1. Additional Results on LSP dataset

We provide additional quantitative results on LSP dataset using person-centric (PC) and observer-centric (OC) evaluation settings.

1.1. LSP Person-Centric (PC)

First, detailed performance analysis is performed when evaluating various parameters of AFR-CNN and results are reported using PCK [13] evaluation measure. Then, performance of the proposed AFR-CNN and Dense-CNN part detection models is evaluated using strict PCP [4] measure.

Detailed AFR-CNN performance analysis (PCK). Detailed parameter analysis of AFR-CNN is provided in Tab. 1 and results are reported using PCK evaluation measure. Respecting parameters for each experiment are shown in the first column and parameter differences between the neighboring rows in the table are highlighted in bold. Re-scoring the 2000 DPM proposals using AFR-CNN with AlexNet [8] leads to 56.9% PCK. This is achieved using basis scale 1 (≈ head size) of proposals and training with initial learning rate (lr) of 0.001 for 80k iterations, after which lr is reduced by 0.1, for a total number of 140k SGD iterations. In addition, bounding box regression and default IoU threshold of 0.5 for positive/negative label assignment [5] have been used. Extending the regions by 4x increases the performance to 65.1% PCK, as it incorporates more context including the information about symmetric body parts and allows to implicitly encode higher-order body part relations into the part detector. No improvements observed for larger scales. Increasing lr to 0.003, lr reduction step to 160k and training for a larger number of iterations (240k) improves the results to 67.4, as higher lr allows for more significant updates of model parameters when finetuned on the task of human body part detection. Increasing the number of training examples by reducing the training IoU threshold to 0.4 results into slight performance improvement (68.8 vs. 67.4% PCK). Further increasing the number of training samples by horizontally flipping each image and performing translation and scale jittering of the ground truth training samples improves the performance to 69.6% PCK and 42.3% AUC. The improvement is more pronounced for smaller distance thresholds (42.3 vs. 40.9% AUC): localization of body parts is improved due to the increased number of jittered samples that significantly overlap with the ground truth. Further increasing the lr, lr reduction step and total number of iterations altogether improves the performance to 72.4% PCK, and very minor improvements are observed when training longer. All results above are achieved by finetuning the AlexNet architecture from the ImageNet model on the MPII training set. Further finetuning the MPII-finetuned model on the LSP training set increases the performance to 77.9% PCK, as the network learns LSP-specific image representations. Using the deeper VGG [14] architecture improves over more shallow AlexNet (77.9 vs. 72.4% PCK, 50.0 vs. 44.6% AUC). Finetuning VGG on LSP achieves remarkable 82.8% PCK and 57.0% AUC. Strong increase in AUC (57.0 vs. 50%) characterizes the improvement for smaller PCK evaluation thresholds. Switching off bounding box regression results into performance drop (81.3% PCK, 53.2% AUC) thus showing the importance of the bounding box regression for better part localization. Overall, we demonstrate that proper adaptation and tweaking of the state-of-the-art generic object detector FR-CNN [5] leads to a strong body part detection model that dramatically improves over the vanilla FR-CNN (82.8 vs. 56.9% PCK, 57.8 vs. 35.9% AUC) and significantly outperforms the state of the art (+9.4% PCK over the best known PCK result [1] and +9.7% AUC over the best known AUC result [15].

Overall performance using PCP evaluation measure.
Performance when using the strict “Percentage of Correct Parts (PCP)” [4] measure is reported in Tab. 2. In contrast to PCK measure evaluating the accuracy of predicting body joints, PCP metric measures the accuracy of predicting body part sticks. AFR-CNN achieves 78.3% PCP. Similar to PCK results, DeepCut SP AFR-CNN slightly improves over unary alone, as it enforces more consistent predictions of body part sticks. Using more general multi-person DeepCut MP AFR-CNN model results into similar performance, which shows the generality of DeepCut MP method. DeepCut SP Dense-CNN slightly improves over Dense-CNN alone (84.3 vs. 83.9% PCP) achieving the best PCP result on LSP dataset using PC annotations. This is in contrast to PCK results where performance differences DeepCut SP Dense-CNN vs. Dense-CNN alone are minor.

We now compare the PCK results to the state of the art. The DeepCut models outperform all other methods by a large margin. The best known PCK result by Chen&Yuille [1] is outperformed by 10.7% PCP. This is interesting, as their deep learning based method relies on the image conditioned pairwise terms while our approach uses more simple geometric only connectivity. Interestingly, AFR-CNN alone outperforms the approach of Fan et al. [17] (78.3 vs. 70.1% PCP), who build on the previous version of the R-CNN detector [6]. At the same time, the best performing dense architecture DeepCut SP Dense-CNN outperforms [17] by +14.2% PCP. Surprisingly, DeepCut SP Dense-CNN dramatically outperforms the method of Tompson et al. [15] (+17.7% PCP) that also produces dense score maps, but additionally includes multi-scale receptive fields and jointly trains appearance and spatial models in a single deep learning framework. We envision that both advances can further improve the performance of DeepCut models. Finally, all proposed approaches significantly outperform earlier non-deep learning based methods [16, 11] relying on hand-crafted image features.

