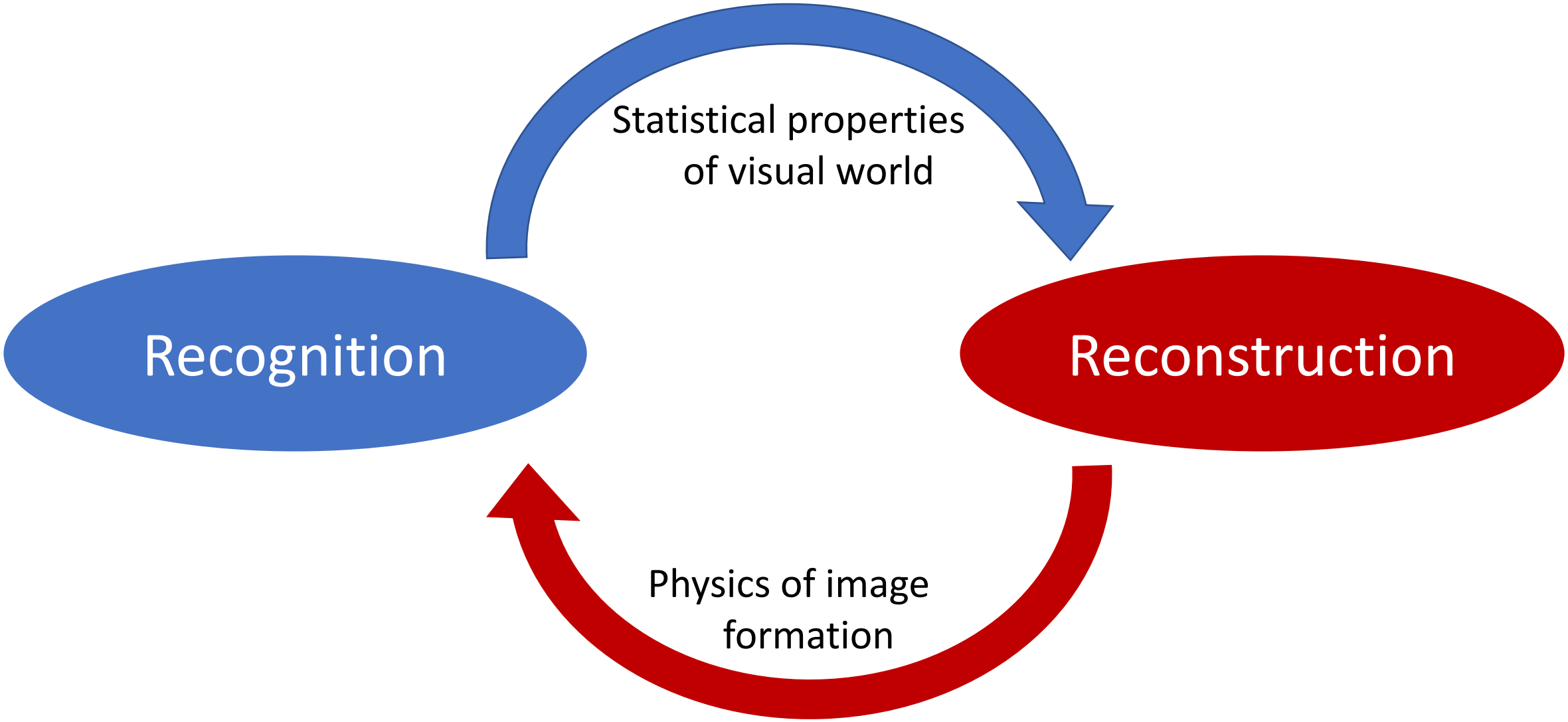
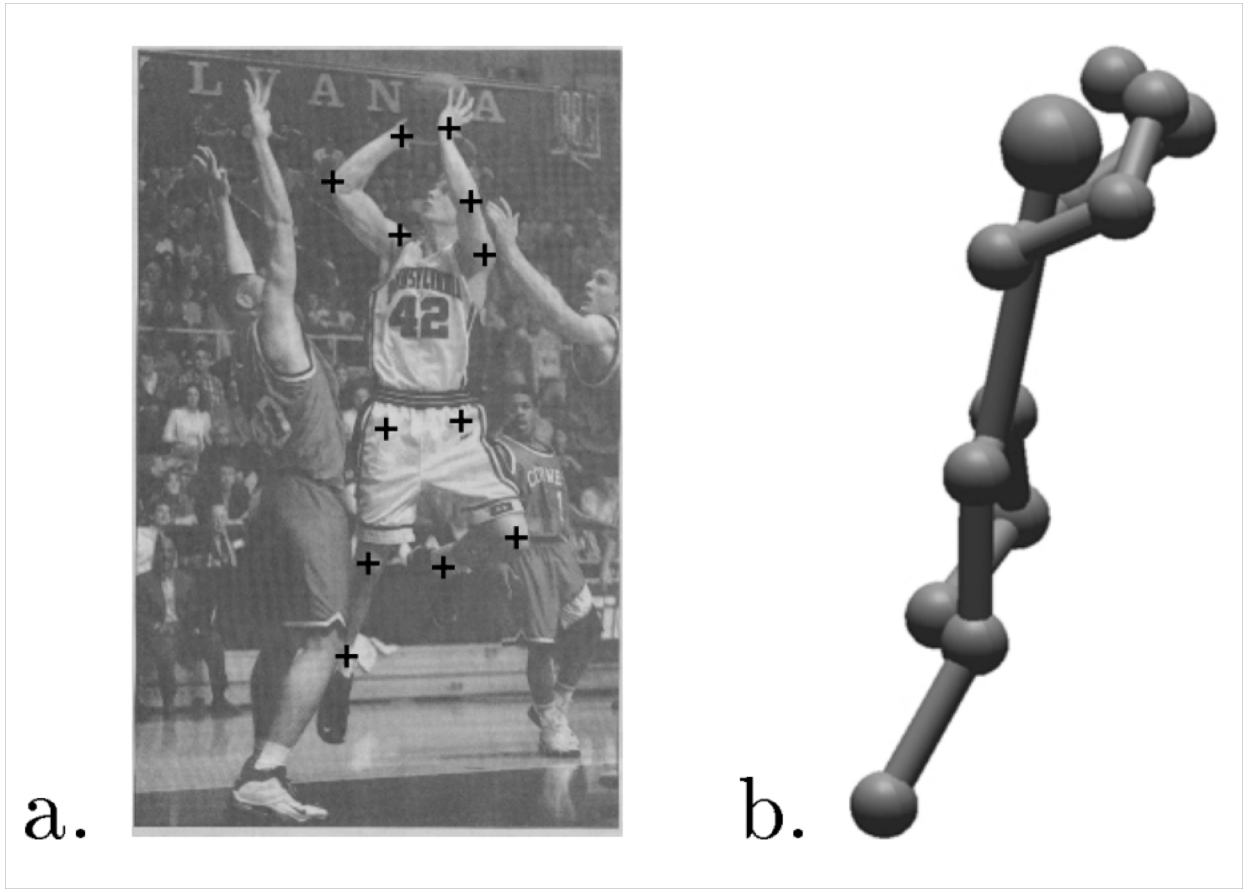
Learning for 3D

Recognition and 3D reasoning



Pose estimation in 3D



Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. C. J. Taylor. In *CVPR,* 2000

Pose estimation in 3D

- Key idea: know relative lengths of each limb4
- Assume *scaled orthographic projection*
  - Valid when variation in depth much smaller than depth

$$x = \frac{X}{Z} \approx \frac{X}{Z_0}$$

$$y = \frac{Y}{Z} \approx \frac{Y}{Z_0}$$

constant

Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. C. J. Taylor. In *CVPR,* 2000

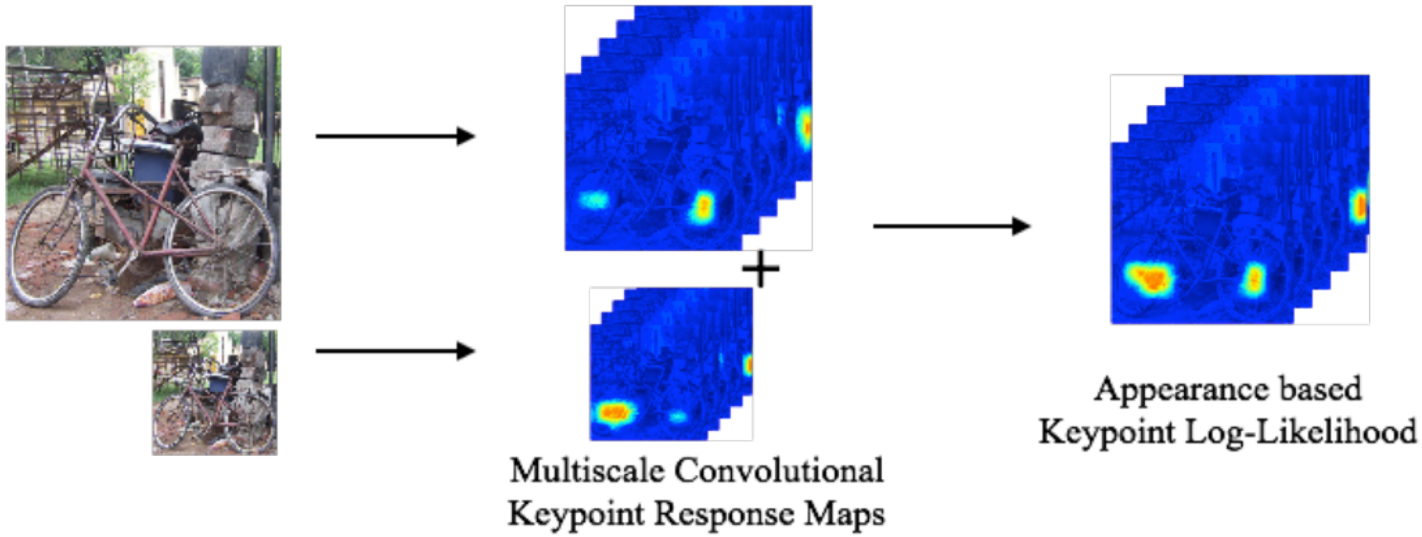
Pose estimation in 3D

$$
\begin{aligned}
l^2 &= (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2 \\
(u_1 - u_2) &= s(X_1 - X_2) \\
(v_1 - v_2) &= s(Y_1 - Y_2) \\
dZ &= (Z_1 - Z_2) \\
\Rightarrow dZ &= \sqrt{l^2 - ((u_1 - u_2)^2 + (v_1 - v_2)^2)/s^2}
\end{aligned}
$$

Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. C. J. Taylor. In *CVPR,* 2000

