Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning

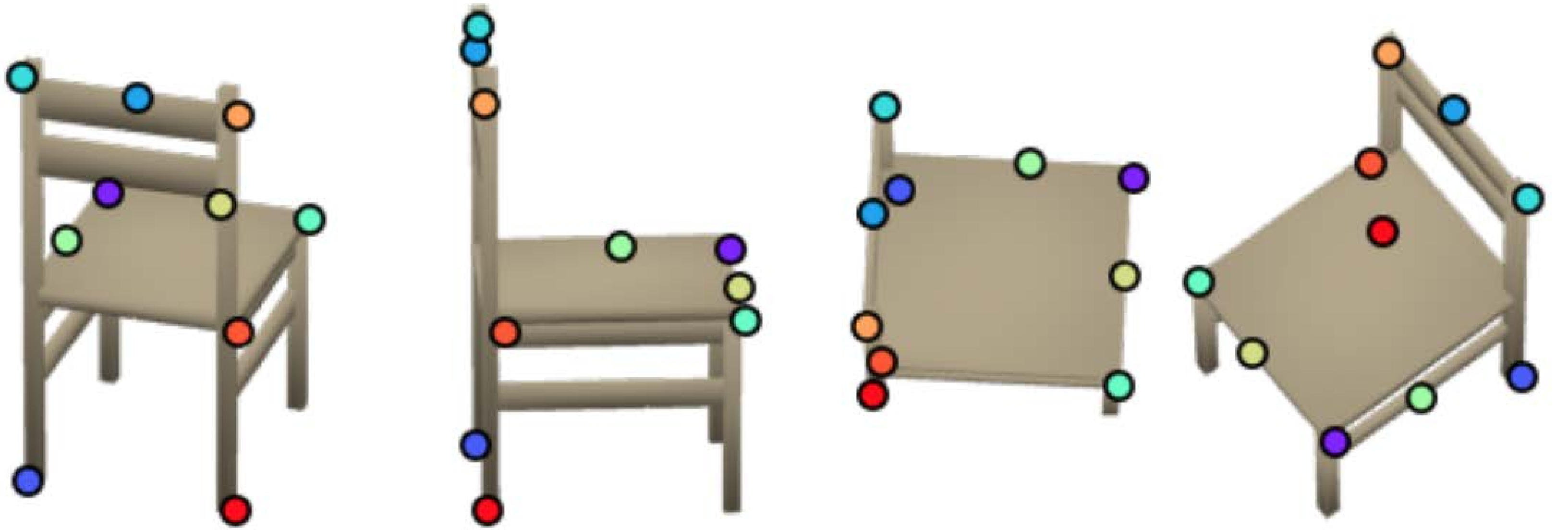
Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, Mohammad Norouzi

