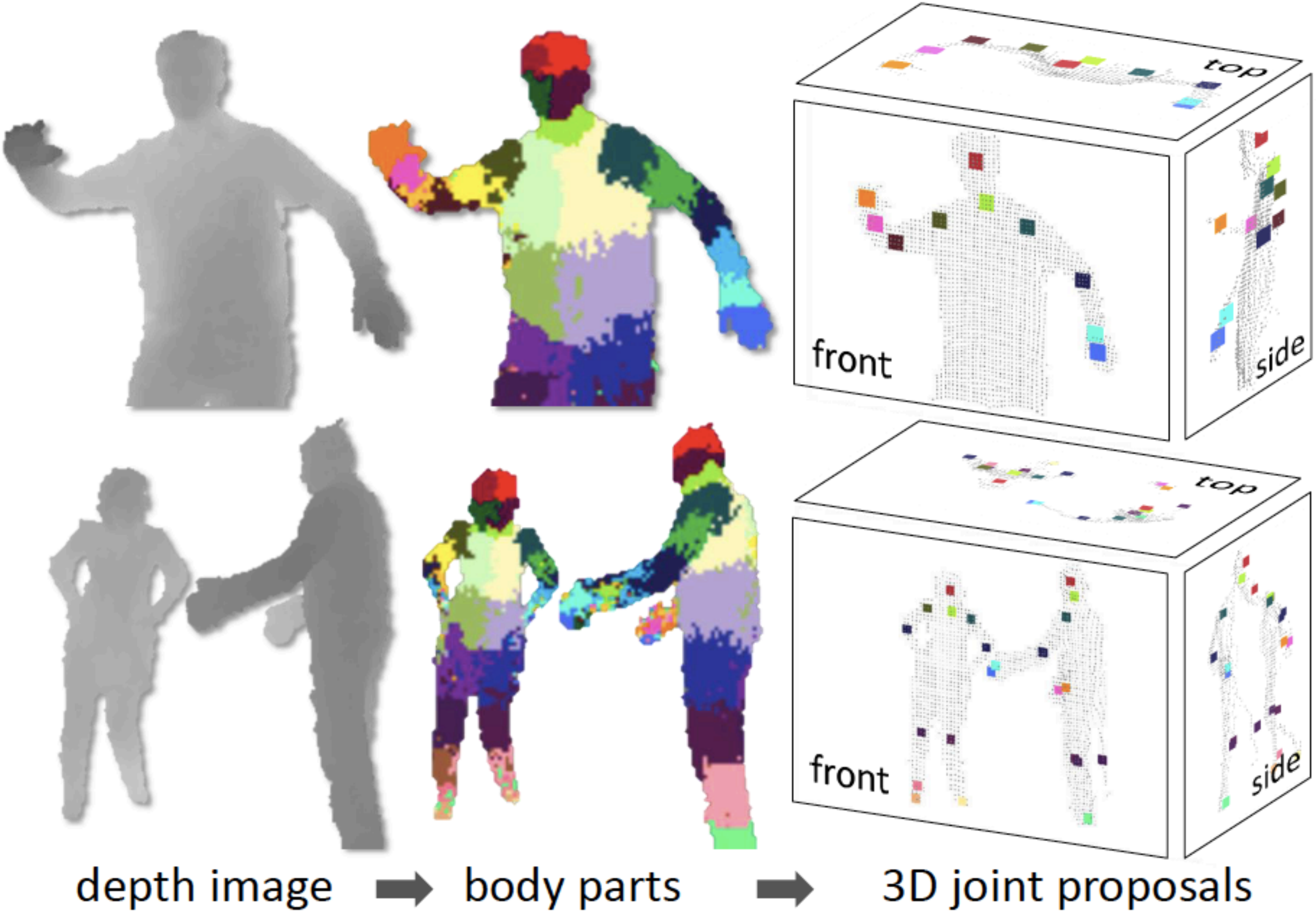
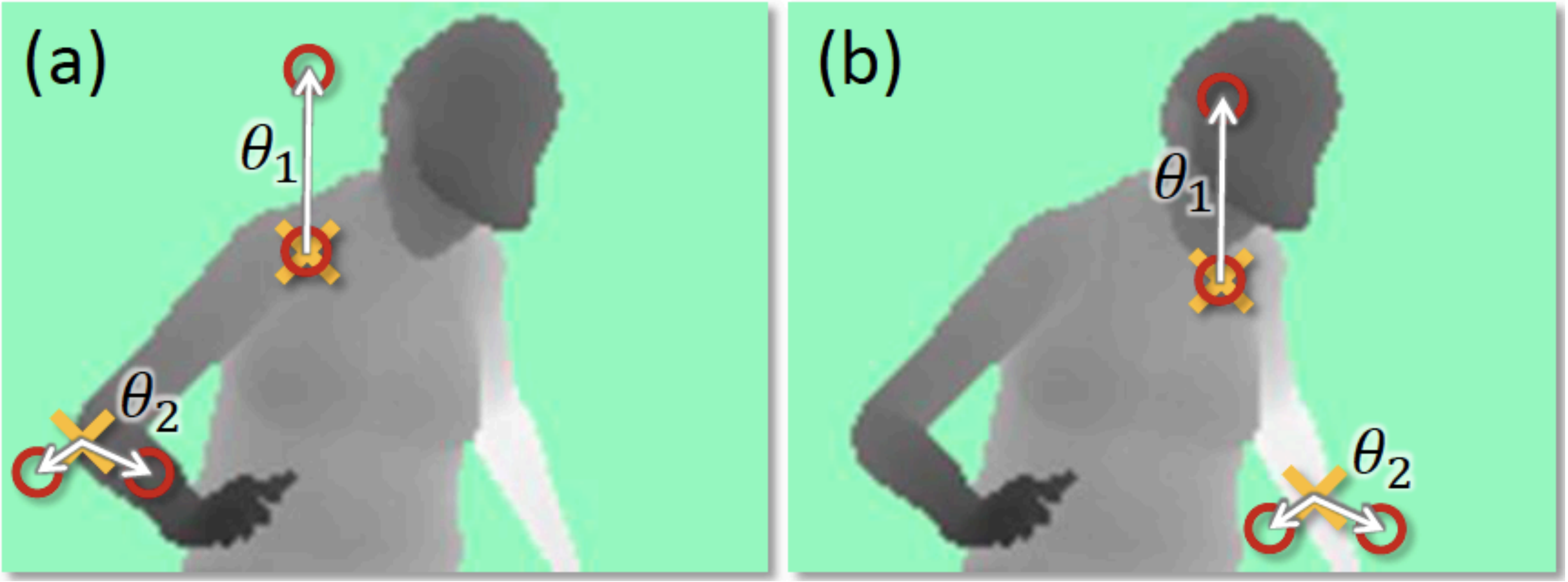
Real-Time Human Pose Recognition in Parts from Single Depth Images

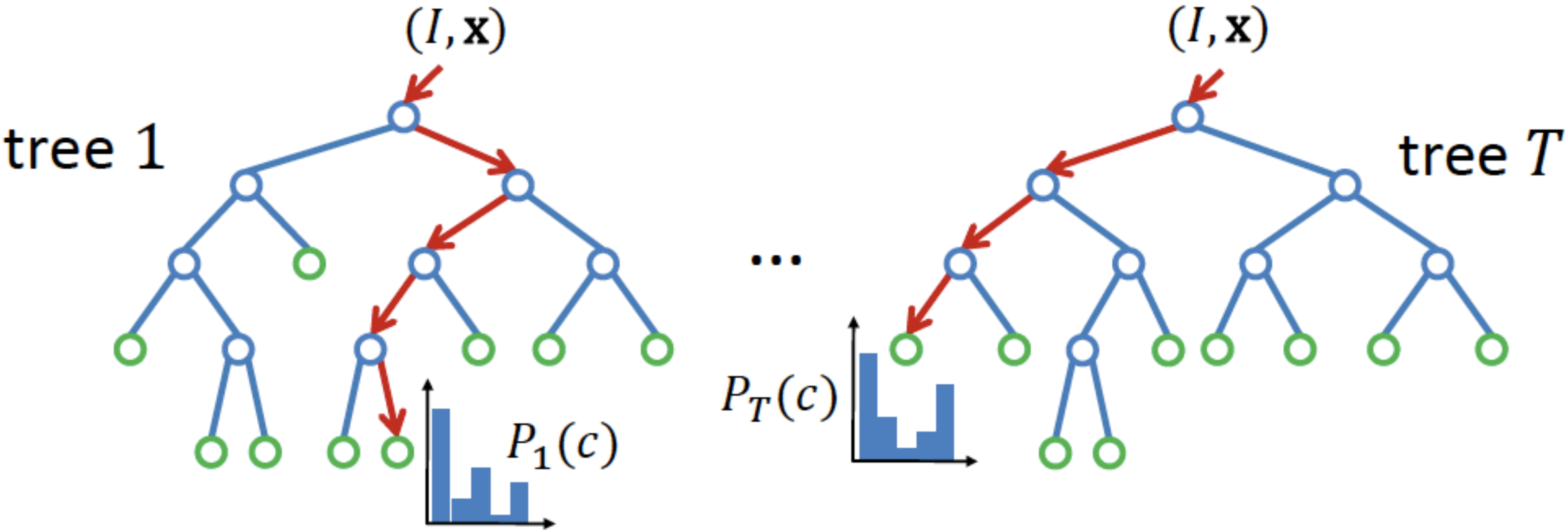

front
top
side
front
top
side
depth image
body parts
3D joint proposals

Features



$$f_\theta(I, \mathbf{x}) = d_I\left(\mathbf{x} + \frac{\mathbf{u}}{d_I(\mathbf{x})}\right) - d_I\left(\mathbf{x} + \frac{\mathbf{v}}{d_I(\mathbf{x})}\right)$$

