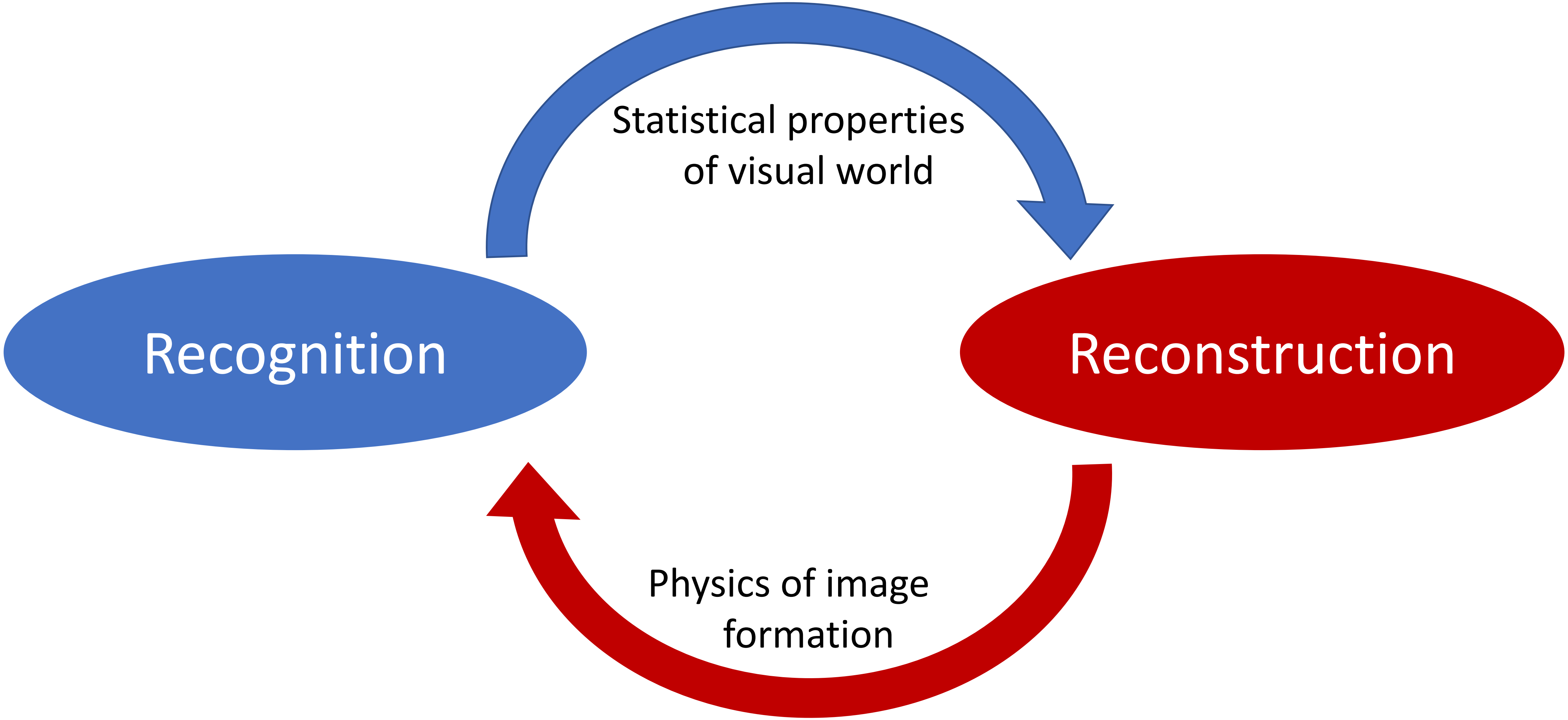
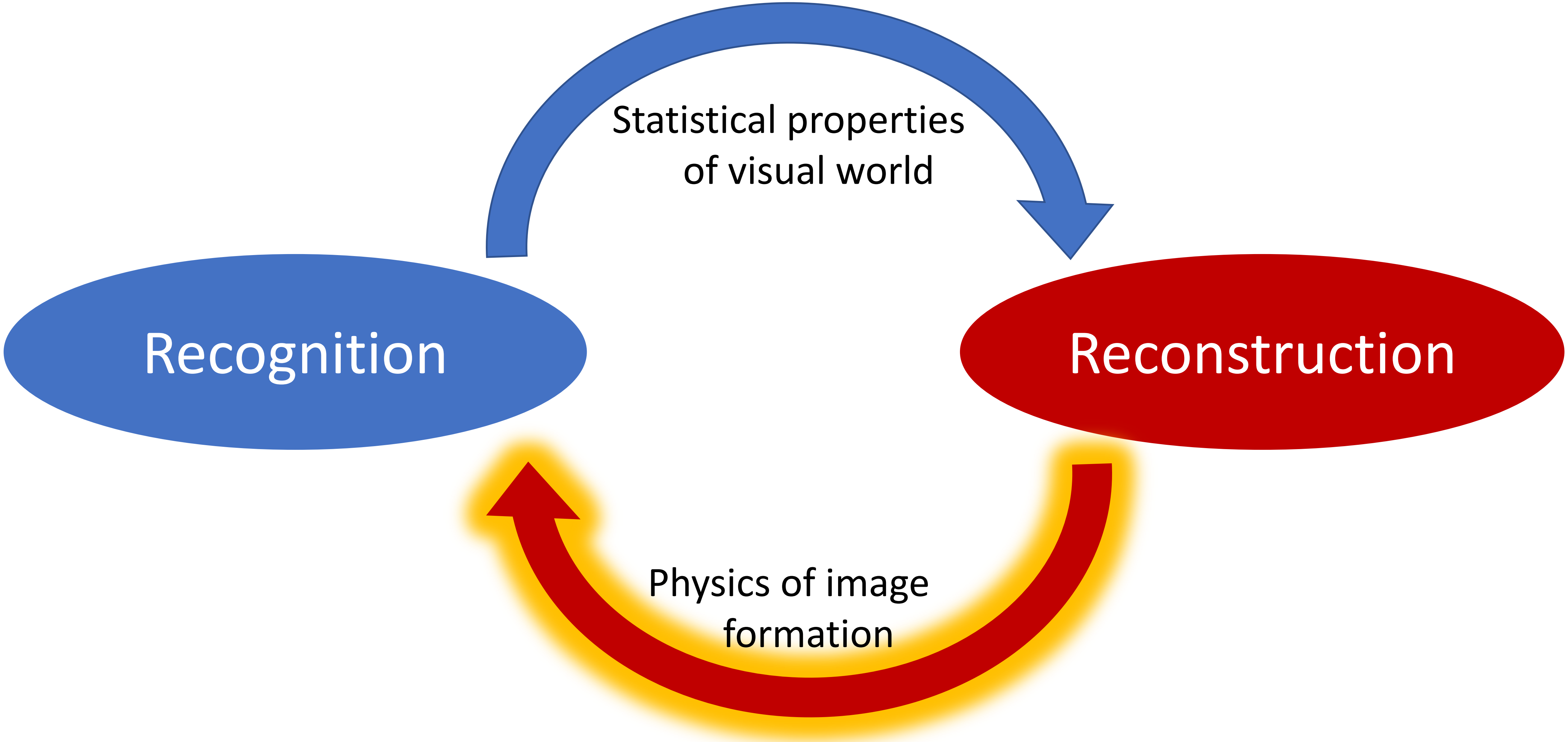
# Learning for 3D