Presentation by Andrew V. Smith

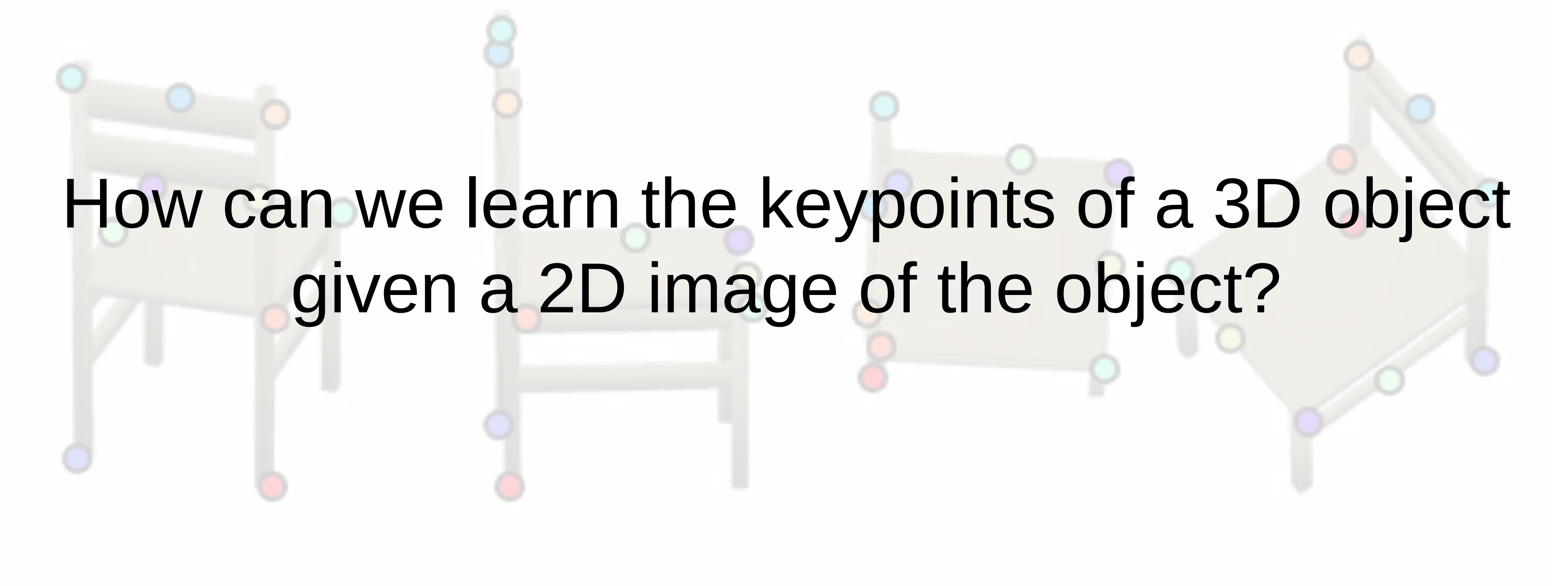
Overview

- Problem Statement
- Solution Overview
- Solution Details
- KeypointNet Architecture
- Testing Methodology
- Evaluation of Results

Problem

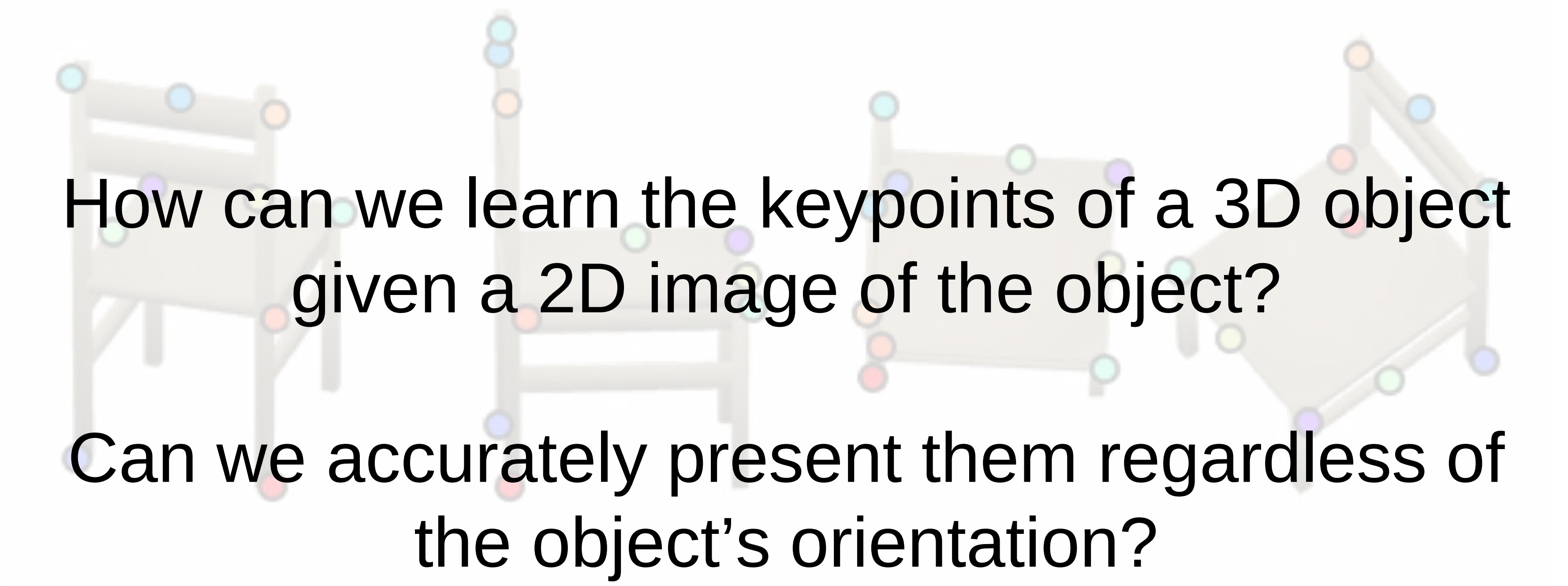


Problem



How can we learn the keypoints of a 3D object given a 2D image of the object?