Randomized Decision Forests



$$P(c|I, \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} P_t(c|I, \mathbf{x})$$

Joint Position Proposals

1. Classified pixels are backprojected into the scene
2. Mean shift is used to find density modes for each part
3. Detected modes are on the surface of the object and are pushed back by a learned offset
4. Modes are scored using the sum of weights of all pixels reaching the mode

$$f_c(\hat{\mathbf{x}}) \propto \sum_{i=1}^{N} w_{ic} \exp\left(-\left\|\frac{\hat{\mathbf{x}} - \hat{\mathbf{x}}_i}{b_c}\right\|^2\right)$$

$$w_{ic} = P(c|I, \mathbf{x}_i) \cdot d_I(\mathbf{x}_i)^2$$

Parameters

1. 3 trees
2. 20 deep
3. 300k training images
4. 2000 training pixels per image
5. 2000 candidate features and 50 cnadidate thresholds pre feature.

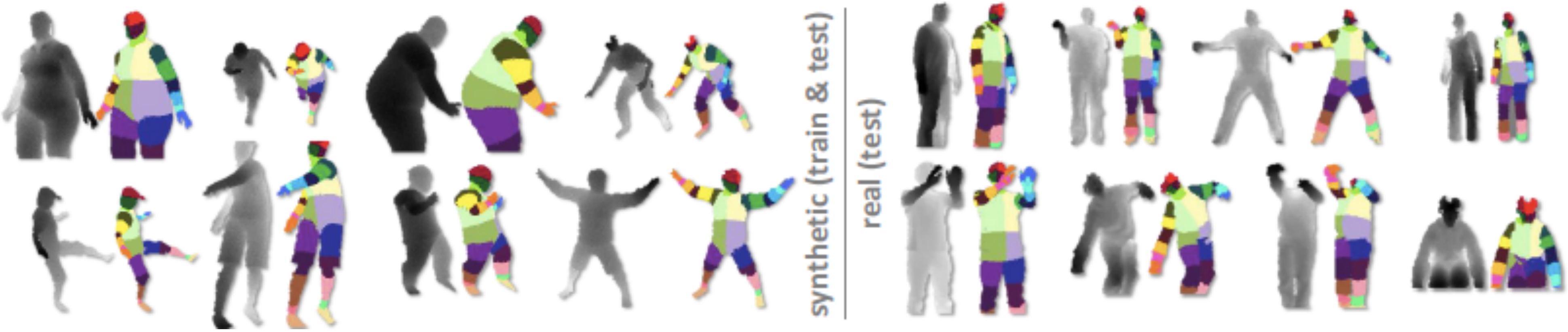
Training time: 1 day on a 1000 core cluster (using 1M images)