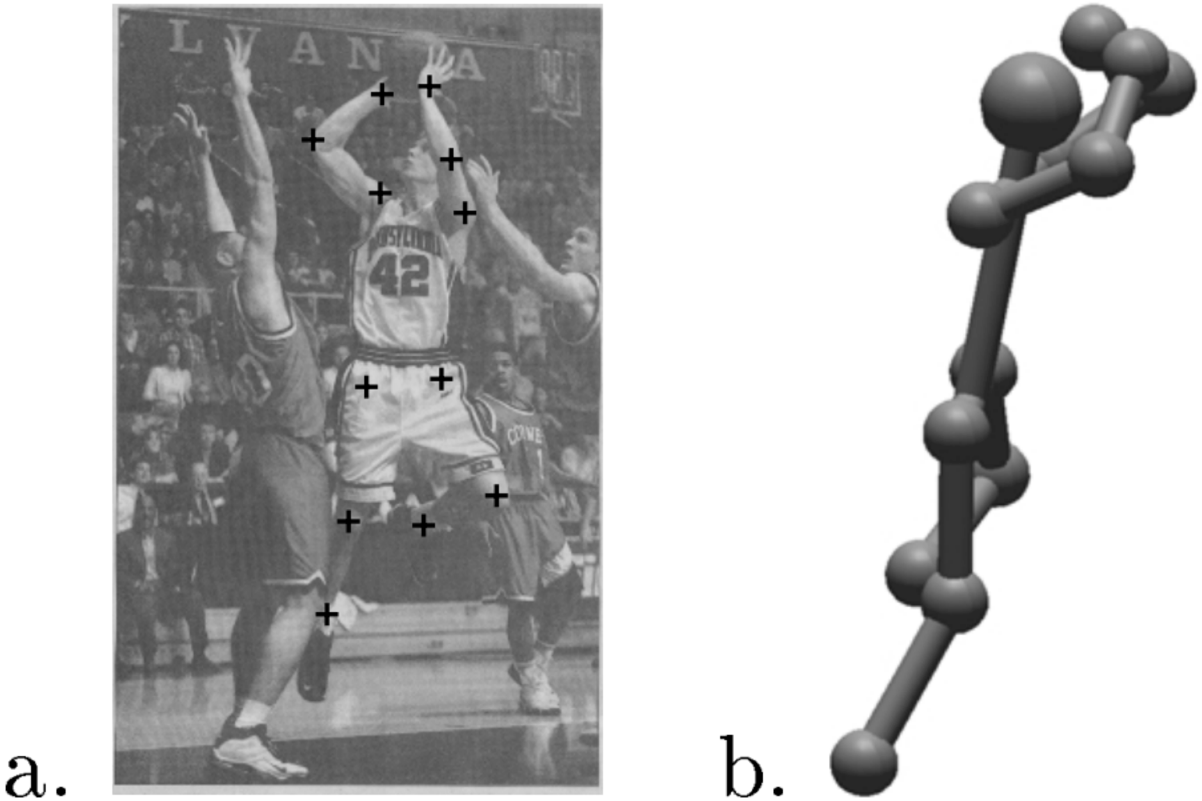
# Recognition and 3D reasoning



# 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

$$x = \frac{X}{Z} \approx \underbrace{\frac{X}{Z_0}}_{\text{constant}}$$

$$y = \frac{Y}{Z} \approx \underbrace{\frac{Y}{Z_0}}_{\text{constant}}$$

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

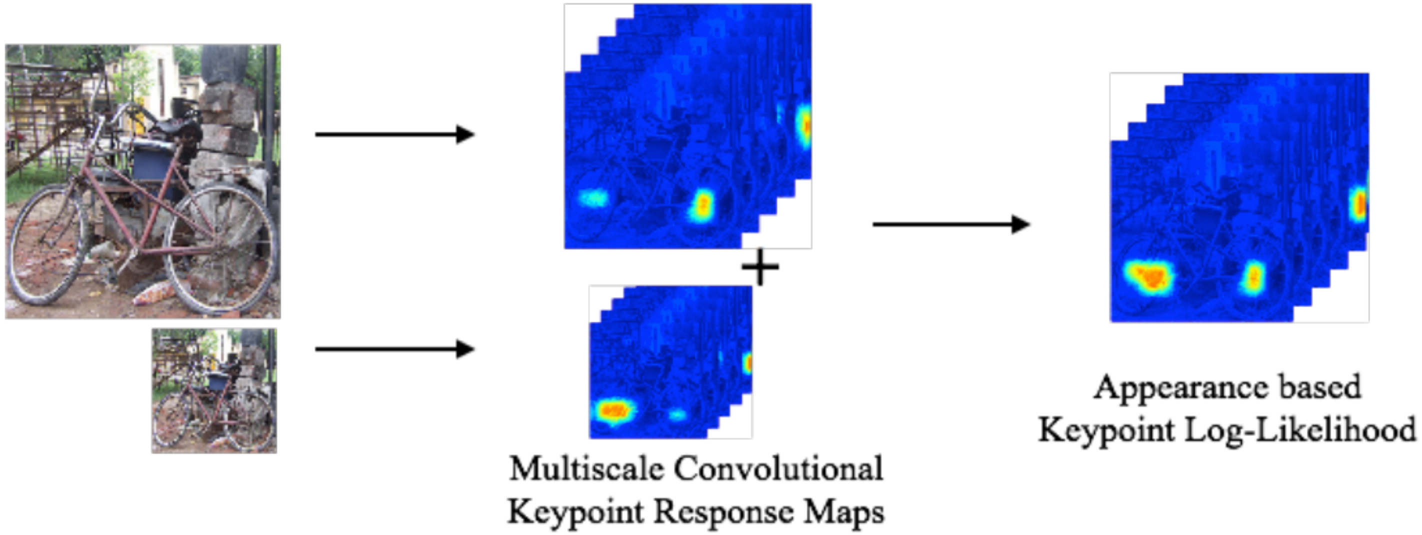
# Pose estimation in 3D

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

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

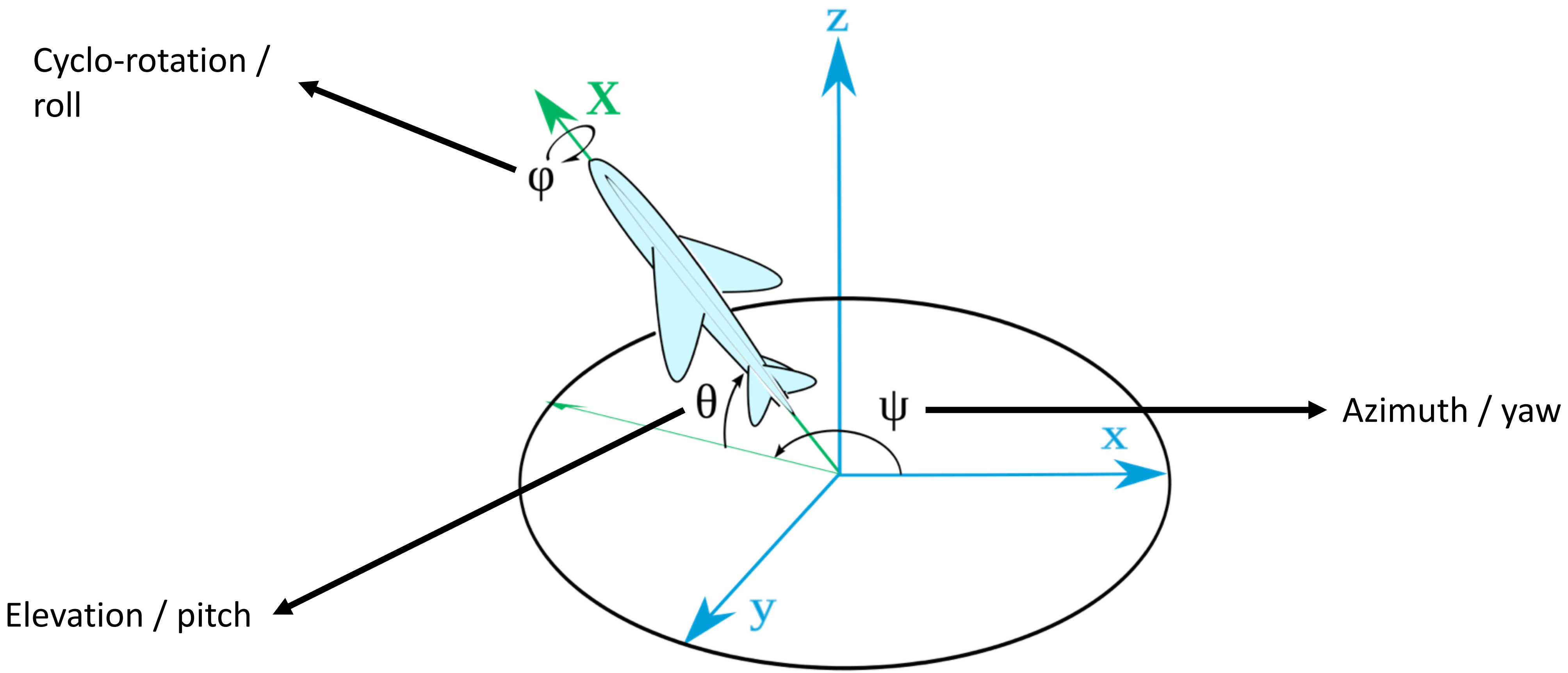
# Pose estimation for rigid objects

# Pose estimation for rigid objects



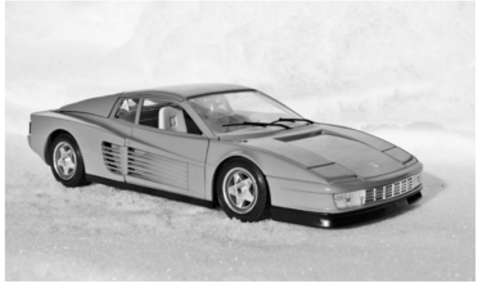

Viewpoints and keypoints. S. Tulsiani and J. Malik. In *CVPR,* 2015

# Pose estimation for rigid objects



# Viewpoint-conditioned pose

Viewpoint prediction



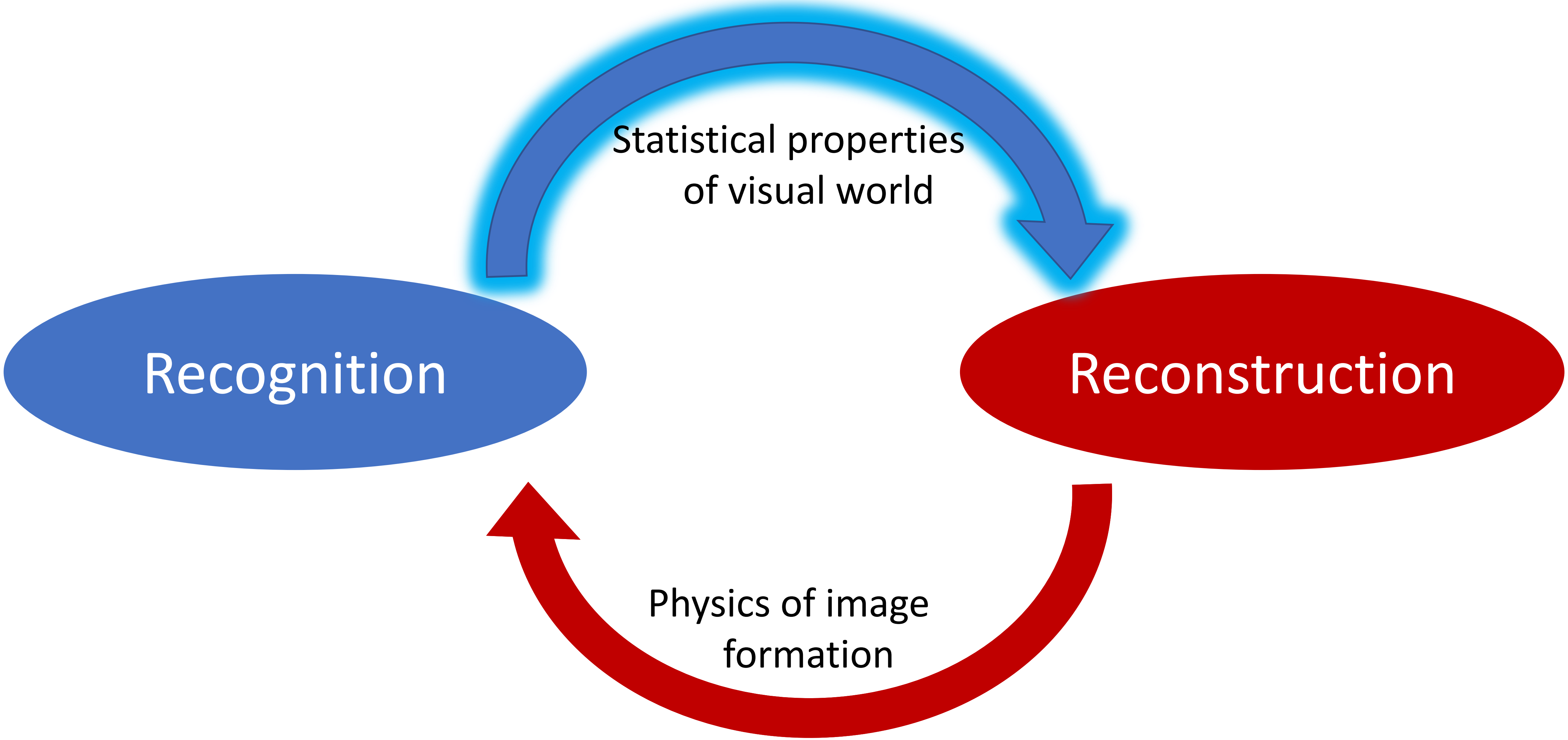
Viewpoints and keypoints. S. Tulsiani and J. Malik. In *CVPR,* 2015

