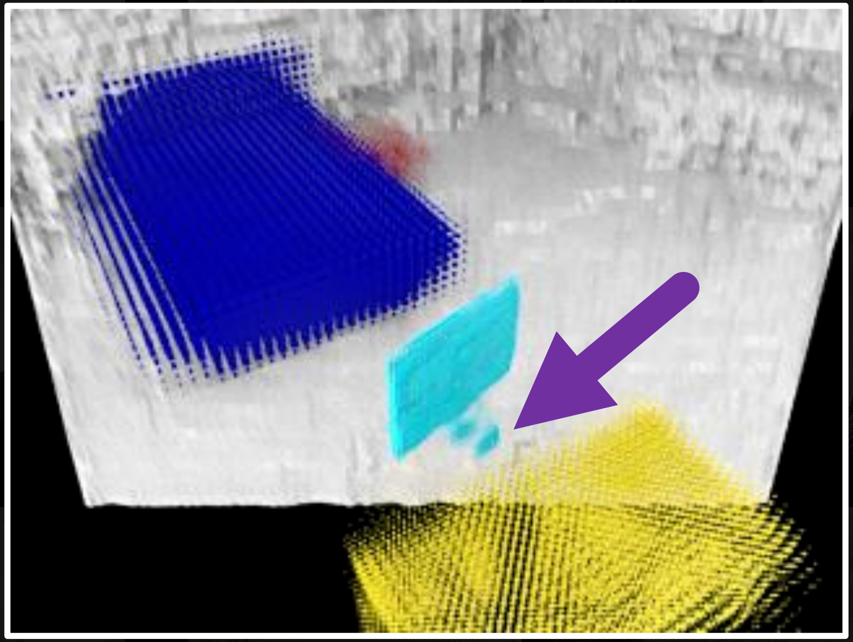




SunCG: Song et al. CVPR 2017, rendered by Zhang et al. CVPR 2017

Results (SUNCG)

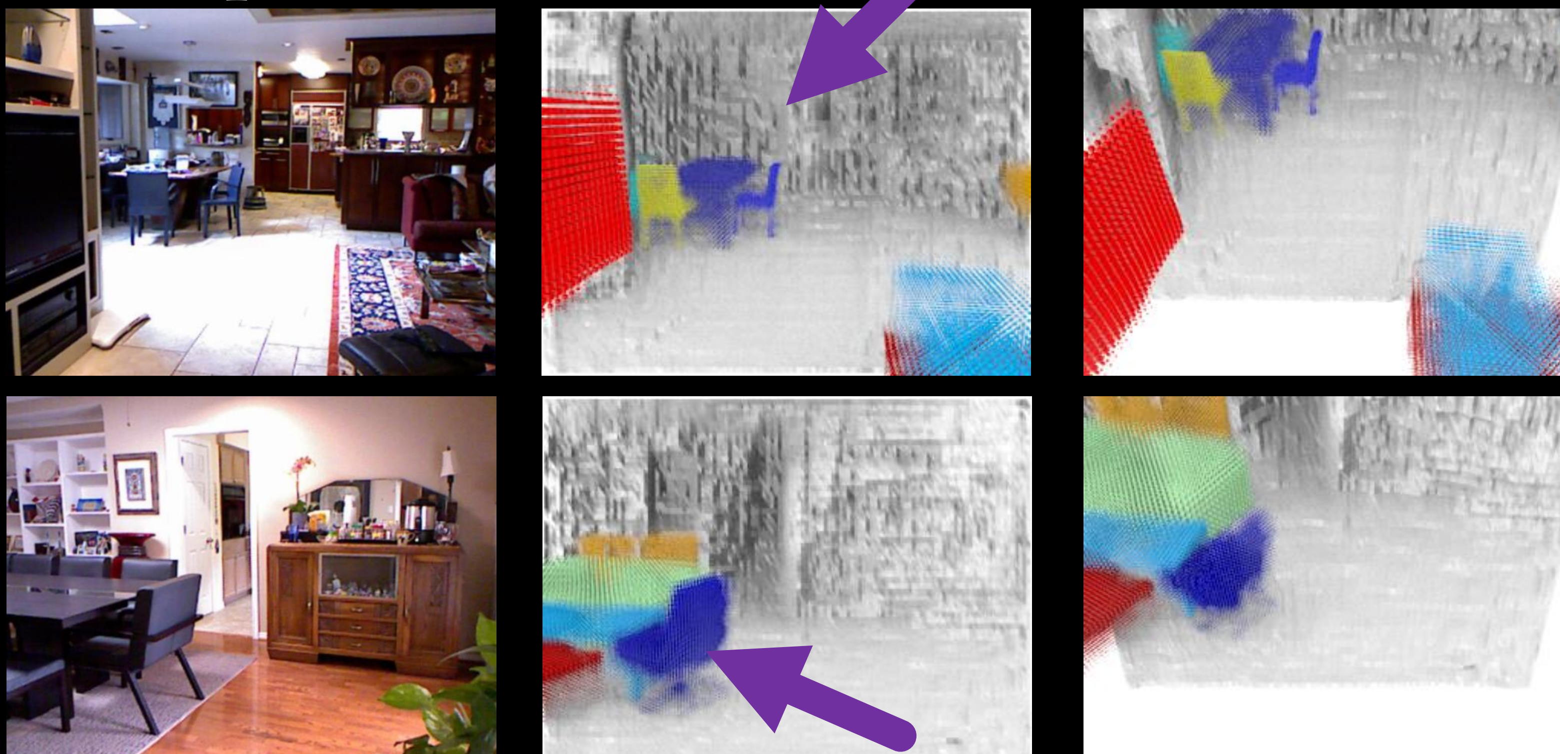
Input Ground-Truth Prediction



SunCG: Song et al. CVPR 2017, rendered by Zhang et al. CVPR 2017

Results (NYUv2)

Input | Prediction | Other View



NYUv2: Silberman et al. ECCV 2012

Representational Benefits

Input RGB Image

Output