Problem



How can we learn the keypoints of a 3D object given a 2D image of the object?

Can we accurately present them regardless of the object’s orientation?

Problem

- Find *optimal set* of 3D keypoints for a downstream task, without keypoint ground truth
- Formulate losses for each keypoint detection
  - **Multi-View Consistency loss**
  - **Relative Pose Estimation Loss**

KeypointNet: The Goal (Testing)

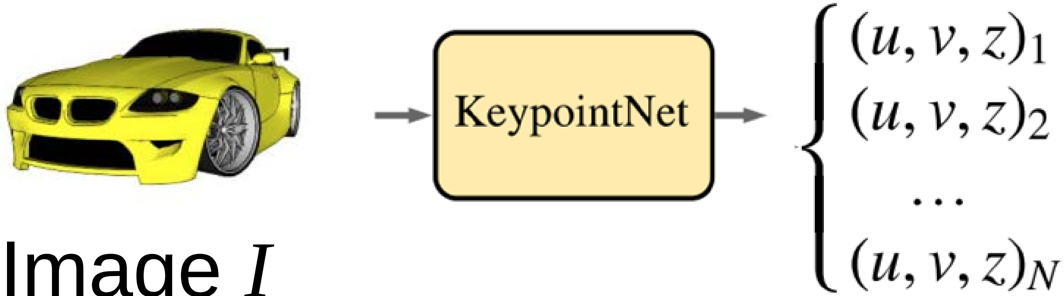


Image I

KeypointNet: The Setup (Training)

Image I

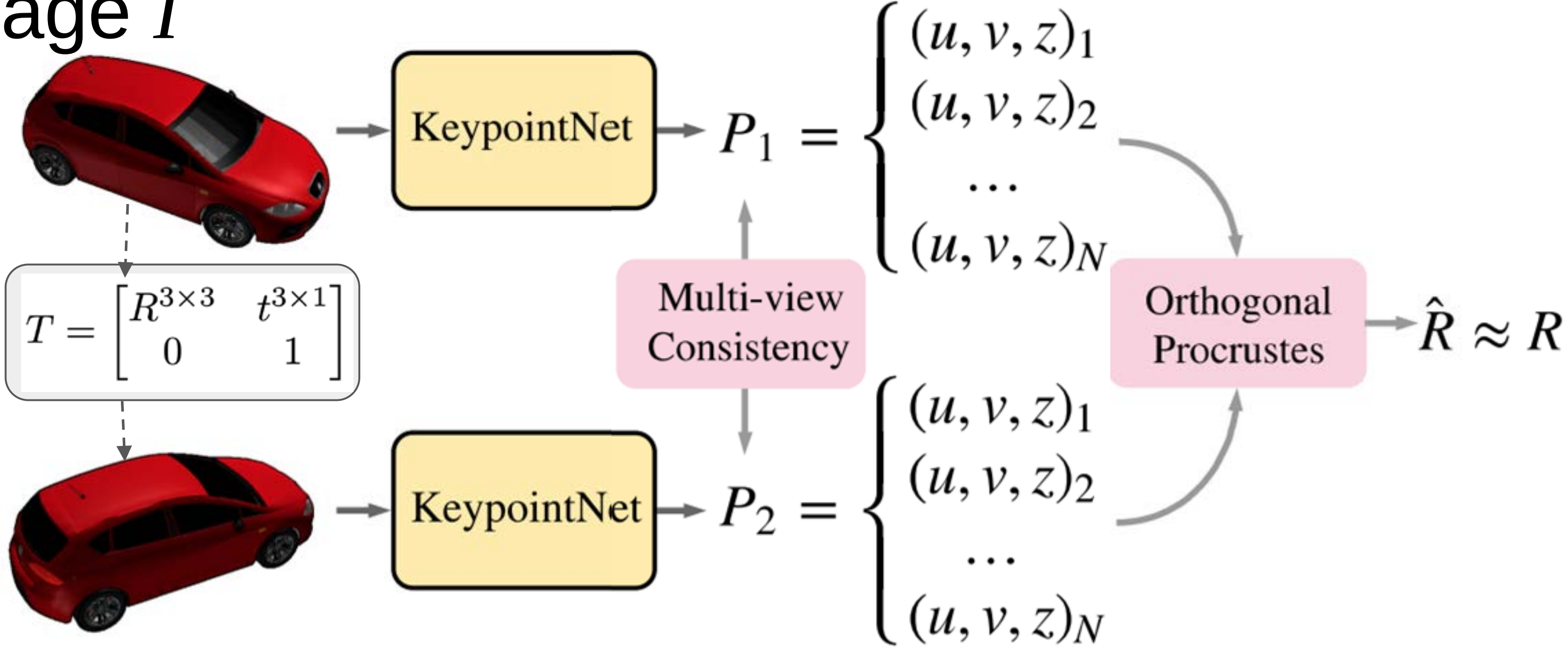
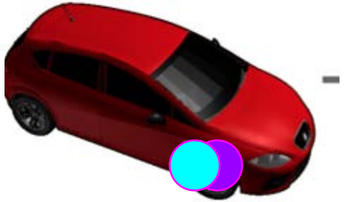