# Viewpoint-conditioned pose

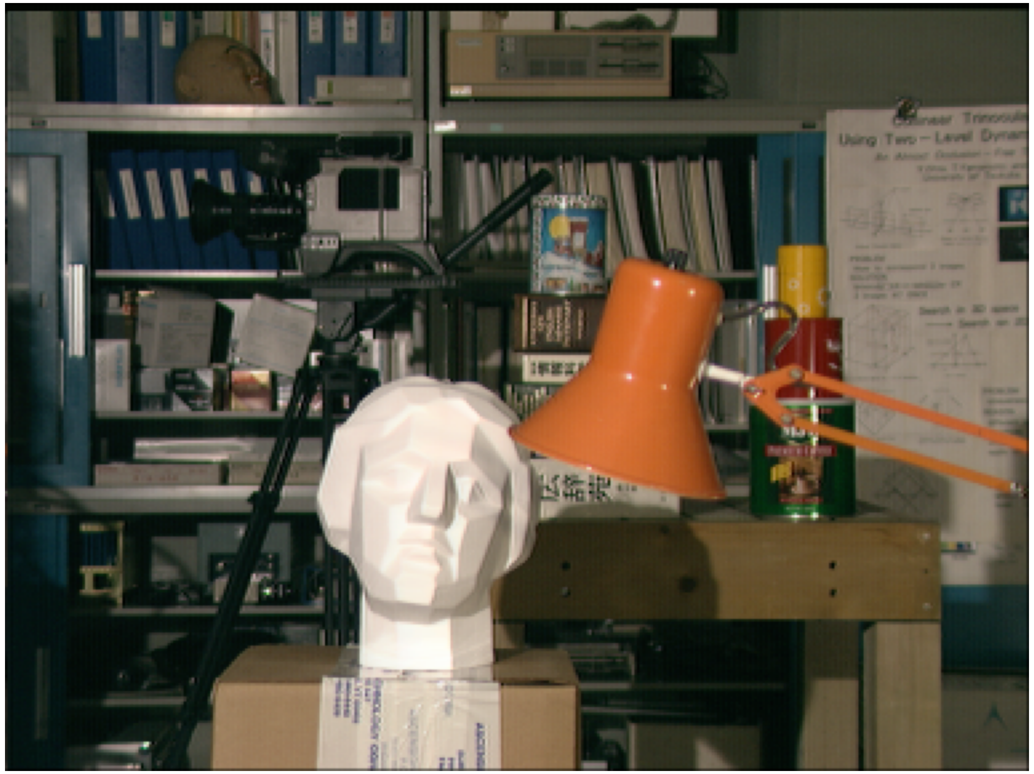
Viewpoint
prediction

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

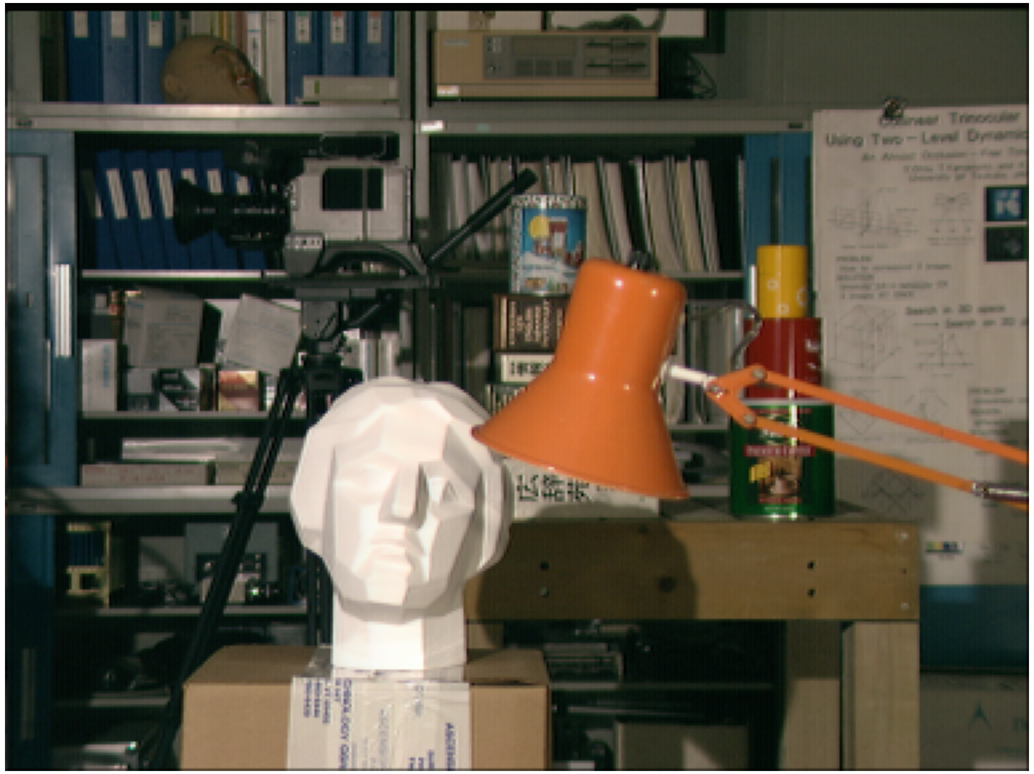
# Recognition and 3D reasoning



# Disparity estimation



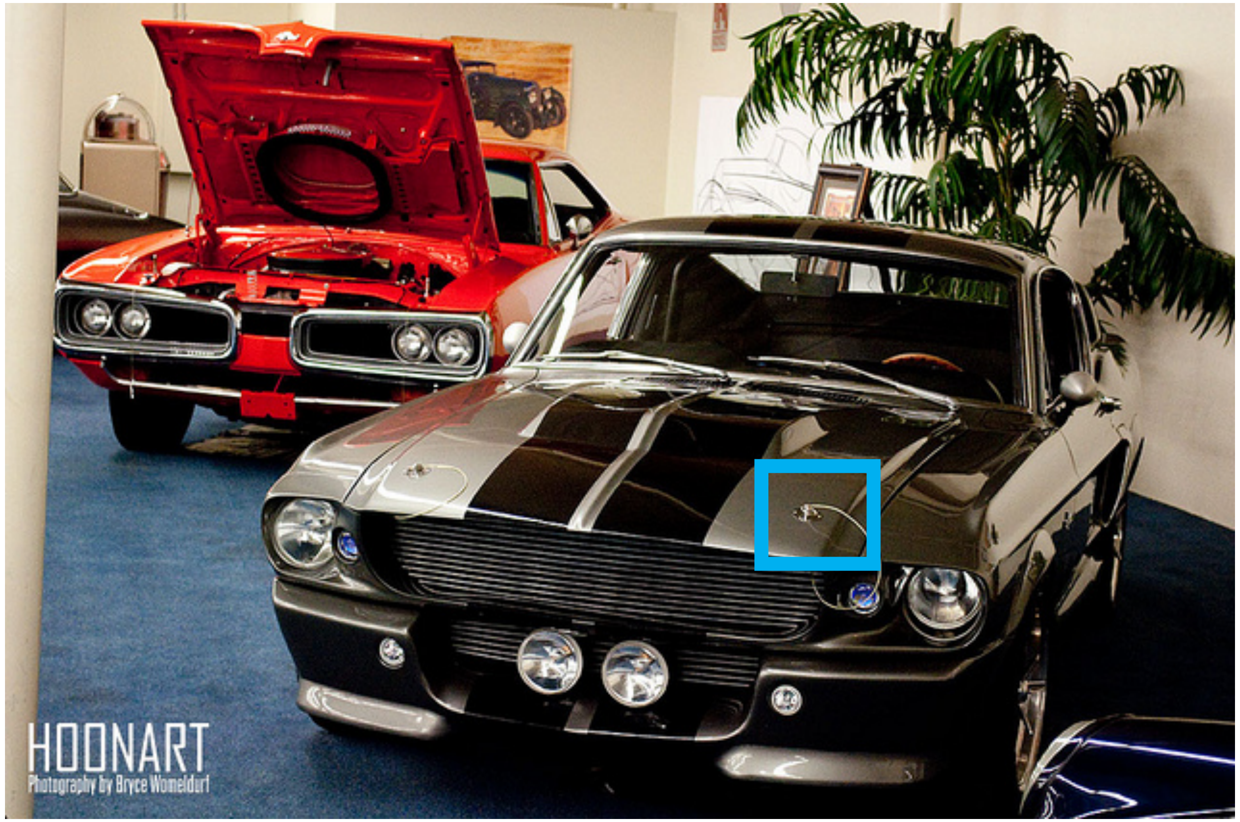
# Disparity estimation



# 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

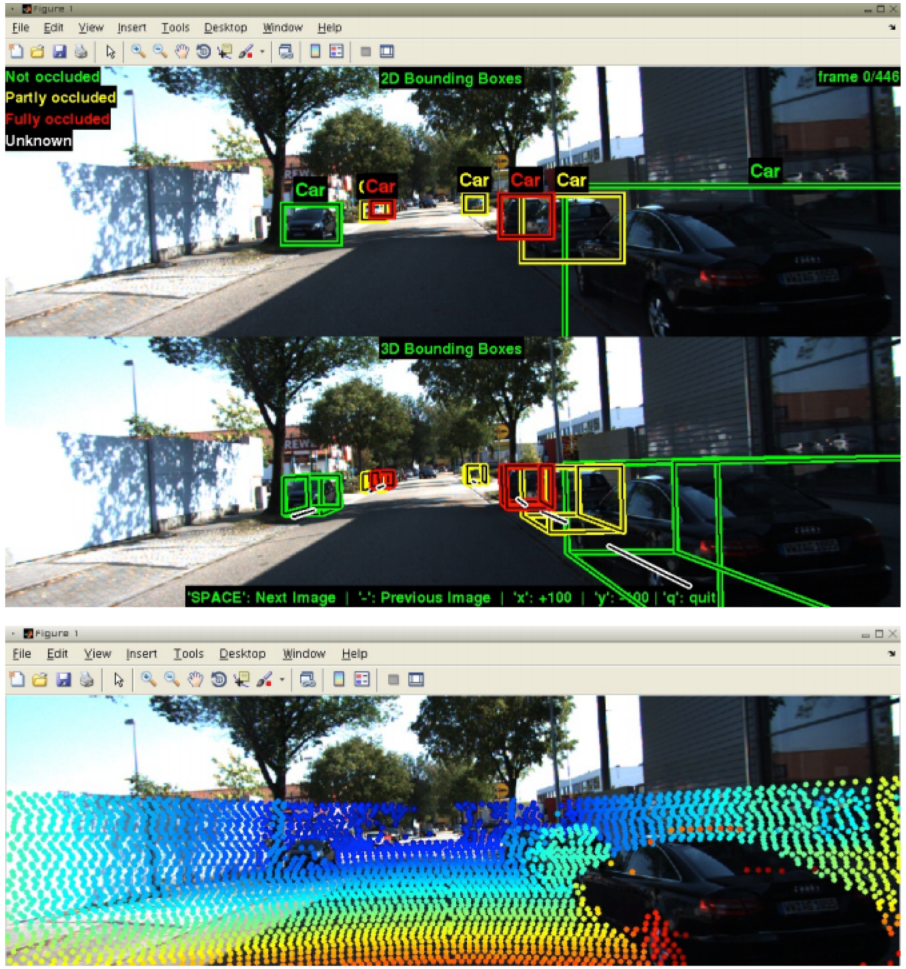
- Idea: learn to compute patch similarity?

