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 

\[ (u, v, z)_1 \]
\[ (u, v, z)_2 \]
\[ \ldots \]
\[ (u, v, z)_N \]
KeypointNet: The Setup (Training)

Image $I$

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

KeypointNet $\rightarrow P_1 = \begin{cases} (u, v, z)_1 \\ (u, v, z)_2 \\ \vdots \\ (u, v, z)_N \end{cases}$

Multi-view Consistency

Orthogonal Procrustes

$\hat{R} \approx R$

Image $I'$

KeypointNet $\rightarrow P_2 = \begin{cases} (u, v, z)_1 \\ (u, v, z)_2 \\ \vdots \\ (u, v, z)_N \end{cases}$
Multi-view Consistency Loss

\[
T = \begin{bmatrix}
R^{3 \times 3} & t^{3 \times 1} \\
0 & 1
\end{bmatrix}
\]

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

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

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

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

\[ \begin{align*}
&X \text{ and } X' \in \mathbb{R}^{3 \times N} \quad X \equiv [X_1, \ldots, \hat{X}_N] \\
&U, \Sigma, V^\top = \text{SVD}(\tilde{X} \tilde{X}'^\top).
\end{align*} \]

\[ \hat{R} = V \text{ diag}(1, 1, \ldots, \det(VU^\top))U^\top \]

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

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

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

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

12 D-Conv, 64 out

D-Conv, 2N out
Overall Architecture Recap

Image $I$

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

$P_1 = \{(u, v, z)_1, (u, v, z)_2, \ldots, (u, v, z)_N\}$

Multi-view Consistency

Image $I'$

$P_2 = \{(u, v, z)_1, (u, v, z)_2, \ldots, (u, v, z)_N\}$

Orthogonal Procrustes

$\hat{R} \approx R$
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

<table>
<thead>
<tr>
<th>Method</th>
<th>Cars</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>3D-SE</td>
<td>Mean</td>
<td>Median</td>
<td>3D-SE</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>a) Supervised</td>
<td>16.268</td>
<td>5.583</td>
<td>0.240</td>
<td>18.350</td>
<td>7.168</td>
<td>0.233</td>
<td>21.882</td>
<td>8.771</td>
</tr>
<tr>
<td>b) Supervised with</td>
<td>13.961</td>
<td>4.475</td>
<td>0.197</td>
<td>17.800</td>
<td>6.802</td>
<td>0.230</td>
<td>20.502</td>
<td>8.261</td>
</tr>
<tr>
<td>pretrained O-Net</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) Ours with pretrained O-Net</td>
<td>13.500</td>
<td>4.418</td>
<td>0.165</td>
<td>18.561</td>
<td>6.407</td>
<td>0.223</td>
<td>14.238</td>
<td>5.607</td>
</tr>
<tr>
<td>d) <strong>Ours</strong></td>
<td>11.310</td>
<td>3.372</td>
<td>0.171</td>
<td>17.330</td>
<td>5.721</td>
<td>0.230</td>
<td>14.572</td>
<td>5.420</td>
</tr>
</tbody>
</table>

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!