Human Token Identification

- Select informative visual tokens with the attention scores of keypoint tokens
- Attention value determines how much information of each visual token is fused into the output

\[
\text{Softmax}(\frac{q_i K_j}{\sqrt{D}}) V_i = \alpha^i V_i
\]

Summary of cross-view Fusion Strategies

- Global Fusion: each pixel in each view calculates attention with respect to all \(n\) pixels of other \(m-1\) views, \(O(m^2)\)
- Epipolar-based Fusion: each pixel in each view calculates attention with \(k\) pixels along the corresponding epipolar lines of other \(m-1\) views, \(O(m^nk)\)

**Human area fusion (ours):** dense global attention among \(k\) human foreground pixels of \(m\) views, \(O(m^2k^2)\)

Results

- selected tokens after each HTI module
- human areas are gradually refined as the network deepens

Results on COCO

- PPT achieves significant acceleration while matching its accuracy on 2D pose estimation
- Pruning background tokens doesn’t hurt the accuracy
- Attention among foreground tokens is sufficient

<table>
<thead>
<tr>
<th>Method</th>
<th>(\Delta\text{PAPS})</th>
<th>(\Delta\text{GLOPAP})</th>
<th>(\Delta\text{AP})</th>
<th>(\Delta\text{AP}_{25})</th>
<th>(\Delta\text{AP}_{50})</th>
<th>(\Delta\text{AP}_{75})</th>
</tr>
</thead>
<tbody>
<tr>
<td>SingleBaseline-Res50</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>SingleBaseline-Res101</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-3 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-2 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-1 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-1 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>(\Delta\text{PAPS})</th>
<th>(\Delta\text{GLOPAP})</th>
<th>(\Delta\text{AP})</th>
<th>(\Delta\text{AP}_{25})</th>
<th>(\Delta\text{AP}_{50})</th>
<th>(\Delta\text{AP}_{75})</th>
</tr>
</thead>
<tbody>
<tr>
<td>SingleBaseline-Res50</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>SingleBaseline-Res101</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-3 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-2 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-1 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Results on Human3.6M

- Human area fusion is better than global attention in both accuracy and efficiency for multi-view pose estimation
- Pruning background tokens doesn’t hurt the accuracy
- Attention among foreground tokens is sufficient

<table>
<thead>
<tr>
<th>Method</th>
<th>(\Delta\text{PAPS})</th>
<th>(\Delta\text{GLOPAP})</th>
<th>(\Delta\text{AP})</th>
<th>(\Delta\text{AP}_{25})</th>
<th>(\Delta\text{AP}_{50})</th>
<th>(\Delta\text{AP}_{75})</th>
</tr>
</thead>
<tbody>
<tr>
<td>SingleBaseline-Res50</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>SingleBaseline-Res101</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-3 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-2 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-1 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>(\Delta\text{PAPS})</th>
<th>(\Delta\text{GLOPAP})</th>
<th>(\Delta\text{AP})</th>
<th>(\Delta\text{AP}_{25})</th>
<th>(\Delta\text{AP}_{50})</th>
<th>(\Delta\text{AP}_{75})</th>
</tr>
</thead>
<tbody>
<tr>
<td>SingleBaseline-Res50</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>SingleBaseline-Res101</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-3 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-2 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Efficient-Pose-1 [5]</td>
<td>9.0</td>
<td>2.4</td>
<td>6.6</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Table 1. Results on COCO validation dataset. The input size is 256 × 192. GLOPAPs\* means the GLOPAPs for the transformers following only the equations from [9], as our method only focus on accelerating the transformers.

Table 4. 2D pose estimation on Human3.6M. The metric is JHR on original image. All inputs are resized to 256 × 256, \(4 \times\) mean the number of views used in cross-view fusion step. The FLOPs is the total computation for each view and cross-view fusion.

Reference


Contact: haoyum3@uci.edu, Paper ID: 46