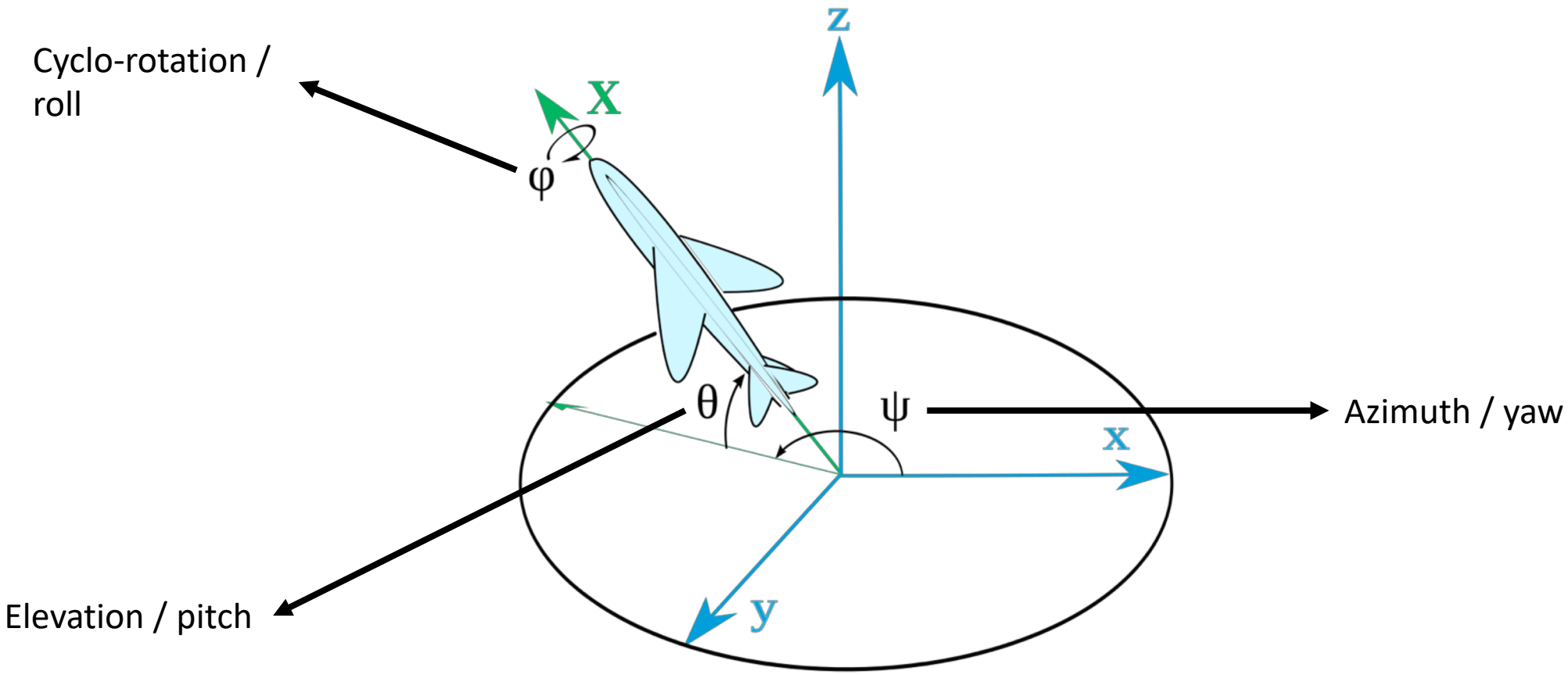
Pose estimation for rigid objects

Pose estimation for rigid objects



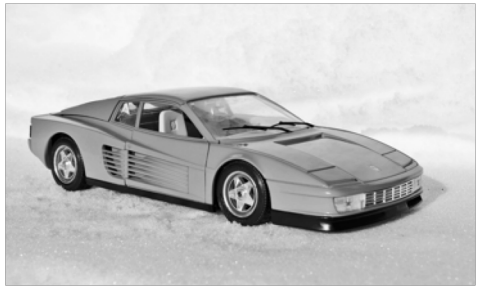
Viewpoints and keypoints. S. Tulsiani and J. Malik. In *CVPR,* 2015

