Holistic, Instance-level Human Parsing

Qizhu Li*, Anurag Arnab*, Philip Torr

* Indicates equal contribution by the authors

05 September 2017
1. Objective
1. Objective

*Instance-aware* body part segmentation of humans
1. Objective

Instance-level human segmentation

Category-level body part segmentation
2. Methodology

1. Do category-level body part segmentation

2. Detect humans

3. Use the instance-level segmentation module to assign instance labels.
2.1 Methodology: Instance-level segmentation module

\[
E(V = v) = - \sum_i^{N} \ln(w_1 \psi_{Box}(v_i) + w_2 \psi_{Global}(v_i) + \epsilon) + \sum_{i<j}^{N} \psi_{Pairwise}(v_i, v_j)
\]

Where \( V = \{V_1, V_2, \ldots V_N\} \) is a multinormial variable at all \( N \) pixels,
\( V_i \in \{1, 2, \ldots D\} \times \{1, 2, \ldots P\} \cup \{0, 0\} \)
\( E(V = v) \) is the energy of \( V \) taking a particular value \( v \)
2.1.1 Methodology: Box term

- **Input 1**: human detections:
  - scores $s_i$ and bounding boxes $B_i$ for $1 \leq i \leq D$

- **Input 2**: semantic segmentation network output:
  - Feature map $Q$ with $P + 1$ channels ($P = 6$ here)

- **Output**:

\[
\psi_{Box} \left( V_k = (i, j) \right) = \begin{cases} 
  s_i Q_k(j), & k \in B_i \\
  0, & k \notin B_i 
\end{cases}
\]

for $(i, j) \in \{1, 2, \ldots, D\} \times \{1, 2, \ldots, P\}$
Input 1: Human detections

Input 2: Semantic body part segmentation probability maps

Output: Box terms
### 2.1.2 Methodology: Global term

- **Input:** semantic segmentation network output:
  - Feature map $Q$ with $P + 1$ channels ($P = 6$ here)

- **Output:**

$$
\psi_{Global}(V_k = (i, j)) = Q_k(j)
$$
for $(i, j) \in \{1, 2, \ldots D\} \times \{1, 2, \ldots P\} \cup \{0, 0\}$
Input 1: Number of human detections

4

Input 2: Semantic body part segmentation probability maps

Output: Global terms
2.2 Methodology: Loss function

• Observation: permuting the label IDs in an instance segmentation ground truth produces an equally valid ground truth.

Figure 1. Permutations of ground truth labels are equally valid
2.2 Methodology: Loss function

- We match ground truth \( y \) to prediction \( p \) before we carry out loss calculation.

- Matched ground truth is given by:
  \[
  y^* = \arg\max_{z \in \pi(y)} \text{IoU}(z, p)
  \]

- Then cross-entropy loss is calculated as per normal

---

(a) Prediction  (a) Matched GT  (a) Ground Truth

Figure 2. Ground truth is matched to our prediction before calculating loss.
2.3 Methodology:
Obtaining segmentation at other granularities

For each pixel we predict part instance label $(i, j)$
i.e. part $j$ of person $i$

- (i) Instance segmentation of human
- (j) Semantic segmentation of body parts
## 3.1 Results:
Part instance segmentation

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU Threshold</th>
<th>$AP_{vol}^r$</th>
<th>$AP_{vol}^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNC</td>
<td>38.8</td>
<td>28.1</td>
<td>19.3</td>
</tr>
<tr>
<td>Ours, piecewise, box term only</td>
<td>38.0</td>
<td>27.4</td>
<td>16.7</td>
</tr>
<tr>
<td>Ours, piecewise</td>
<td>38.8</td>
<td>28.5</td>
<td>17.6</td>
</tr>
<tr>
<td>Ours, end-to-end</td>
<td>39.0</td>
<td>28.6</td>
<td>17.4</td>
</tr>
<tr>
<td>Ours, piecewise, box term only, OHEM</td>
<td>38.7</td>
<td>28.9</td>
<td>17.5</td>
</tr>
<tr>
<td>Ours, piecewise, OHEM</td>
<td>39.7</td>
<td>29.7</td>
<td>18.7</td>
</tr>
<tr>
<td>Ours, end-to-end, OHEM</td>
<td><strong>40.6</strong></td>
<td><strong>30.4</strong></td>
<td>19.1</td>
</tr>
</tbody>
</table>