Image I’

Multi-view Consistency Loss

$$T = \begin{bmatrix} R^{3\times3} & t^{3\times1} \\ 0 & 1 \end{bmatrix}$$

$$\pi([x, y, z, 1]^\top) = \left[\frac{fx}{z}, \frac{fy}{z}, z, 1\right]^\top = [u, v, z, 1]^\top$$



$$[\hat{u}, \hat{v}, \hat{z}, 1]^\top \sim \pi T \pi^{-1}([u, v, z, 1]^\top)$$

$$[\hat{u'}, \hat{v'}, \hat{z'}, 1]^\top \sim \pi T^{-1} \pi^{-1}([u', v', z', 1]^\top)$$

$$L_{\text{con}} = \frac{1}{2N} \sum_{i=1}^{N} \left\| [u_i, v_i, u'_i, v'_i]^\top - [\hat{u'_i}, \hat{v'_i}, \hat{u}_i, \hat{v}_i]^\top \right\|^2$$

Relative Pose Estimation Loss

$X \text{ and } X' \in \mathbb{R}^{3\times N}$ $\qquad$ $\hat{X} \equiv [X_1, \ldots, X_N]$

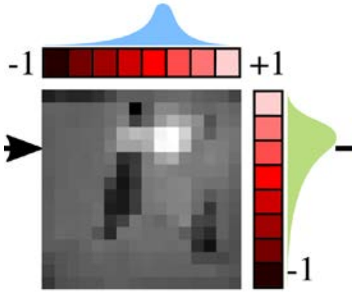
$$U, \Sigma, V^\top = \mathrm{SVD}(\tilde{X}\tilde{X}'^\top).$$

$$\hat{R} = V \,\mathrm{diag}(1, 1, \ldots, \det(VU^\top))U^\top$$

$$L_{\text{pose}} = 2 \arcsin\left(\frac{1}{2\sqrt{2}}\left\|\hat{R} - R\right\|_F\right)$$

