Deep Learning in 3D

CSE P576

Vitaly Ablavsky

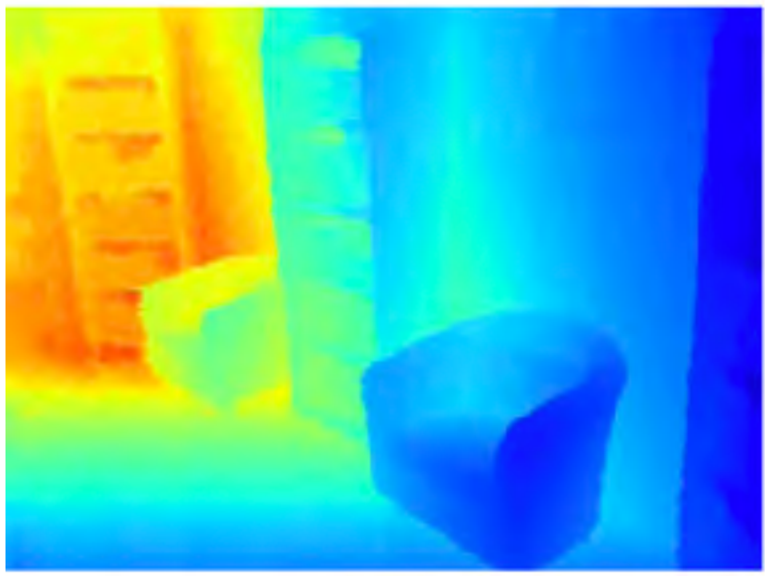
These slides were developed by Dr. Matthew Brown for CSEP576 Spring 2020 and adapted (slightly) for Fall 2021
credit → Matt
blame → Vitaly

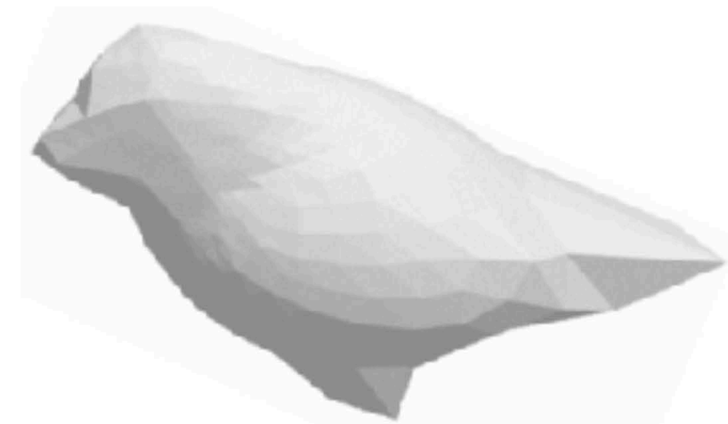
Deep Learning in 3D

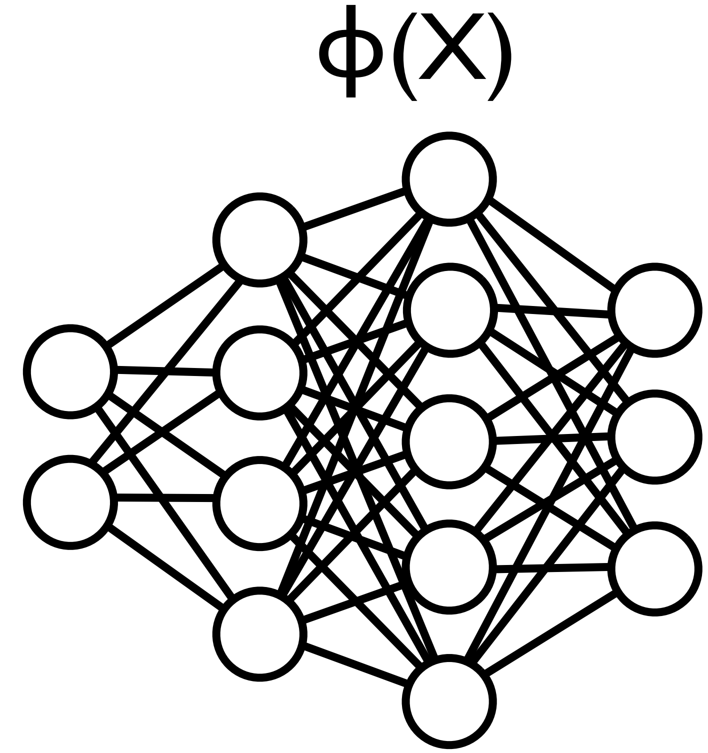
- We'll focus on predicting 3D from one or more image
- Supervision: depth, mesh, silhouettes, view supervision
- Representations: Depth, Points, Meshes, Voxels, SDFs
- Neural Scene Representation and Rendering

3D Representation

- Many ways to represent objects in 3D



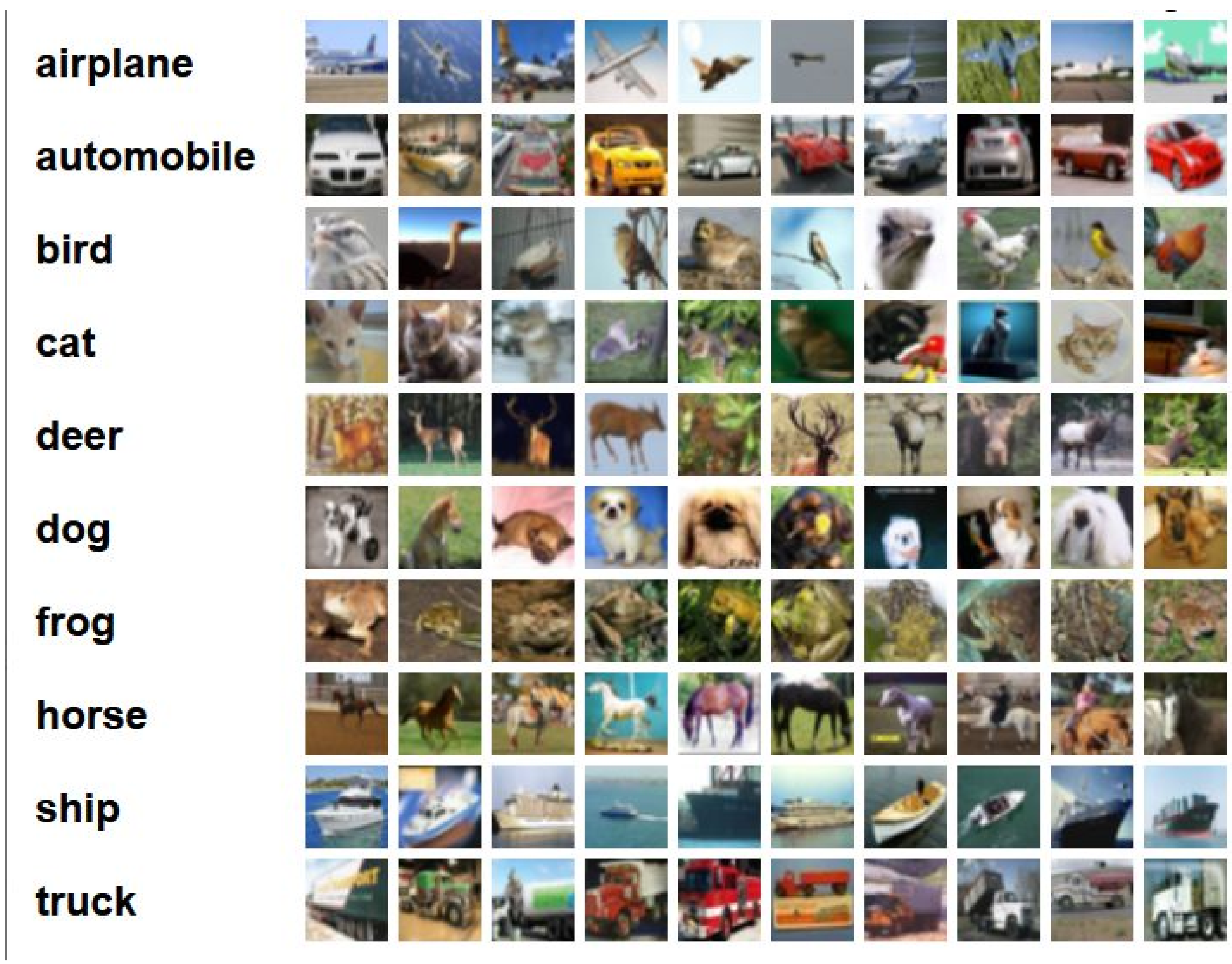


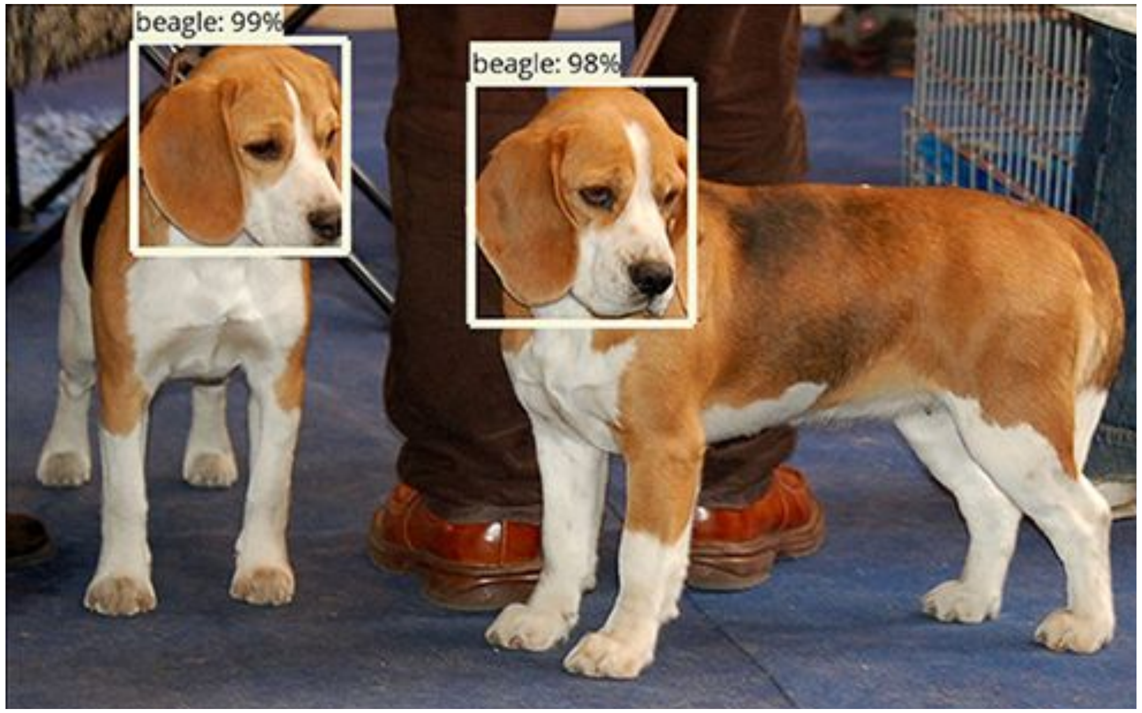
Learning in 3D
Is a Different Learning Task

Previous Lectures

Whole-image classification



Object detection

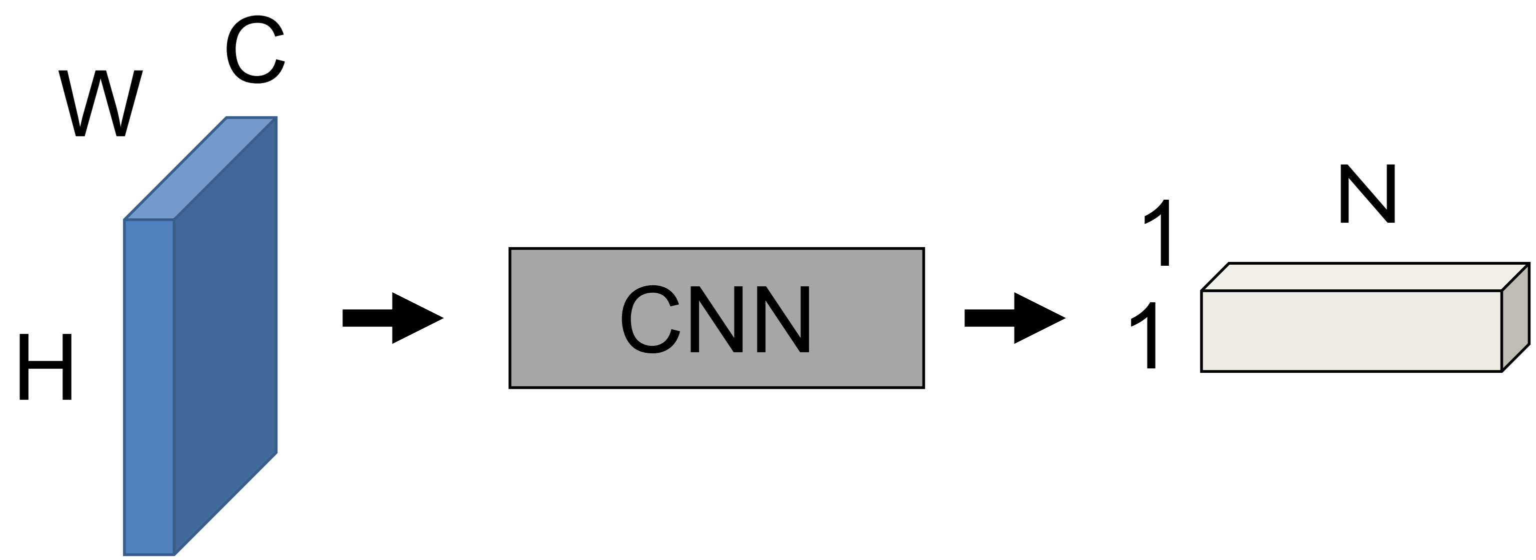


Pixel Labelling

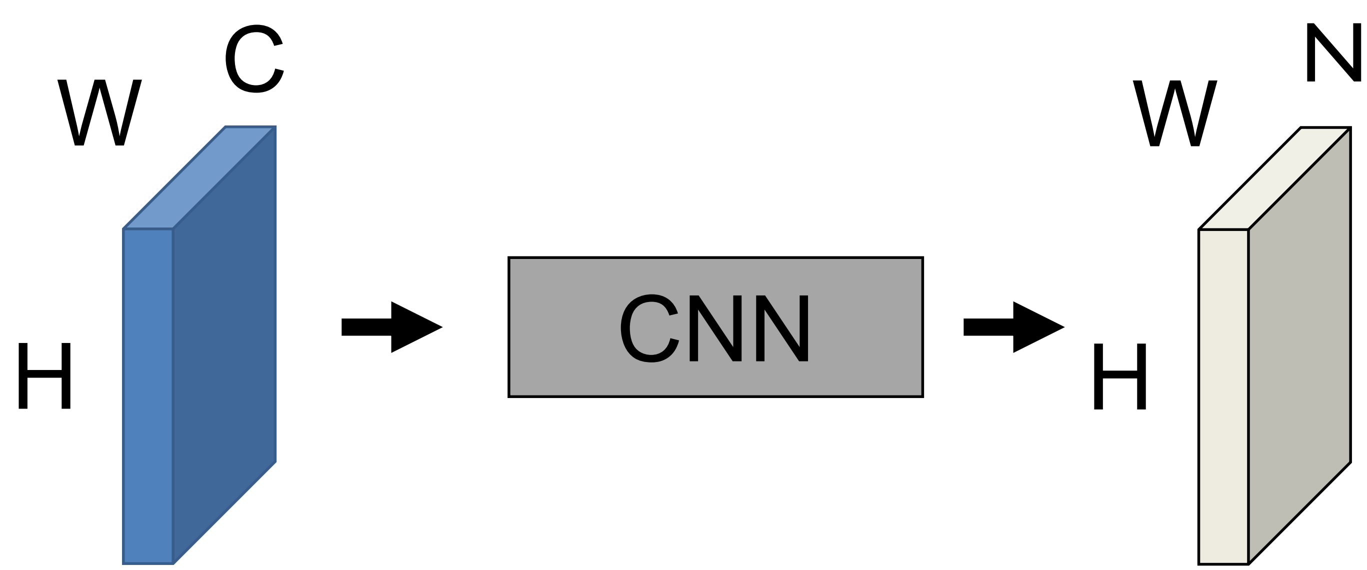
- Per-Pixel Regression + Classification, Examples, Architectures
- Depth Estimation: direct vs self supervised, pretraining
- Super-Resolution, Colorization, Image Translation

Pixel vs Image Labelling

- Image labelling, e.g., classification (N class scores per image)