# The KITTI Dataset and Benchmark

Depth from here

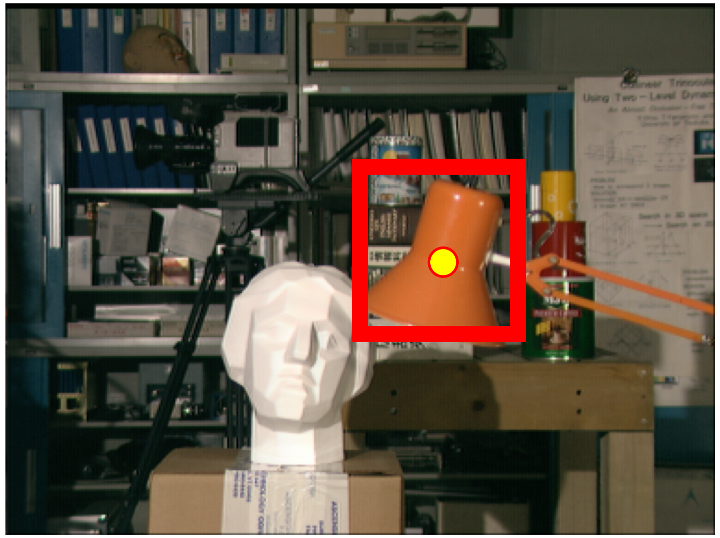

Stereo pair from here

# The KITTI Dataset and Benchmark

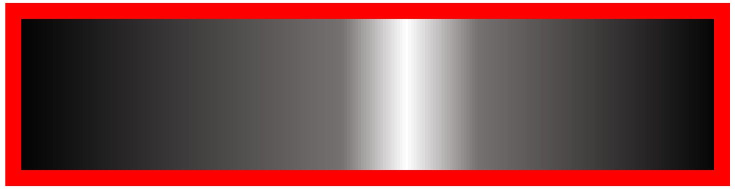



Vision meets Robotics: The KITTI Dataset. Andreas Geiger, Philip Lenz, Christoph Stiller and Raquel Urtasun. In *IJRR,* 2013.

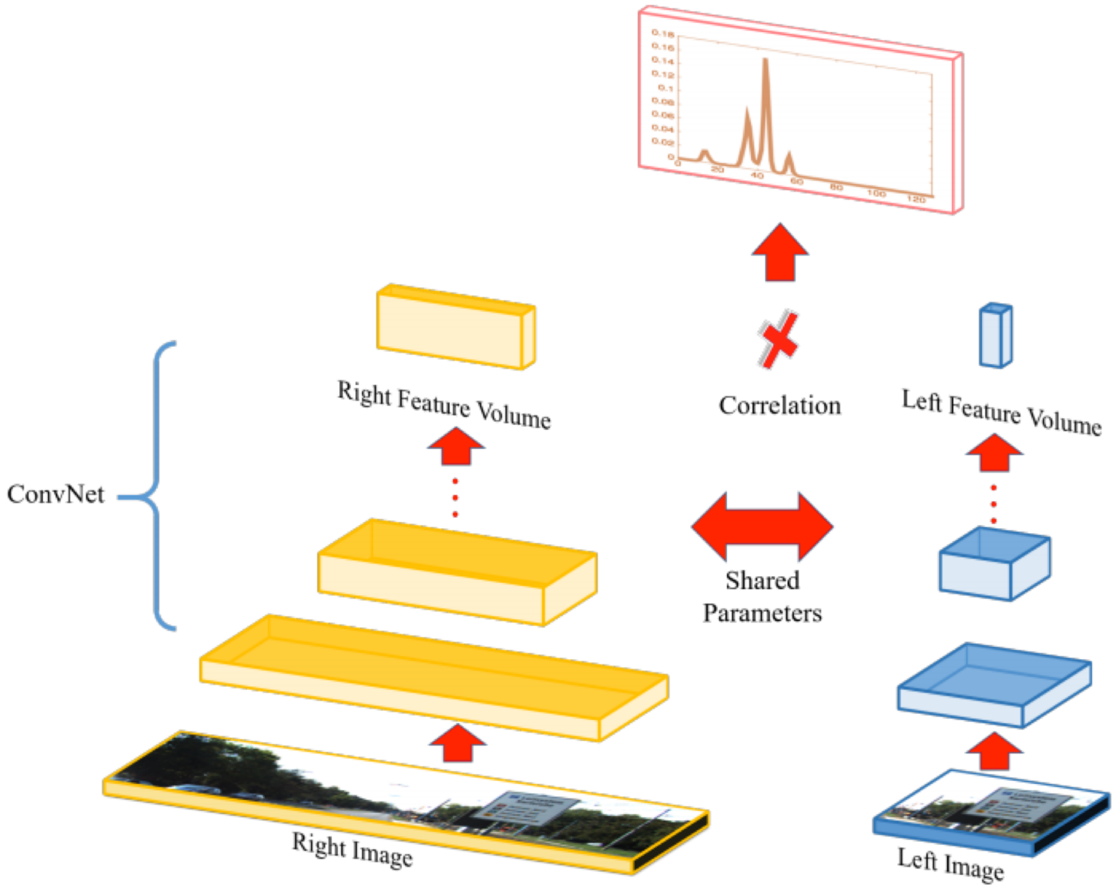
# Learning patch similarity for disparity estimation