Regarding Keypoints

$$[u_i, v_i]^\top = \sum_{u,v} [u \cdot g_i(u,v), v \cdot g_i(u,v)]^\top$$



$$z_i = \sum_{u,v} d_i(u,v) g_i(u,v)$$

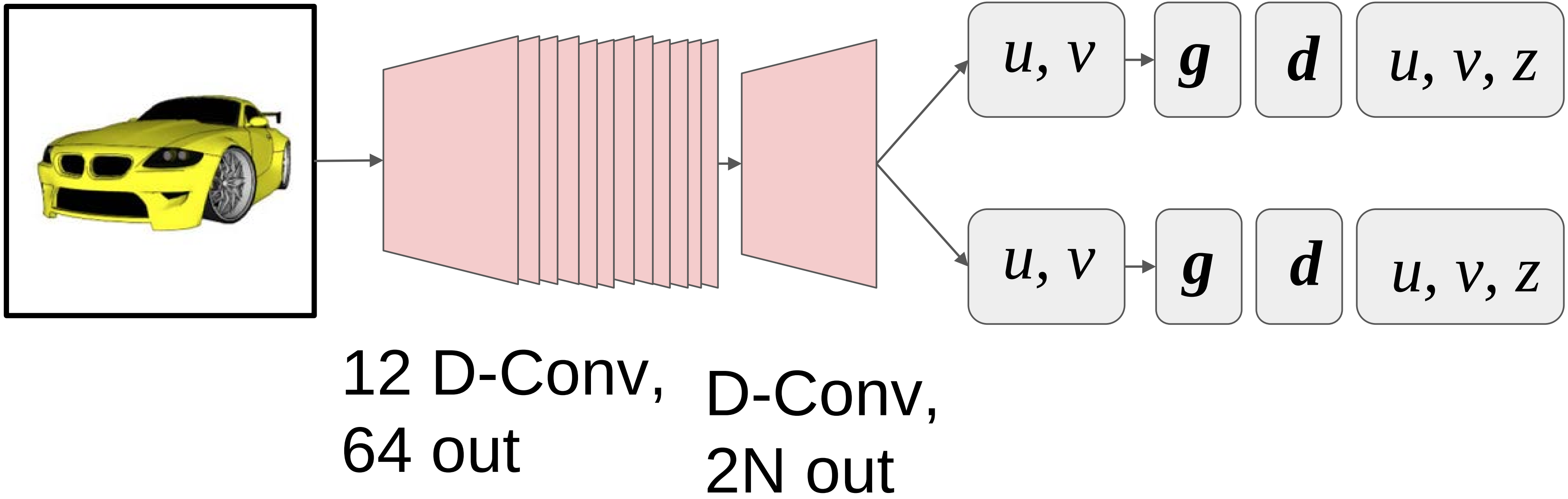
Regarding Keypoints (p. 2)

$$L_{\text{sep}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \max\left(0, \delta^2 - \|X_i - X_j\|^2\right)$$

$$L_{\text{obj}} = \frac{1}{N} \sum_{i=1}^{N} -\log \sum_{u,v} b(u,v) g_i(u,v)$$

$$L_{\text{var}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{u,v} g_i(u,v) \left\| [u,v]^\top - [u_i, v_i]^\top \right\|^2$$