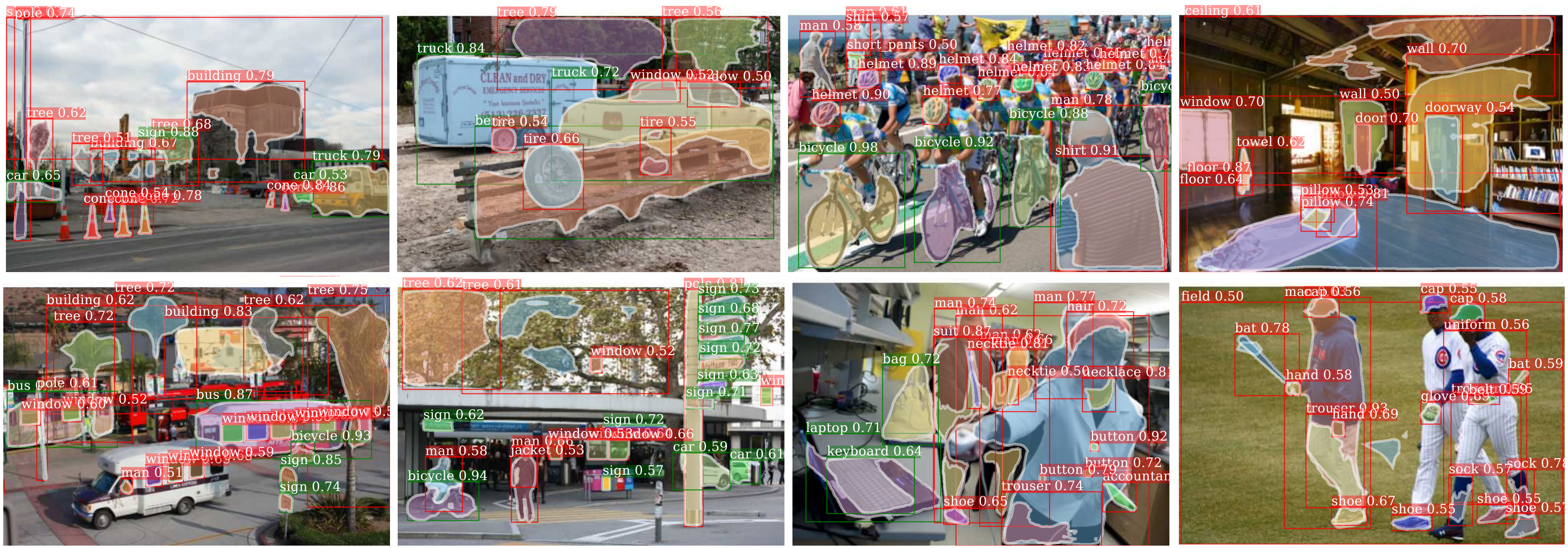
- Pixel labelling, e.g., segmentation, depth estimation, superres, (N class scores, depth, RGB value etc. per pixel)



[David Fouhey]

Segmentation

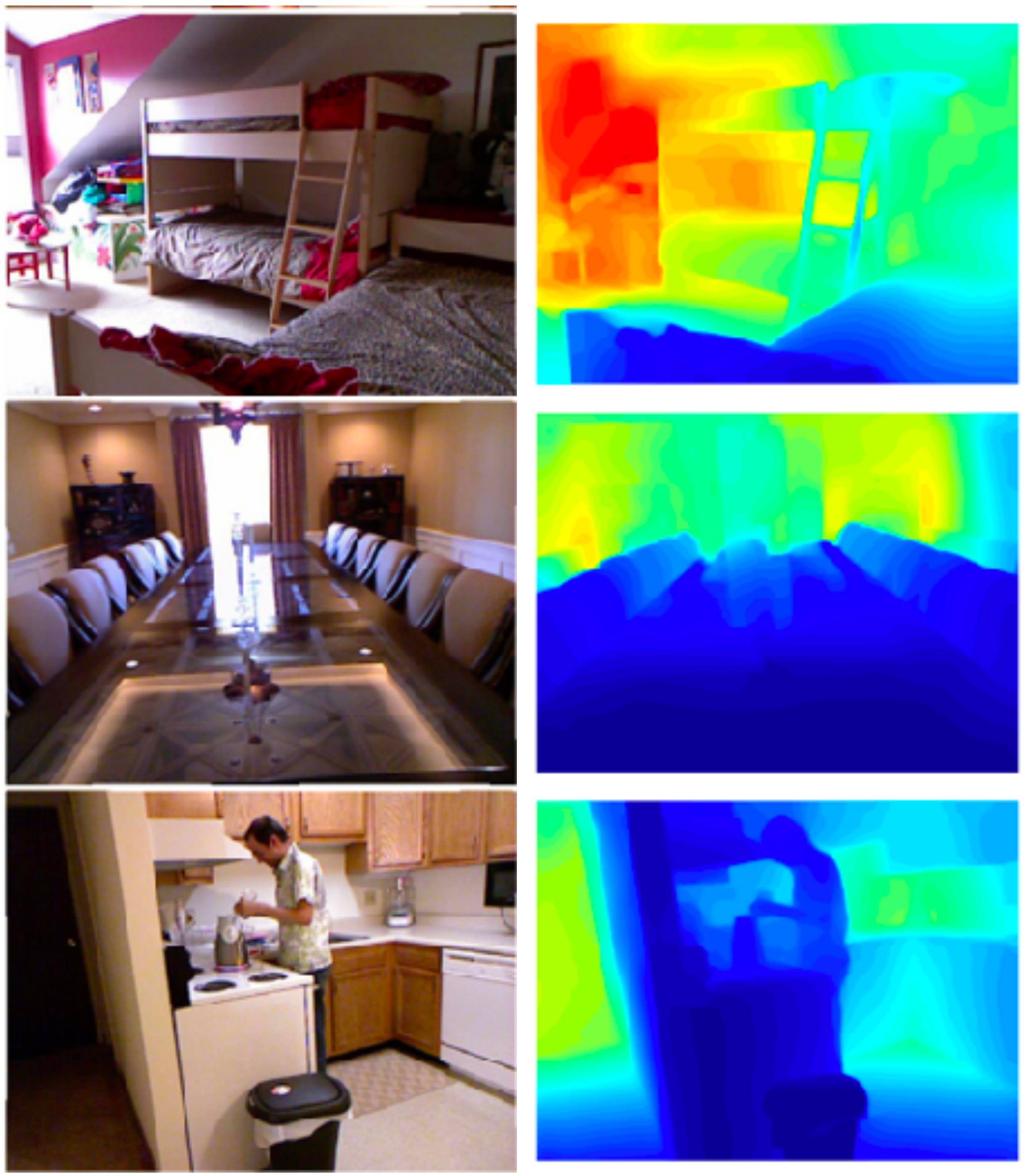
- Predict object identity and/or category per pixel



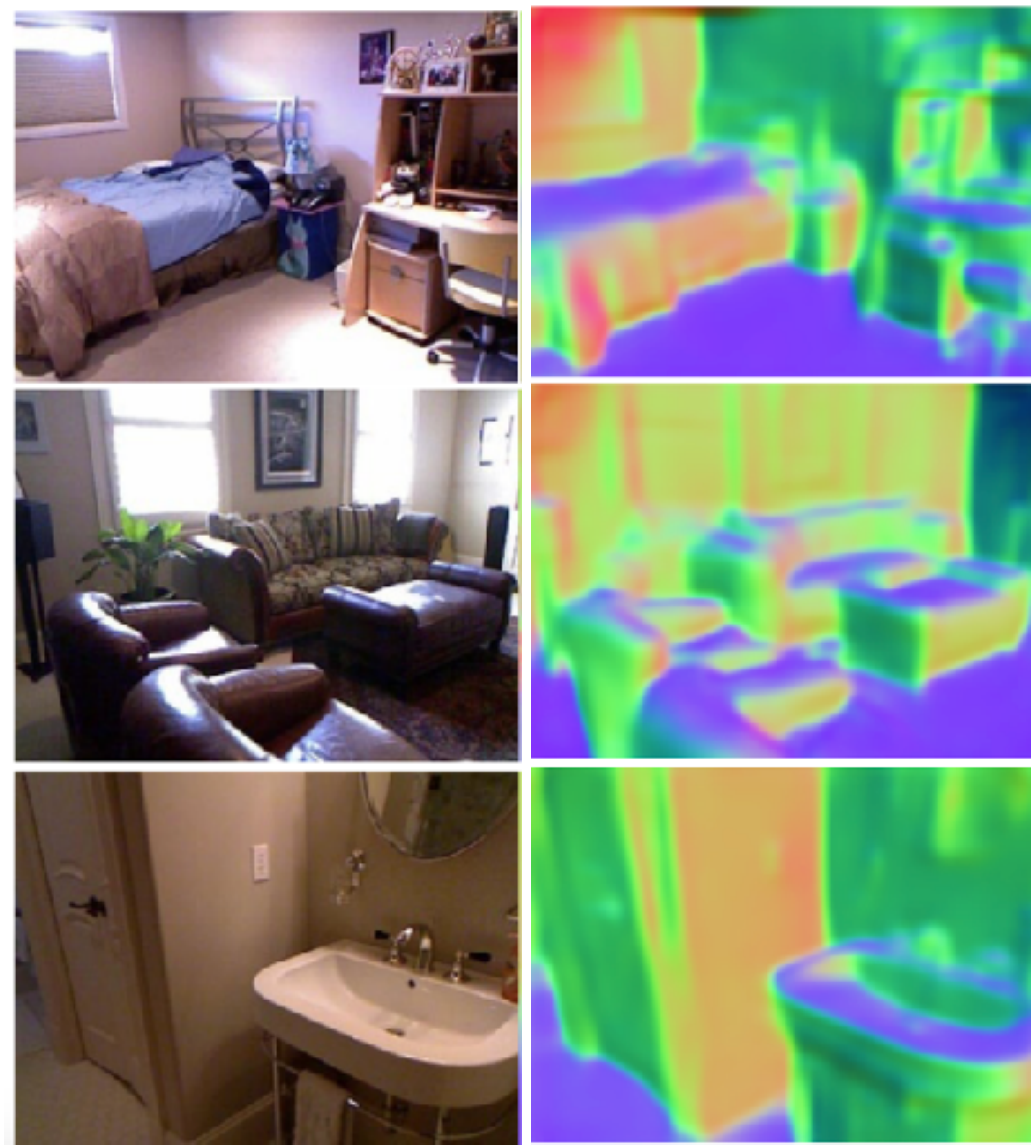
[Hu et al 2017]

Depth + Normals Estimation

- Predict depth or surface normal per pixel, given RGB input



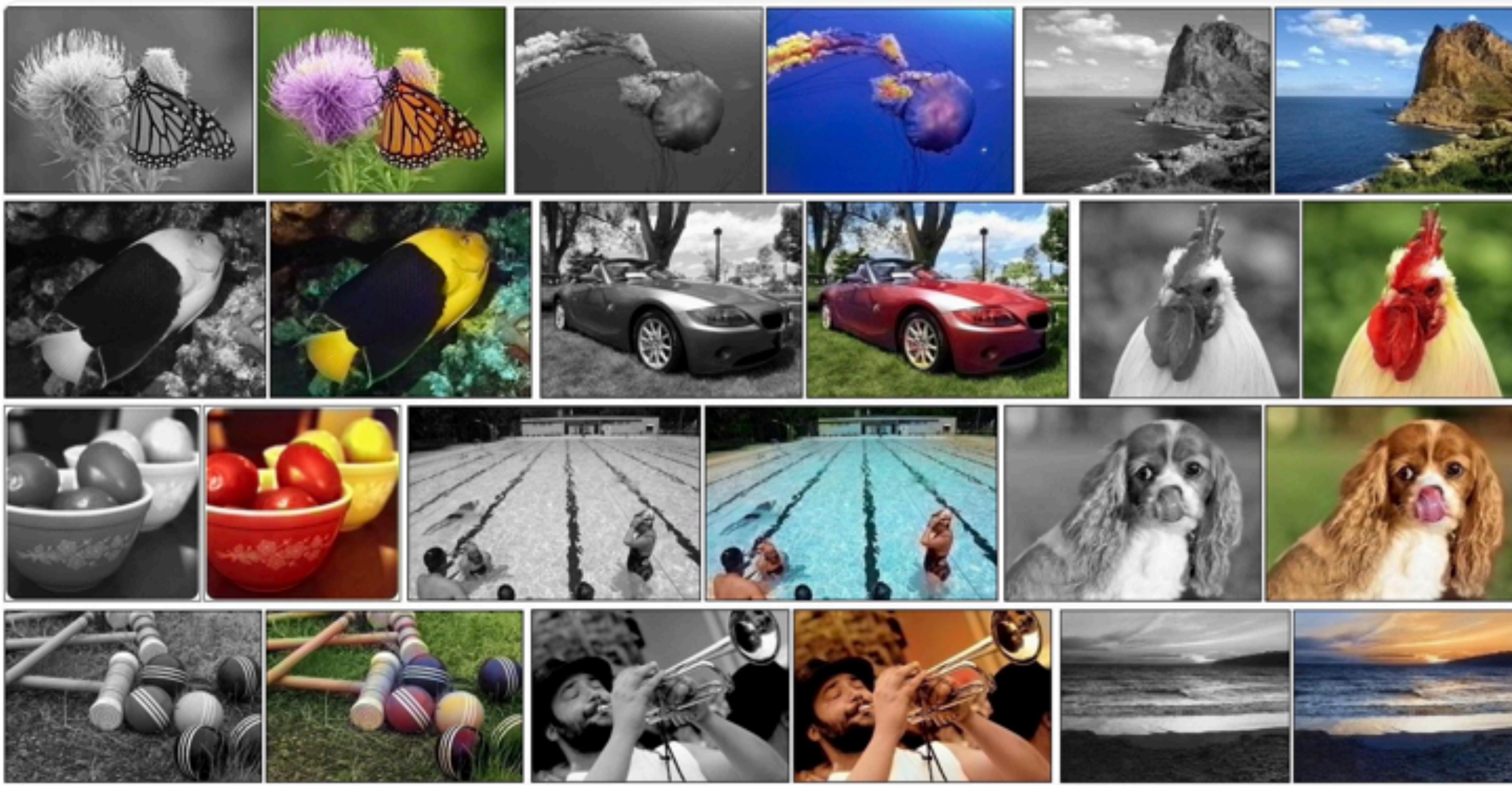
[Alhashim Wonka 2019]



[Eigen Fergus 2015]

Image Colorization

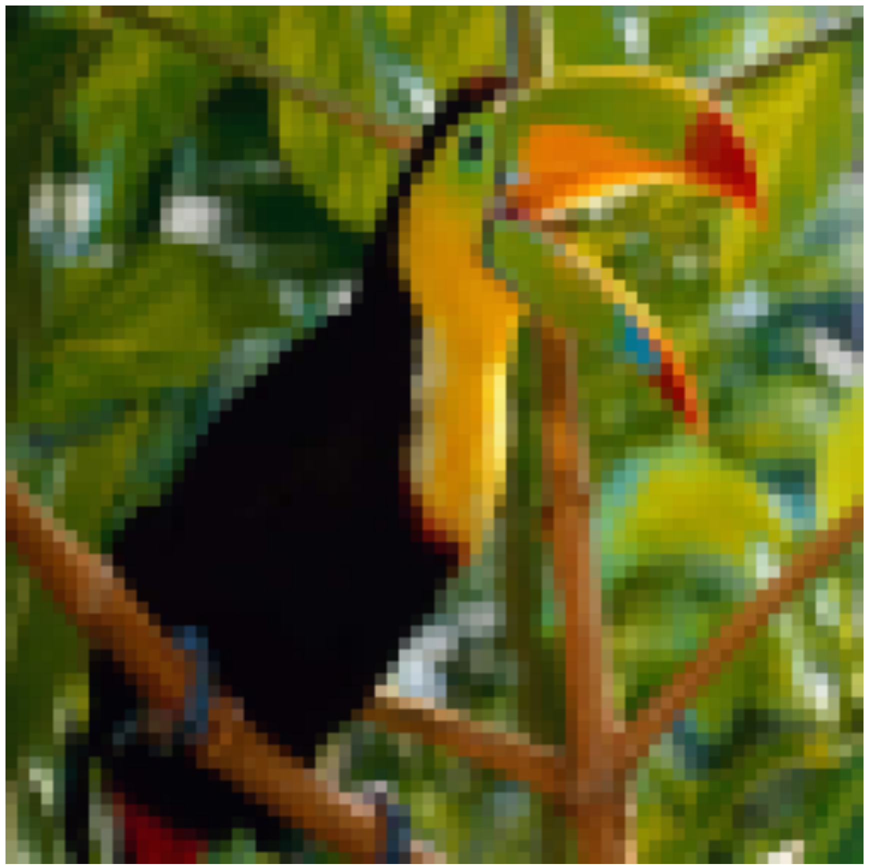
- Predict color per pixel, given grayscale input