Table 2. Ablation study and comparison of $AP_{vol}^r$ at various thresholds to MNC on Pascal Person-Parts test set. $AP_{vol}^r = \frac{1}{9} \sum_{t=1}^{9} AP_{t/10}^r$. 
3.1 Results:
Part instance segmentation

Success cases...
3.1 Results:
Part instance segmentation

Failure cases...
3.1 Results:
Part instance segmentation

Comparison to MNC...
### 3.2 Results:
Human instance segmentation

<table>
<thead>
<tr>
<th>Methods</th>
<th>IoU Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>SDS</td>
<td>47.9</td>
</tr>
<tr>
<td>Chen et al.</td>
<td>48.3</td>
</tr>
<tr>
<td>PFN</td>
<td>48.4</td>
</tr>
<tr>
<td>Arnab et al.</td>
<td>58.6</td>
</tr>
<tr>
<td>R2-IOS</td>
<td>60.4</td>
</tr>
<tr>
<td>Arnab et al.</td>
<td>65.6</td>
</tr>
<tr>
<td>Ours, piecewise</td>
<td>64.0</td>
</tr>
<tr>
<td>Ours, end-to-end</td>
<td><strong>70.2</strong></td>
</tr>
</tbody>
</table>

Table 3. Comparison of $AP^r$ at various thresholds for instance-level human segmentation on the VOC 2012 validation set. $AP^r_{vol} = \frac{1}{9} \sum_{t=1}^{9} AP^r_{t/10}$. 
3.2 Results:
Human instance segmentation

Comparison to MNC...
(We run the public MNC model on Pascal Person-Parts test set)
3.3 Results:
Semantic segmentation of body parts

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deeplab</td>
<td>53.0</td>
</tr>
<tr>
<td>Attention</td>
<td>56.4</td>
</tr>
<tr>
<td>HAZN</td>
<td>57.5</td>
</tr>
<tr>
<td>LG-LSTM</td>
<td>58.0</td>
</tr>
<tr>
<td>Graph LSTM</td>
<td>60.2</td>
</tr>
<tr>
<td>Deeplab-v2</td>
<td>64.9</td>
</tr>
<tr>
<td>RefineNet</td>
<td>68.6</td>
</tr>
<tr>
<td>Ours, pre-trained</td>
<td>65.9</td>
</tr>
<tr>
<td>Ours, final network</td>
<td>66.3</td>
</tr>
</tbody>
</table>

Table 4. Comparison of semantic part segmentation results on the Pascal Person-Parts test set.
The End

Thank you!
Appendix
Methodology:
Category-level segmentation module

Figure 2. Comparison of our category-level segmentation module to Deeplab-v2
Methodology: Category-level segmentation module

<table>
<thead>
<tr>
<th></th>
<th>Test IoU [%]</th>
<th>Memory [GB]</th>
<th>Time [s] (fps [s⁻¹])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deeplab-v2</td>
<td>64.4</td>
<td>9.5</td>
<td>0.396 (2.5)</td>
</tr>
<tr>
<td>Deeplab-v2+CRF</td>
<td>64.9</td>
<td>11.2</td>
<td>0.960 (1.0)</td>
</tr>
<tr>
<td>Ours</td>
<td>65.9</td>
<td>4.3</td>
<td>0.255 (3.9)</td>
</tr>
</tbody>
</table>

Table 1. Comparison of our category-level segmentation module to Deeplab-v2. Tests done on the Pascal Person-Parts dataset. Memory and time requirements are for a single forward pass of the network.
Experiments

Training steps:

1. Pretrain the **semantic segmentation network** on all VOC 2012 train and SBD images minus VOC 2012 val and Pascal Person-Parts test to learn the VOC 21 classes.

2. Finetune the model on Pascal Person-Parts training set to predict the 7 body part classes (including the background class).

3. Train a **human detector** on VOC 07+12 trainval images minus VOC 2012 val and Pascal Person-Parts test images. We use the publicly available R-FCN framework.

4. Finetune the **full instance model** end-to-end.
(a) Input Image

(b) Semantic Segmentation

(c) Instance Human Segmentation

(d) Instance Part Segmentation