Keypoint Net: Architecture



Overall Architecture Recap

Image I

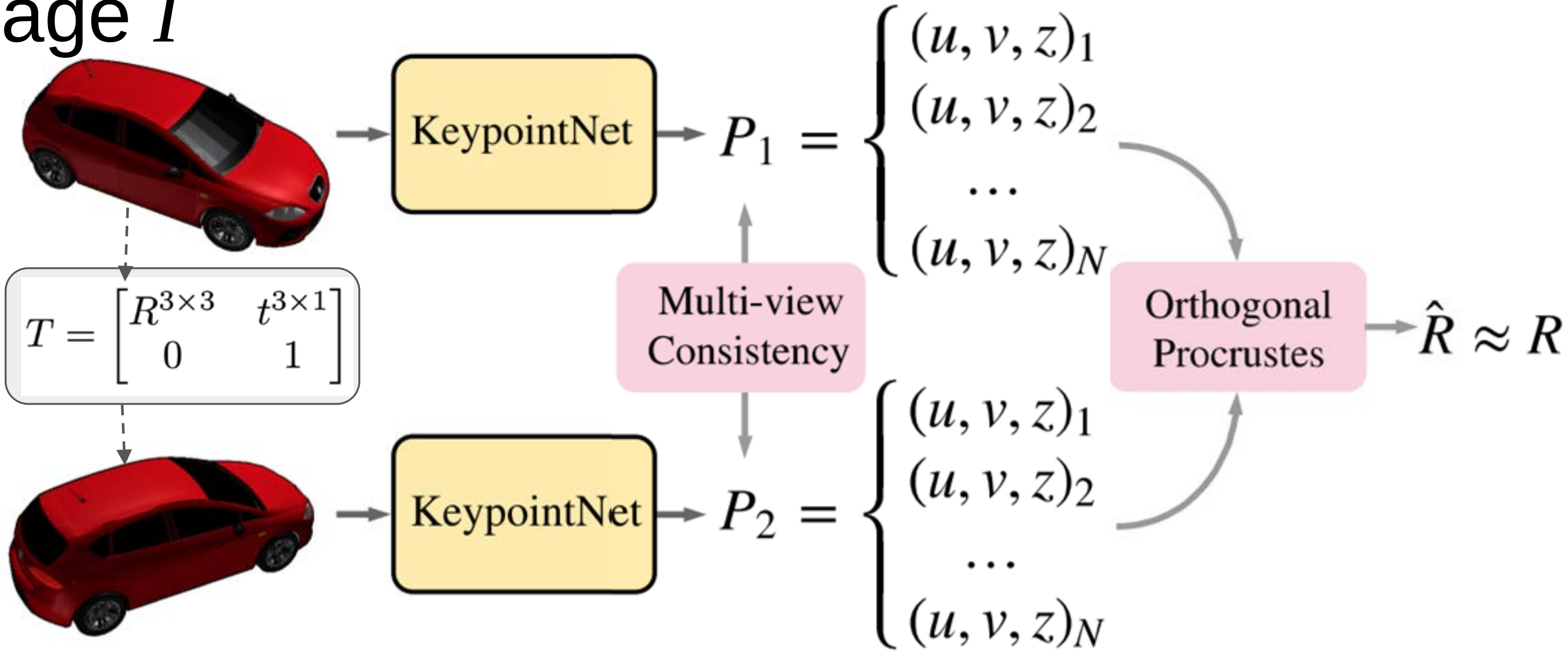


Image I’

Testing Methodology

Models from ShapeNet

- 100 pairs {I, I’} per model

Test against Supervised

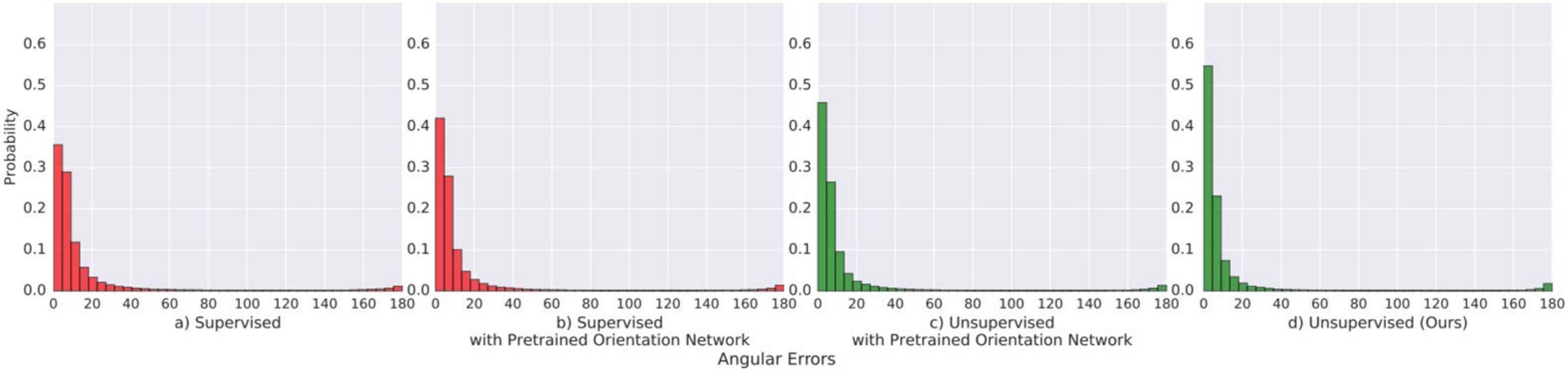
- Keypoints given from MTurk
- Orientation flags during training
- Angular Distance error & 3D standard error