Pose estimation for rigid objects



Viewpoint-conditioned pose

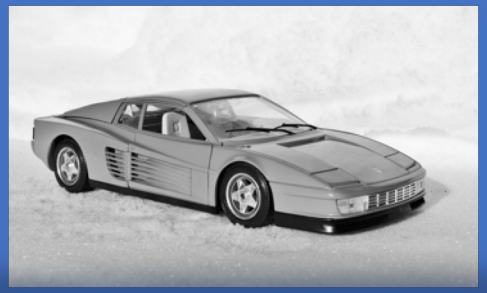
Viewpoint prediction



Viewpoints and keypoints. S. Tulsiani and J. Malik. In *CVPR,* 2015

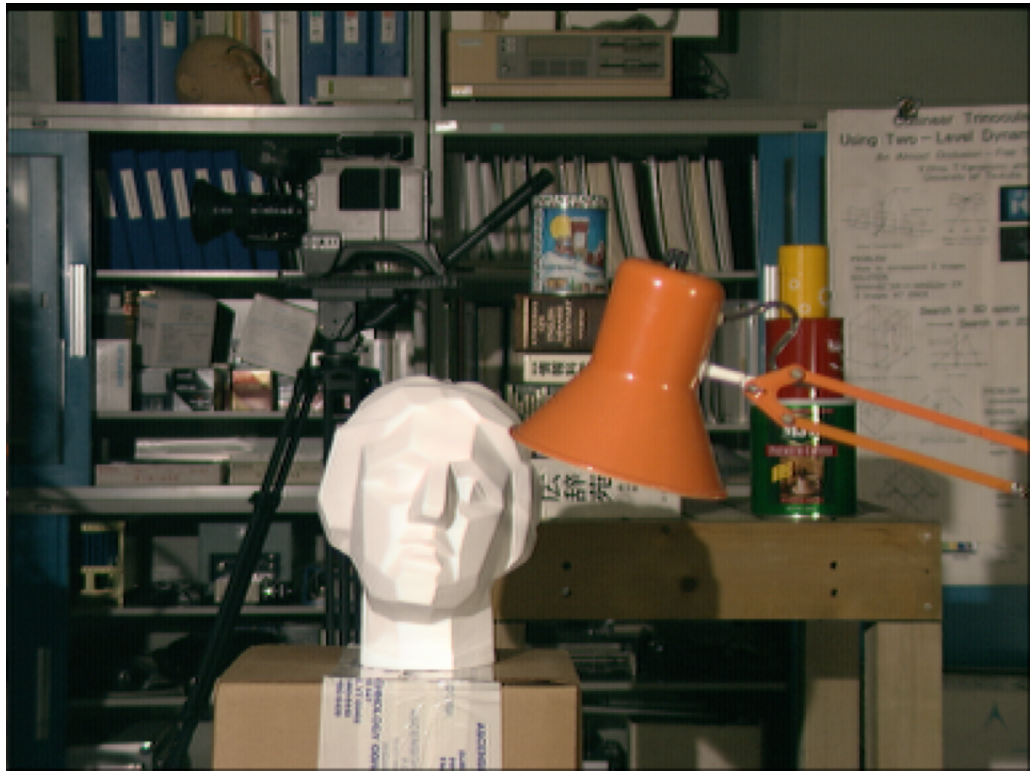
Viewpoint-conditioned pose

Viewpoint prediction

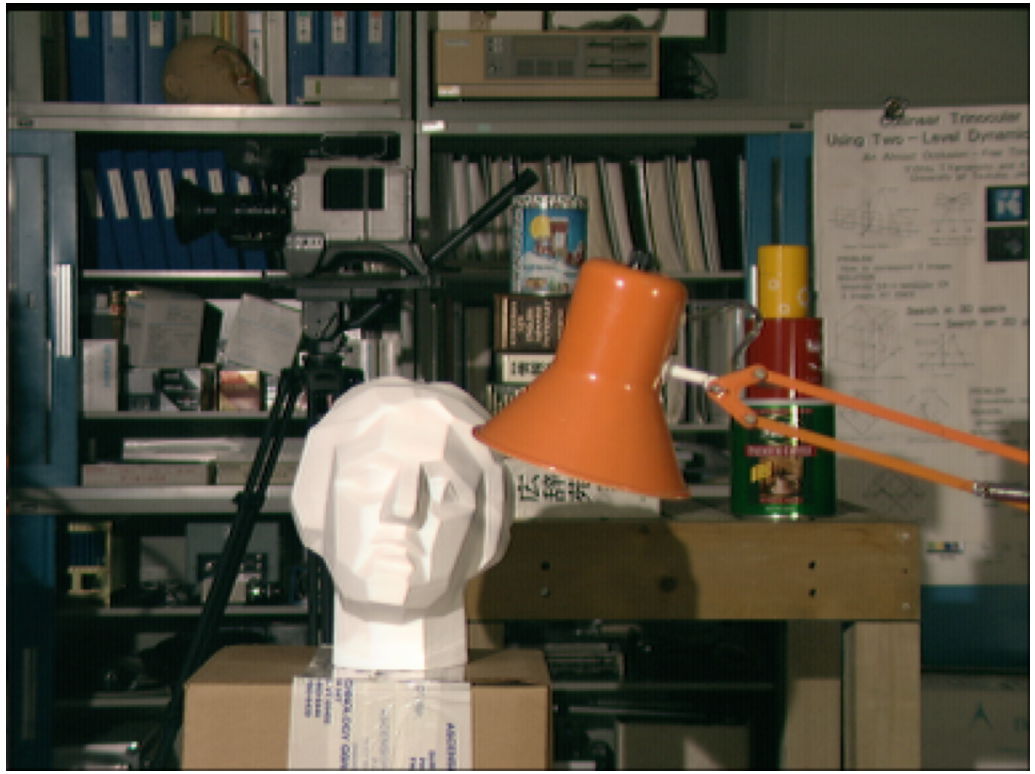


Viewpoints and keypoints. S. Tulsiani and J. Malik. In *CVPR,* 2015

Disparity estimation



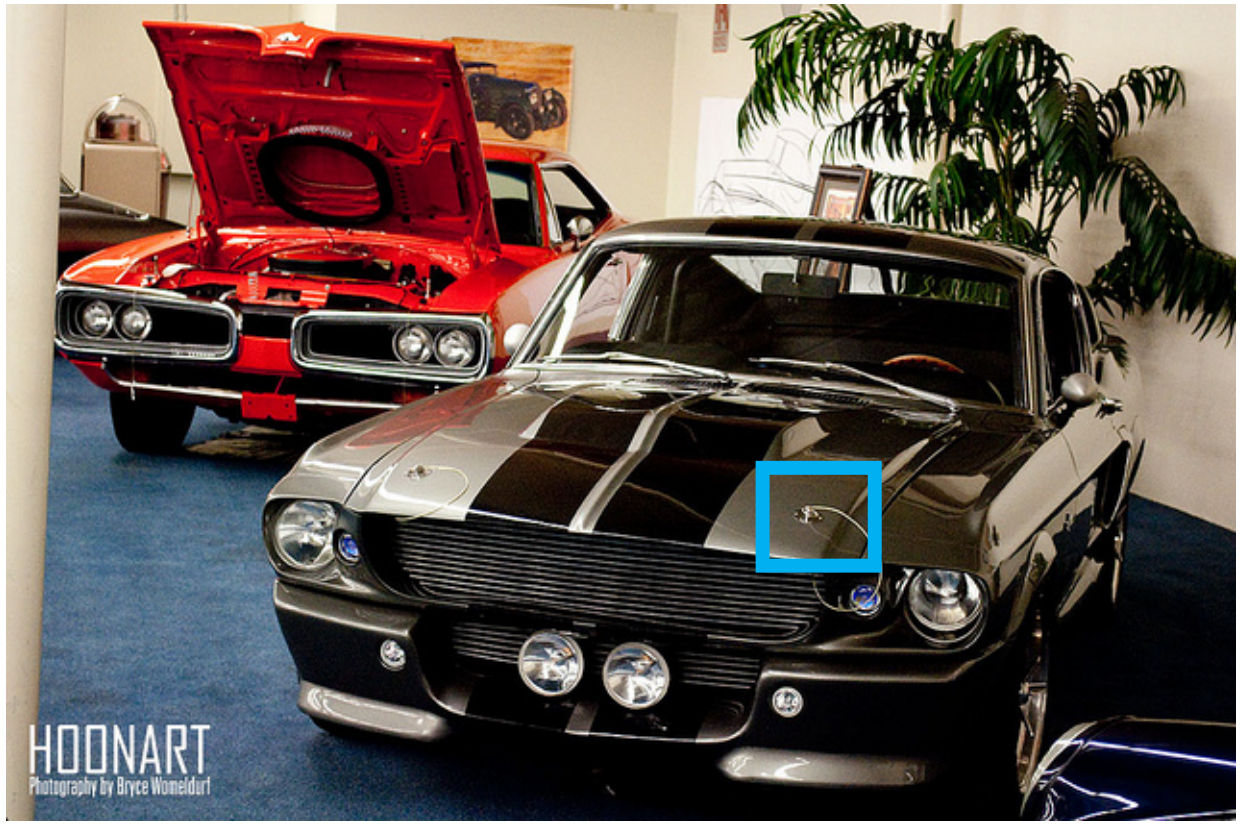
Disparity estimation



Disparity estimation

- Goal:
  - Assign disparity value to each pixel
- Basic idea:
  - Disparity image should be *smooth*
- Energy minimization
  - min E(d), where d is disparity image
  - $E(d) = E_{data}(d) + E_{smoothness}(d)$
- $E_{data}(d)$: scores based on NCC (for example)
- $E_{smoothness}(d) = \sum_{i,j} \rho(d(i,j) - d(i,j+1)) + \rho(d(i,j) - d(i+1,j))$

Measuring patch similarity is hard



Measuring patch similarity is hard

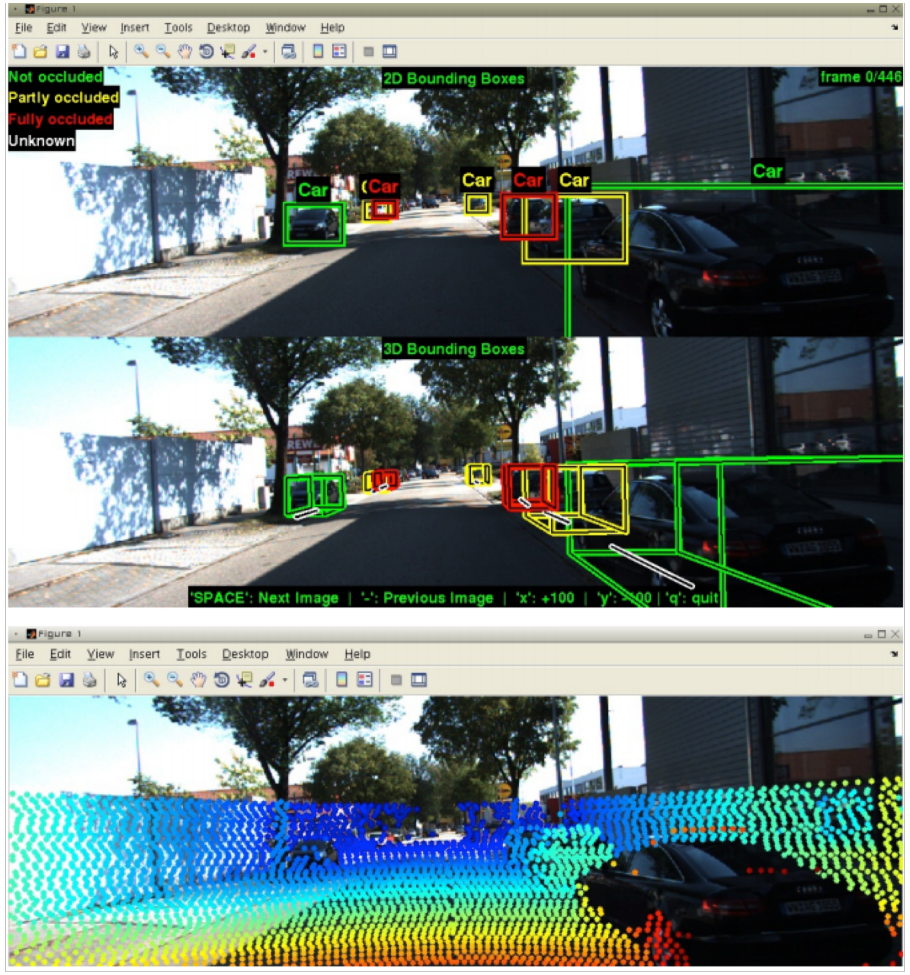
- Idea: learn to compute patch similarity?

The KITTI Dataset and Benchmark

Depth from here

Stereo pair from here

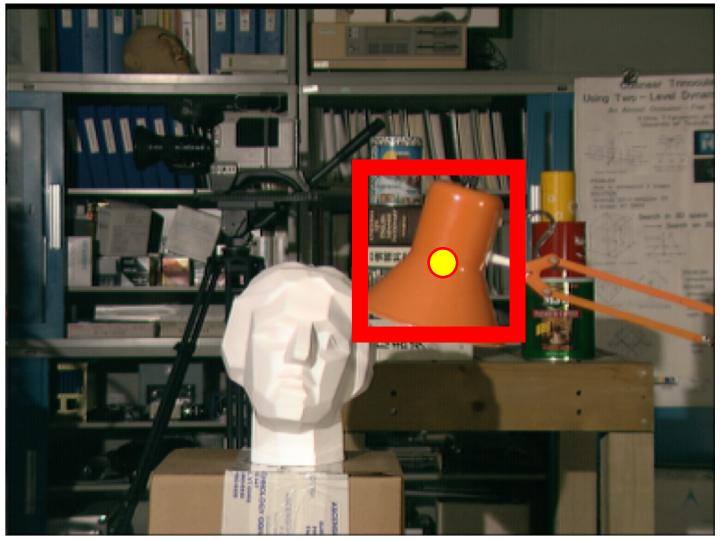
The KITTI Dataset and Benchmark



Vision meets Robotics: The KITTI Dataset. Andreas Geiger, Philip Lenz, Christoph Stiller and Raquel Urtasun. In *IJRR,* 2013.

Learning patch similarity for disparity estimation

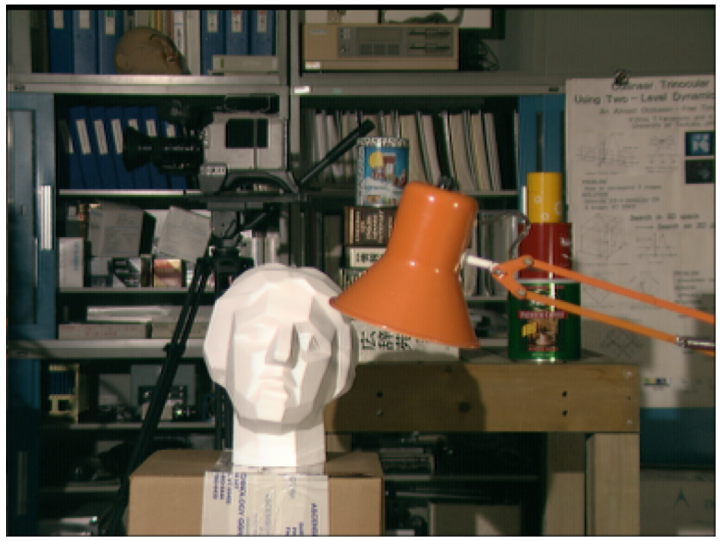
- Given a pixel in image 1, and corresponding row in image 2, get matching location



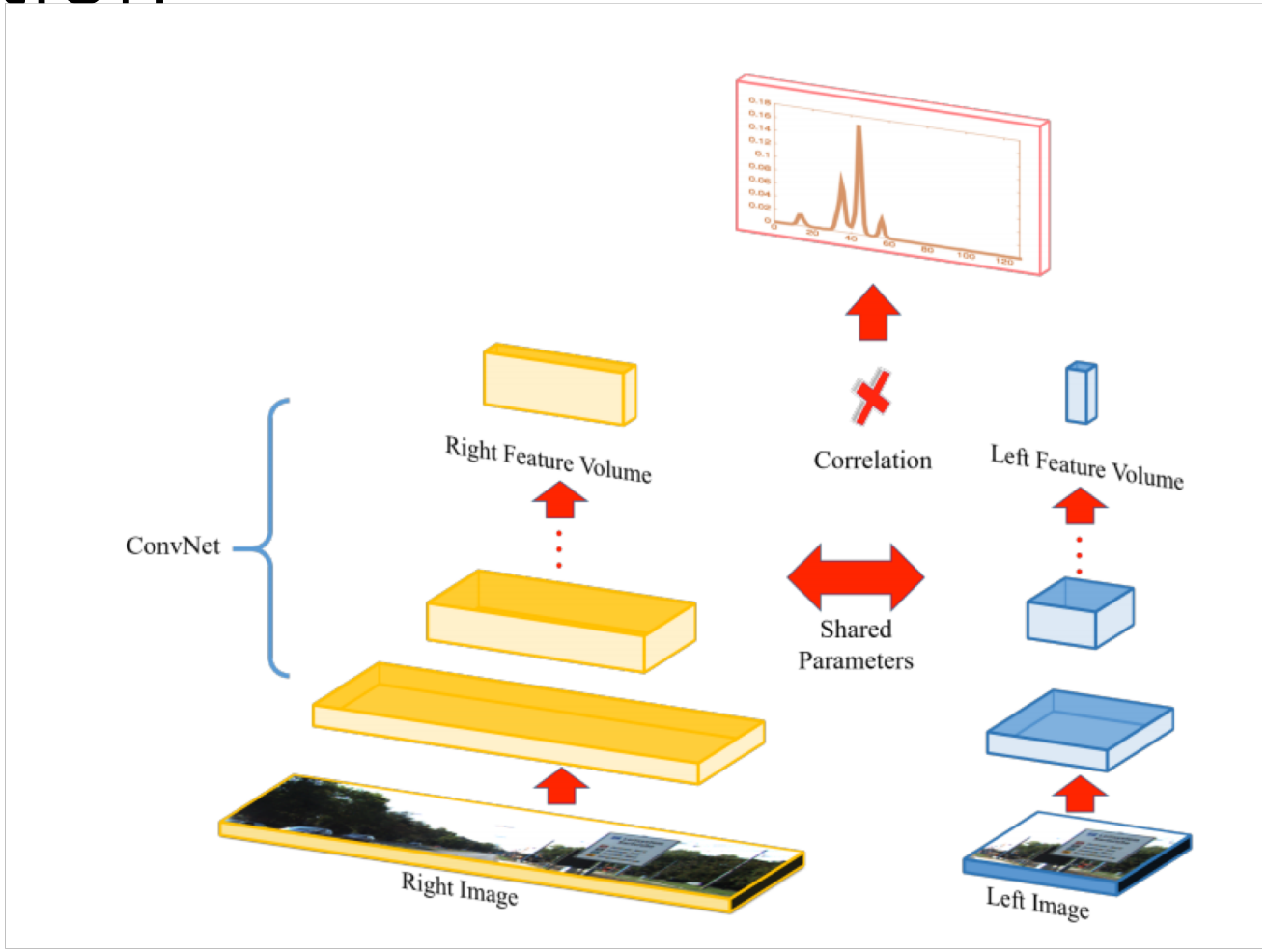
Learning patch similarity for disparity estimation