# Learning patch similarity for disparity estimation



# 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

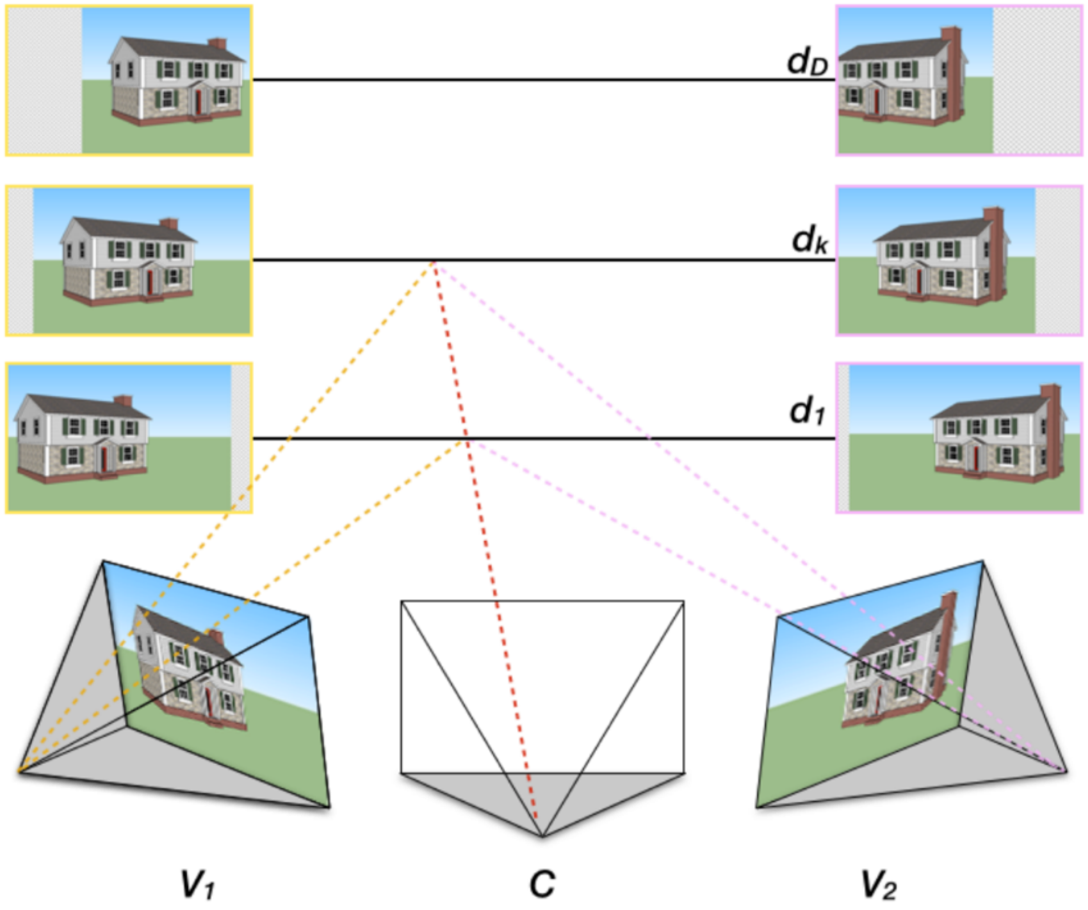
- Given scenes with >=3 rectified views
- Use 2 views to produce depth
  - Compute scores for each disparity
  - Match pixel to pixel with best disparity
  - Disparity = 1/depth
- Use depth to produce 3$^{rd}$ view
- Use reconstruction error

# Plane-sweep stereo

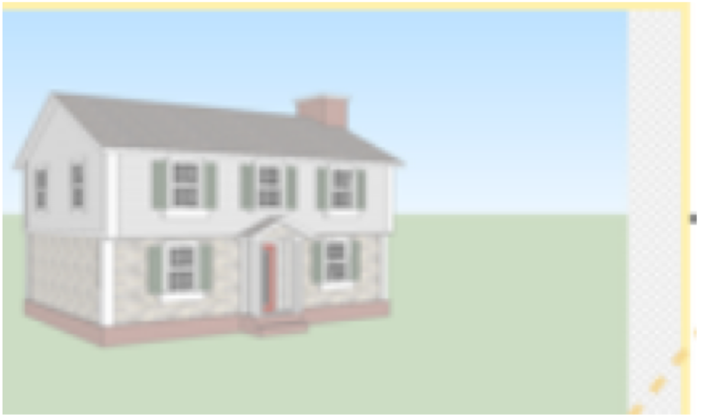
- Stereo till now:
  - Go *pixel by pixel*
  - For each pixel, *compute score for each disparity*
- Plane-sweep:
  - Go *disparity-by-disparity (or depth-by-depth)*
  - For each disparity, *compute scores for all pixels*

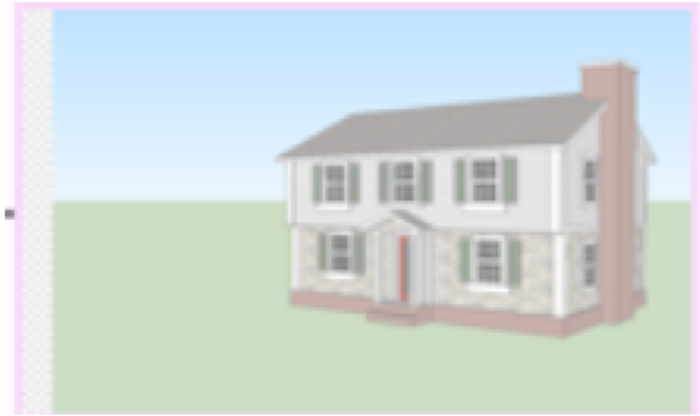
# Plane-sweep stereo

- 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

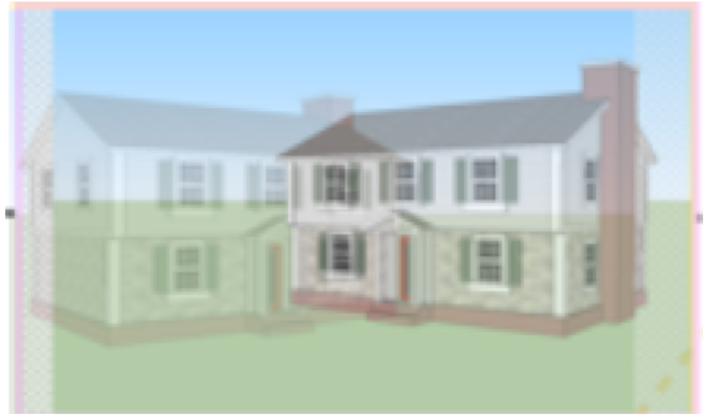




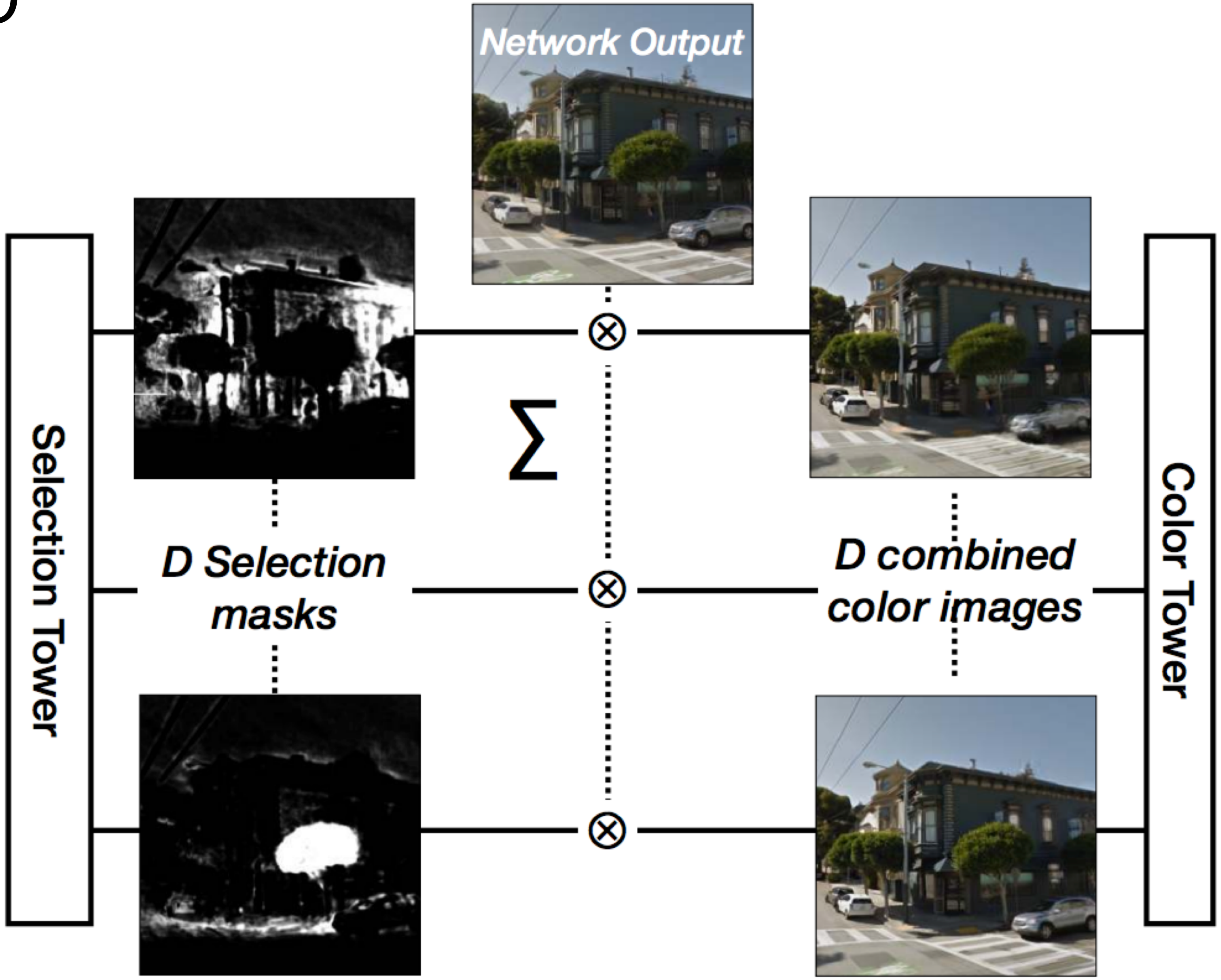
If depth is correct, appearance should match!