Quantitative Results

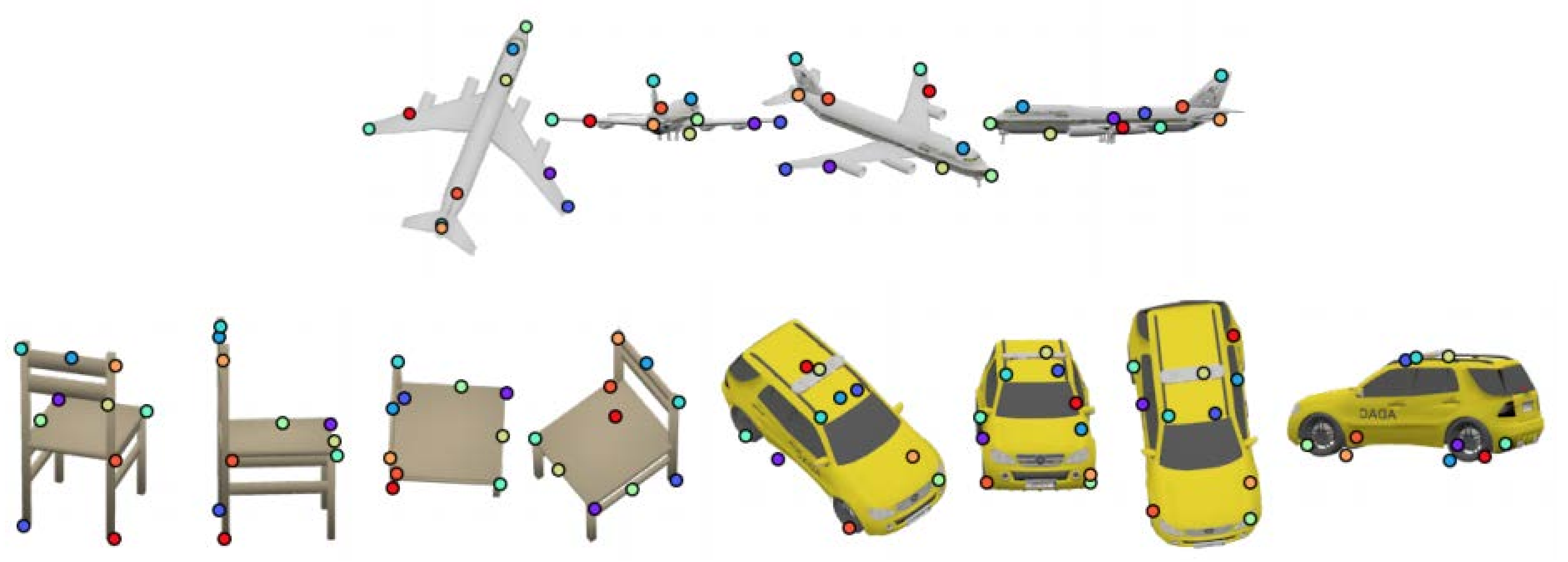
Method	Cars			Planes			Chairs		
	Mean	Median	3D-SE	Mean	Median	3D-SE	Mean	Median	3D-SE
a) Supervised	16.268	5.583	0.240	18.350	7.168	0.233	21.882	8.771	0.269
b) Supervised with pretrained O-Net	13.961	4.475	0.197	17.800	6.802	0.230	20.502	8.261	0.248
c) Ours with pretrained O-Net	13.500	4.418	0.165	18.561	6.407	0.223	14.238	5.607	0.203
d) **Ours**	11.310	3.372	0.171	17.330	5.721	0.230	14.572	5.420	0.196

Mean, Median Angular Distance Errors; and 3D standard errors reported. Lower is better.

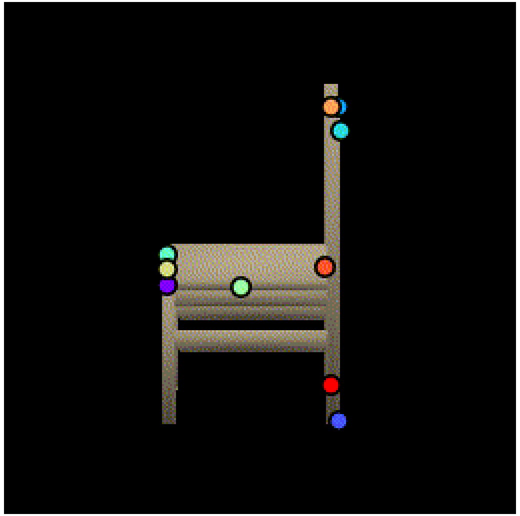
Quantitative Results (p. 2)



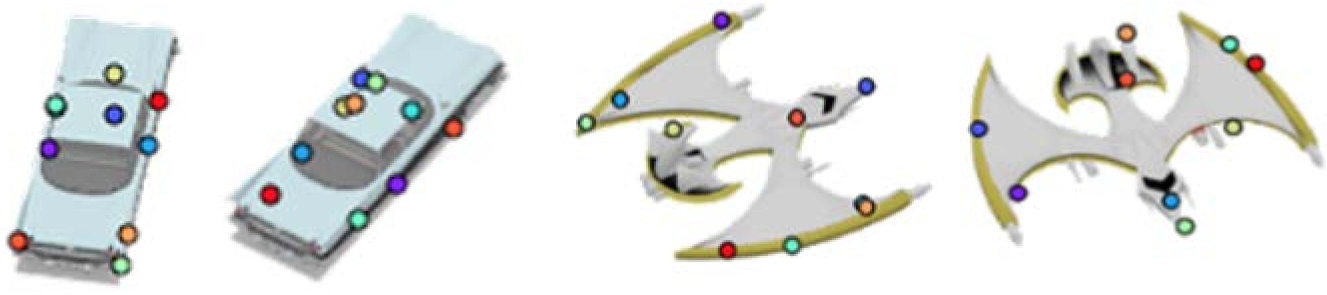
Qualitative Results



Qualitative Results (p. 2)