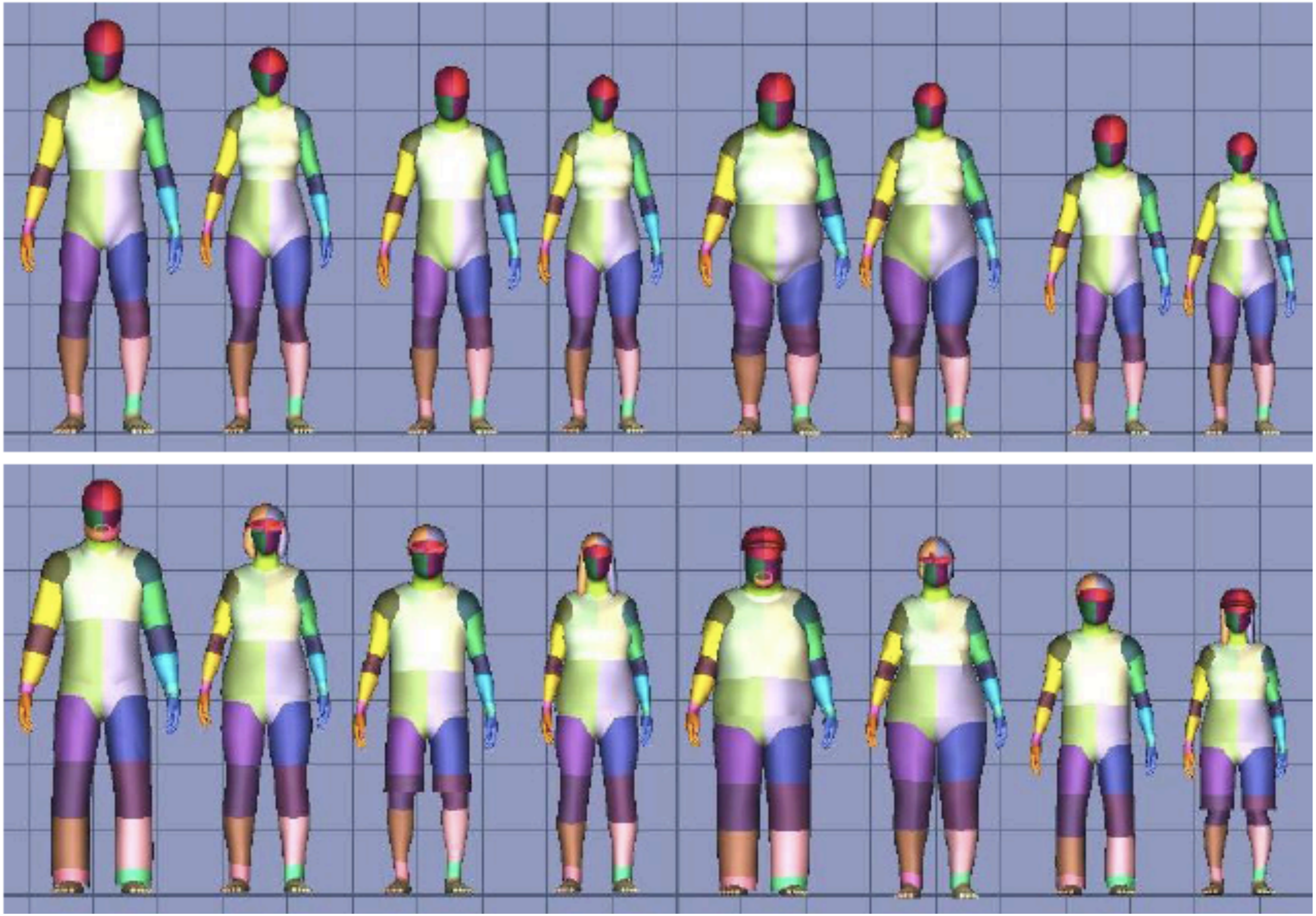
Dataset

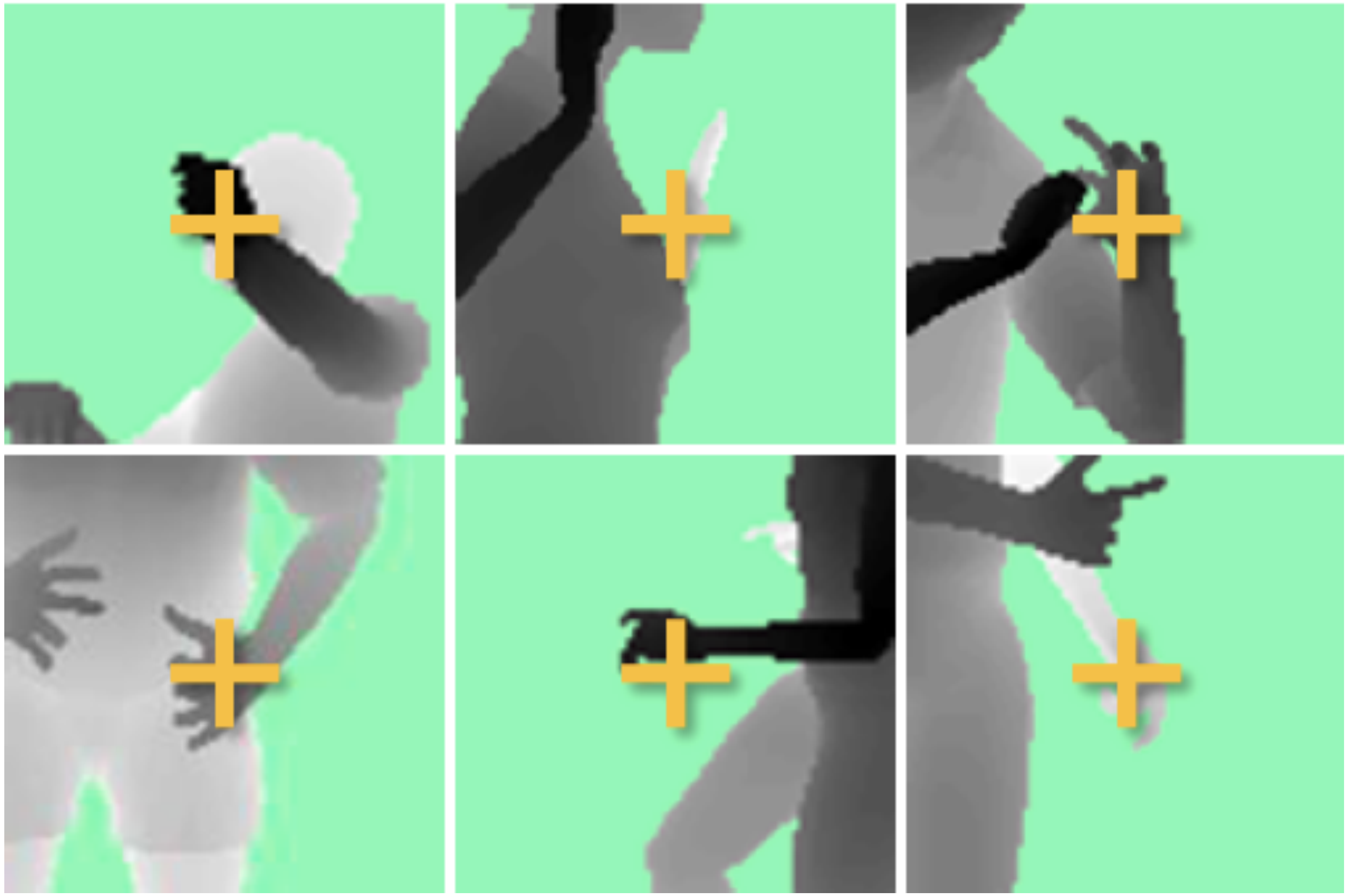
A synthetic dataset is used for training (and test):

1. 500k mocap frames reduced to 100k dissimilar frames
2. Mocap retargetted to 15 base meshes
3. Part labels specified in texture map (31 parts used)
4. Some randomization is applied

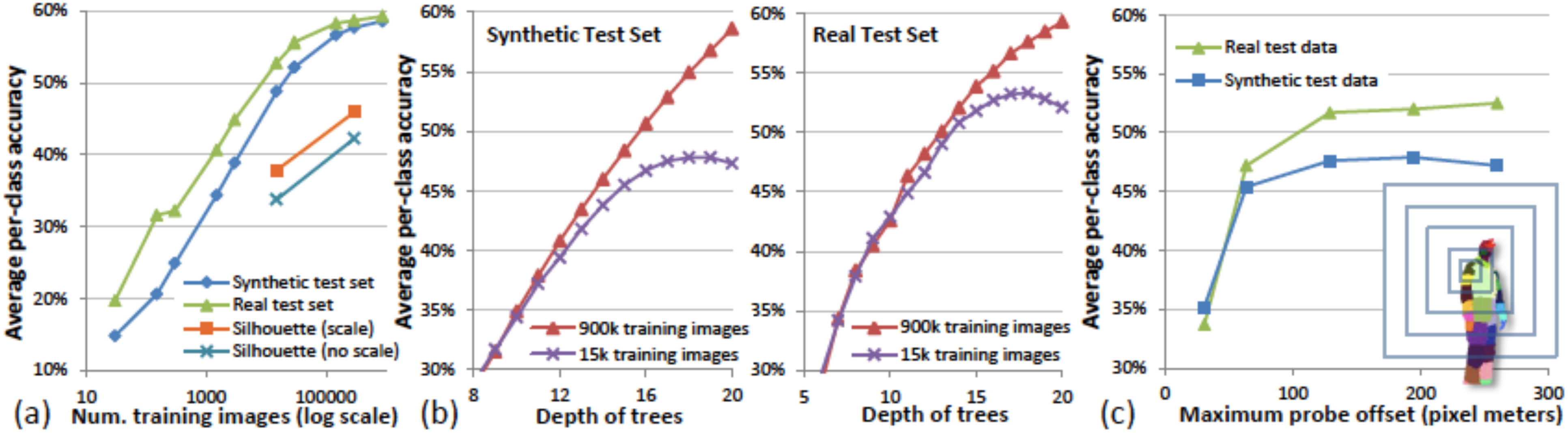
A real hand labeled dataset with 8808 depth images from 15 subjects is also used for test only.