# Plane-sweep stereo

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



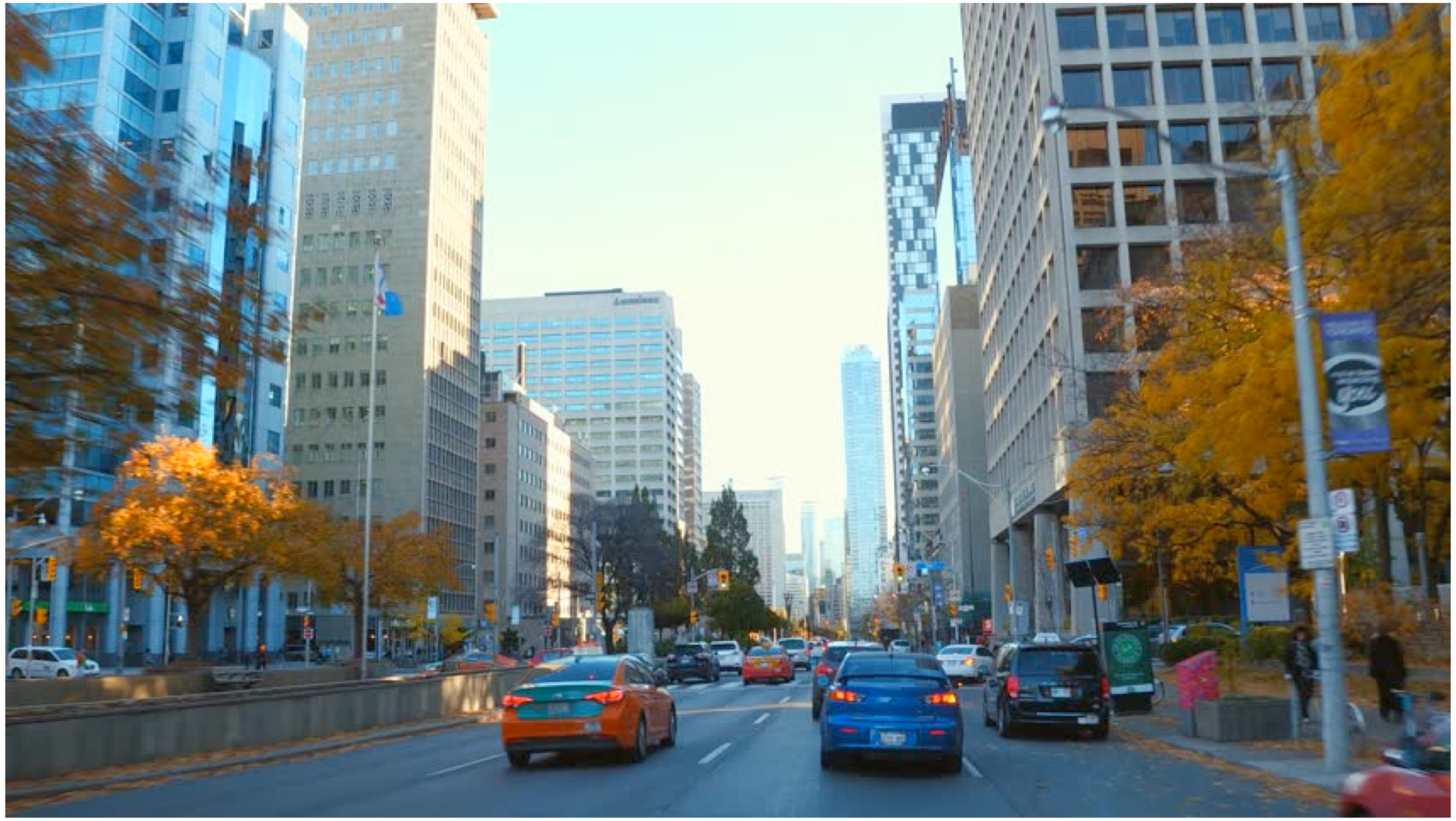
# Deep Stereo




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?

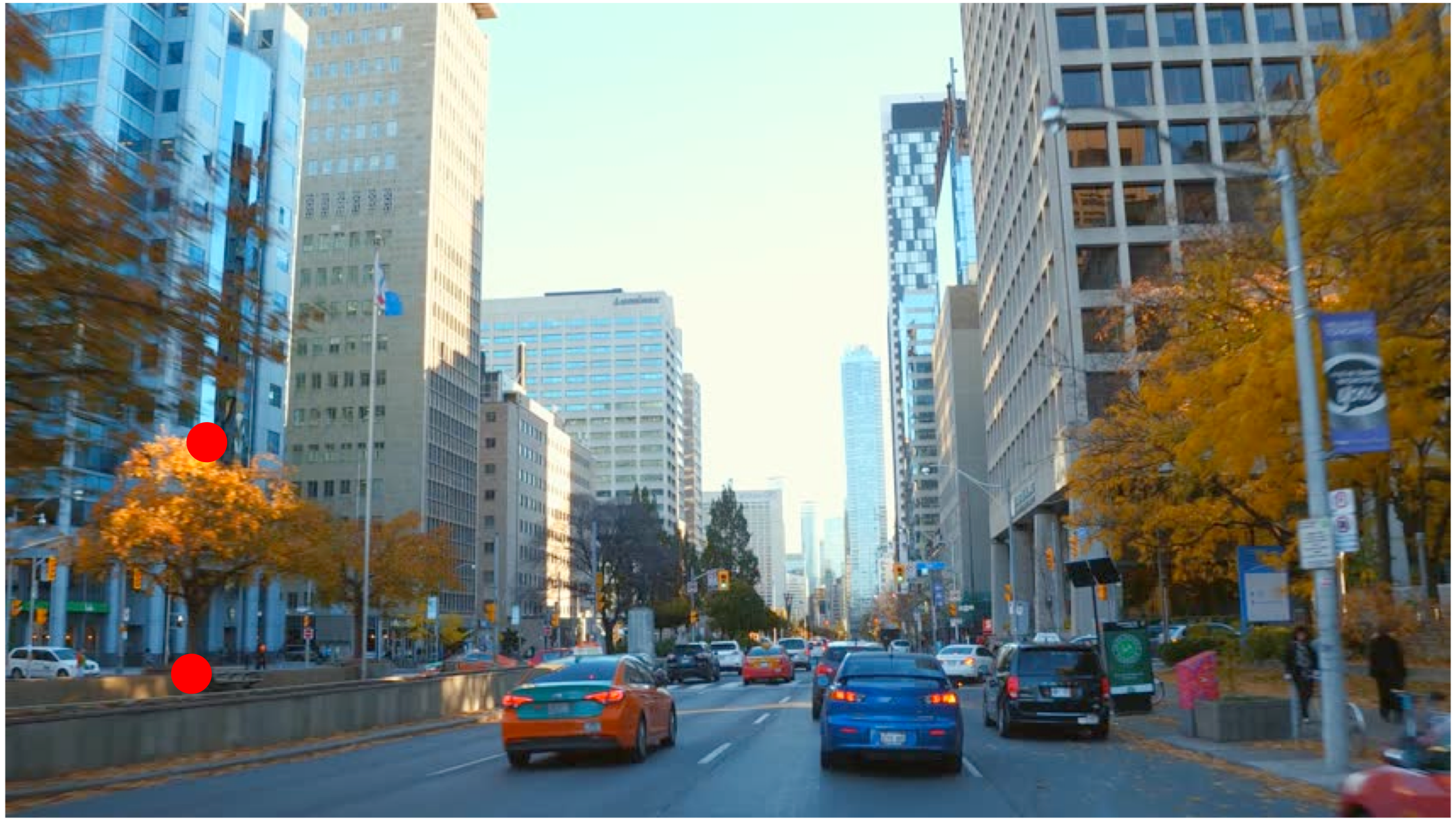


# Estimating depth from a single image

- Why is this even possible?

