Unsupervised Learning of Object Landmarks through Conditional Image Generation

Tomas Jakab* ¹, Ankush Gupta* ¹, Hakan Bilen², Andrea Vedaldi¹

¹ VGG, University of Oxford, ² University of Edinburgh
* equal contribution

arxiv.org/abs/1806.07823
Our goal

Learn semantically meaningful landmarks without any manual annotations
Motivation
Why to learn landmarks?

Low dimensional object representation
Interpretable
Why unsupervised?

Reduce dependency on expensive manual annotations

Leverage vast amount of videos available online
Method
Intuition
Intuition

new pose
Intuition

new pose

head

right shoulder

\[
\begin{bmatrix}
  x_1 \quad y_1 \\
  \vdots \\
  x_k \quad y_k
\end{bmatrix}
\]
Intuition

new pose

head

right shoulder

\[
\begin{bmatrix}
x_1, y_1 \\
\vdots \\
x_k, y_k
\end{bmatrix}
\]
Model

goal: reconstruct target
Model

\[ \mathbf{x} \rightarrow \Psi \rightarrow \hat{\mathbf{x}}' \]

\[ \mathbf{x}' \rightarrow \mathbf{y}' \]

goal: reconstruct target
Model

appearance
encoding

\( X \)

\( \Psi \)

\( \hat{X}' \)

\( X' \)

\( \Phi \)

\( K \)

\( -1 \)

\( +1 \)

\( [x_1, y_1] \)

\( \cdots \)

\( [x_k, y_k] \)

\( y' \)
Model

unsupervised keypoint extraction
Loss

\[ L \approx \text{reconstruction loss} \]
Bottleneck

heatmap for each keypoint
Bottleneck for each heatmap row-wise & column-wise sum for each heatmap.

heatmap for each keypoint

row-wise & column-wise sum
Bottleneck

for each heatmap

2 vectors

row-wise & column-wise sum

heatmap for each keypoint
Bottleneck

For each heatmap:
- Row-wise & column-wise sum of 2 vectors
- Apply softmax

Heatmap for each keypoint
Bottleneck

For each heatmap, row-wise & column-wise sum of 2 vectors applies softmax weighted sum of a grid.

Heatmap for each keypoint

Row-wise & column-wise sum

Weighted sum of a grid

$K$

$[x_1, y_1]$

$[x_k, y_k]$

$\vdots$
Bottleneck

For each heatmap row-wise & column-wise sum 2 vectors apply softmax weighted sum of a grid draw Gaussians
Bottleneck

for each heatmap

row-wise & column-wise sum

2 vectors

apply softmax

weighted sum of a grid

draw Gaussians

end-to-end differentiable

y'

heatmap for each keypoint

K

$[x_1, y_1]$

$[x_k, y_k]$
Bottleneck

\[ x \xrightarrow{\Phi} \Psi \xrightarrow{\text{model}} \hat{x}' \]

\[ x' \xrightarrow{K} y' \]
Results
Human faces

unsupervised landmarks
Human faces

unsupervised landmarks
Human faces

unsupervised landmarks
Human faces

unsupervised landmarks
Human faces

unsupervised landmarks

linear regression

regressed landmarks
### Human faces

<table>
<thead>
<tr>
<th>Method</th>
<th>$K$</th>
<th>MAFL</th>
<th>AFLW</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFAN</td>
<td>15.84</td>
<td>10.94</td>
<td></td>
</tr>
<tr>
<td>TCDCN</td>
<td>7.95</td>
<td>7.65</td>
<td></td>
</tr>
<tr>
<td>Cascaded CNN</td>
<td>9.73</td>
<td>8.97</td>
<td></td>
</tr>
<tr>
<td>RAR</td>
<td>–</td>
<td>7.23</td>
<td></td>
</tr>
<tr>
<td>MTCNN</td>
<td>5.39</td>
<td>6.90</td>
<td></td>
</tr>
<tr>
<td>Thewlis [1]</td>
<td>50</td>
<td>6.67</td>
<td>10.53</td>
</tr>
<tr>
<td>Thewlis [2] (frames)</td>
<td>–</td>
<td>5.83</td>
<td>8.80</td>
</tr>
<tr>
<td>Zhang [3] w/ equiv.</td>
<td>30</td>
<td>3.16</td>
<td><strong>6.58</strong></td>
</tr>
<tr>
<td>w/o equiv.</td>
<td>30</td>
<td>8.42</td>
<td>–</td>
</tr>
<tr>
<td>Ours</td>
<td>30</td>
<td>3.23</td>
<td>7.20</td>
</tr>
<tr>
<td>Ours selfsup.</td>
<td>30</td>
<td><strong>3.08</strong></td>
<td>6.98</td>
</tr>
</tbody>
</table>

- **Supervised methods**
- **Unsupervised methods**
- Uses equivariance
Human pose

Unsupervised landmarks on Human3.6m
Human pose

Regressed landmarks on BBCPose
3D objects smallNORB

invariance to 3D pose, lighting and object shape
Disentangling style and geometry
Model freeze parameters
Street numbers

appearance

geometry

reconstruction
Human faces

appearance

geometry

reconstruction
Human pose

appearance

geometry

reconstruction
Related work


Conditional Image Generation for Learning the Structure of Visual Objects

Tomas Jakab*¹, Ankush Gupta*¹, Hakan Bilen², Andrea Vedaldi¹

¹ VGG, University of Oxford, ² University of Edinburgh

* equal contribution

arxiv.org/abs/1806.07823