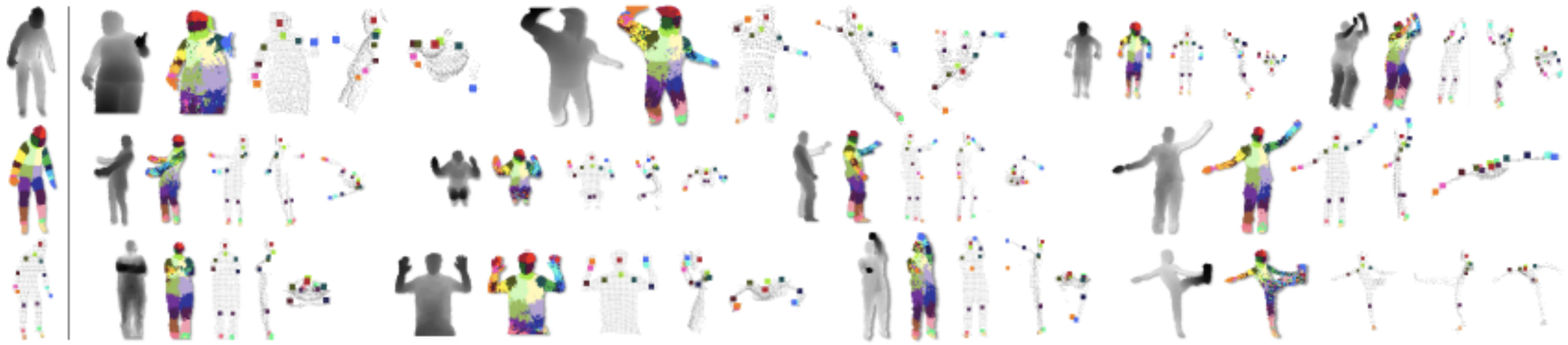




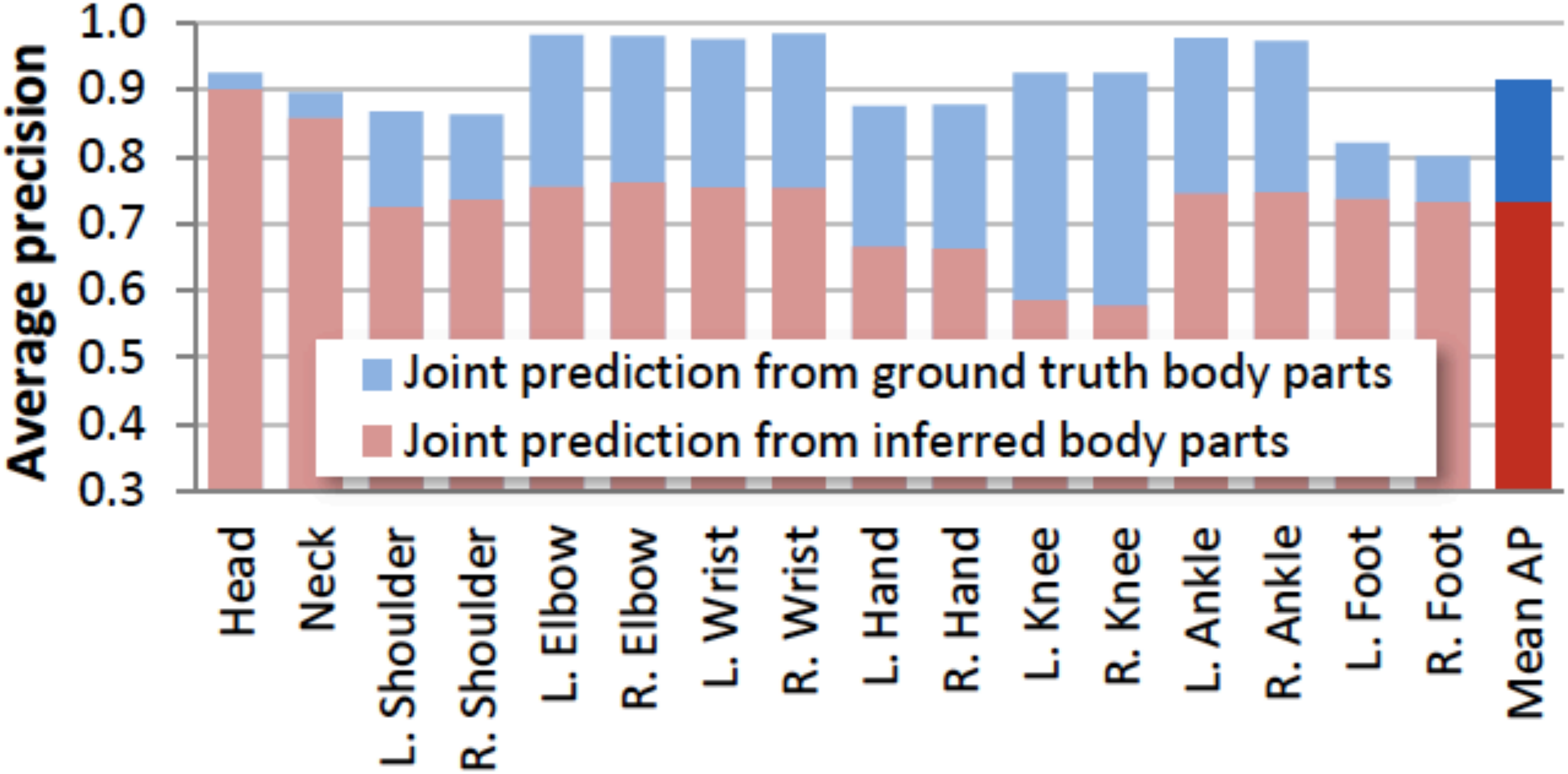


Results


60%
50%
40%
30%
20%
10%
Average per-class accuracy
Synthetic test set
Real test set
Silhouette (scale)
Silhouette (no scale)
10
1000
100000
(a)
Num. training images (log scale)
60%
55%
50%
45%
40%
35%
30%
Average per-class accuracy
Synthetic Test Set
900k training images
15k training images
8
12
16
20
(b)
Depth of trees
60%
55%
50%
45%
40%
35%
30%
Real Test Set
900k training images
15k training images
5
10
15
20
Depth of trees
60%
55%
50%
45%
40%
35%
30%
Average per-class accuracy
Real test data
Synthetic test data
0
100
200
300
(c)
Maximum probe offset (pixel meters)



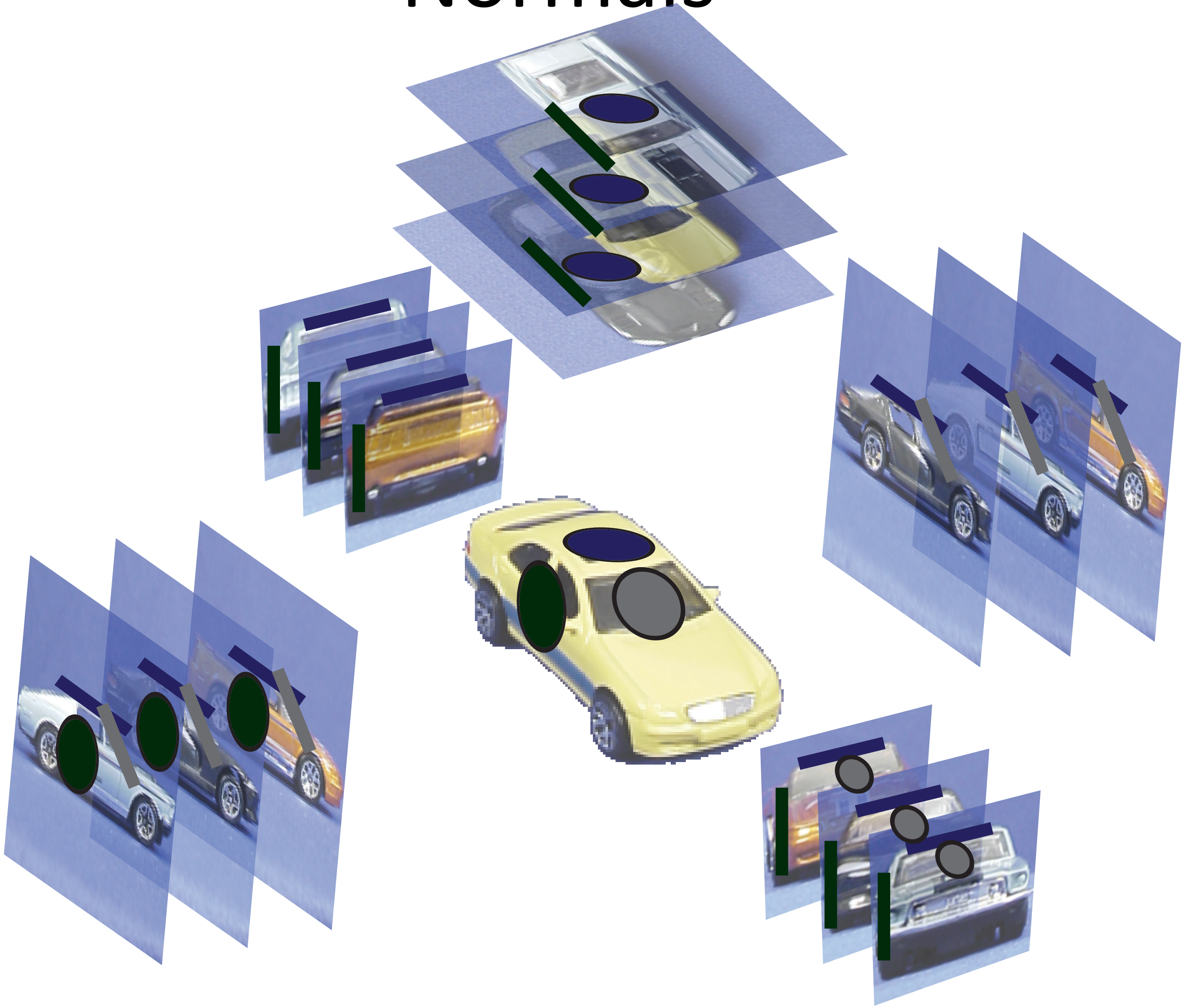
Results



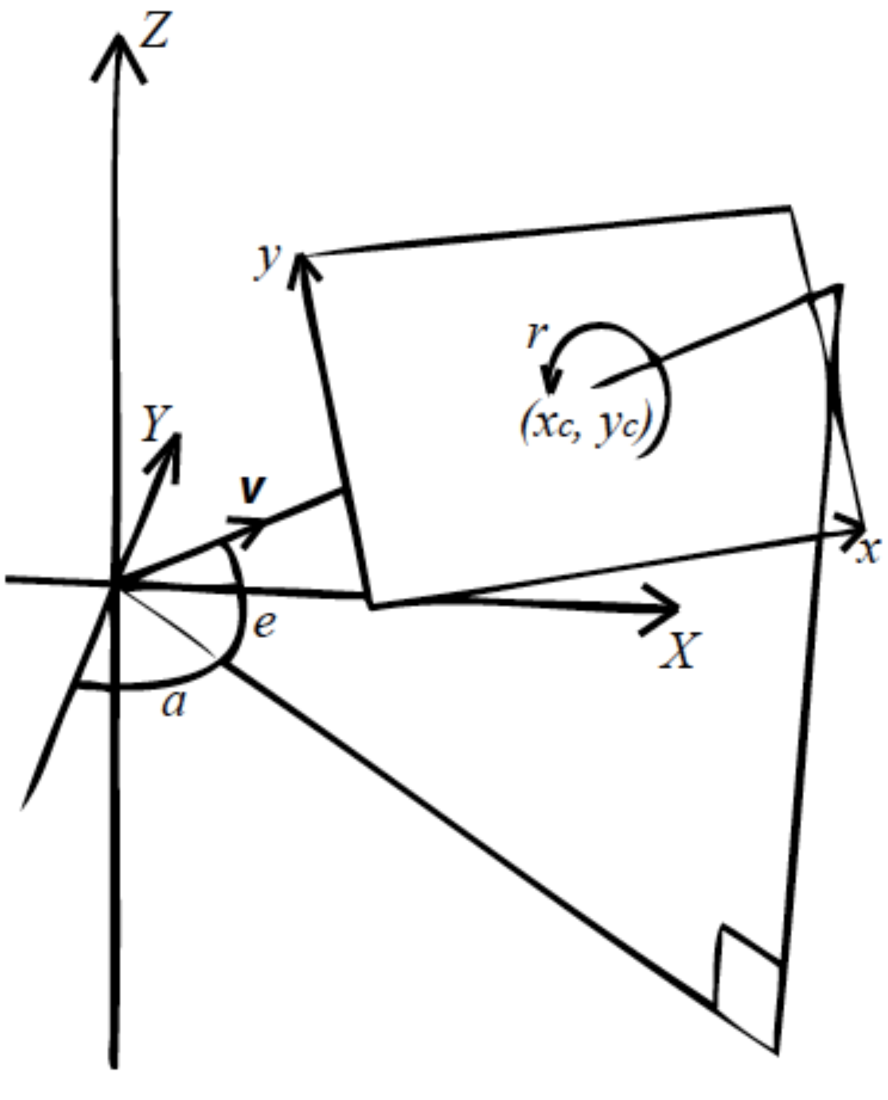
video

Multi-View Object Recognition

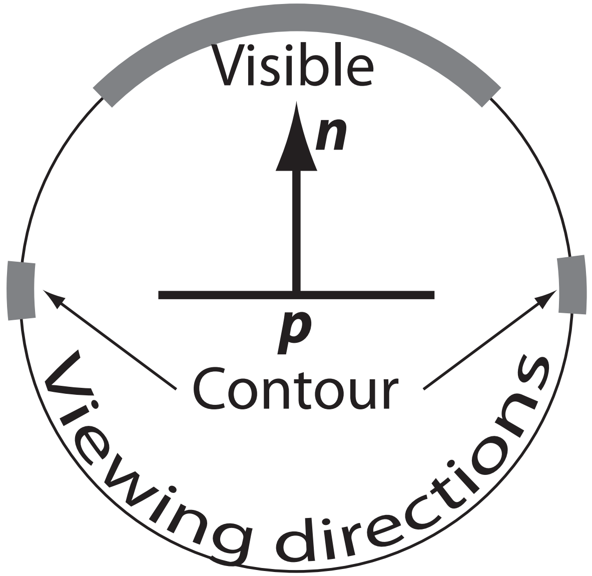
A Representation Based on Surface Normals