Required Output:

Input:



Learning patch similarity for disparity estimation



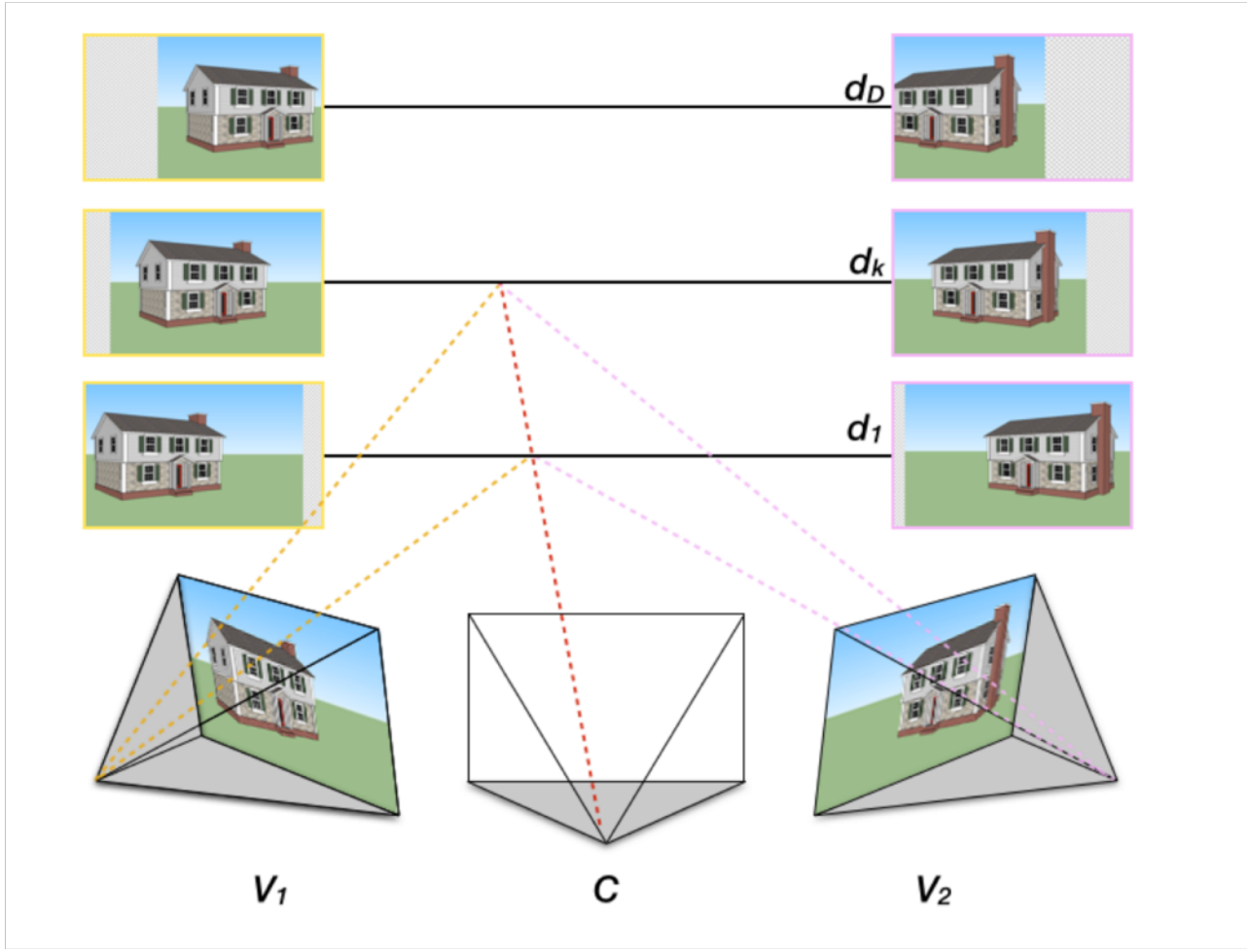
Efficient Deep Learning for Stereo Matching. Wenjie Luo, Alexander G. Schwing, Raquel Urtasun. In *CVPR,* 2016

Learning stereo without depth supervision

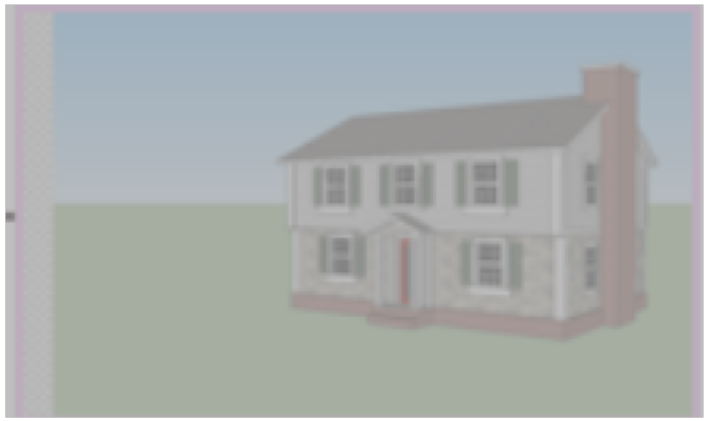
- Given scenes with >=3 rectified views
- Use 2 views to produce depth
  - Compute scores for each disparity
  - Match pixel to pixel with best disparity
  - Disparity = 1/depth
- Use depth to produce 3rd view
- Use reconstruction error

Plane-sweep stereo

- For every possible depth value d:
  - Assume every pixel in middle image has the same depth d
  - Use d to compute disparity to left and right image
  - *Reproject* left and right images to center coordinate system using disparity
  - *Compute score* for all pixels



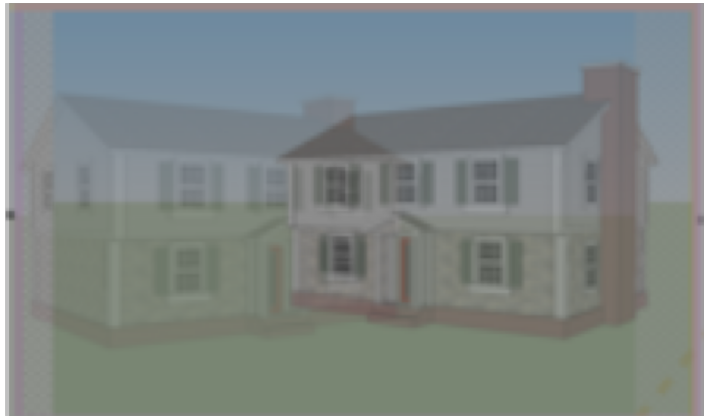
Plane-sweep stereo



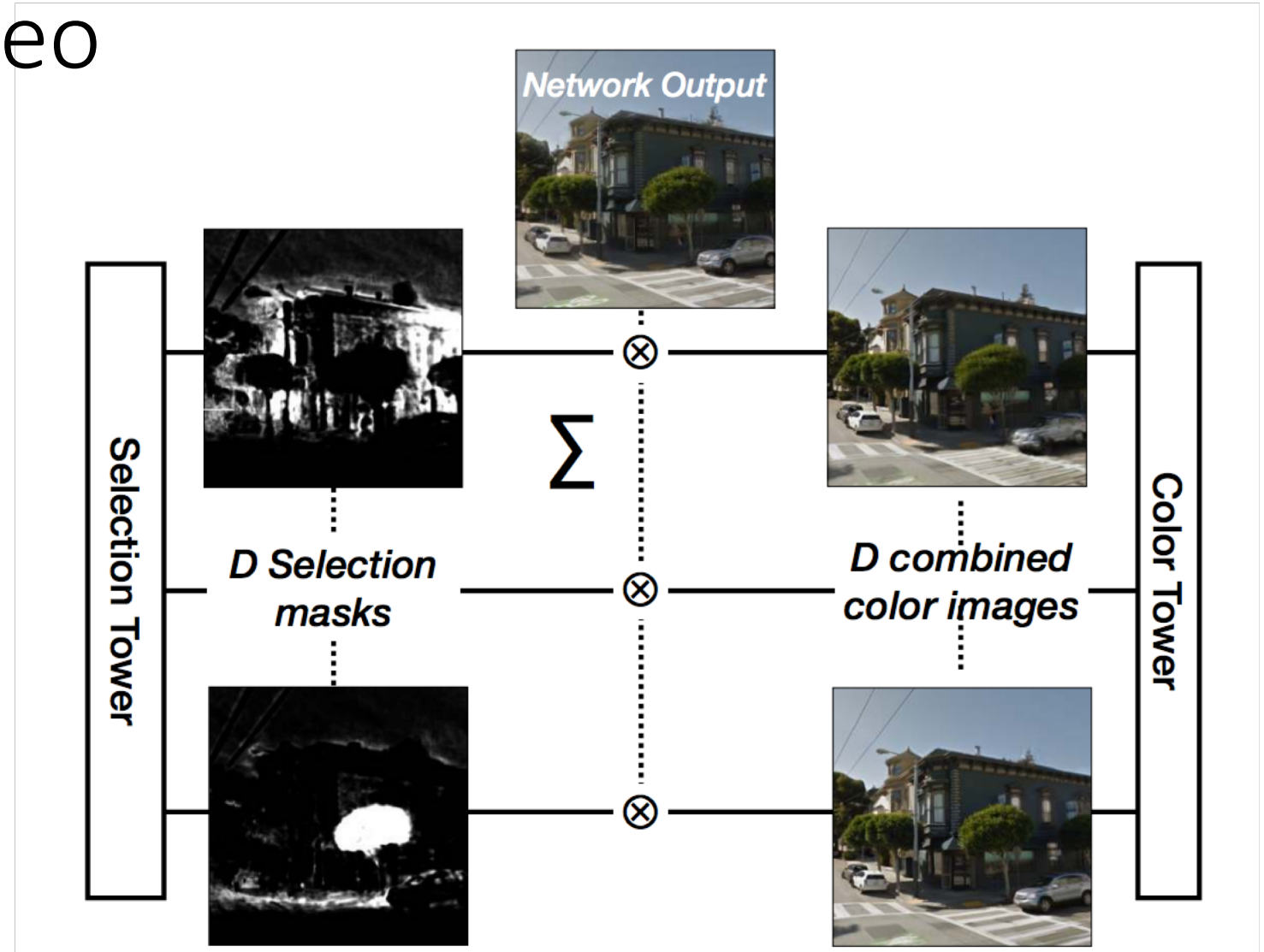
If depth is correct, appearance should match!