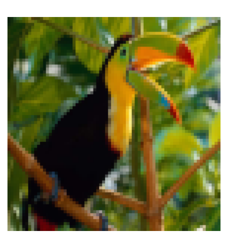
[Zhang et al. 2016]

Super-Resolution

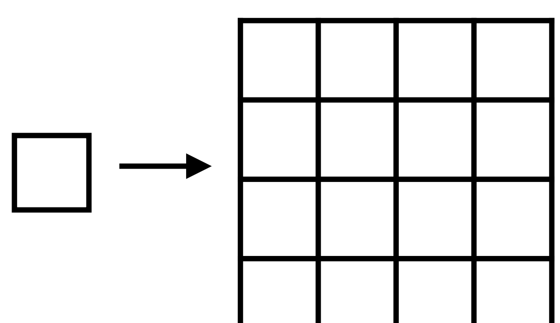
- Predict high resolution RGB, given low resolution RGB input



4 x downsampled

real size = 

bicubic upsample

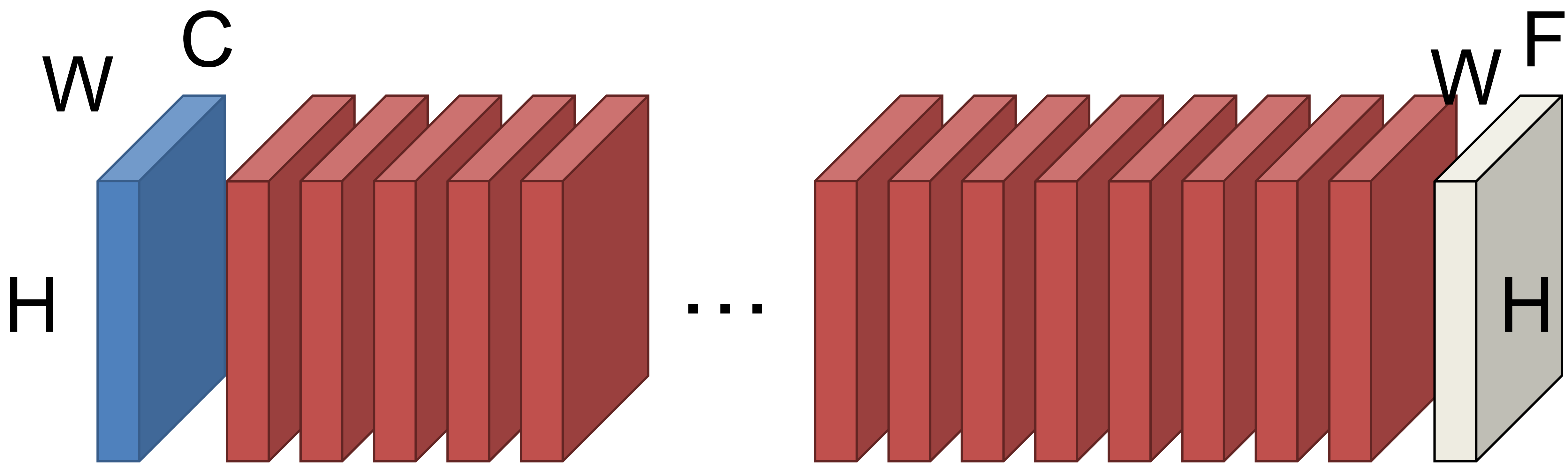


4 x superresolution

1 pixel → 16 pixels

[Ledig et al. 2017]

Why Not Stack Convolutions?



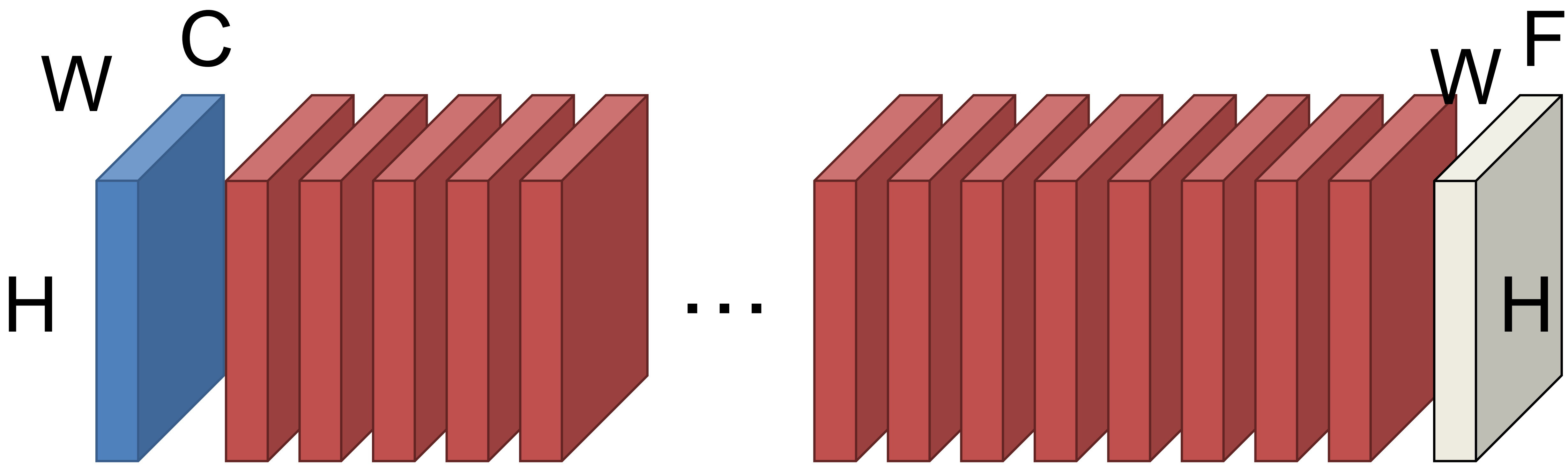
n 3x3 convs have a receptive field of 2n+1 pixels

How many convolutions until >=200 pixels?

100

[David Fouhey]

Why Not Stack Convolutions?



Suppose 200 3x3 filters/layer, H=W=400

Storage/layer/image: 200 * 400 * 400 * 4 bytes = 122MB

Uh oh!*

*100 layers, batch size of 20 = 238GB of memory!

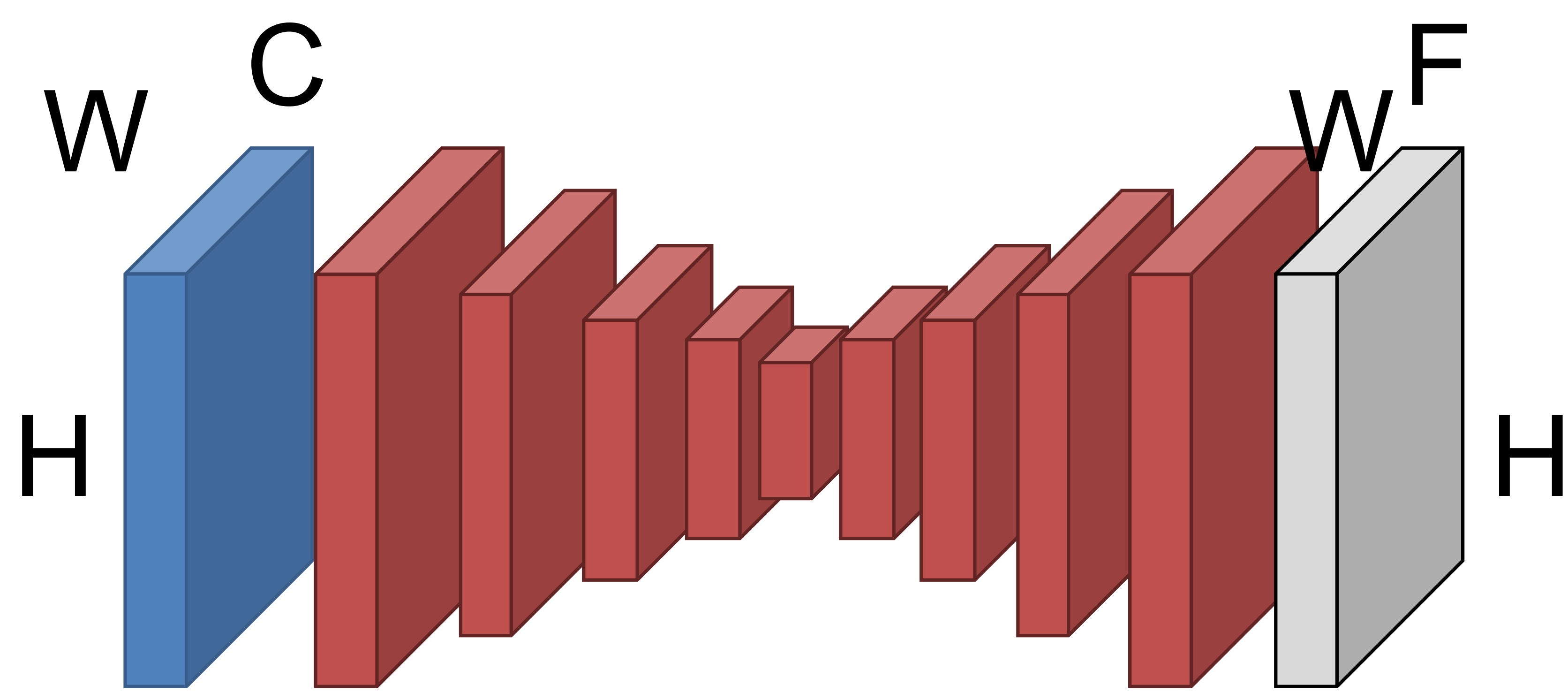
[David Fouhey]

Encoder-Decoder

Key idea: First **downsample** towards middle of network. Then **upsample** from middle.

How do we downsample?

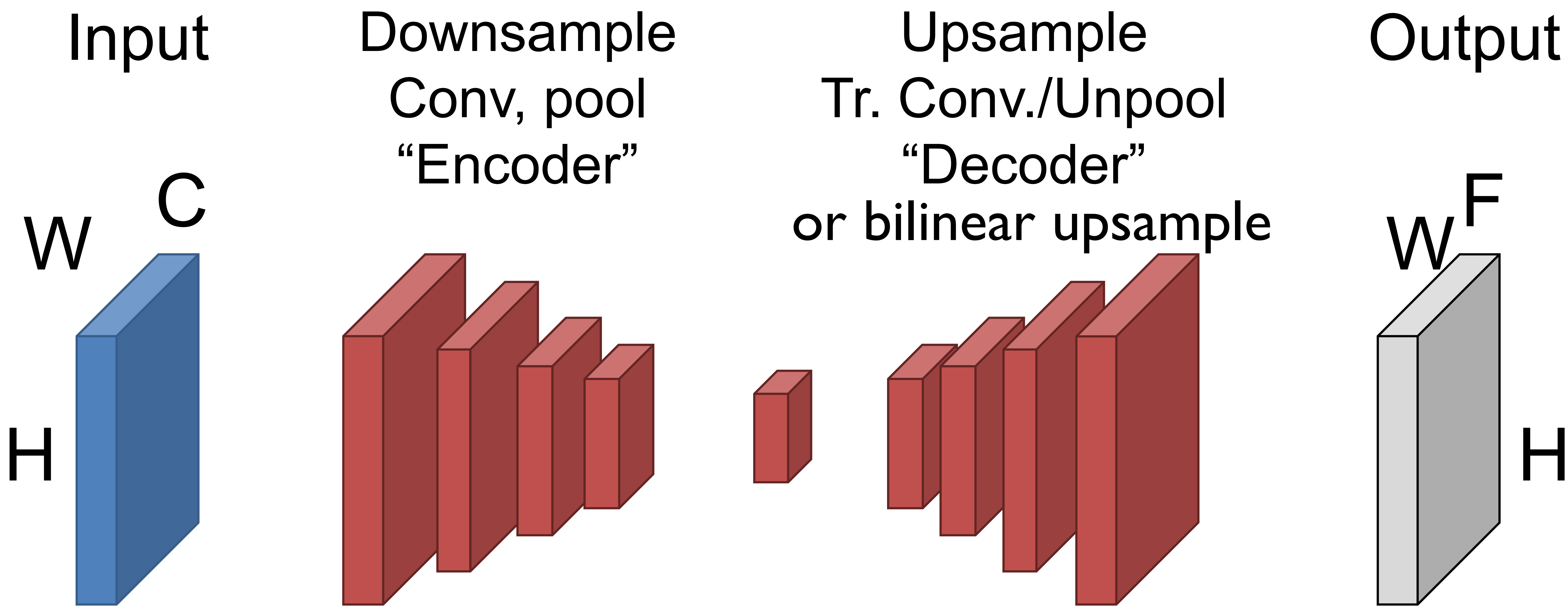
Convolutions, pooling



[David Fouhey]

Putting it Together

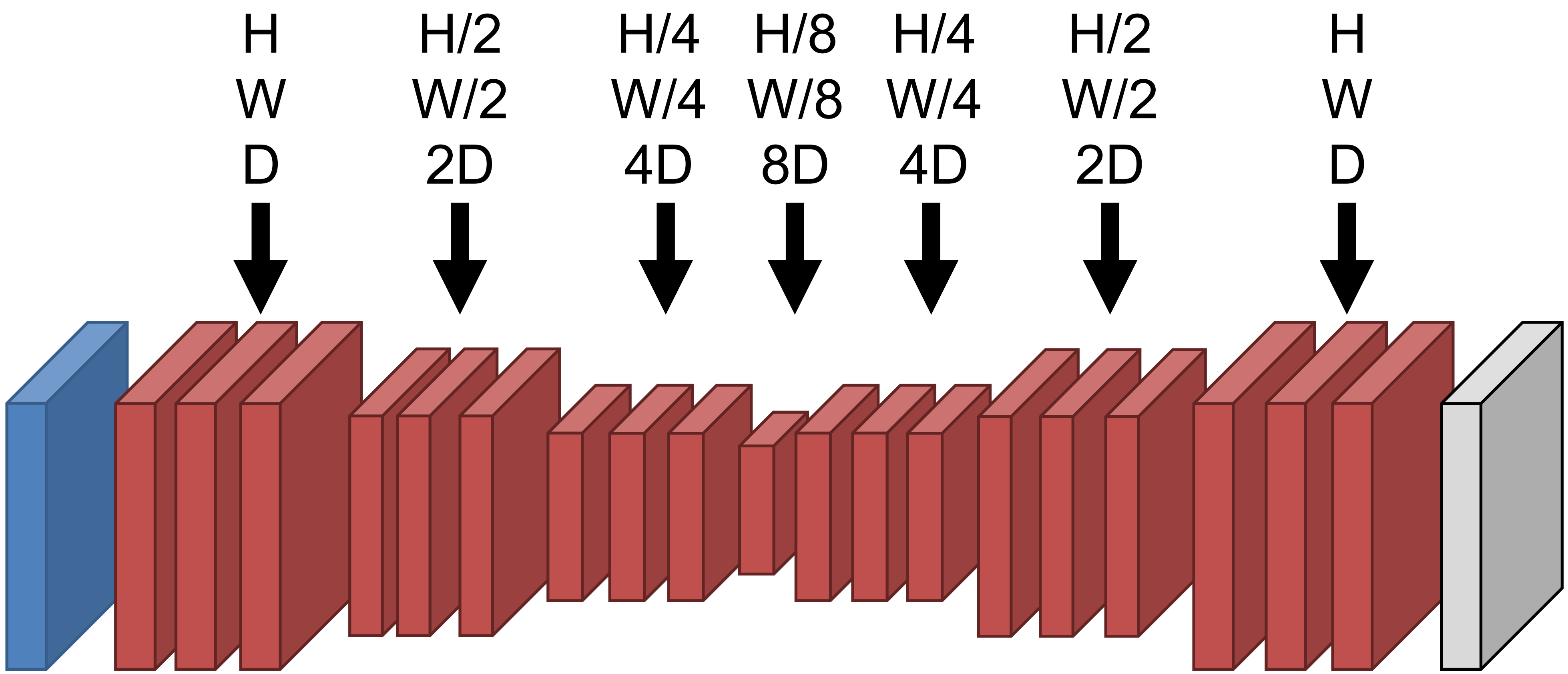
Convolutions + pooling downsample/compress/encode
Transpose convs./unpoolings upsample/uncompress/decode



[David Fouhey]