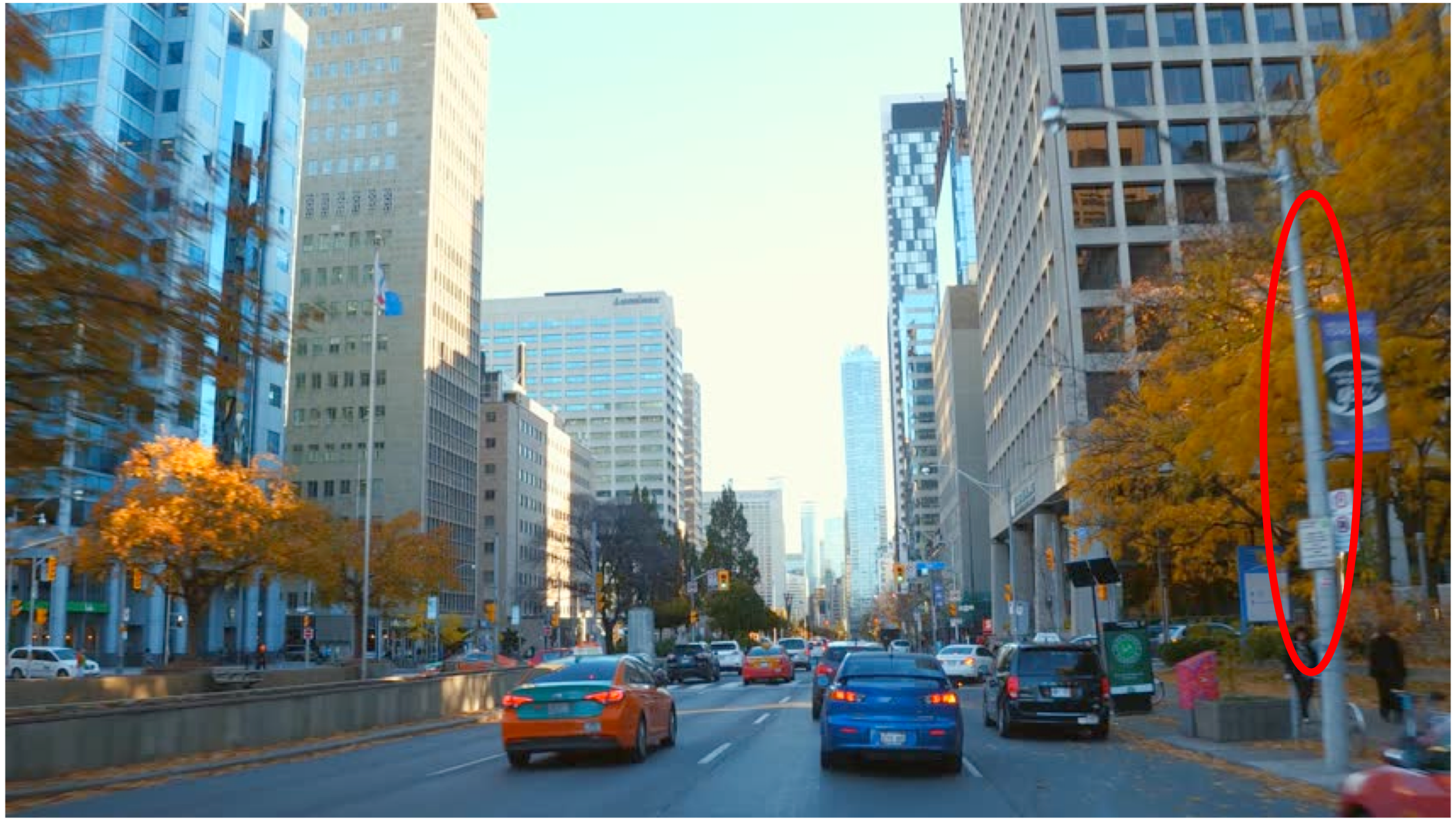
# Estimating depth from a single image

- Why is this even possible?



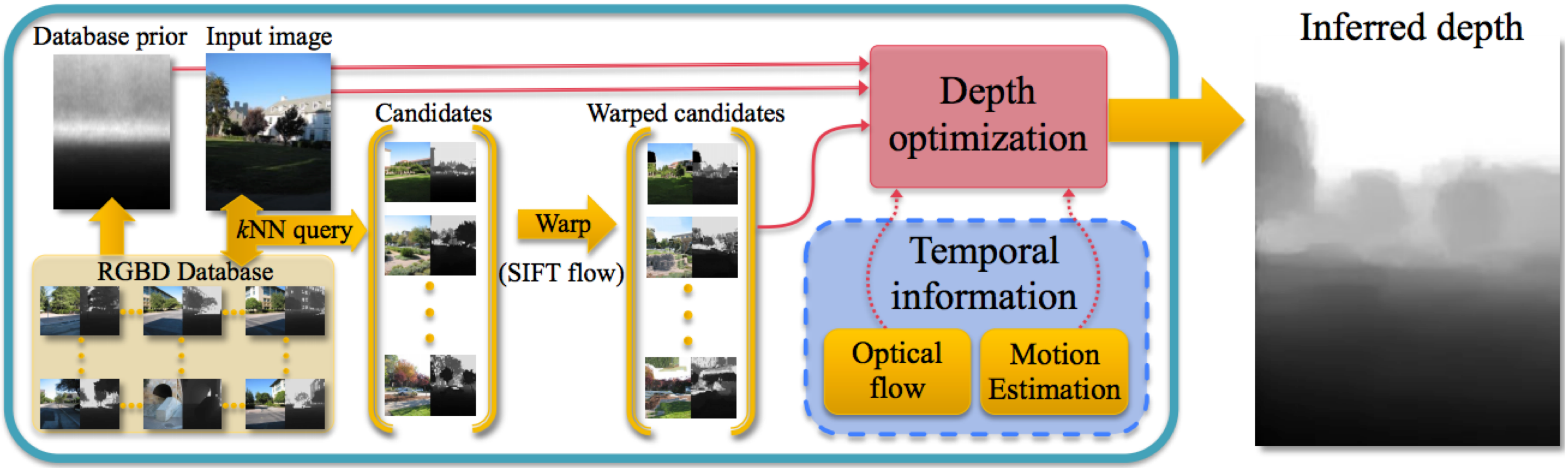
# Estimating depth from a single image

- Why is this even possible?



# Estimating depth from a single image

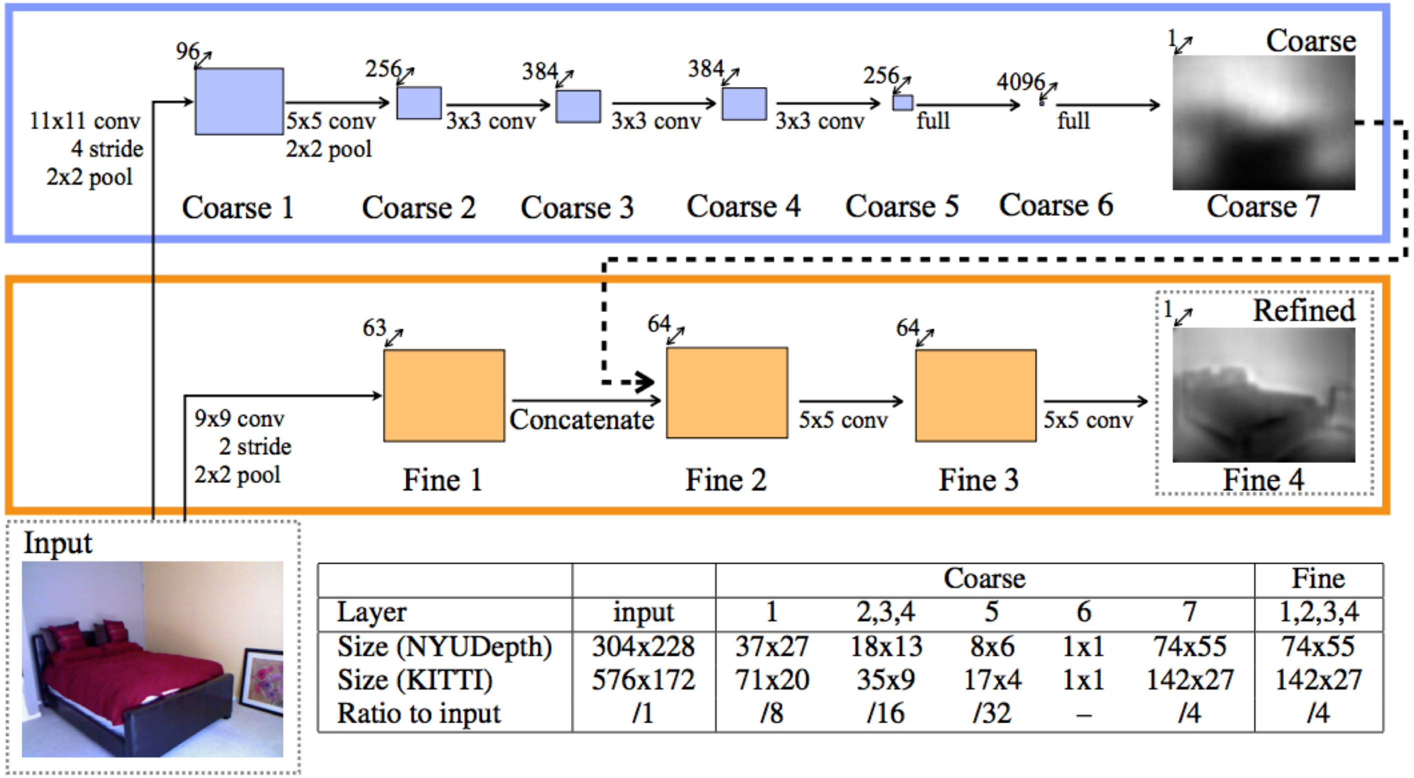
# Estimating depth from a single image




DepthTransfer: Depth Extraction from Video Using Non-parametric Sampling. Kevin Karsch, Ce Liu, Sing Bing Kang. TPAMI 2013.