Plane-sweep stereo

- Score for pixel only depends on the two patches at that pixel
- Can be computed using *convolutions*



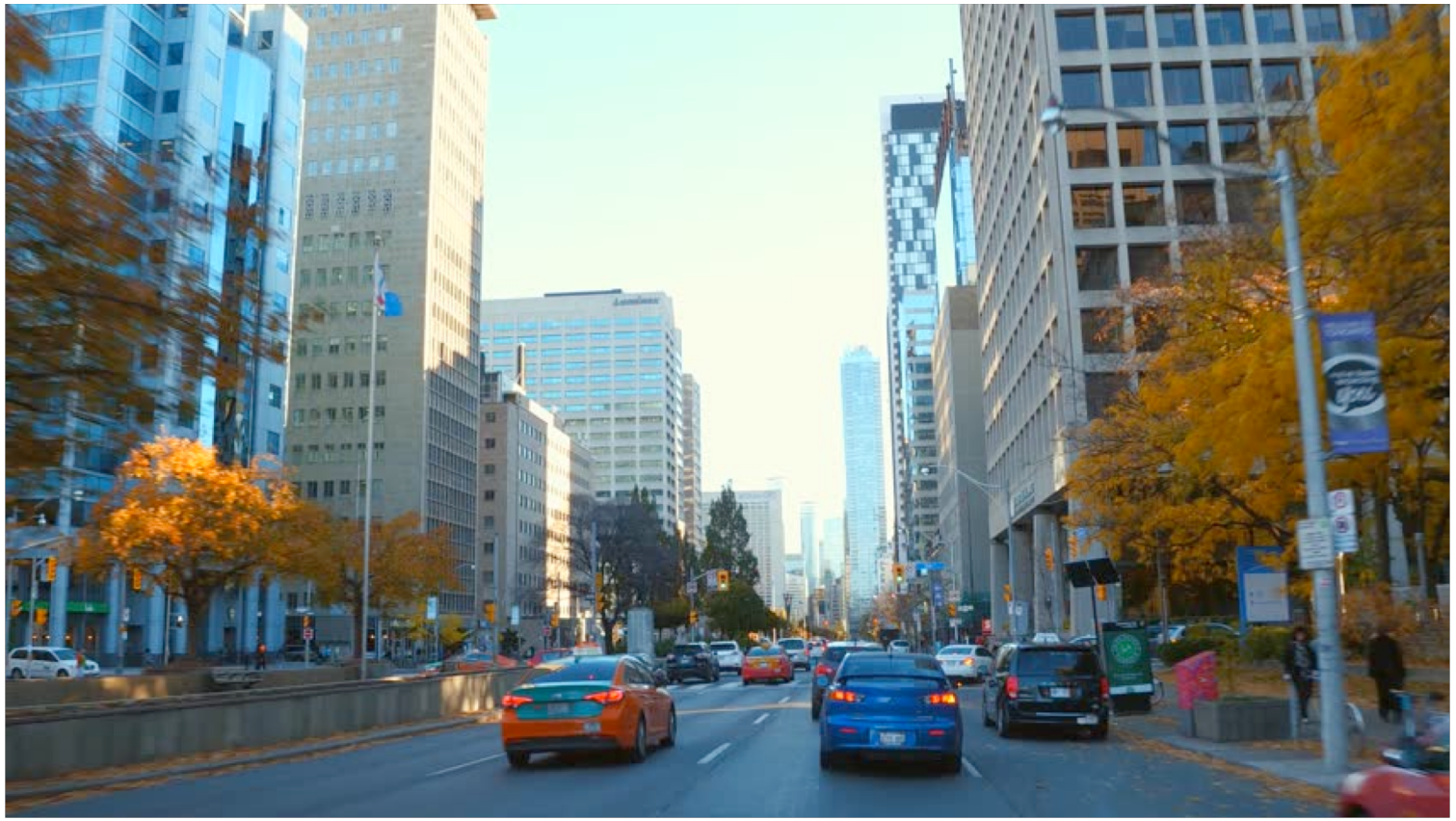
Deep Stereo



DeepStereo: Learning to Predict New Views from the World's Imagery. John Flynn, Ivan Neulander, James Philbin, Noah Snavely. *CVPR,* 2016

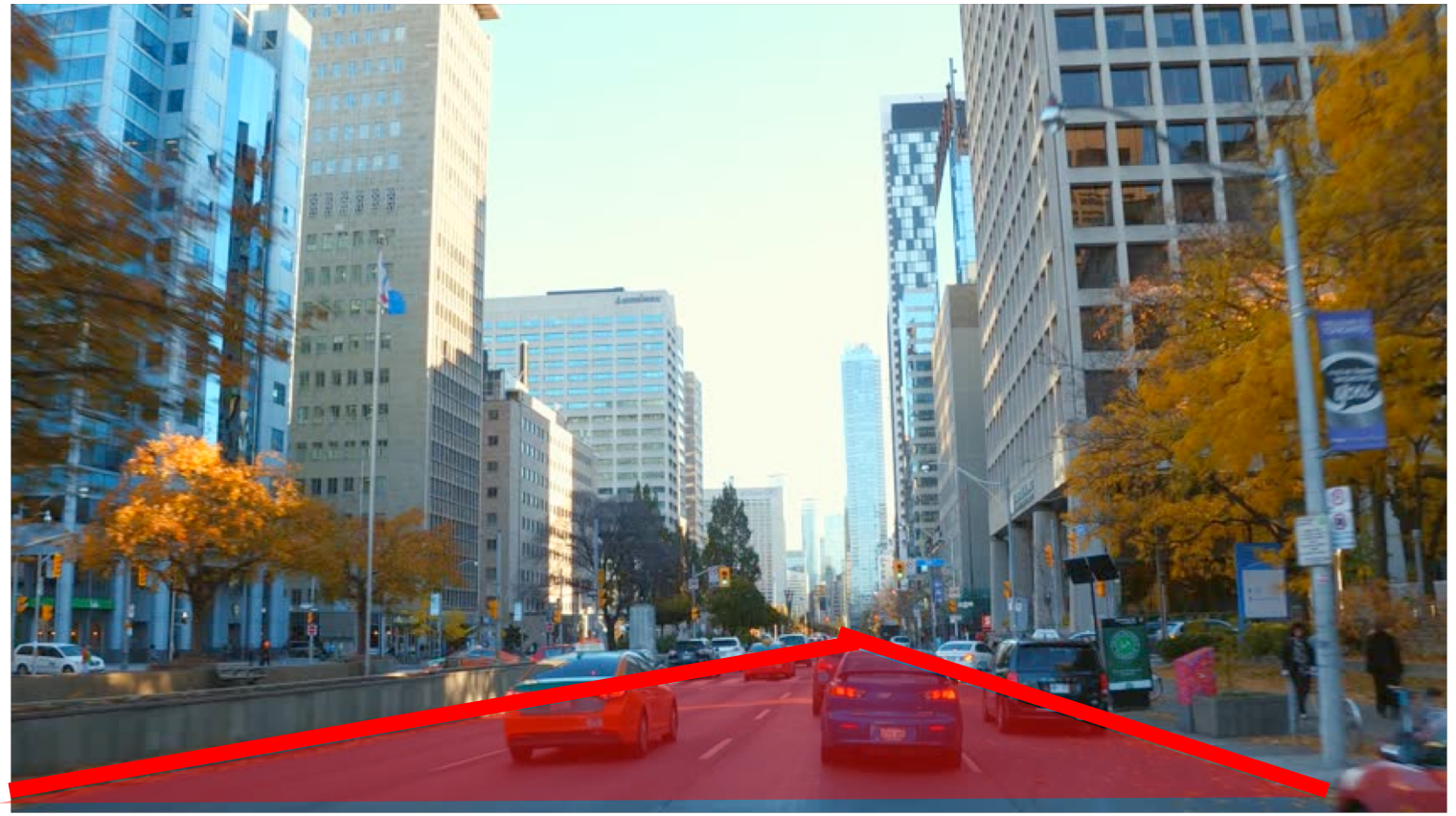
Estimating depth from a single image

- Why is this even possible?



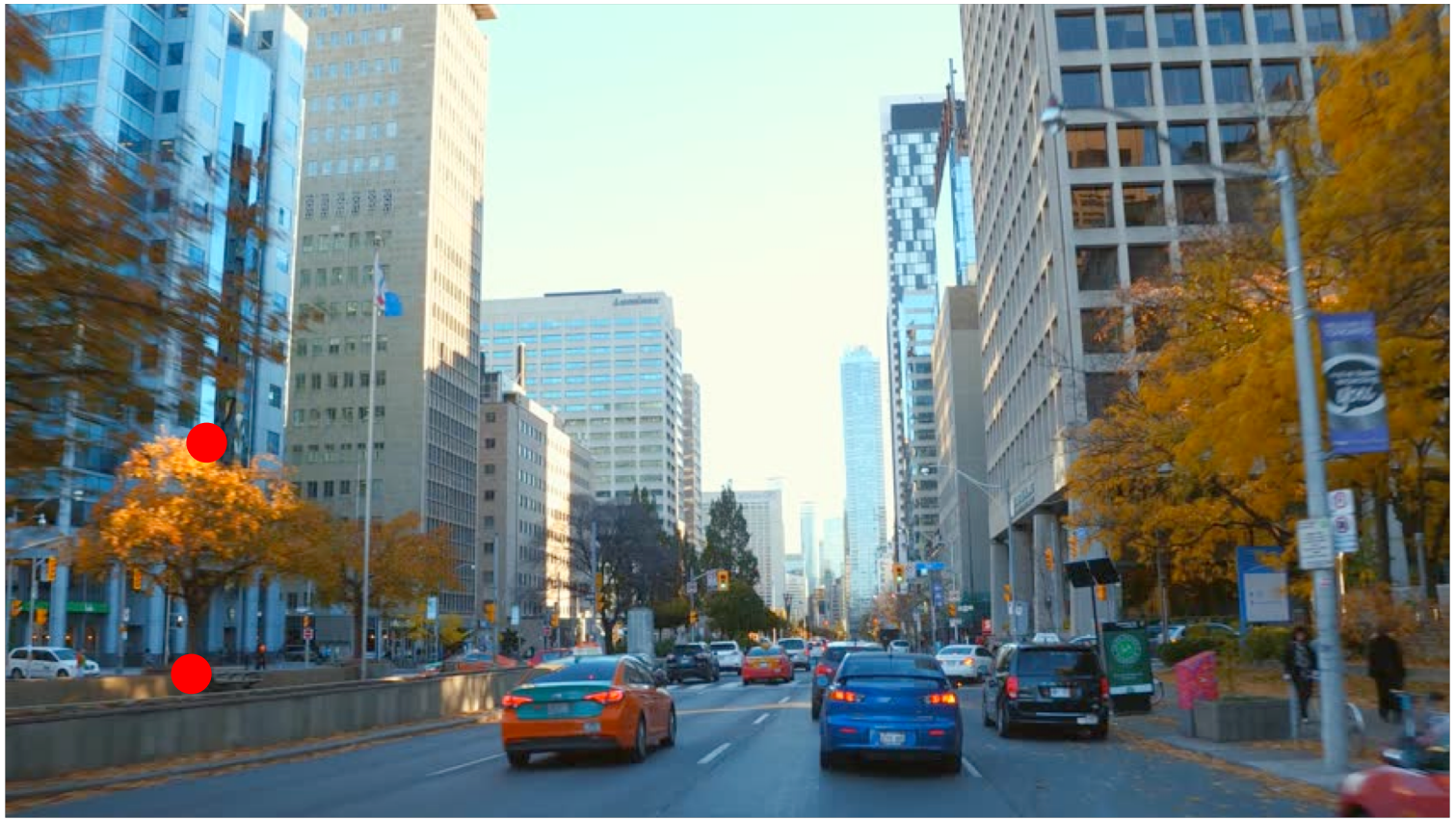
Estimating depth from a single image

- Why is this even possible?