Putting It Together – Block Sizes

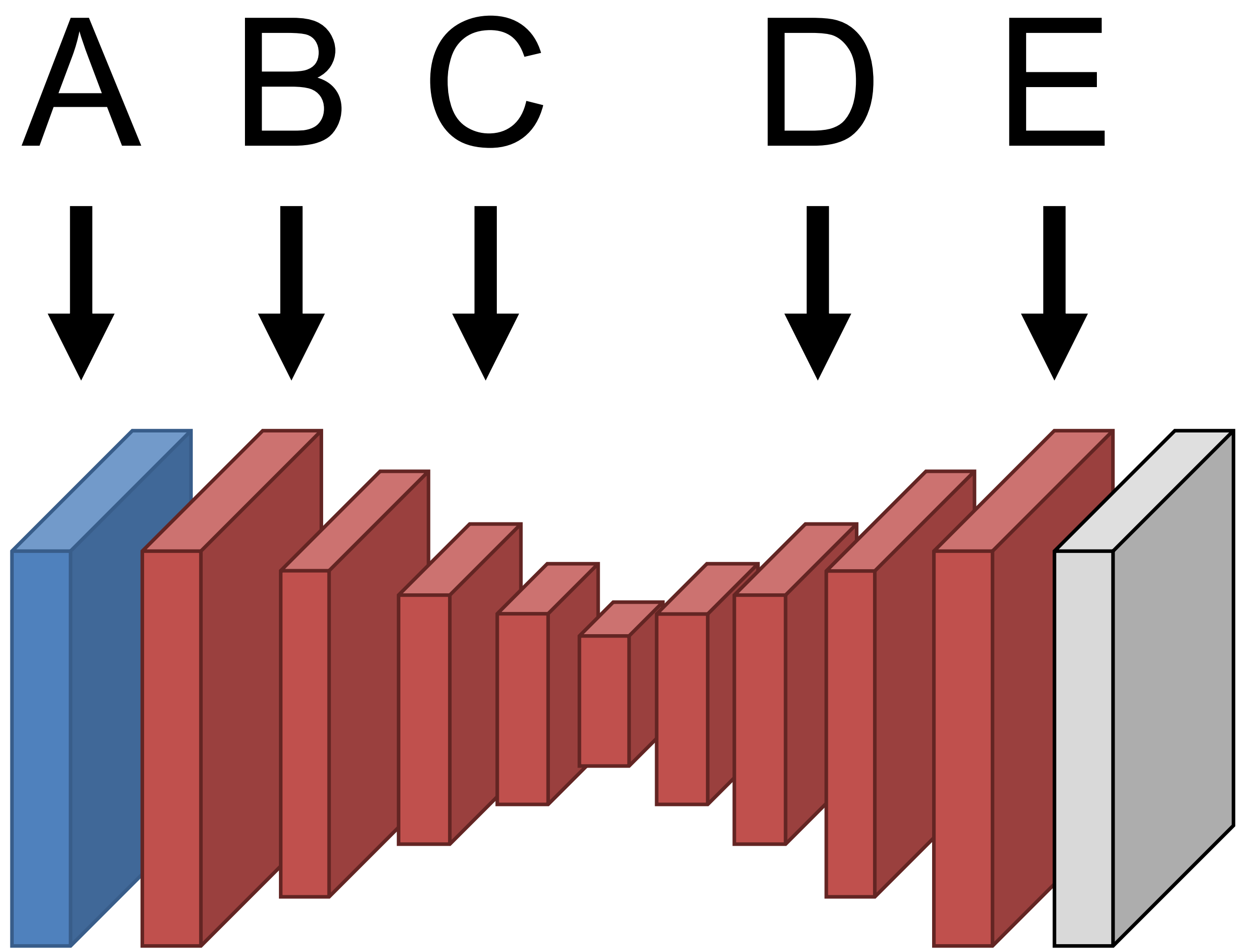
- Often multiple layers at each spatial resolution.
  - Often halve spatial resolution and double feature depth every few layers

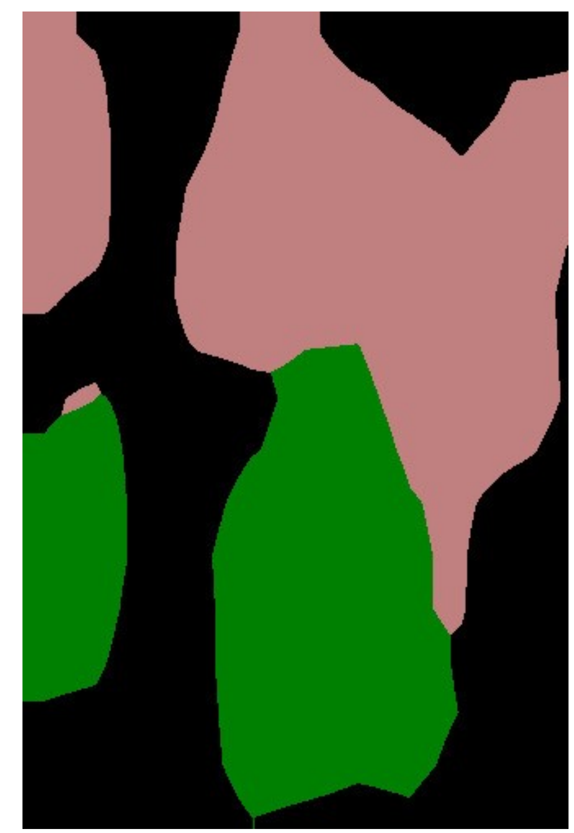


[David Fouhey]

Missing Details

Where is the useful information about the high-frequency details of the image?





Result from Long et al. *Fully Convolutional Networks For Semantic Segmentation*. CVPR 2014

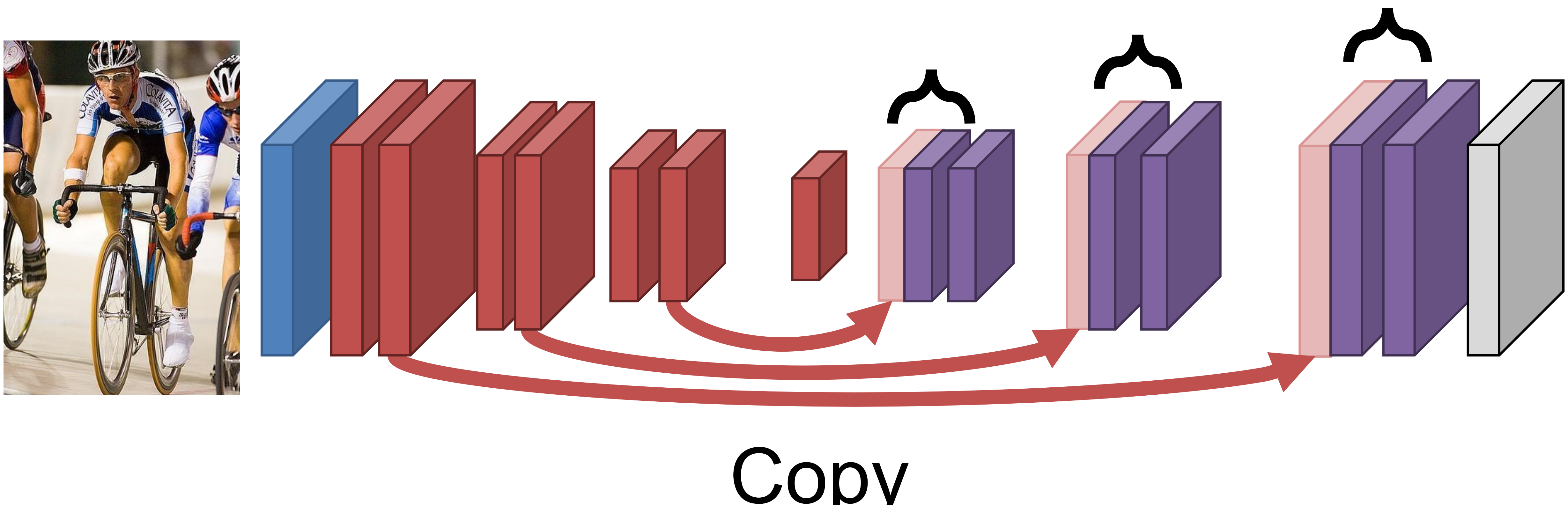
[David Fouhey]

Missing Details

How do you send details forward in the network?

You copy the activations forward.

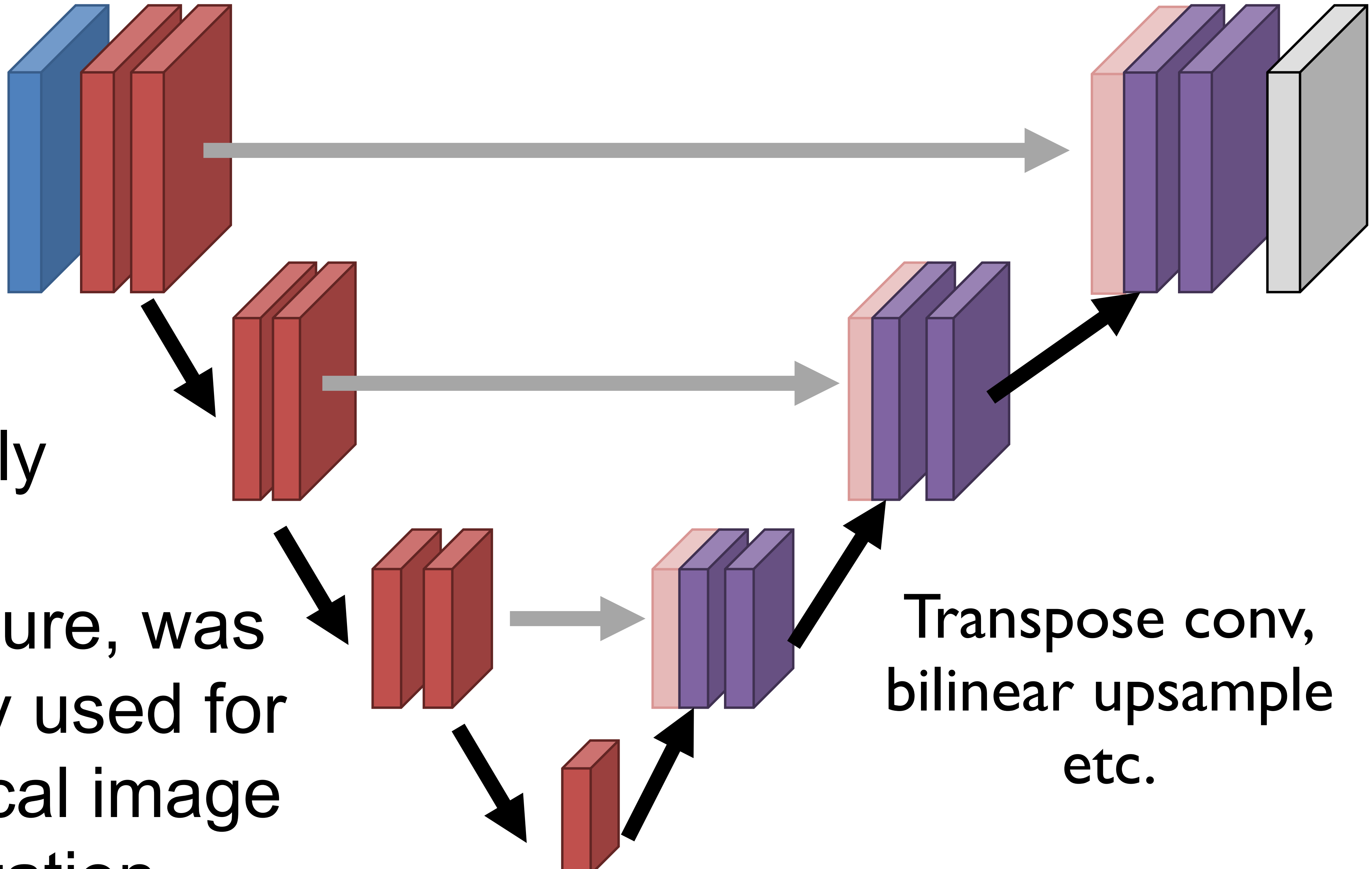
Subsequent layers at the same resolution figure out how to fuse things.



Result from Long et al. *Fully Convolutional Networks For Semantic Segmentation*. CVPR 2014

[David Fouhey]

U-Net



Extremely popular architecture, was originally used for biomedical image segmentation.

Transpose conv, bilinear upsample etc.

Ronneberger et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015

[David Fouhey]

Single-View Depth Estimation

[T. Zhou, A. Geiger]

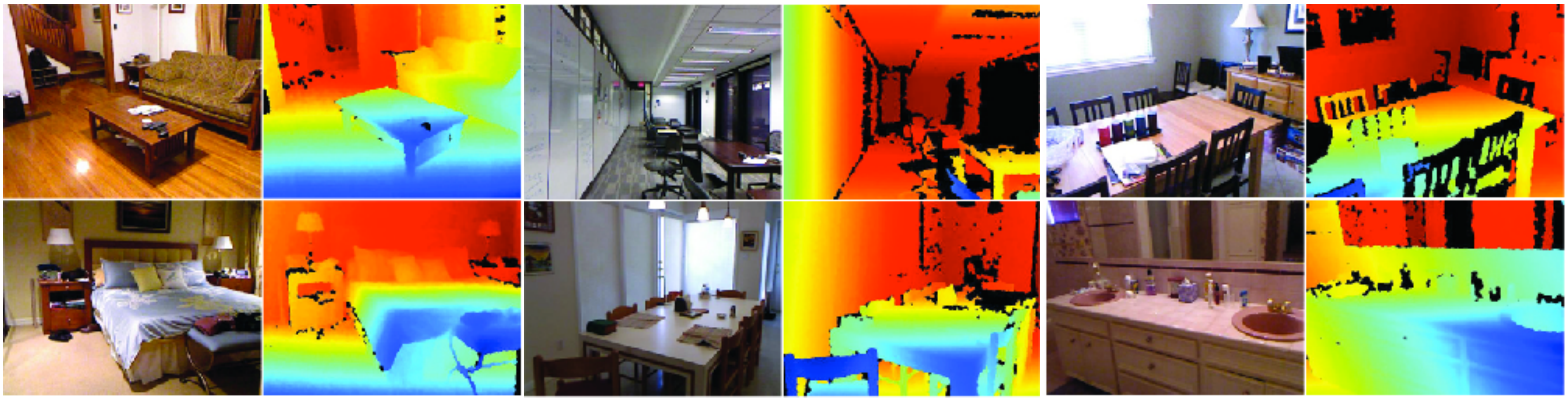
Single-View Depth Estimation

[T. Zhou, A. Geiger]

Single-View Depth Estimation

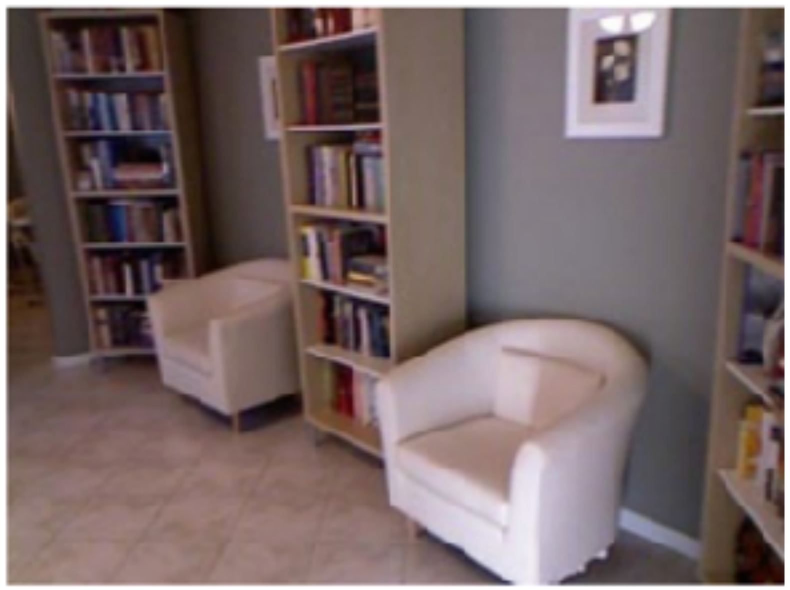
[T. Zhou, A. Geiger]