# 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

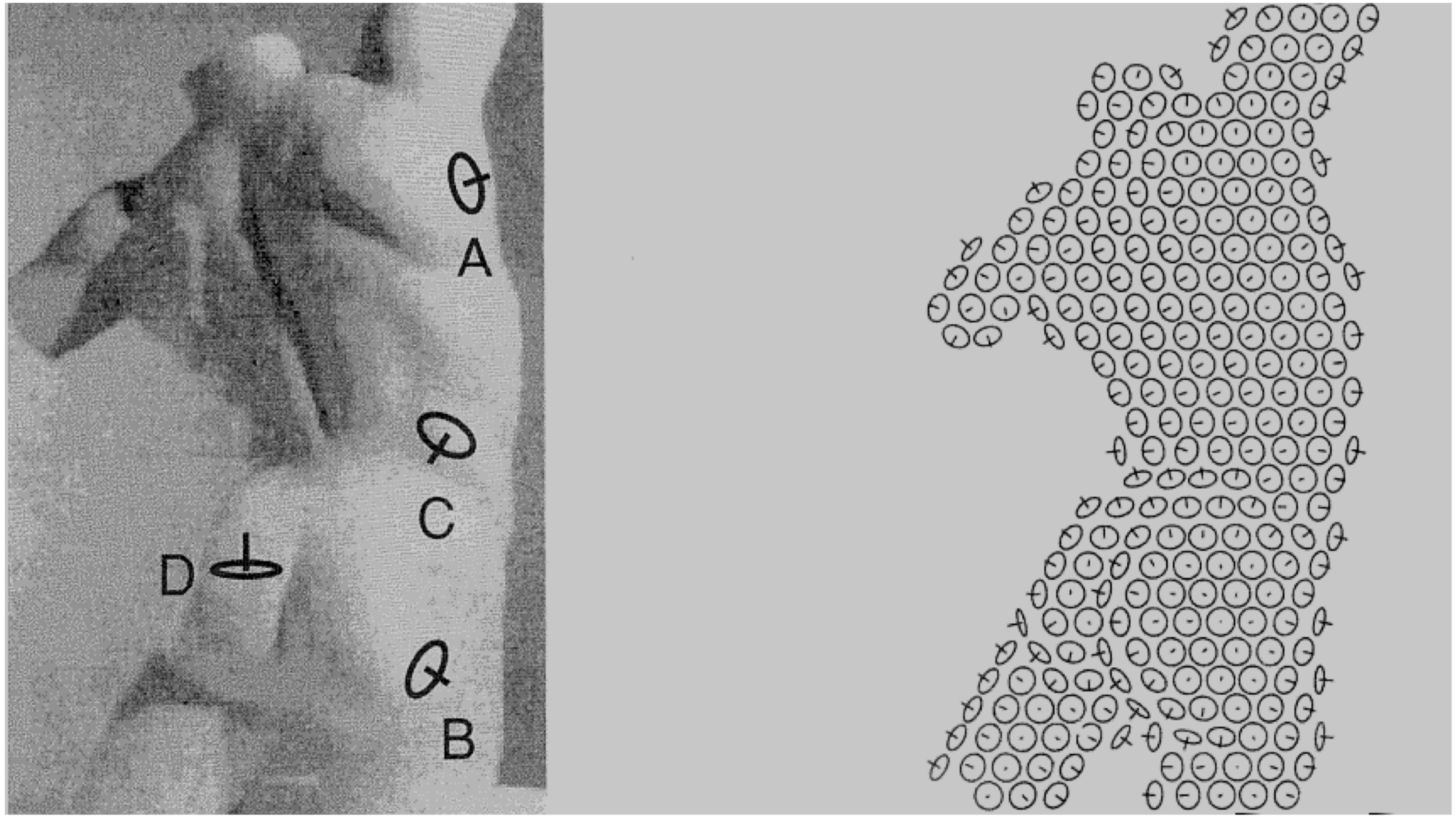
# Metric depth is a bad target

- 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

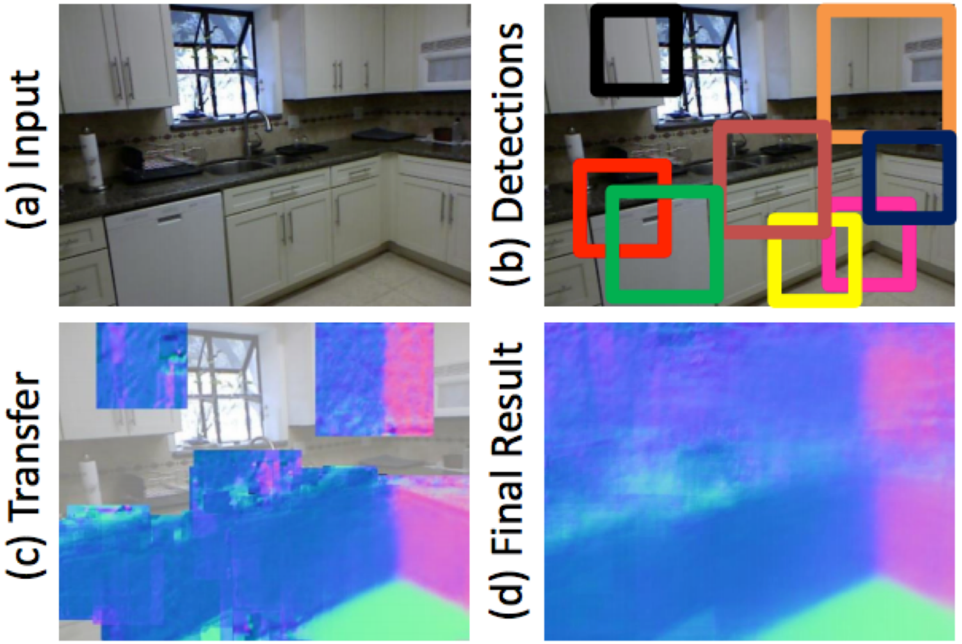
# Humans perceive surface normals, not just depth, through a combination of various pictorial cues




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

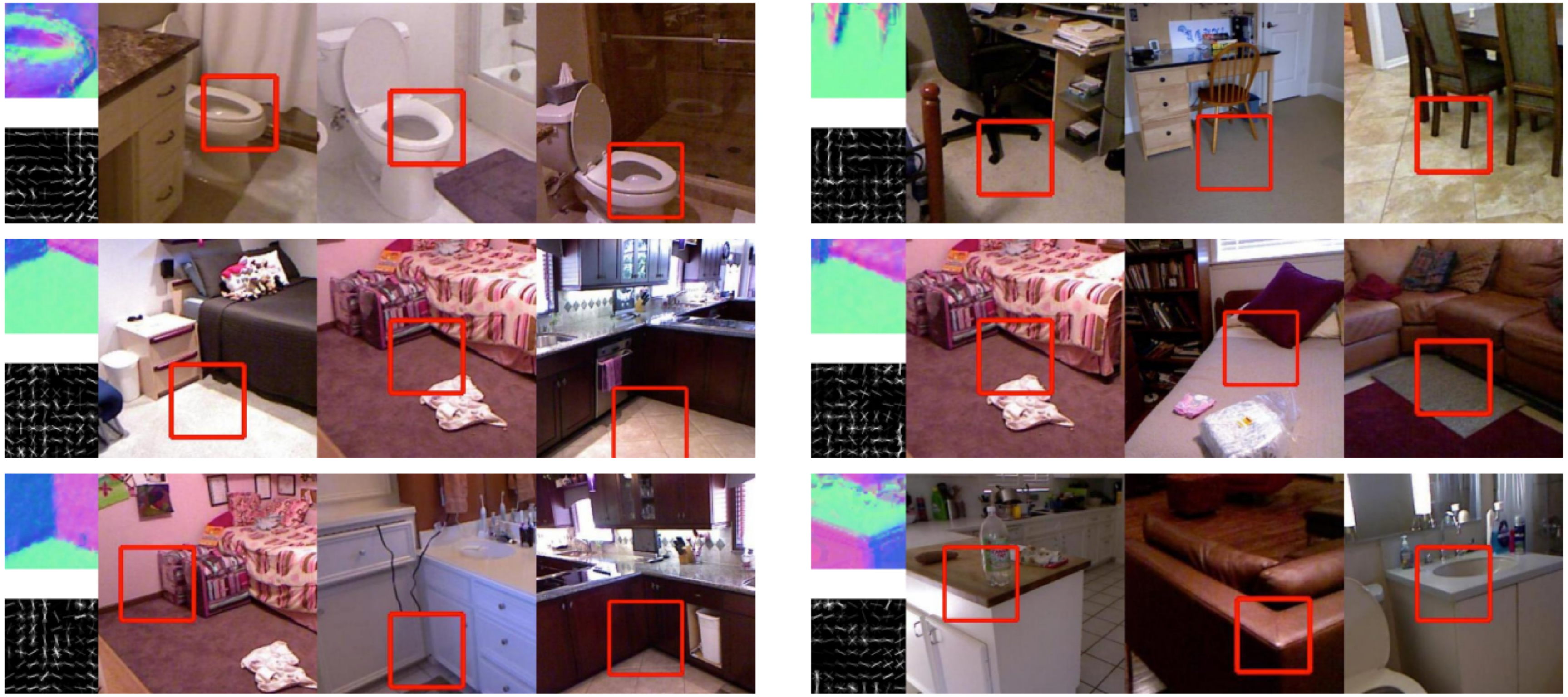
Slide credit: Jitendra Malik

# Estimating normals from a single image

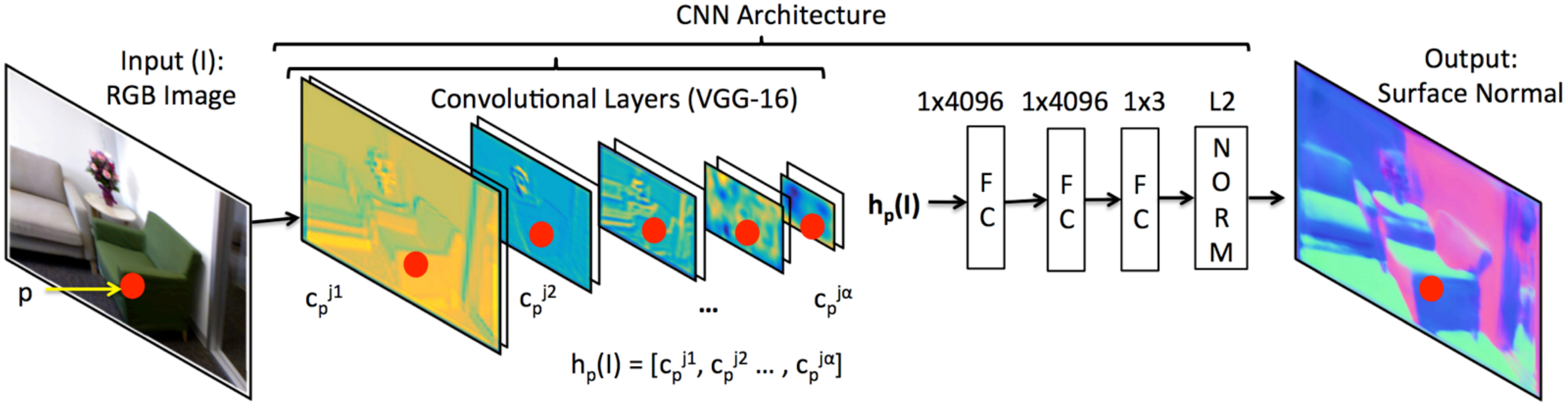



Data-Driven 3D Primitives for Single Image Understanding. David F. Fouhey, Abhinav Gupta, Martial Hebert. In *ICCV* 2013.

# 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