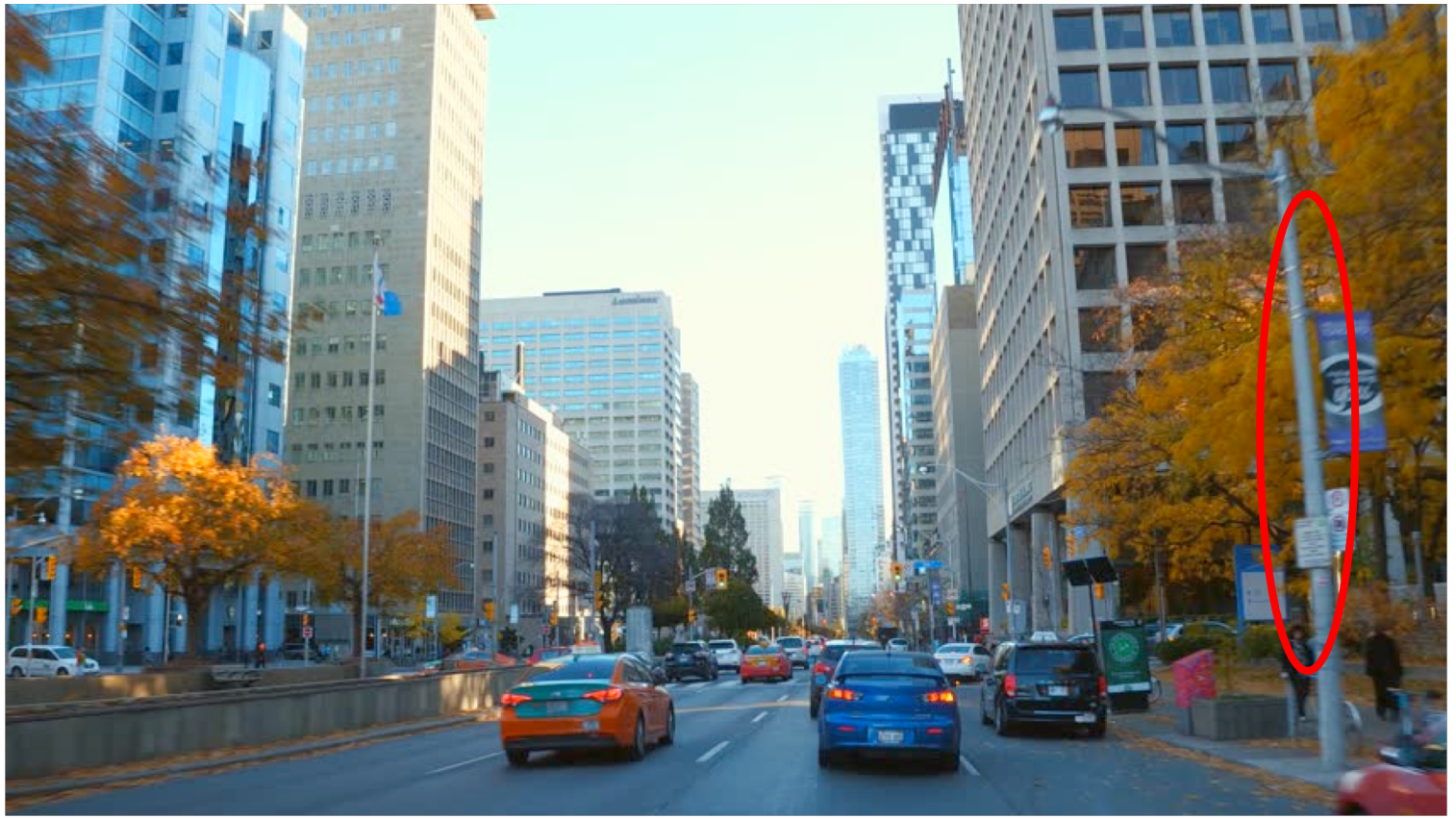
Estimating depth from a single image

- Why is this even possible?



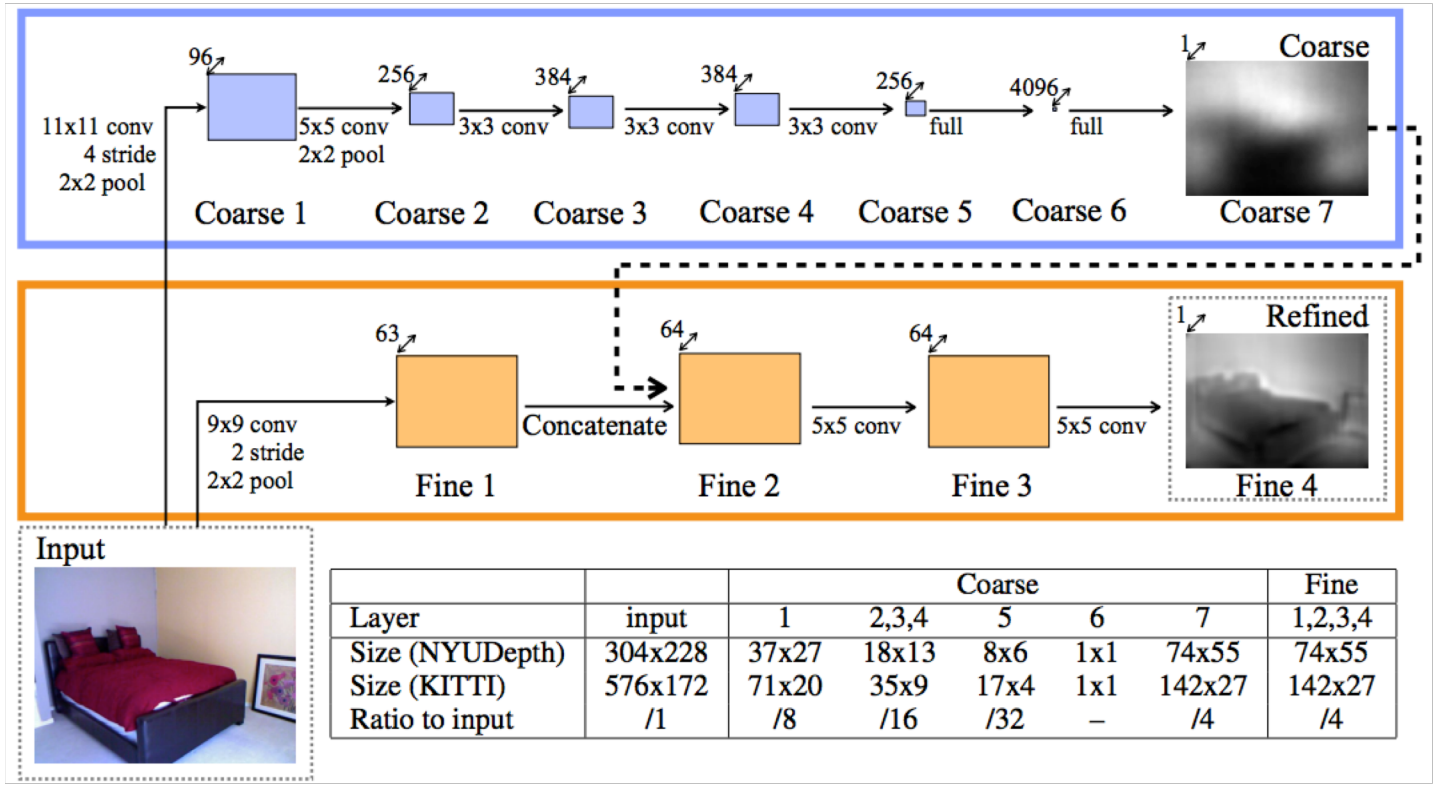
Estimating depth from a single image

- Why is this even possible?



Estimating depth from a single image

- Yet another image-to-image translation
- Again, resolution issues



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. David Eigen, Christian Puhrsch, Rob Fergus. In *NIPS,* 2014

Metric depth is a bad target

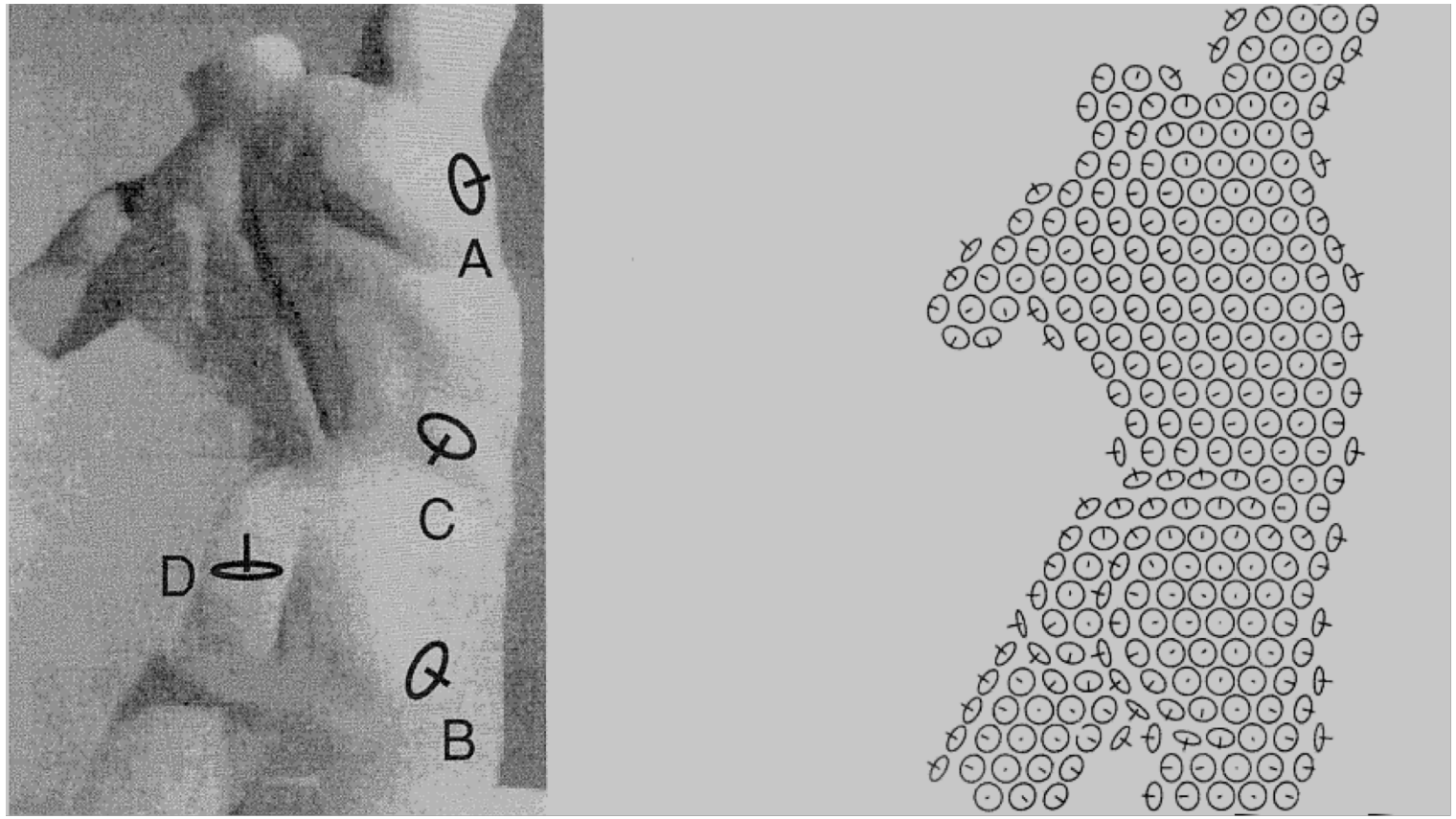
Metric depth is a bad target

- Only relative depths matter
- Only logarithmic scales matter

$$D(y, y^*) = \frac{1}{n^2} \sum_{i,j} ((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*))^2$$