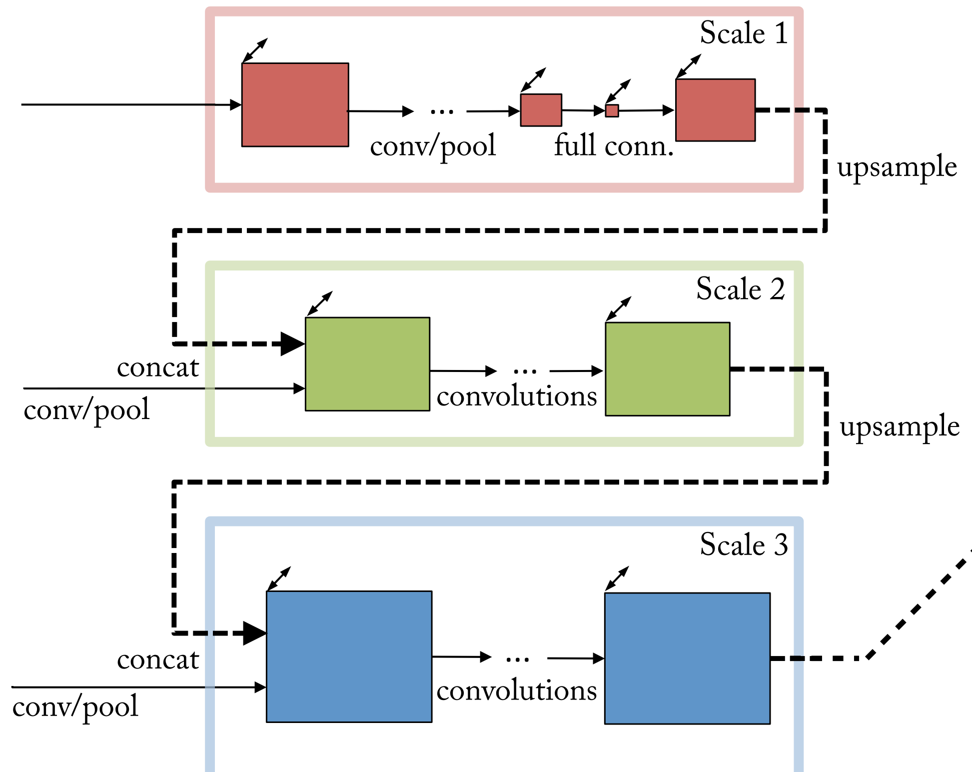
NYU Depth v2 Dataset

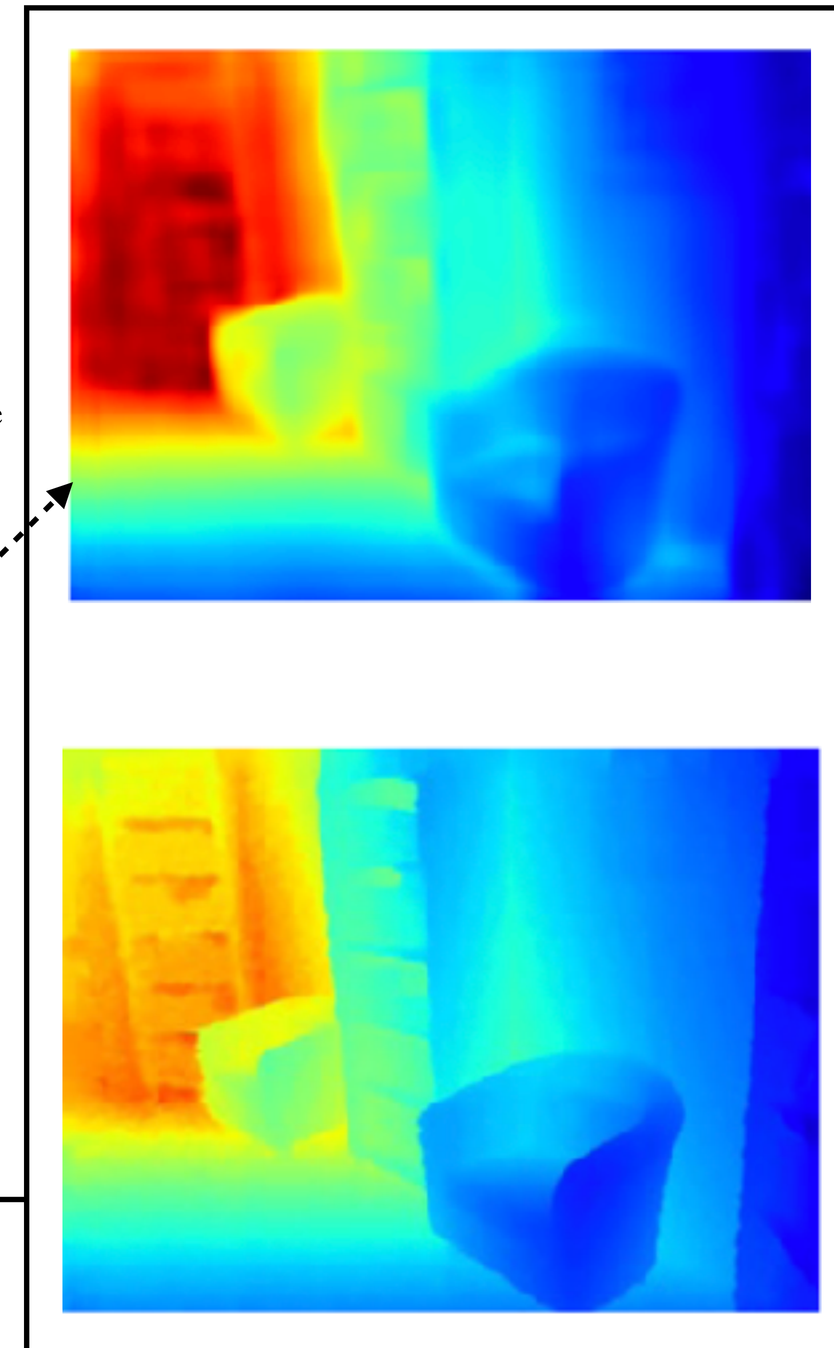


- 400K RGBD frames captured using Microsoft Kinect
- ~1500 have segmentation labels (26 classes) as well
- The dataset has depth holes, note offset between RGB and NIR cameras, and NIR dot projector, also raw RGB + D frames are not synchronized
- Synchronized and filled subset of 50K images by [Alhashim Wonka 2018] — see Project 4 description
- Limited to indoor scenes due to active NIR illumination

NYU Depth Estimation







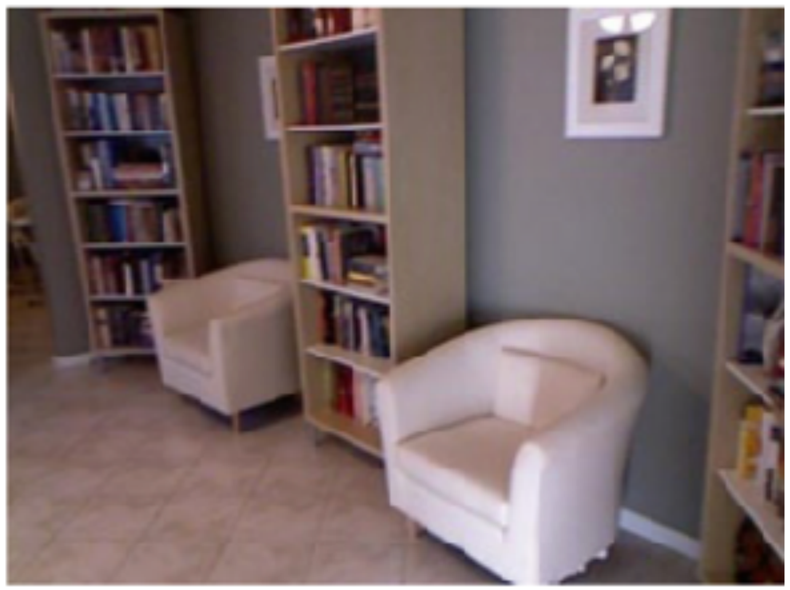
multi-scale
architecture

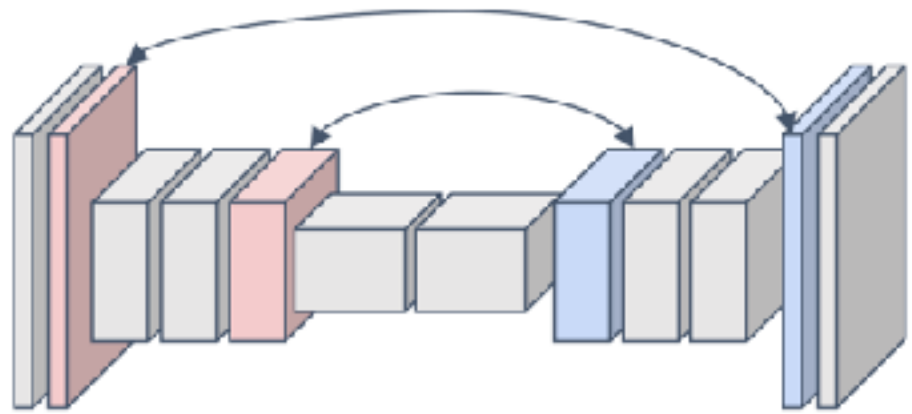
Loss,
e.g., L2

Direct supervision
via Kinect RGB+D

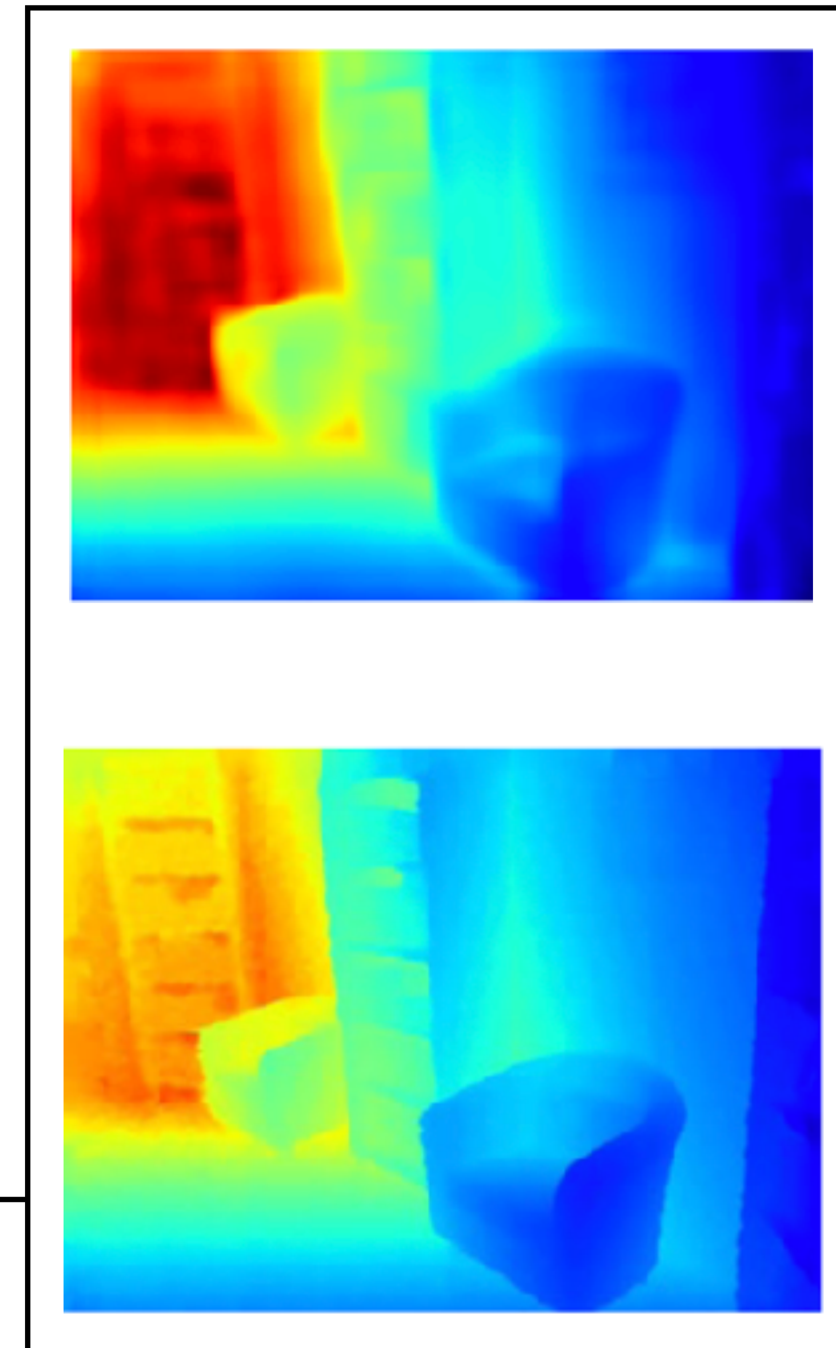
[Eigen Fergus 2015]