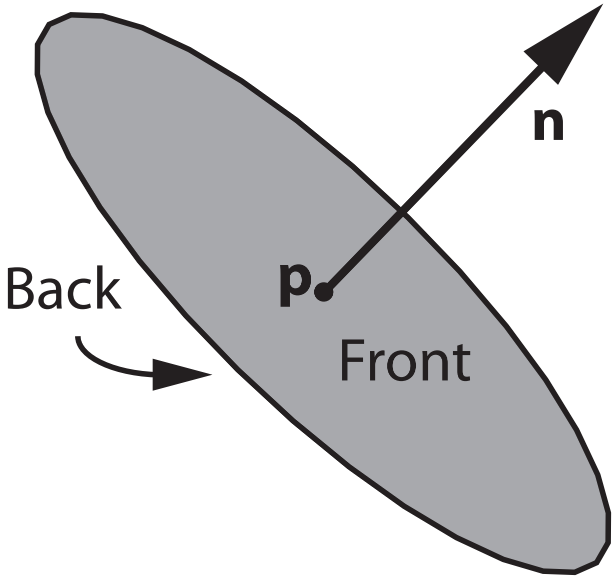


Camera Model

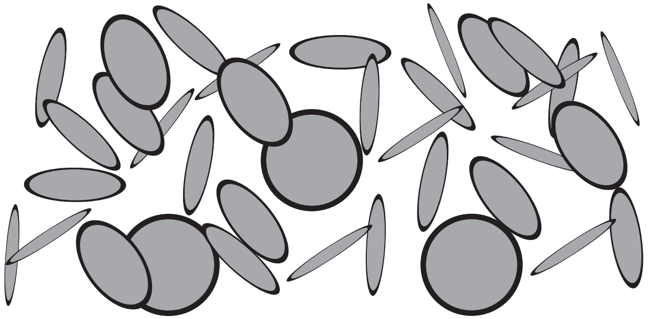


Surfels

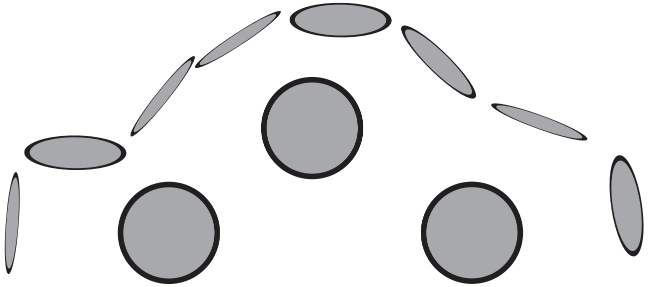




Training



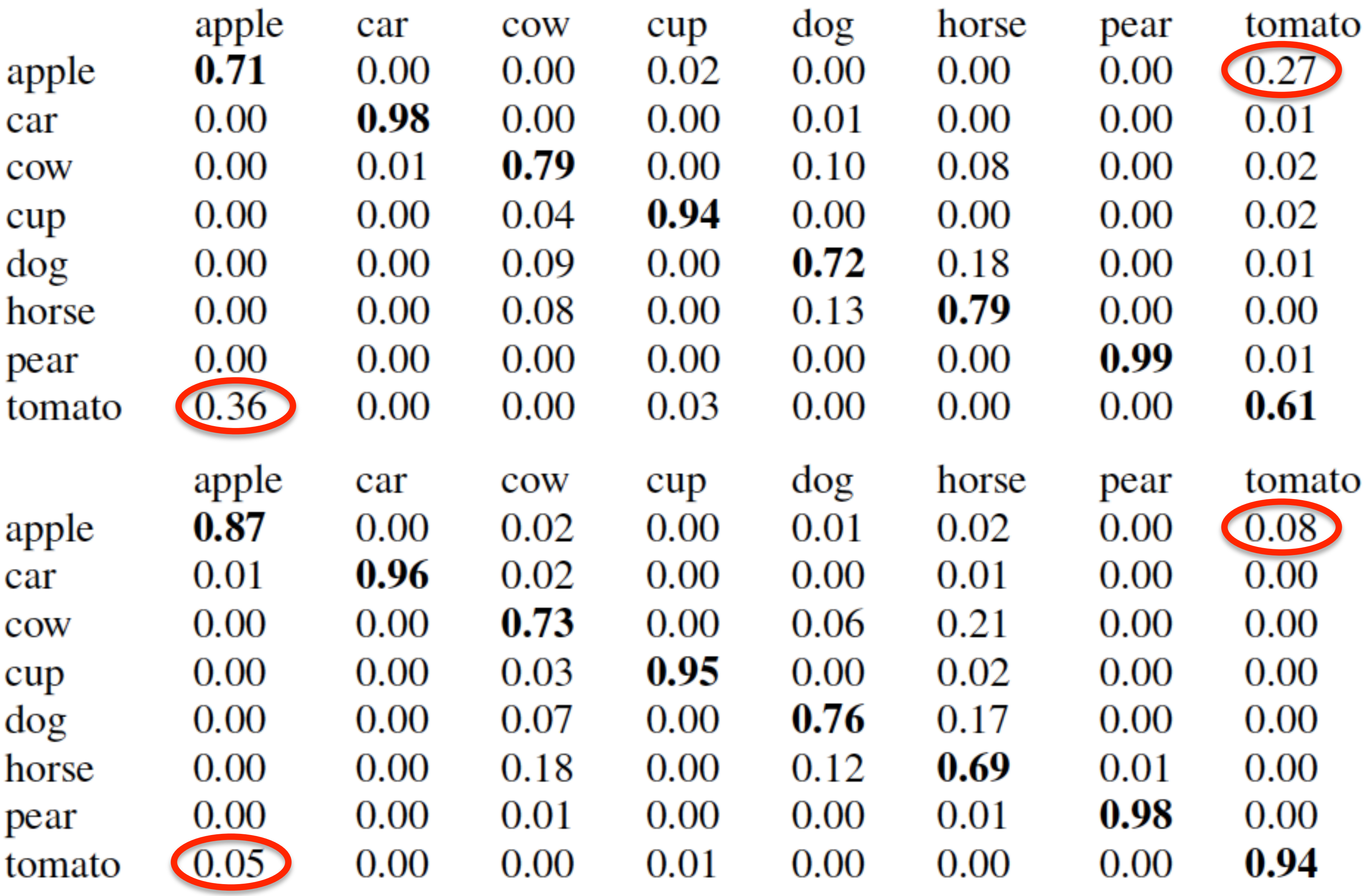
Before



After

Experiments – ETH80 dataset

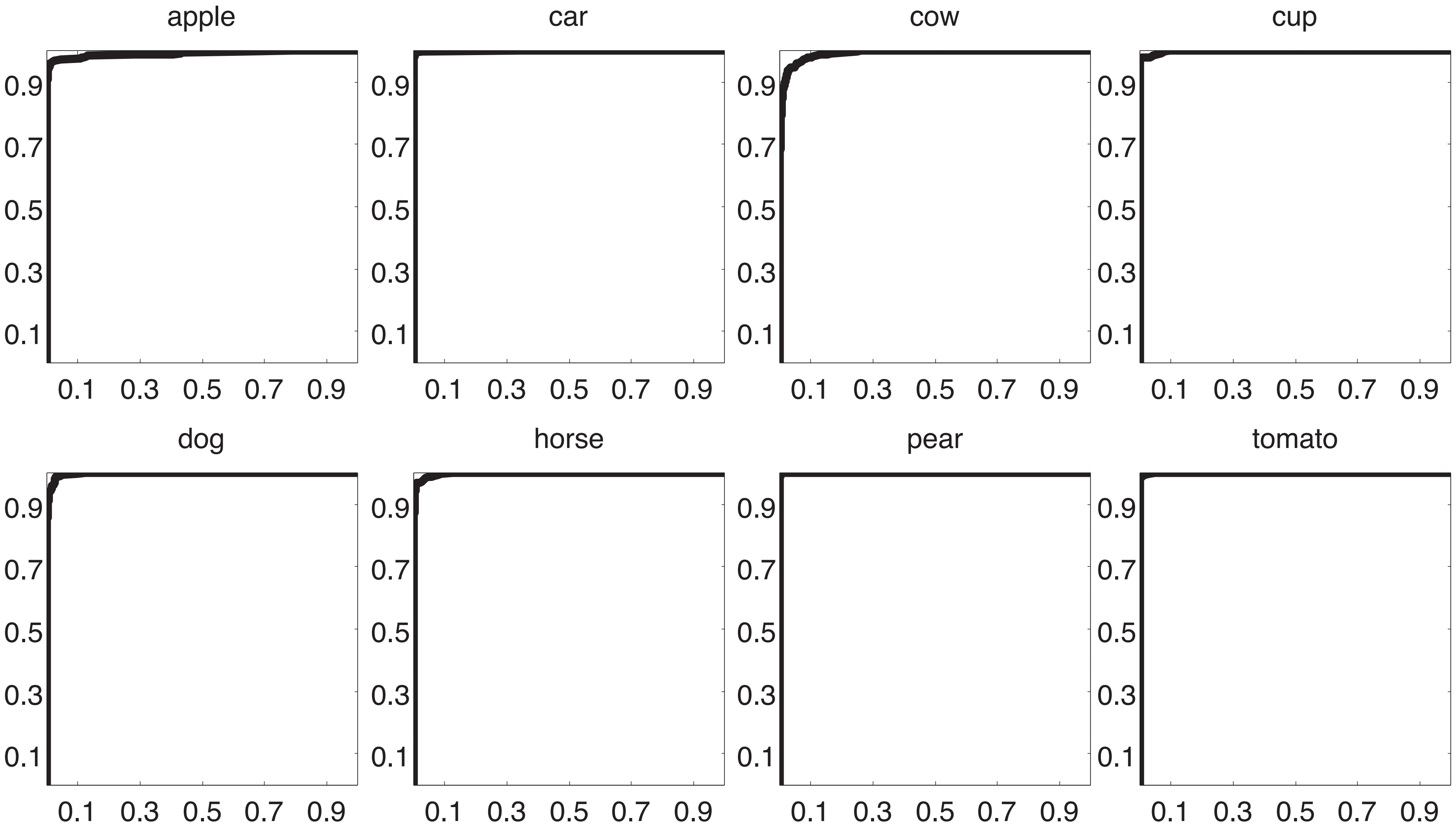
Categorization



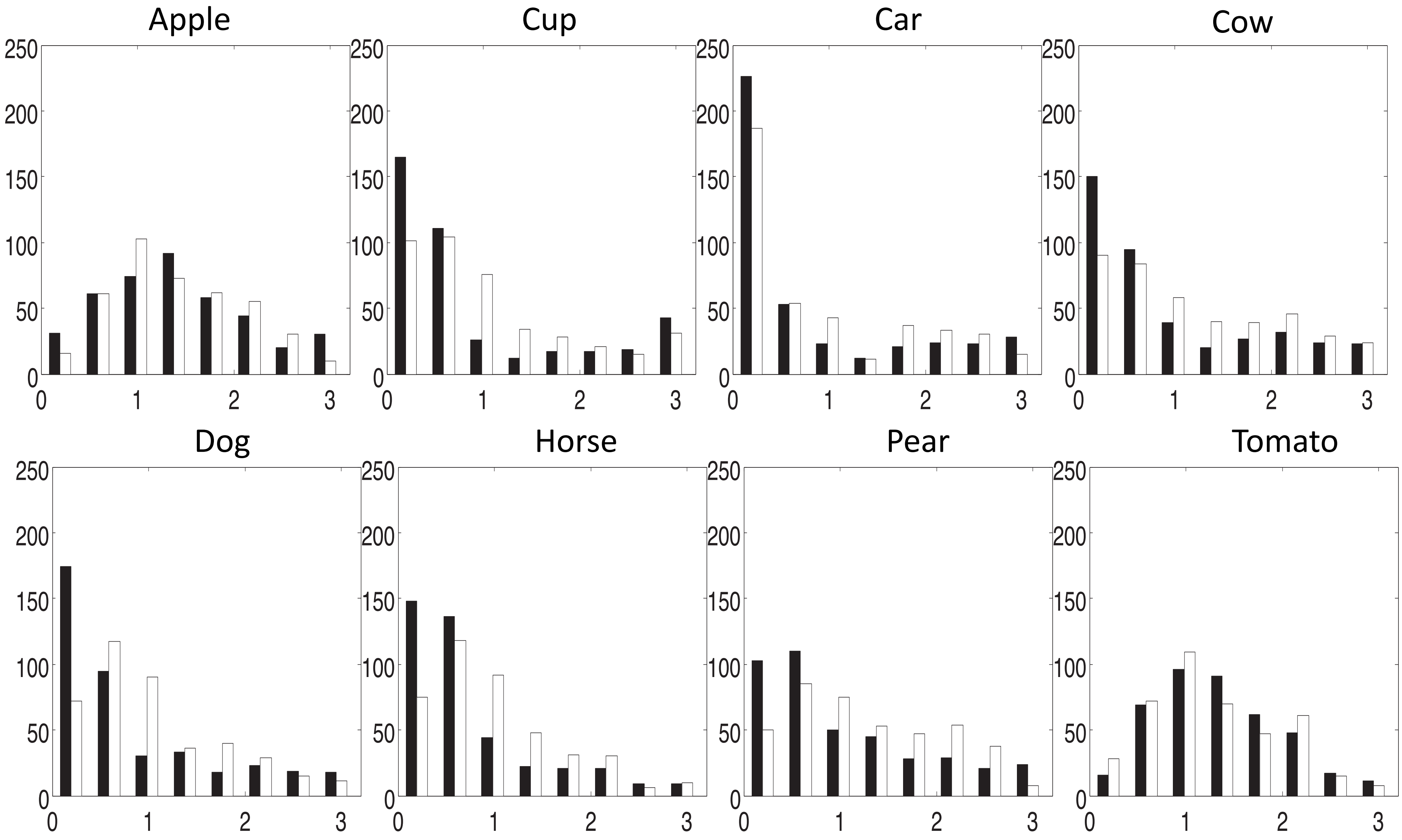
	apple	car	cow	cup	dog	horse	pear	tomato
apple	**0.71**	0.00	0.00	0.02	0.00	0.00	0.00	0.27
car	0.00	**0.98**	0.00	0.00	0.01	0.00	0.00	0.01
cow	0.00	0.01	**0.79**	0.00	0.10	0.08	0.00	0.02
cup	0.00	0.00	0.04	**0.94**	0.00	0.00	0.00	0.02
dog	0.00	0.00	0.09	0.00	**0.72**	0.18	0.00	0.01
horse	0.00	0.00	0.08	0.00	0.13	**0.79**	0.00	0.00
pear	0.00	0.00	0.00	0.00	0.00	0.00	**0.99**	0.01
tomato	0.36	0.00	0.00	0.03	0.00	0.00	0.00	**0.61**

	apple	car	cow	cup	dog	horse	pear	tomato
apple	**0.87**	0.00	0.02	0.00	0.01	0.02	0.00	0.08
car	0.01	**0.96**	0.02	0.00	0.00	0.01	0.00	0.00
cow	0.00	0.00	**0.73**	0.00	0.06	0.21	0.00	0.00
cup	0.00	0.00	0.03	**0.95**	0.00	0.02	0.00	0.00
dog	0.00	0.00	0.07	0.00	**0.76**	0.17	0.00	0.00
horse	0.00	0.00	0.18	0.00	0.12	**0.69**	0.01	0.00
pear	0.00	0.00	0.01	0.00	0.00	0.01	**0.98**	0.00
tomato	0.05	0.00	0.00	0.01	0.00	0.00	0.00	**0.94**

