Team MSRA Keypoints Detection

Bin Xiao\textsuperscript{1}, Dianqi Li\textsuperscript{2}, Ke Sun\textsuperscript{1}, Lei Zhang\textsuperscript{2}, Jingdong Wang\textsuperscript{1}

\textsuperscript{1}Microsoft Research Asia \textsuperscript{2}Microsoft
Outline

• Top-Down Pipeline
• Human Pose Estimation
  • General pipeline for human pose estimation
  • Our simple baseline method for human pose estimation
• Experiment Results
Top-Down Pipeline

• Person Detector
  • Re-implement Mask-RCNN
    • Backbone: Xception
    • Powered by FPN and Deformable Convolution
    • Data: COCO only
  • Performance (on COCO test-dev dataset):
    • Box AP (person): 60.9
    • Box AR (person): 72.59

• Simple Baseline Network for Human Pose Estimation

Outline

• Top-Down Pipeline
• Human Pose Estimation
  • General pipeline for human pose estimation
  • Our simple baseline method for human pose estimation
• Experiment Results
General Pipeline for Human Pose Estimation

256x256 → CNN Network → 64x64
State-of-the-art networks

Human Pose Estimation

• How good could a simple method be?
Our simple baseline

- **Contraction**
- **Expansion**

256x256 → 8x8 → 64x64

- **Backbone Network**
- **Deconvolution Head**

Deconvolution Module

L2 Loss
## Summary

<table>
<thead>
<tr>
<th></th>
<th>Stacked Hourglass</th>
<th>CPN</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilinear Up-sampling</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Deconvolution layers</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Skip layer feature concatenation</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Multi-stage architecture</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Multi-supervision</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>
Outline

• Top-Down Pipeline
• Human Pose Estimation
  • General pipeline for human pose estimation
  • Our simple baseline method for human pose estimation
• Experiment Results
## Results on COCO validation dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Input Size</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-stage Hourglass</td>
<td>-</td>
<td>256x192</td>
<td>66.9</td>
</tr>
<tr>
<td>CPN</td>
<td>ResNet-50</td>
<td>256x192</td>
<td>68.6</td>
</tr>
</tbody>
</table>
| Ours            | ResNet-50    | 256x192      | 70.4 | 3.5
### COCO Submission

<table>
<thead>
<tr>
<th>Method</th>
<th>AP(COCO test-dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Simple Baseline Network($ResNet50$)</td>
<td>70.2</td>
</tr>
<tr>
<td>+$256 \times 192 \rightarrow 384 \times 288$</td>
<td>71.3↑ 1.2</td>
</tr>
<tr>
<td>+$ResNet50 \rightarrow ResNet152$</td>
<td>73.7↑ 2.4</td>
</tr>
<tr>
<td>+Training with External Dataset</td>
<td>75.4↑ 1.7</td>
</tr>
<tr>
<td>+Models Ensemble</td>
<td>76.5↑ 1.1</td>
</tr>
</tbody>
</table>
Summary

- A simple and effect baseline method is proposed.
- State-of-the-art results are achieved.
- We hope such baseline network would benefit the field by easing the idea development and evaluation.
Human Pose Estimation

The sequences are from PoseTrack dataset.
Teams

Bin Xiao  Dianqi Li  Ke Sun  Lei Zhang  Jingdong Wang
Thank you!

Code and Models is available at https://github.com/Microsoft/human-pose-estimation.pytorch