NYU Depth Estimation





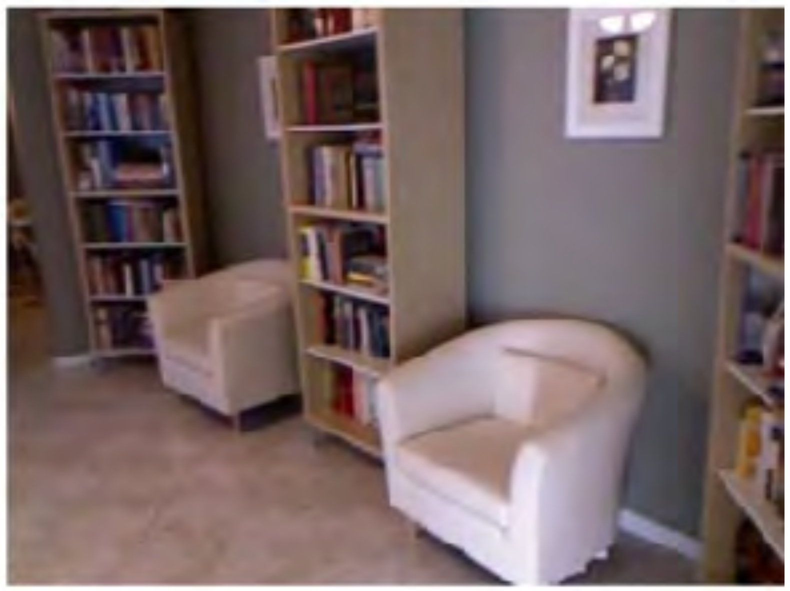
U-Net with skip connections

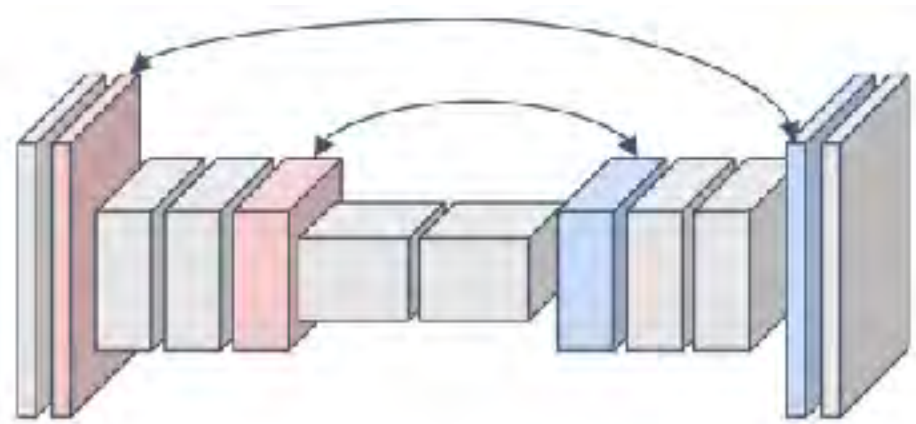


Direct supervision via Kinect RGB+D

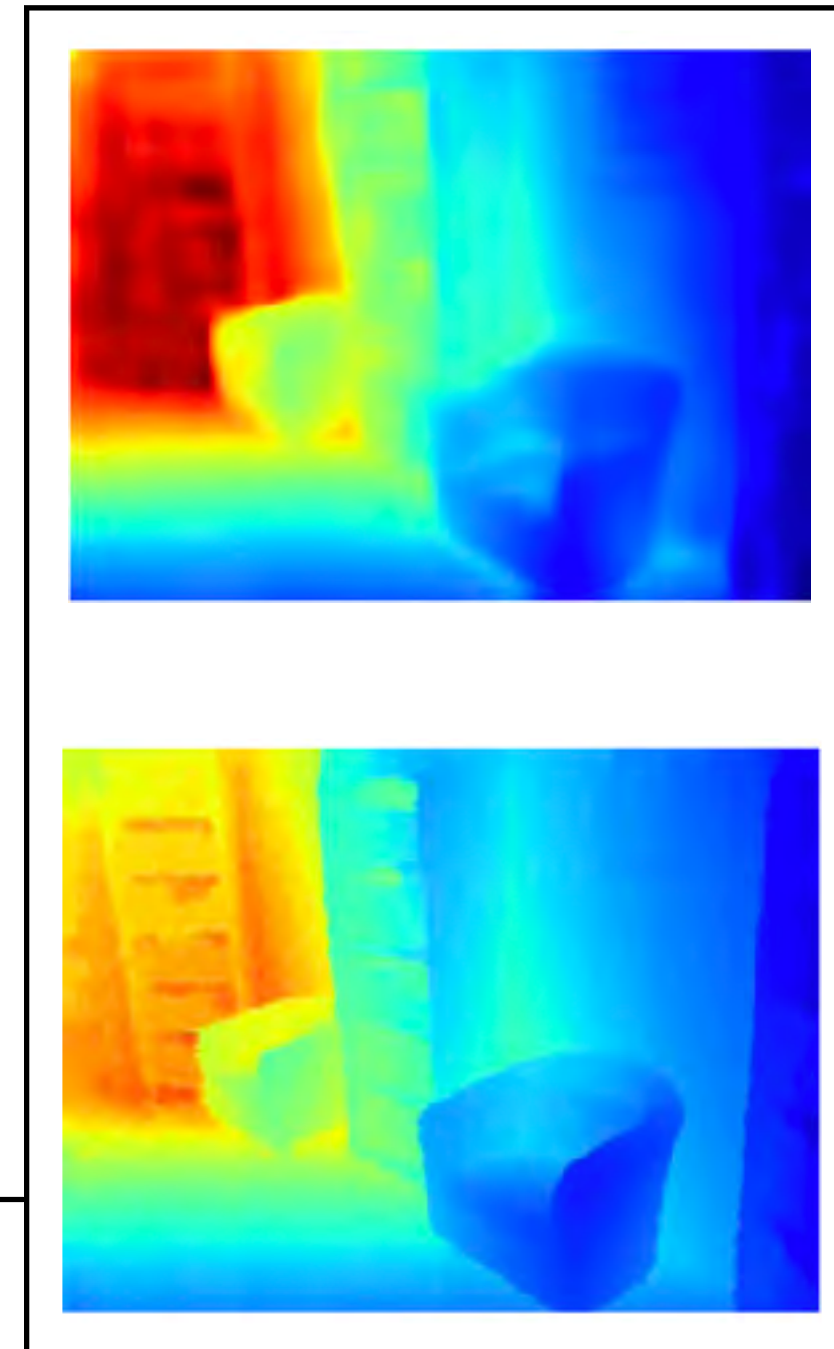
Loss, e.g., L2

Single-View Depth Estimation





U-Net with skip connections

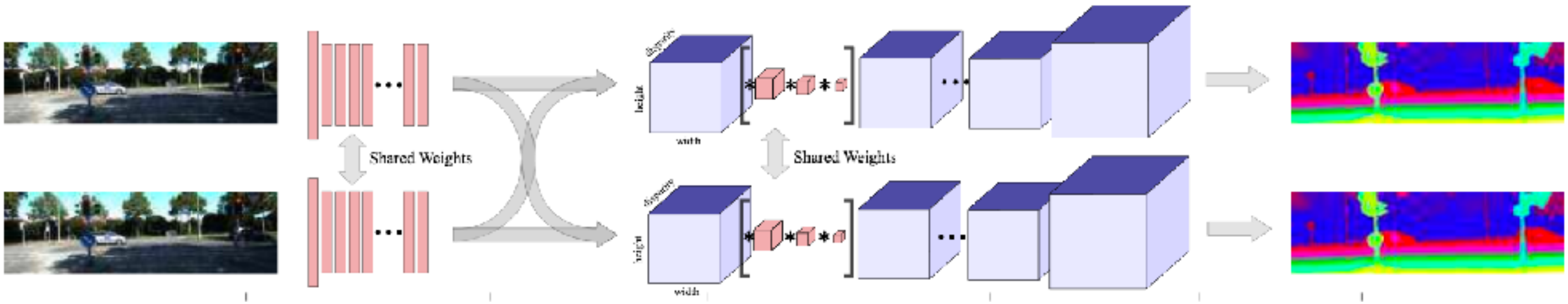


Direct supervision via Kinect RGB+D

Loss, e.g., L2

2-view Stereo

- Form HxWxD=disparity volume and use 3D convolution



Extract features at each pixel using 2D CNN

Form volume by sliding features from 2nd image at D disparities

Perform 3D convolution on feature volume

Treat output as disparity cost volume and perform soft argmax

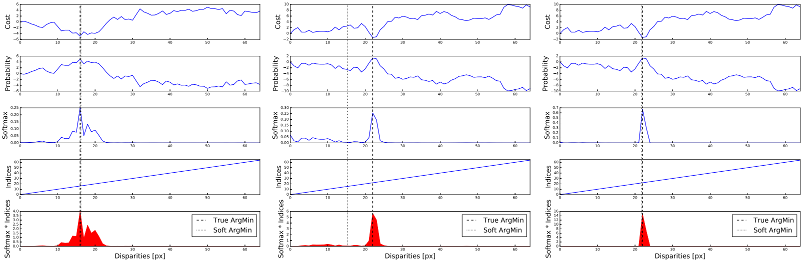
https://www.youtube.com/watch?v=VtAzDS1NLmo [Kendall et al. 2017]

End-to-end Deep Stereo Regression Architecture

	Layer Description	Output Tensor Dim.
	Input image	$H \times W \times C$
	Unary features (section 3.1)	
1	5×5 conv, 32 features, stride 2	$\frac{1}{2}H \times \frac{1}{2}W \times F$
2	3×3 conv, 32 features	$\frac{1}{2}H \times \frac{1}{2}W \times F$
3	3×3 conv, 32 features	$\frac{1}{2}H \times \frac{1}{2}W \times F$
	add layer 1 and 3 features (residual connection)	$\frac{1}{2}H \times \frac{1}{2}W \times F$
4-17	(repeat layers 2,3 and residual connection) $\times$ 7	$\frac{1}{2}H \times \frac{1}{2}W \times F$
18	3×3 conv, 32 features, (no ReLu or BN)	$\frac{1}{2}H \times \frac{1}{2}W \times F$
	Cost volume (section 3.2)	
	Cost Volume	$\frac{1}{2}D \times \frac{1}{2}H \times \frac{1}{2}W \times 2F$
	Learning regularization (section 3.3)	
19	3-D conv, $3 \times 3 \times 3$, 32 features	$\frac{1}{2}D \times \frac{1}{2}H \times \frac{1}{2}W \times F$
20	3-D conv, $3 \times 3 \times 3$, 32 features	$\frac{1}{2}D \times \frac{1}{2}H \times \frac{1}{2}W \times F$
21	From Cost Volume: 3-D conv, $3 \times 3 \times 3$, 64 features, stride 2	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
22	3-D conv, $3 \times 3 \times 3$, 64 features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
23	3-D conv, $3 \times 3 \times 3$, 64 features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
24	From 21: 3-D conv, $3 \times 3 \times 3$, 64 features, stride 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
25	3-D conv, $3 \times 3 \times 3$, 64 features	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
26	3-D conv, $3 \times 3 \times 3$, 64 features	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
27	From 24: 3-D conv, $3 \times 3 \times 3$, 64 features, stride 2	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 2F$
28	3-D conv, $3 \times 3 \times 3$, 64 features	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 2F$
29	3-D conv, $3 \times 3 \times 3$, 64 features	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 2F$
30	From 27: 3-D conv, $3 \times 3 \times 3$, 128 features, stride 2	$\frac{1}{32}D \times \frac{1}{32}H \times \frac{1}{32}W \times 4F$
31	3-D conv, $3 \times 3 \times 3$, 128 features	$\frac{1}{32}D \times \frac{1}{32}H \times \frac{1}{32}W \times 4F$
32	3-D conv, $3 \times 3 \times 3$, 128 features	$\frac{1}{32}D \times \frac{1}{32}H \times \frac{1}{32}W \times 4F$
33	$3 \times 3 \times 3$, 3-D transposed conv, 64 features, stride 2	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 2F$
	add layer 33 and 29 features (residual connection)	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 2F$
34	$3 \times 3 \times 3$, 3-D transposed conv, 64 features, stride 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
	add layer 34 and 26 features (residual connection)	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
35	$3 \times 3 \times 3$, 3-D transposed conv, 64 features, stride 2	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
	add layer 35 and 23 features (residual connection)	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
36	$3 \times 3 \times 3$, 3-D transposed conv, 32 features, stride 2	$\frac{1}{2}D \times \frac{1}{2}H \times \frac{1}{2}W \times F$
	add layer 36 and 20 features (residual connection)	$\frac{1}{2}D \times \frac{1}{2}H \times \frac{1}{2}W \times F$
37	$3 \times 3 \times 3$, 3-D trans conv, 1 feature (no ReLu or BN)	$D \times H \times W \times 1$
	Soft argmin (section 3.4)	
	Soft argmin	$H \times W$

[Kendall et al. 2017]

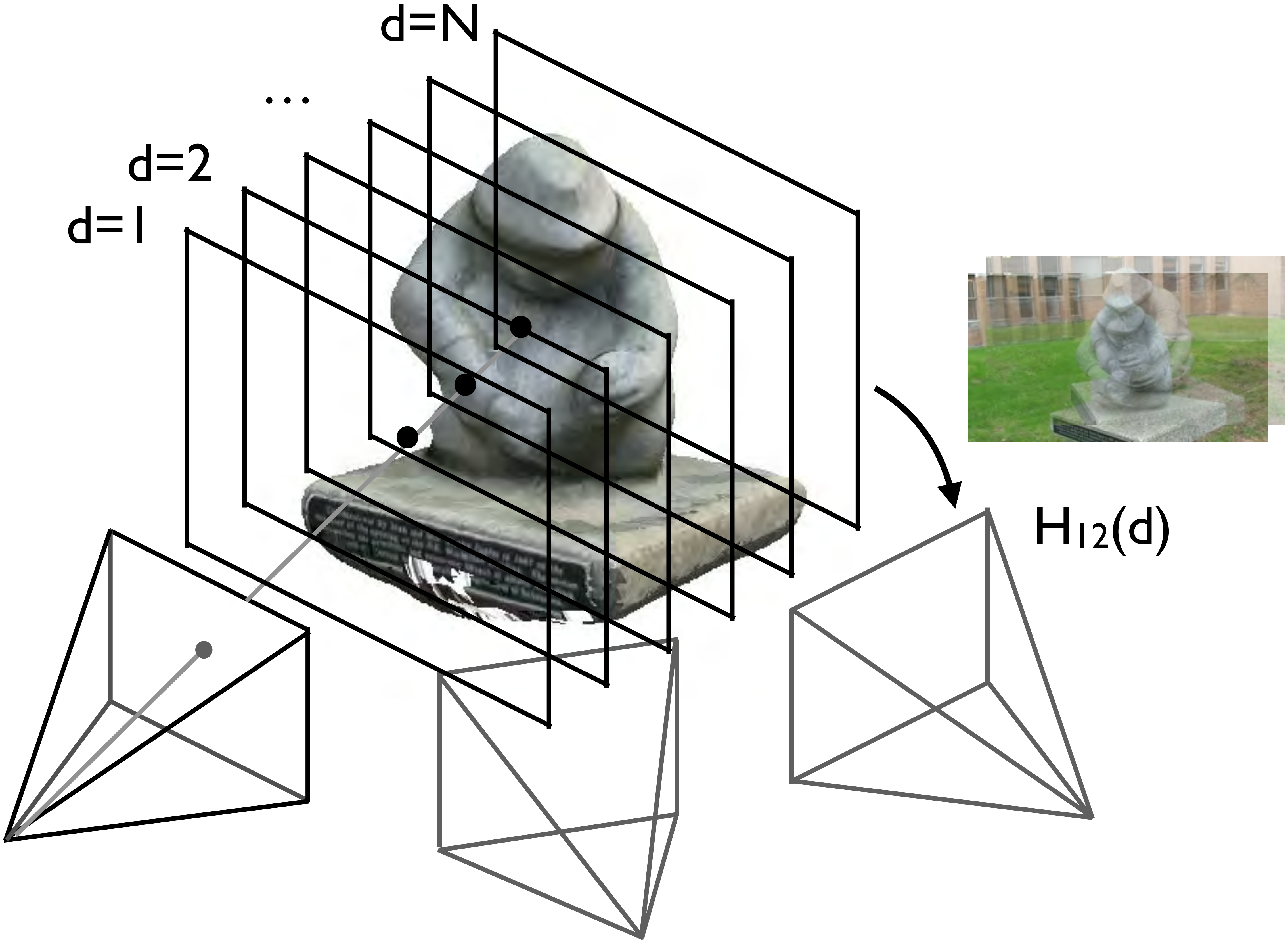
Computing Sub-pixel Disparity



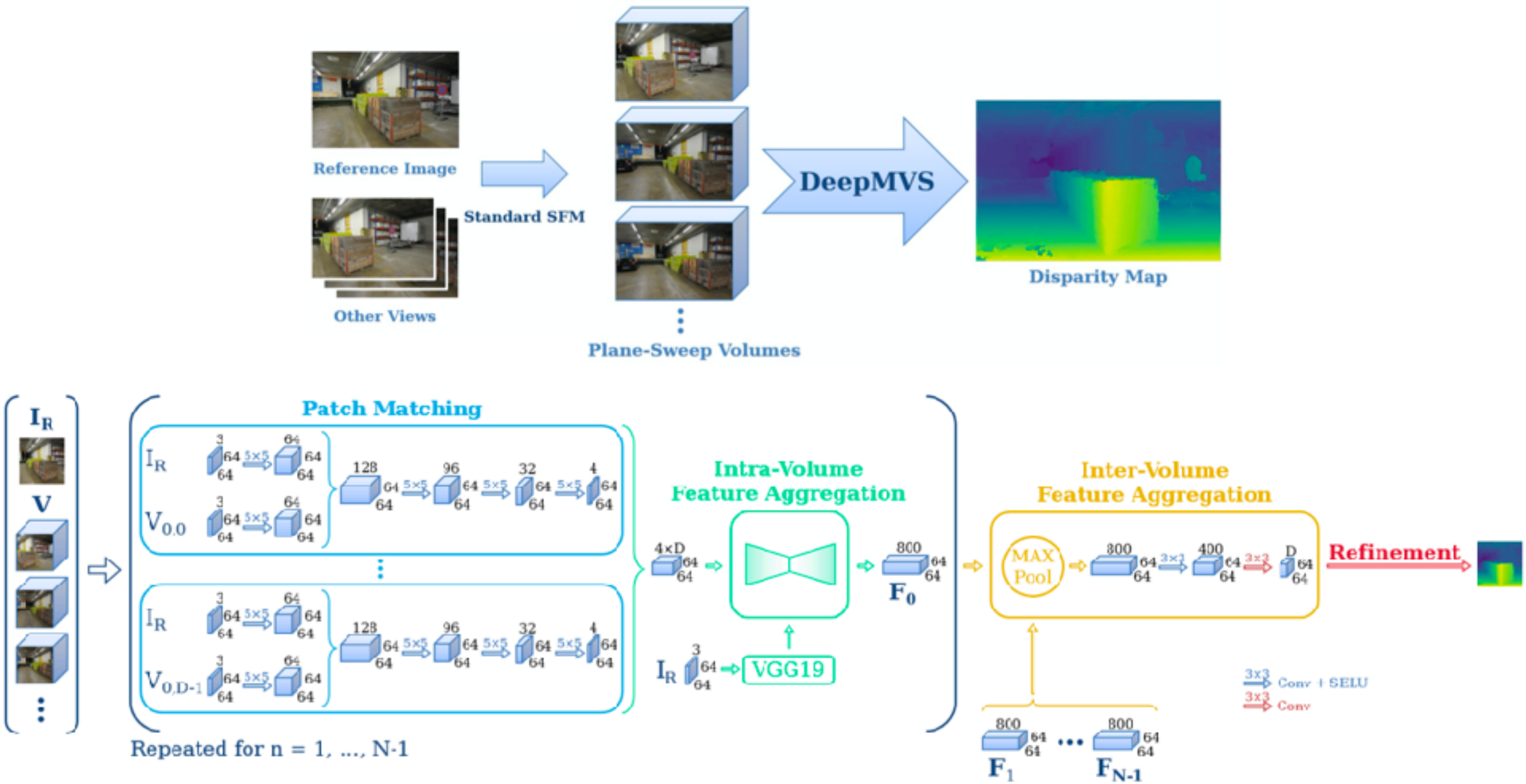
(a) Soft ArgMin (b) Multi-modal distribution (c) Multi-modal distribution with prescaling

[Kendall et al. 2017]

Plane Sweep Stereo
(reminder from Lecture 5)