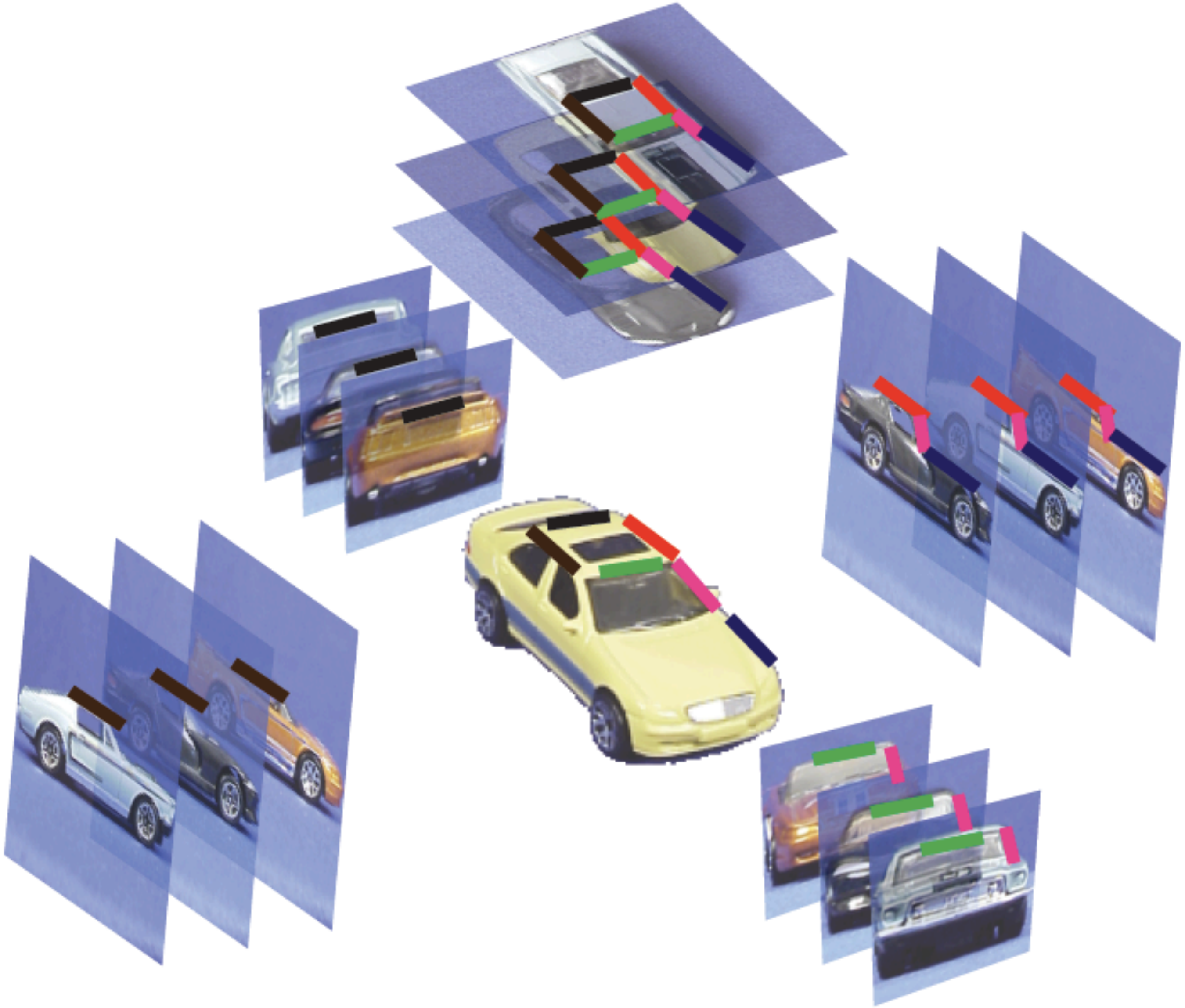
Recognition / Detection



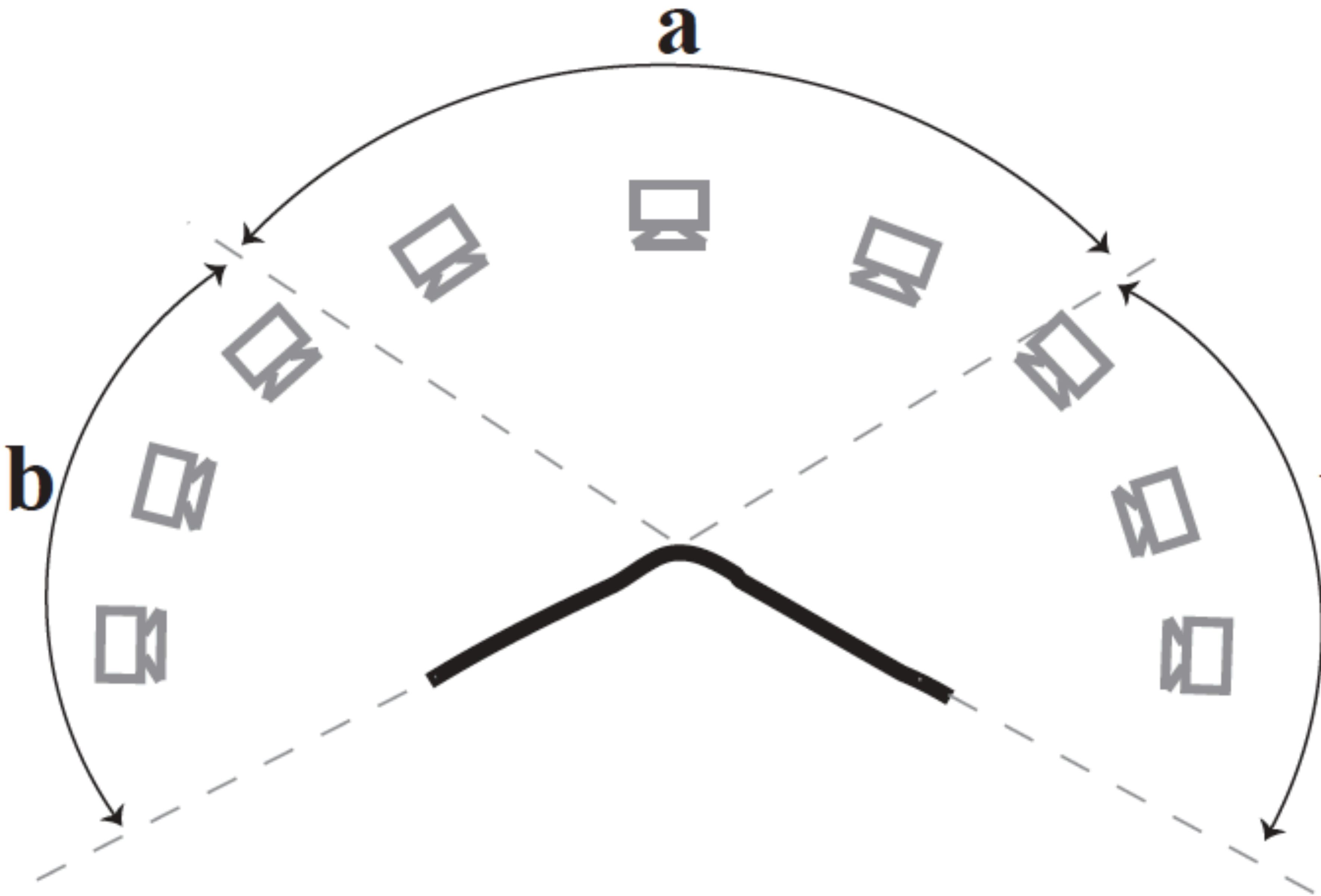
Pose Estimation



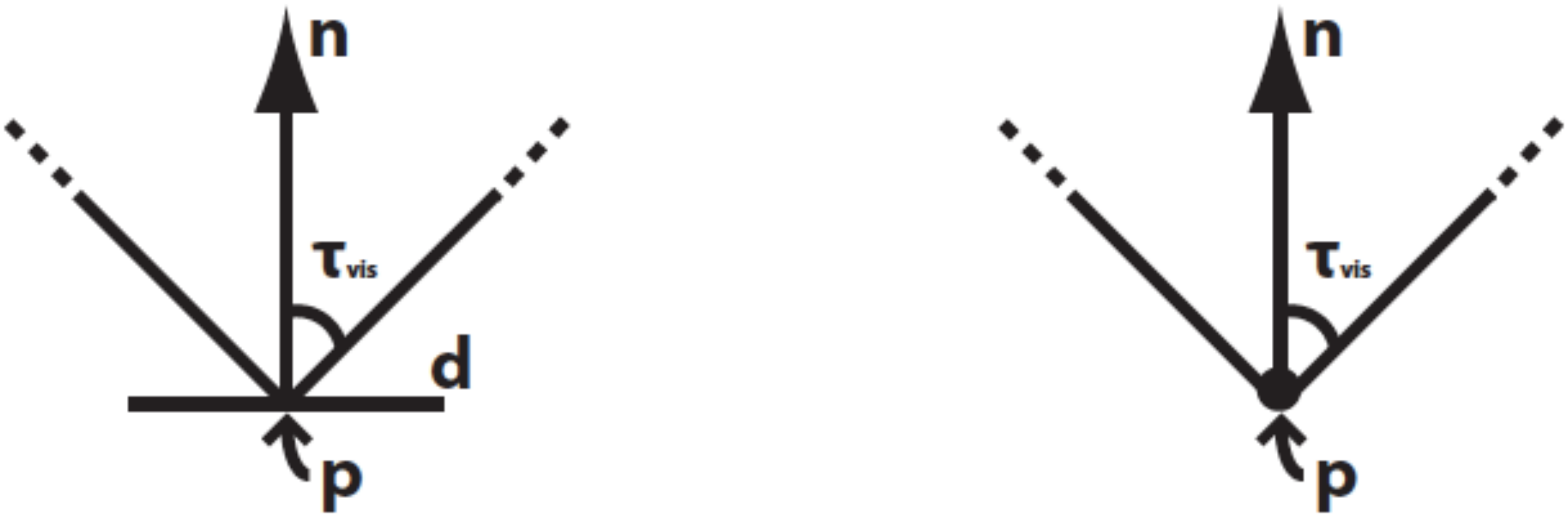
A Representation Based on Surface Curvature