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. David Eigen, Christian Puhrsch, Rob Fergus. In *NIPS,* 2014

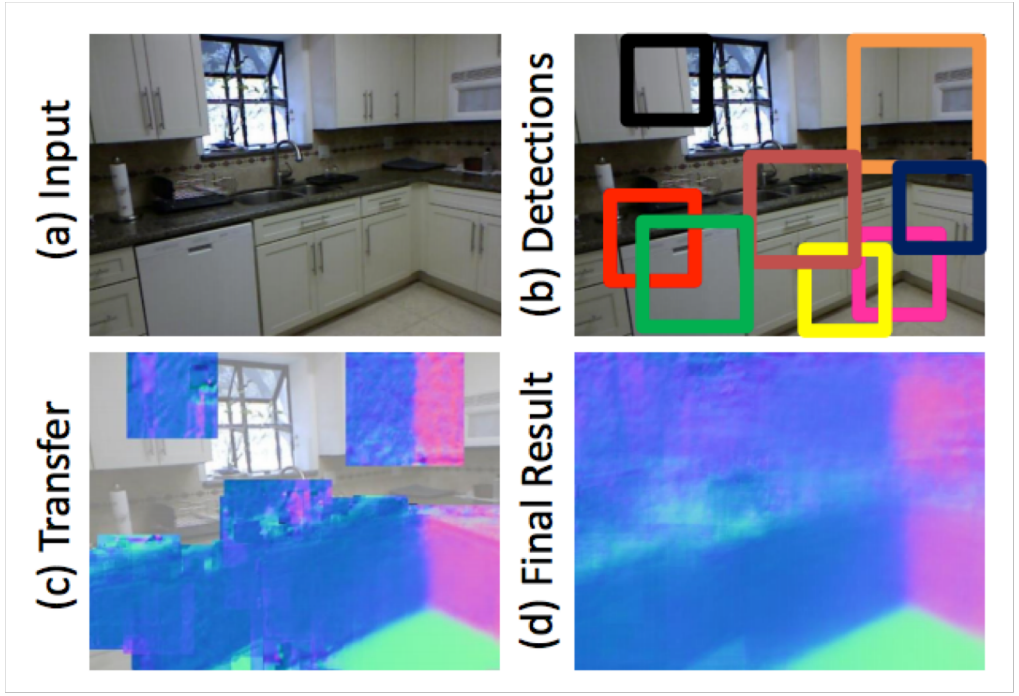
Humans perceive surface normals, not just depth, through a combination of various pictorial cues



Surface perception in pictures. Koenderink, van Doorn and Kappers, 1992

Slide credit: Jitendra Malik

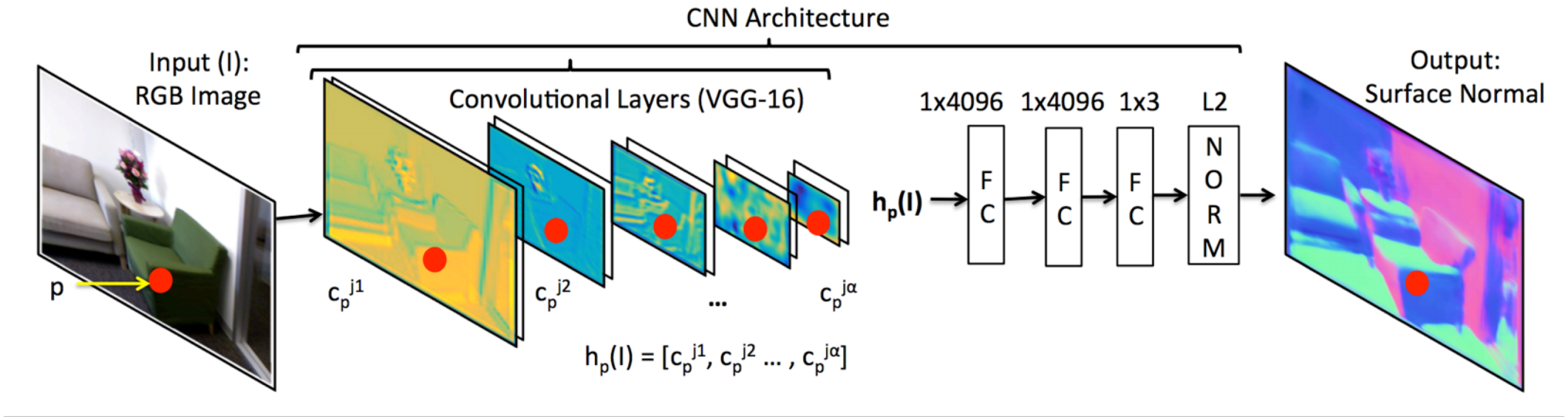
Estimating normals from a single image



Data-Driven 3D Primitives for Single Image Understanding. David F. Fouhey, Abhinav Gupta, Martial Hebert. In *ICCV* 2013.

Estimating normals from a single image

Estimating normals from a single image



Marr Revisited: 2D-3D Alignment via Surface Normal Prediction. Aayush Bansal, Bryan Russell, Abhinav Gupta. In *CVPR,* 2016

Estimating normals from a single image

2.5D vs 3D prediction

- Predicting depth / surface normals for every pixel is not full reconstruction
  - “2.5D reconstruction”
  - Does not contain parts of the scene that are hidden from view
- Can we do full 3D reconstruction?
- Simpler situation: can we do full 3D reconstruction of isolated objects?

Shapenet

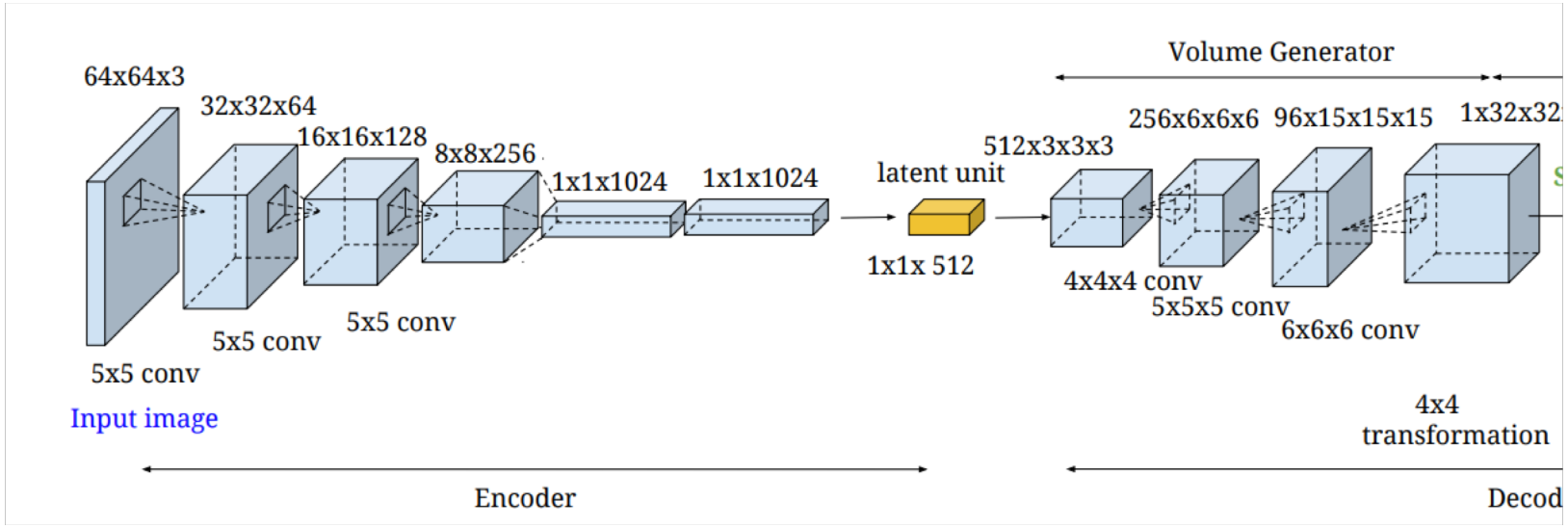
Reconstructing 3D shapes from images using machine learning

- Input:
  - Single image or multiple images of the same object
- Output:
  - 3D shape
- Representation?

Representation of 3D shapes

- Voxel grids
  - Discretize volume into grid cells
  - Identify cells that are occupied by object
- Advantages:
  - Easy representation for ML: analog of pixels
- Disadvantages:
  - Memory-inefficient
  - Difficult to capture surface

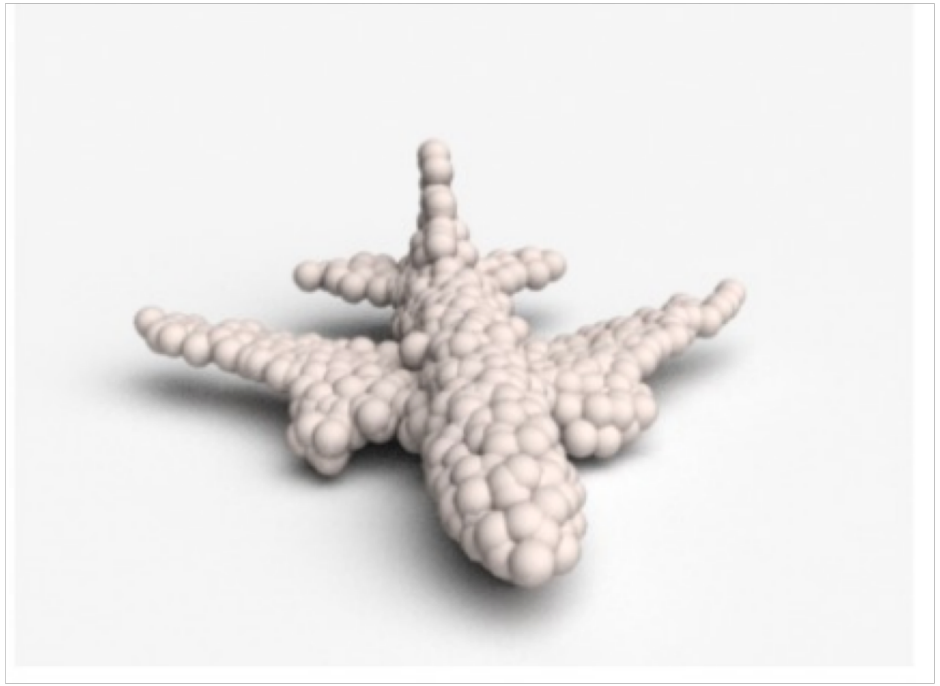
Architectures for generating voxel grids



1. Choy, Christopher B., et al. "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction." *European conference on computer vision*. Springer, Cham, 2016.
2. Yan, Xinchen, et al. "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision." *Advances in Neural Information Processing Systems*. 2016.