Multi-view Stereo

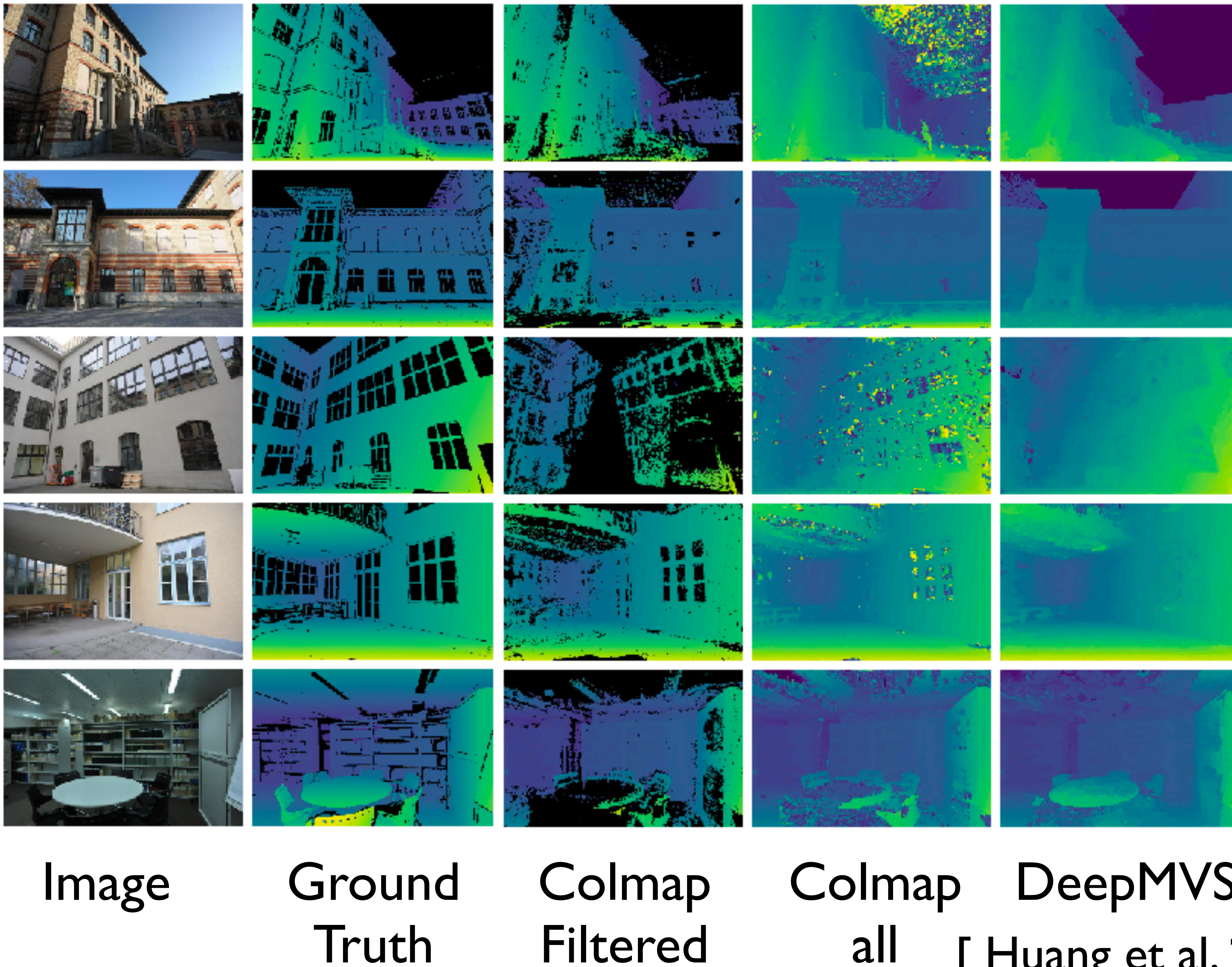


Compare patches in ref image to plane sweep volumes from other images

Perform intra and inter-volume aggregation of features

[DeepMVS, Huang et al. 2018]

DeepMVS: Results



[Huang et al. 2018]

DeepMVS: Ablation Studies

Components	Geo. error	Pho. error
Pretraining	0.051	0.242
+ U-net	0.043	0.230
+ U-net + VGG	0.040	0.226
+ U-net + VGG + DenseCRF	**0.036**	**0.224**
+ U-net + VGG + DenseCRF − MVS-SYNTH	0.037	0.225

[Huang et al. 2018]

DeepMVS: Progressive Improvement

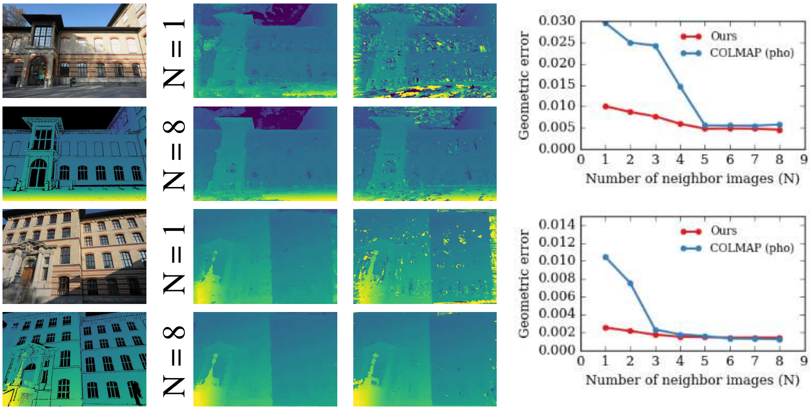


Image / ground truth | Our result | COLMAP [33] (unfiltered) | Geometric Errors

[Huang et al. 2018]

3D Shape Representations: Point Cloud

- Represent shape as a set of P points in 3D space
- (+) Can represent fine structures without huge numbers of points
- ( ) Requires new architecture, losses, etc
- (-) Doesn't explicitly represent the surface of the shape: extracting a mesh for rendering or other applications requires post-processing

Fan et al, "A Point Set Generation Network for 3D Object Reconstruction from a Single Image", CVPR 2017

Processing Pointcloud Inputs: PointNet

Want to process pointclouds as **sets**: order should not matter

Run MLP on each point

Max-Pool

Fully Connected

Input pointcloud: P x 3

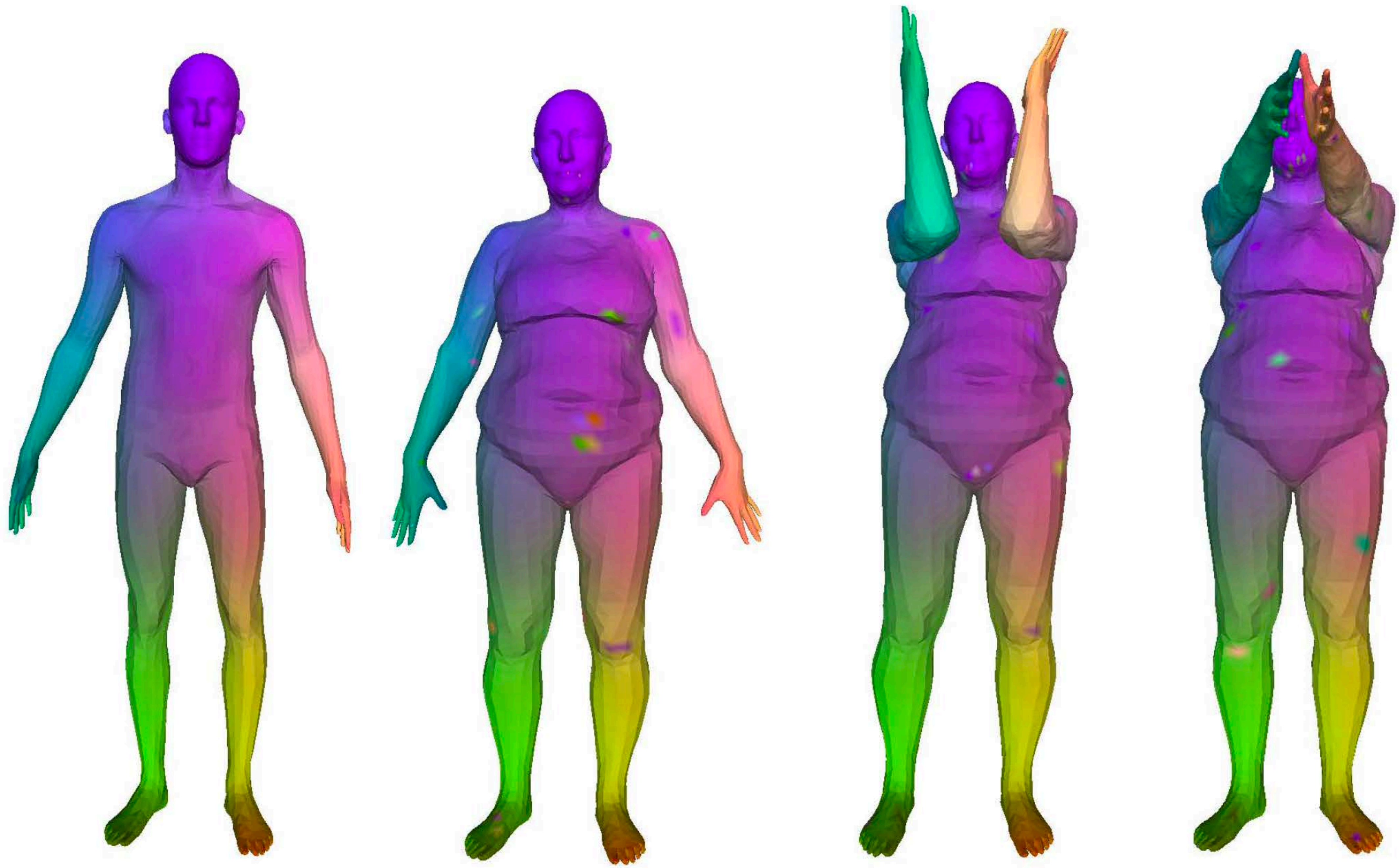
Point features: P x D

Pooled vector: D

Class score: C

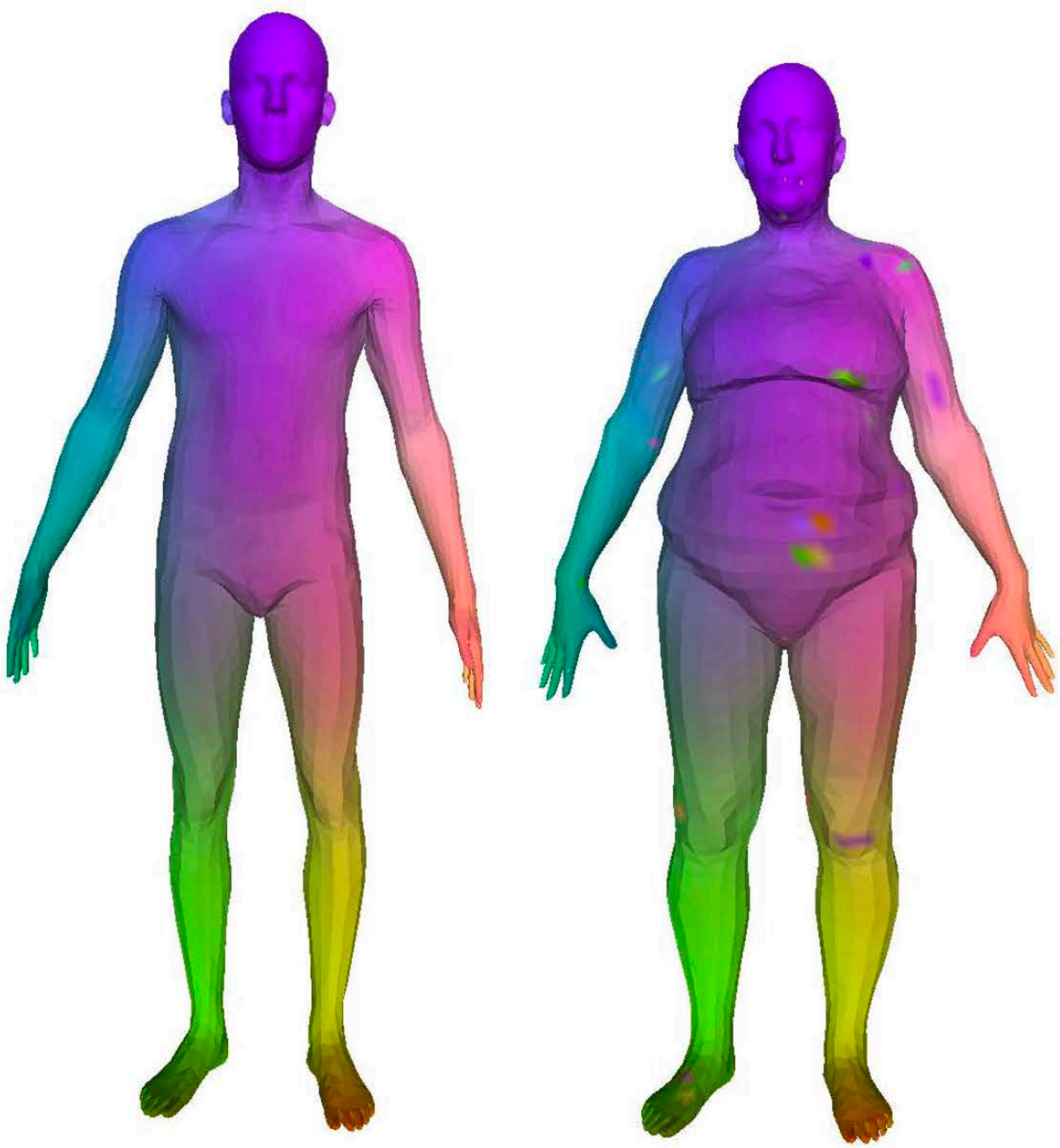
Qi et al, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation", CVPR 2017
Qi et al, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space", NeurIPS 2017

Processing Mesh (and PointCloud): FeaStNet



[Verma, Boyer, and Verbeek CVPR 2018]

FeaStNet: Problem Statement



Vertex-labeling problem:

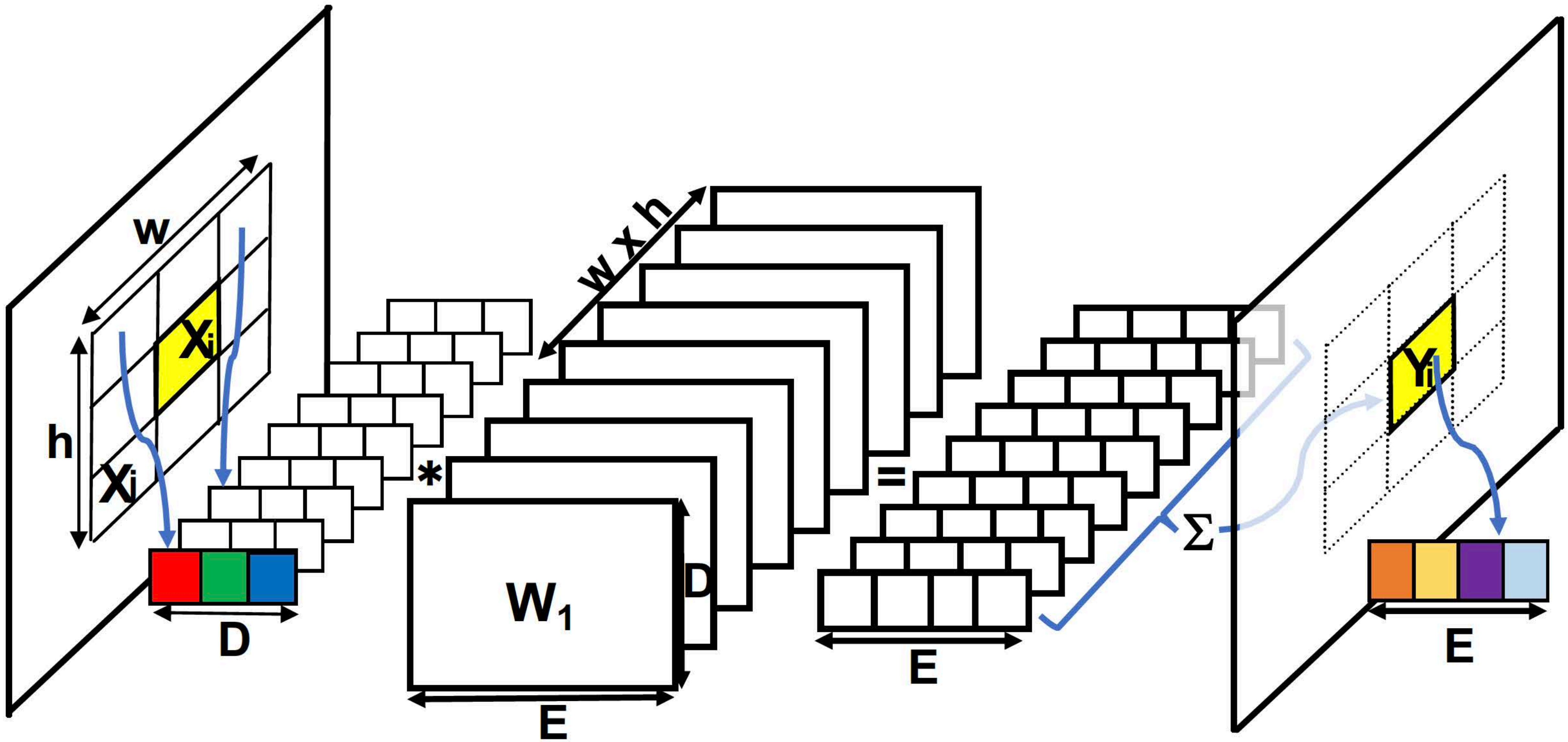
Reference shape: 6,980 vertices

Let each vertex in the reference shape be its own class (label).
Y = {0, ..., 6980-1}