Surface curvature -> image gradient



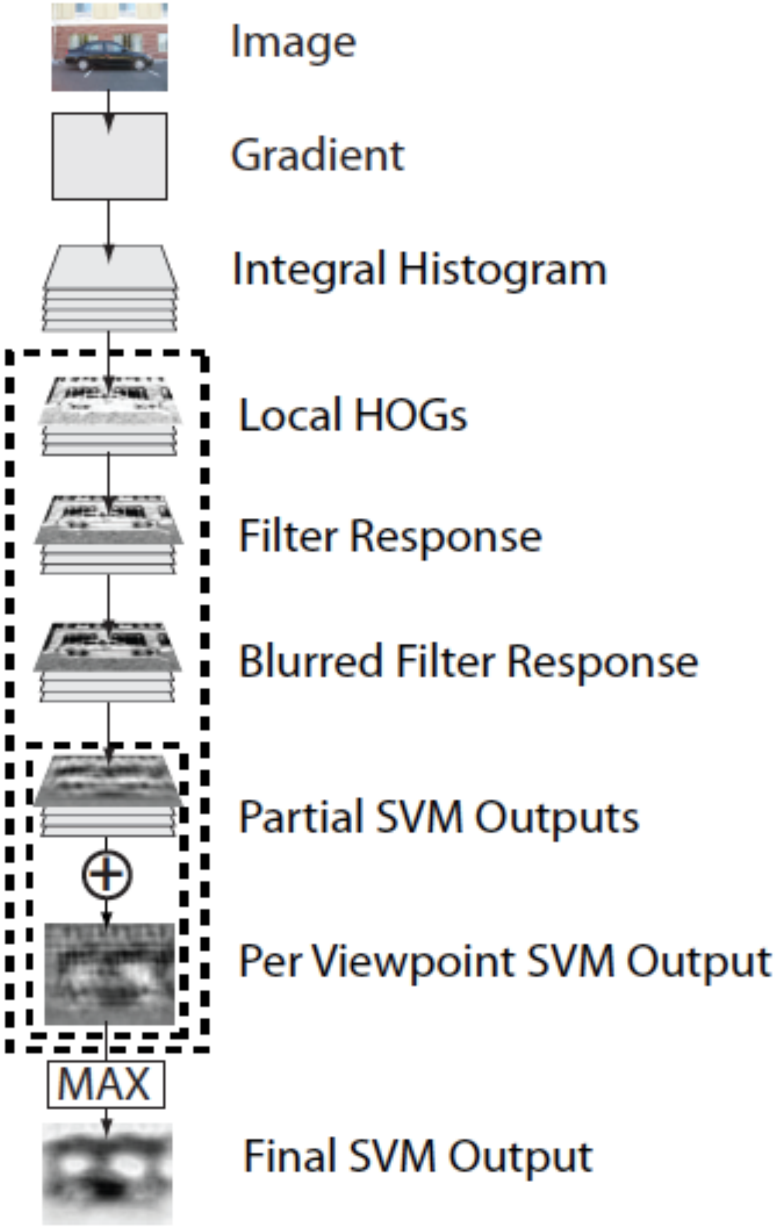
3D Edges



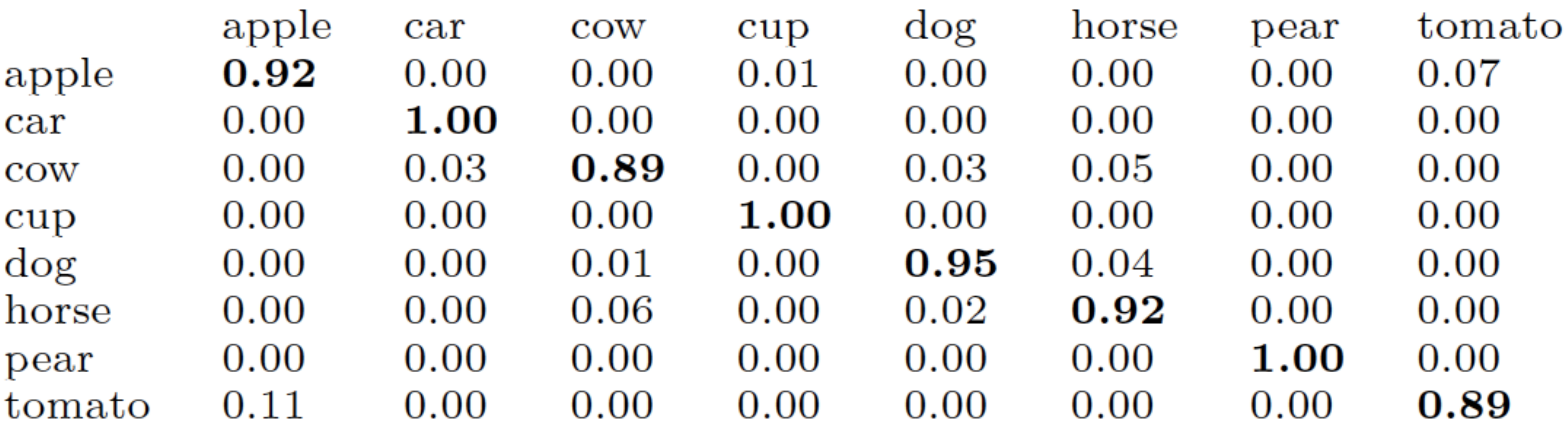
(a)

(b)

Processing pipeline

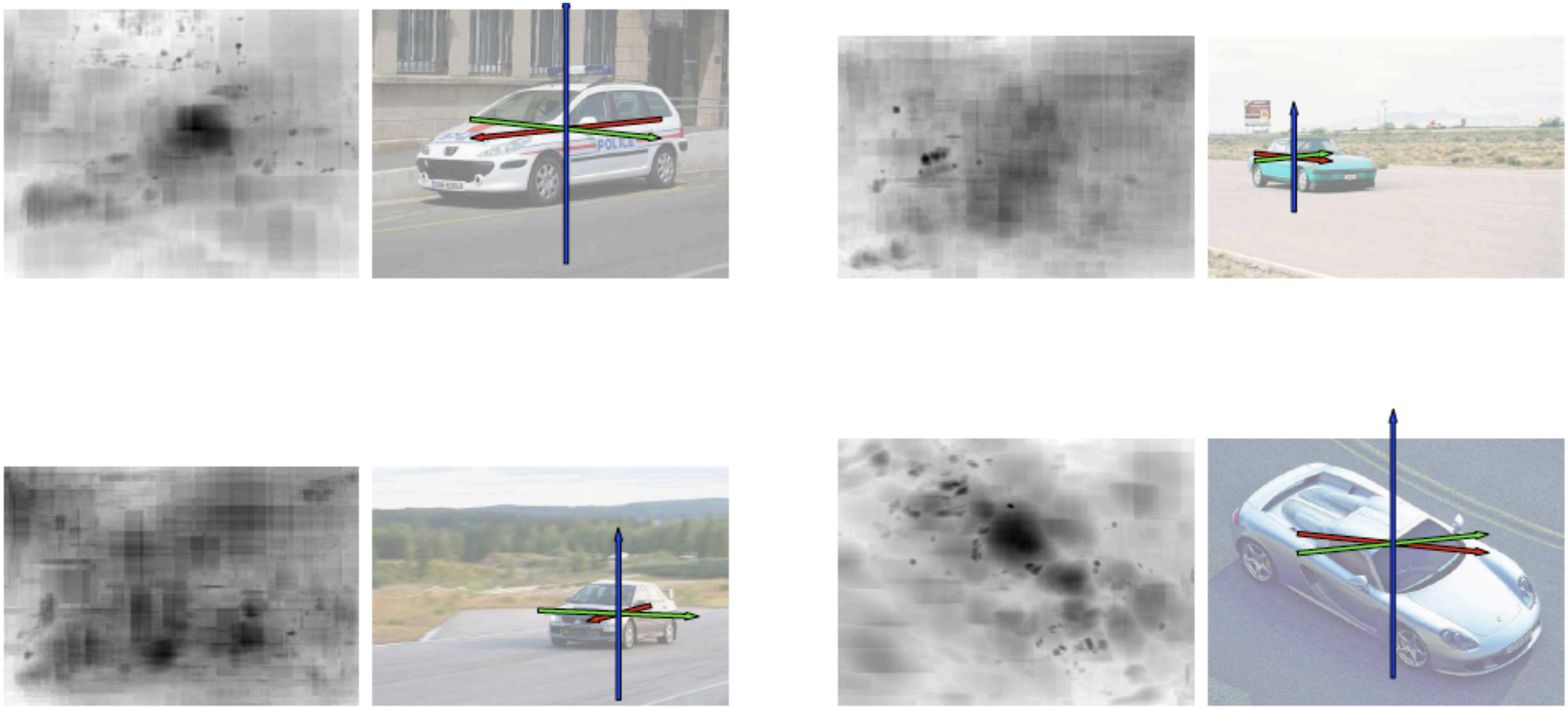


Categorization



	apple	car	cow	cup	dog	horse	pear	tomato
apple	**0.92**	0.00	0.00	0.01	0.00	0.00	0.00	0.07
car	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
cow	0.00	0.03	**0.89**	0.00	0.03	0.05	0.00	0.00
cup	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
dog	0.00	0.00	0.01	0.00	**0.95**	0.04	0.00	0.00
horse	0.00	0.00	0.06	0.00	0.02	**0.92**	0.00	0.00
pear	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00
tomato	0.11	0.00	0.00	0.00	0.00	0.00	0.00	**0.89**

Detections



Detections