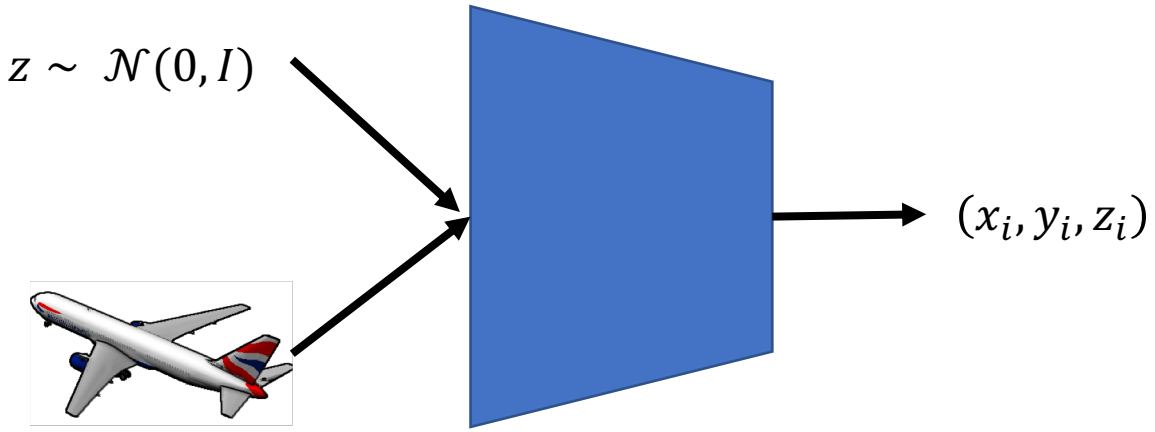
Representation of 3D shapes

- Point clouds
- Each point lies on surface
- Advantages:
  - Common representation produced by sensors (e.g. LiDAR)
  - Sparse, so memory efficient
- Disadvantages:
  - Difficult output to predict: sets
  - Difficult to extract surface



Architecture for generating point clouds

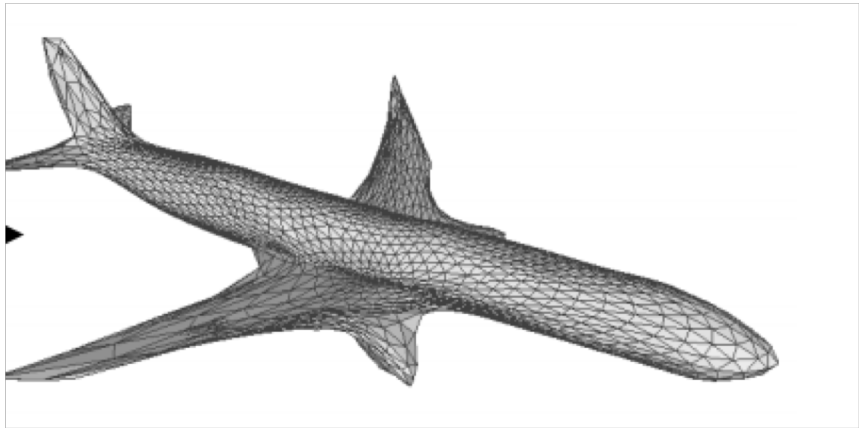
- Not an established answer
- One possibility: cloud of points = *samples* from an underlying distribution
- Generative modeling



Fan, Haoqiang, Hao Su, and Leonidas J. Guibas. "A point set generation network for 3d object reconstruction from a single image." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Representation of 3D shapes

- Meshes
- Advantages
  - Common in graphics
  - Surfaces are triangles in the mesh
  - Sparse representation: memory efficient
  - Can easily encode color, texture, surface normals
- Disadvantages
  - Extremely difficult to predict: graph



Architecture for producing meshes

- Assume connectivity and faces are the same as that of a sphere
- Move only vertices
- Cannot change *topology* of objects

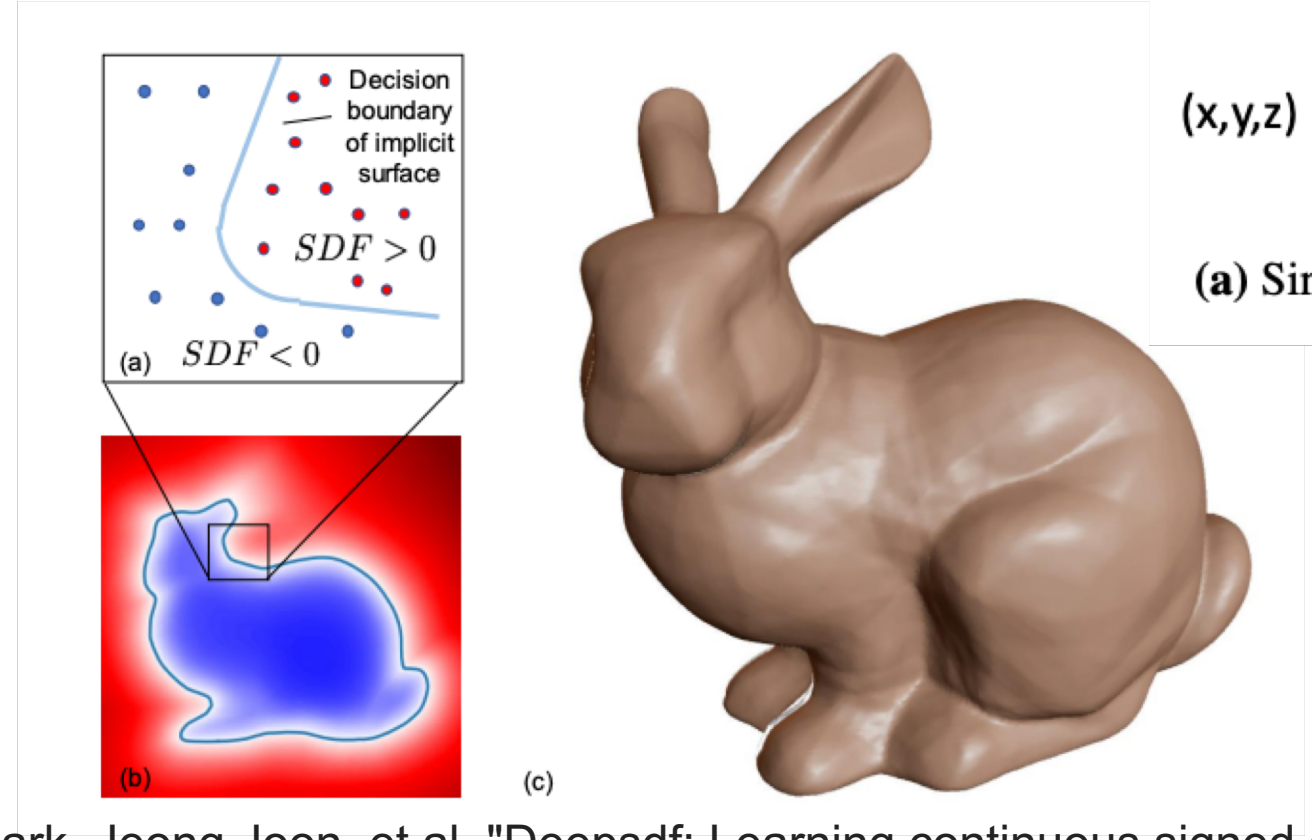
Wang, Nanyang, et al. "Pixel2mesh: Generating 3d mesh models from single rgb images." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

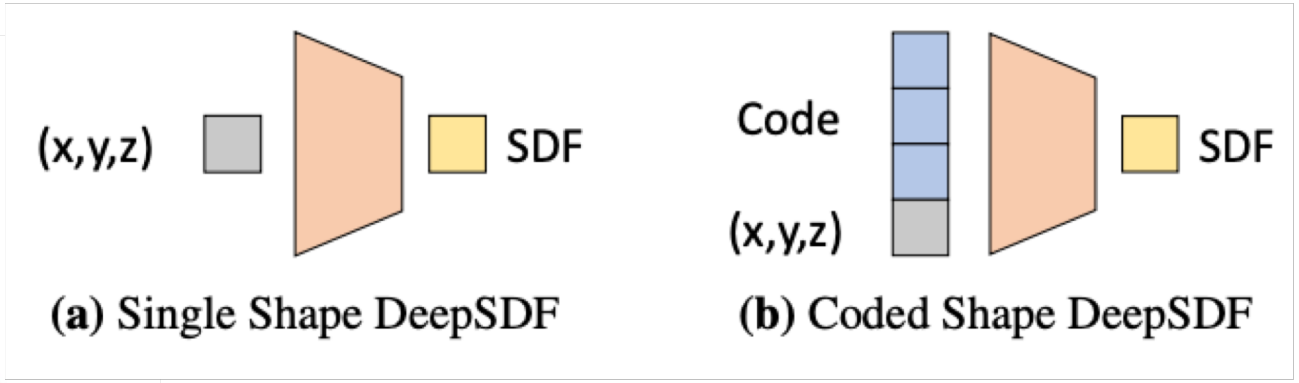
Representation of 3D shapes

- Implicit shapes
- A shape is a *function* that takes (x, y, z) as input and produces as output
  - Boolean on whether it is inside shape or not
  - Real value indicating distance from surface (”signed distance functions”)
- This *function* can be a *neural network*
- Thus each shape is a *neural network*
- Can additionally take e.g. feature vector as input

Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Implicit shapes





Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Supervision?

- Fully supervised [1]
- Supervised with multiple views from *known* cameras [2]
  - Predict shape from one image
  - Project it to other views
  - Ensure *photometric consistency*
- Supervised with multiple views from *unknown* cameras [3]
  - Also jointly learn to predict camera pose

1. Choy, Christopher B., et al. "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction." *European conference on computer vision*. Springer, Cham, 2016.
2. Yan, Xinchen, et al. "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision." *Advances in Neural Information Processing Systems*. 2016.
3. Tulsiani, Shubham, et al. "Multi-view supervision for single-view reconstruction via differentiable ray consistency." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.