For the target shape (on the right), label each vertex using Y

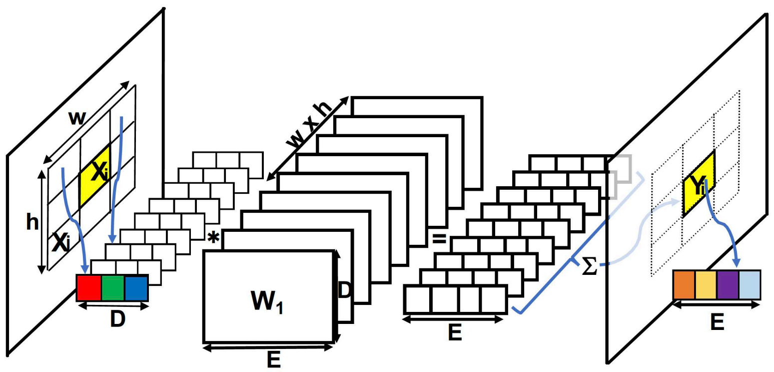
[Verma, Boyer, and Verbeek CVPR 2018]

Rethinking Convolution



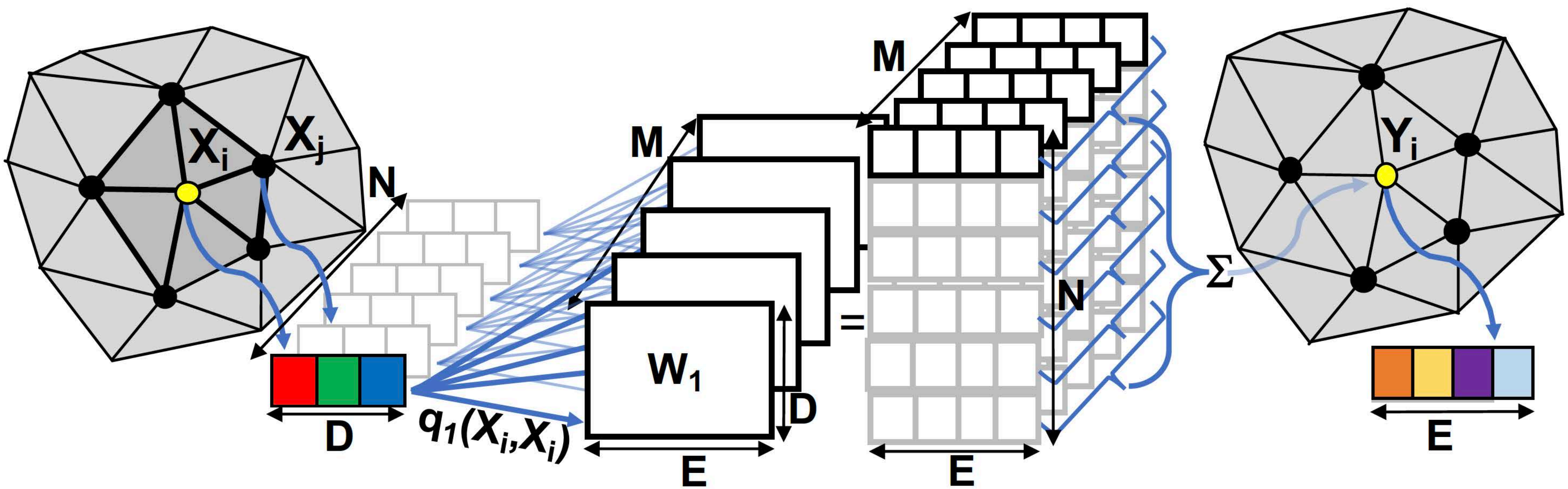
[Verma, Boyer, and Verbeek CVPR 2018]

Generalized Convolution



← convolution on the image lattice

convolution on an arbitrary graph topology



[Verma, Boyer, and Verbeek CVPR 2018]

Generalized Convolution

$$\mathbf{y}_i = \mathbf{b} + \sum_{m=1}^{M} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} q_m(\mathbf{x}_i, \mathbf{x}_j) \mathbf{W}_m \mathbf{x}_j,$$

$$q_m(\mathbf{x}_i, \mathbf{x}_j) \propto \exp\left(\mathbf{u}_m^\top \mathbf{x}_i + \mathbf{v}_m^\top \mathbf{x}_j + c_m\right),$$

with $\sum_{m=1}^{M} q_m(\mathbf{x}_i, \mathbf{x}_j) = 1,$

The only additional parameters w.r.t. a conventional CNN are the vectors $\mathbf{u}_m$, $\mathbf{v}_m$, which contain $2MD$ parameters.

[Verma, Boyer, and Verbeek CVPR 2018]

uses 3D mesh models from IKEA

3D Datasets: Object-Centric

ShapeNet

~50 categories, ~50k 3D CAD models

Standard split has 13 categories, ~44k models, 25 rendered images per model

Many papers show results here

(-) Synthetic, isolated objects; no context

(-) Lots of chairs, cars, airplanes

Pix3D

9 categories, 219 3D models of IKEA furniture aligned to ~17k real images

Some papers train on ShapeNet and show qualitative results here, but use ground-truth segmentation masks

(+) Real images! Context!

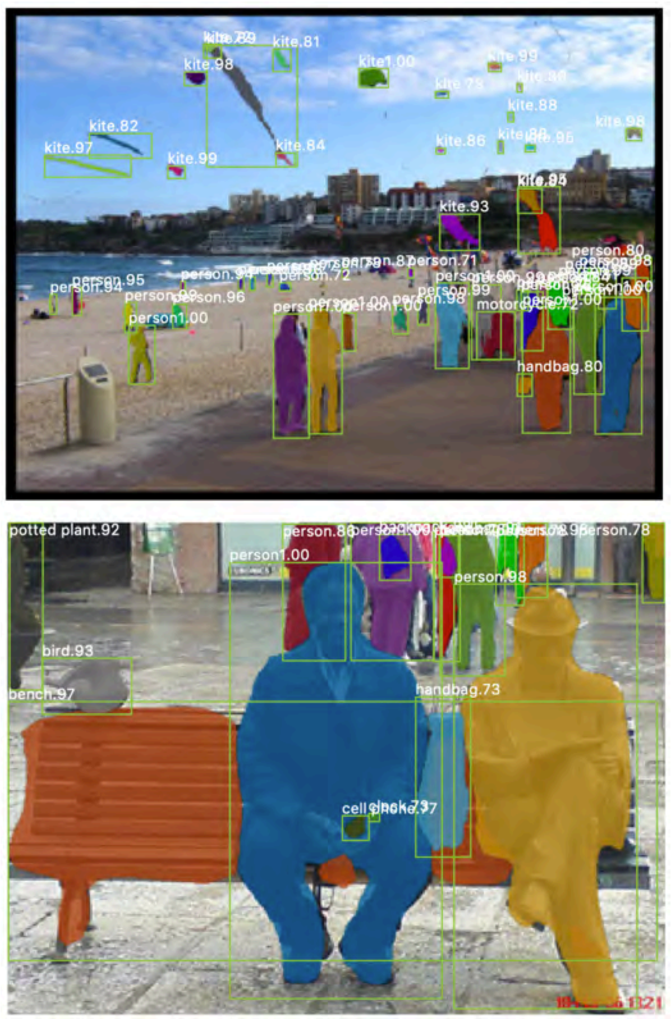
(-) Small, partial annotations – only 1 obj/image

Chang et al, "ShapeNet: An Information-Rich 3D Model Repository", arXiv 2015
Choy et al, "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction", ECCV 2016

Sun et al, "Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling", CVPR 2018

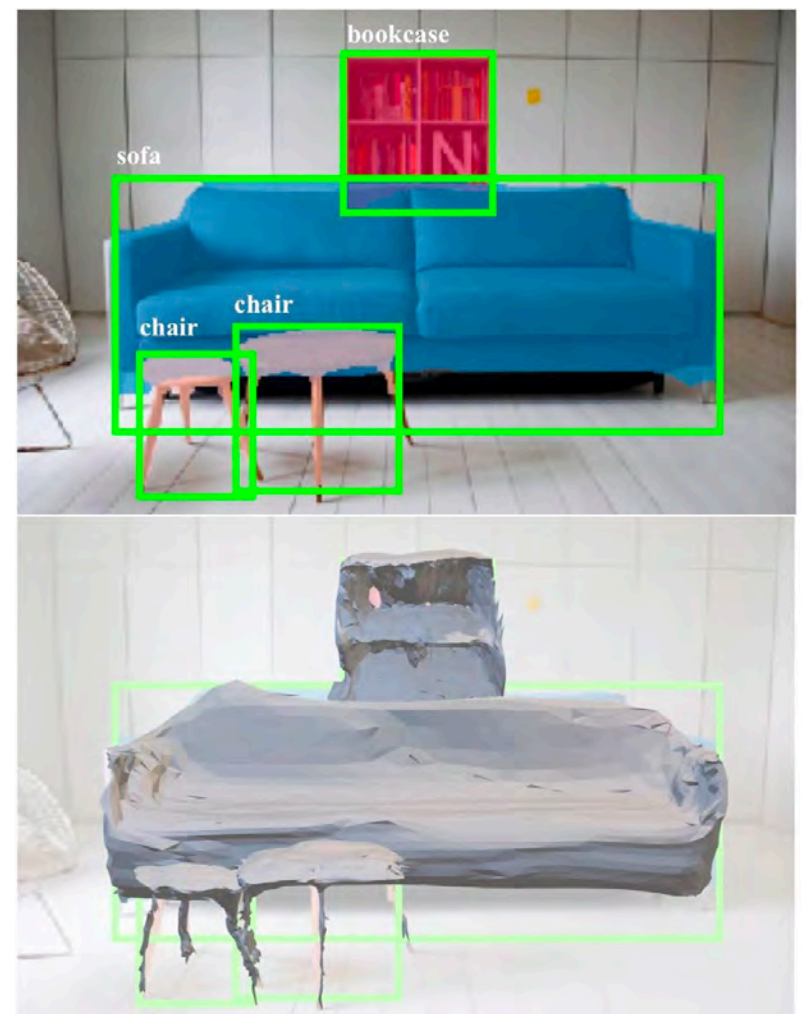
3D Shape Prediction: Mesh R-CNN

Mask R-CNN:
2D Image -> 2D shapes



He, Gkioxari, Dollár, and Girshick, “Mask R-CNN”, ICCV 2017

Mesh R-CNN:
2D Image -> **Triangle Meshes**



Gkioxari, Malik, and Johnson, “Mesh R-CNN”, ICCV 2019

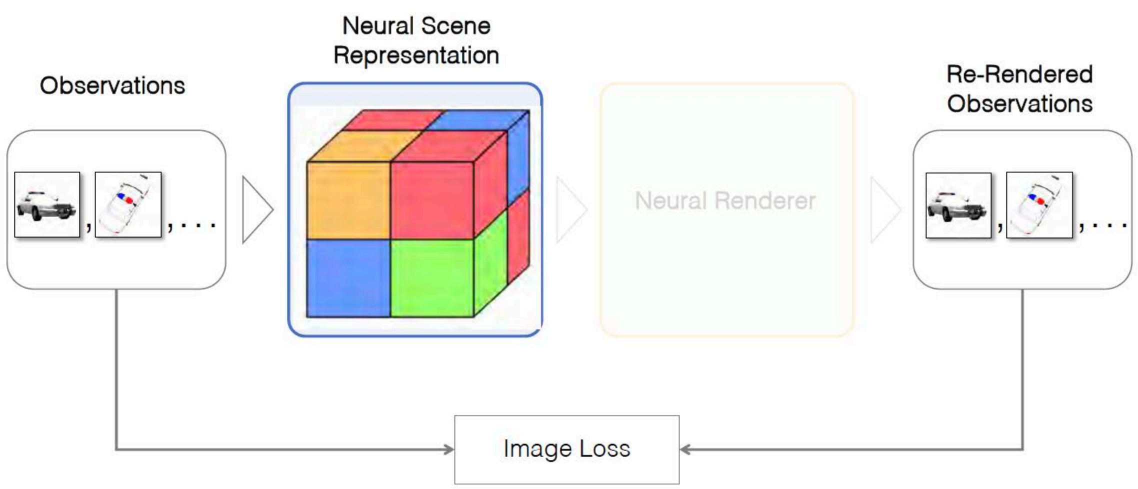
Detect objects and extract silhouettes

Estimate 3D mesh

There Is More To Do in 3D

DeepVoxels

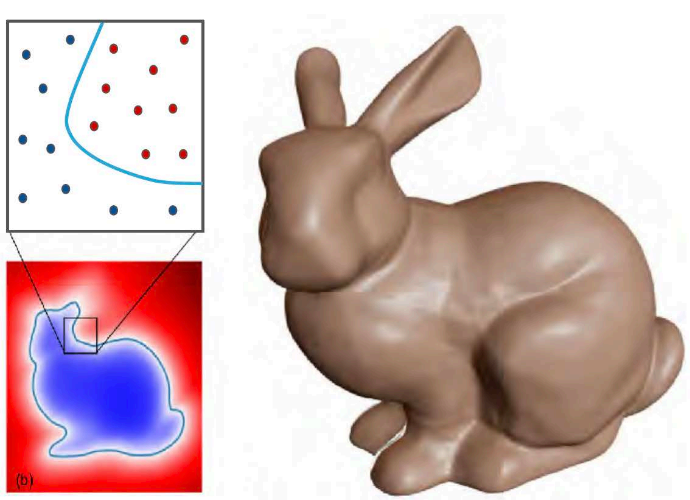
Embedding vector per voxel



Scene represented as an embedding vector per 3D point

DeepSDF

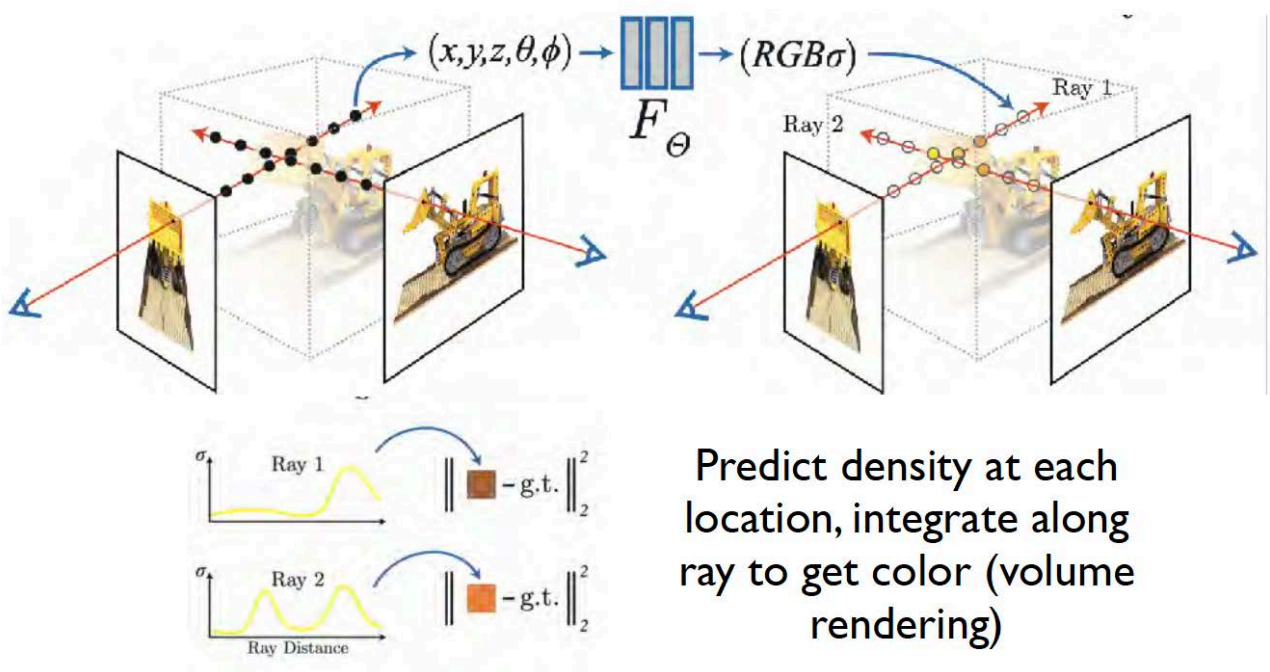
- CPPN for signed distance function, SDF=f(X)



[Slides: Jeong Joon Park]

Neural Radiance Fields

- Another continuous scene representation using a FCN



Predict density at each location, integrate along ray to get color (volume rendering)

We've Reached the End of the Class

But there is so much more to computer vision!

Stay in touch: vxa@uw.edu