Qualitative Results (p. 3, failure cases)



Additional Results (ablation, primary losses)

Additional Results (other testing)

Figure 8: Results on a non-rigidly deformed car.

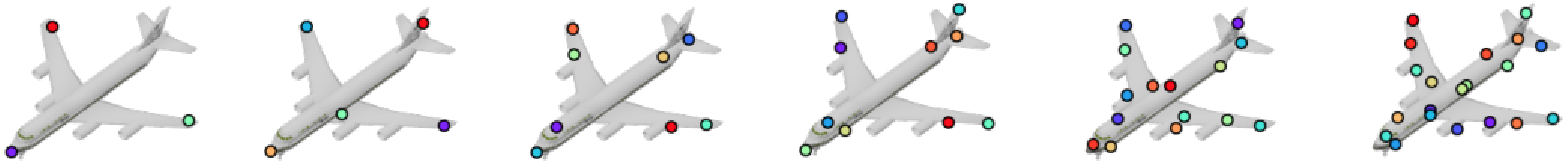
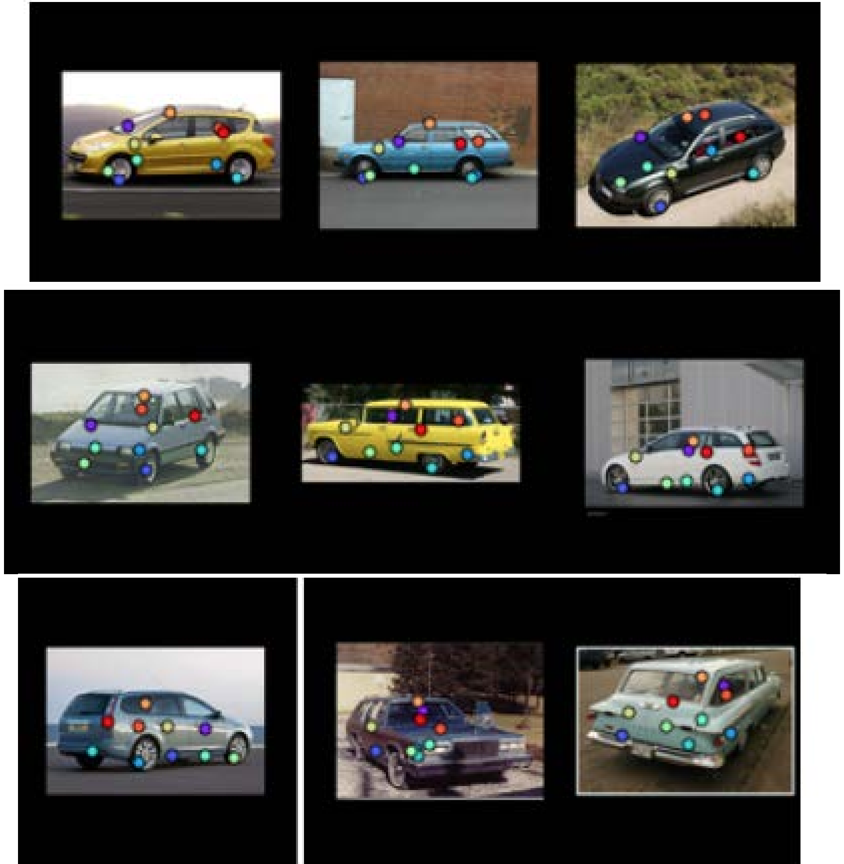


Figure 9: Results using networks trained to predict different numbers of keypoints. (Colors do not correspond across results as they are learned independently.)

Additional Results (proof-of-concept ImageNet)



More Information

https://keypointnet.github.io/

Summary

- Semi-supervised end-to-end keypoint finder
- Combines keypoint and geometry learning in one network
- Outperforms supervised method

Thank you!