1.2. LSP Observer-Centric (OC)

We now evaluate the performance of the proposed part detection models on LSP dataset using the observer-centric (OC) annotations [3]. In contrast to the person-centric (PC) annotations used in all previous experiments, OC annotations do not penalize for the right/left body part prediction flips and count a body part to be the right body part, if it is on the right side of the line connecting pelvis and neck, and a body part to be the left body part otherwise.

Evaluation is performed using the official OC annotations provided by [10, 3]. Prior to evaluation, we first finetune the AFR-CNN and Dense-CNN part detection models from ImageNet on MPII and MPII+LSPET training sets, respectively, (same as for PC evaluation), and then further finetuned the models on LSP OC training set.

**PCK evaluation measure.** Results using OC annotations and PCK evaluation measure are shown in Tab. 3 and in Fig. 1. AFR-CNN achieves 84.2% PCK and 58.1% AUC. This result is only slightly better compared to AFR-CNN evaluated using PC annotations (84.2 vs 82.8% PCK, 58.1 vs. 57.0% AUC). Although PC annotations correspond to a harder task, only small drop in performance when using PC annotations shows that the network can learn to accurately predict person’s viewpoint and correctly label left/right limbs in most cases. This is contrast to earlier approaches based on hand-crafted features whose performance drops much stronger when evaluated in PC evaluation setting (e.g. [11] drops from 71.0% PCK when using OC annotations to 58.0% PCK when using PC annotations). Similar to PC case, Dense-CNN detection model outperforms AFR-CNN (88.2 vs. 84.2% PCK and 65.0 vs. 58.1% AUC). The differences are more pronounced when examining the

<table>
<thead>
<tr>
<th>Setting</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>PCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet scale 1, lr 0.001, lr step 80k, # iter 140k, IoU pos/neg 0.5</td>
<td>82.2</td>
<td>67.0</td>
<td>49.6</td>
<td>45.4</td>
<td>53.1</td>
<td>52.9</td>
<td>48.2</td>
<td>56.9</td>
</tr>
<tr>
<td>AlexNet scale 4, lr 0.001, lr step 80k, # iter 140k, IoU pos/neg 0.5</td>
<td>85.7</td>
<td>74.4</td>
<td>61.3</td>
<td>53.2</td>
<td>64.1</td>
<td>61.3</td>
<td>53.8</td>
<td>56.1</td>
</tr>
<tr>
<td>AlexNet scale 4, lr 0.003, lr step 160k, # iter 240k, IoU pos/neg 0.5</td>
<td>87.5</td>
<td>76.7</td>
<td>64.8</td>
<td>56.0</td>
<td>68.2</td>
<td>68.7</td>
<td>59.6</td>
<td>68.1</td>
</tr>
<tr>
<td>AlexNet scale 4, lr 0.003, lr step 160k, # iter 240k, IoU pos/neg 0.4, data augment</td>
<td>87.8</td>
<td>77.8</td>
<td>66.0</td>
<td>58.1</td>
<td>70.9</td>
<td>69.6</td>
<td>59.8</td>
<td>69.6</td>
</tr>
<tr>
<td>AlexNet scale 4, lr 0.004, lr step 320k, # iter 1M, IoU pos/neg 0.4, data augment</td>
<td>88.1</td>
<td>79.3</td>
<td>68.9</td>
<td>62.6</td>
<td>73.5</td>
<td>69.3</td>
<td>64.7</td>
<td>72.4</td>
</tr>
</tbody>
</table>

Table 1: PCK performance of AFR-CNN (unary) on LSP (PC) dataset. AFR-CNN is finetuned from ImageNet on MPII (lines 1-6, 8), and then finetuned on LSP (lines 7, 9).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>PCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG scale 4, lr 0.003, lr step 160k, # iter 320k, IoU pos/neg 0.4, data augment</td>
<td>91.0</td>
<td>84.2</td>
<td>74.6</td>
<td>67.7</td>
<td>74.6</td>
<td>72.8</td>
<td>77.9</td>
<td>51.6</td>
</tr>
</tbody>
</table>

Table 2: Pose estimation results (PCP) on LSP (PC) dataset.
Table 3: Pose estimation results (PCK) on LSP (OC) dataset.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>PCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR-CNN (unary)</td>
<td>95.3</td>
<td>88.3</td>
<td>78.5</td>
<td>74.2</td>
<td>87.3</td>
<td>84.2</td>
<td>81.2</td>
<td>84.2</td>
</tr>
<tr>
<td>Dense-CNN (unary)</td>
<td>97.4</td>
<td>92.0</td>
<td>83.8</td>
<td>79.0</td>
<td>93.1</td>
<td>88.3</td>
<td>83.7</td>
<td>88.2</td>
</tr>
<tr>
<td>Chen&amp;Yuille [1]</td>
<td>91.5</td>
<td>84.7</td>
<td>70.3</td>
<td>63.2</td>
<td>82.7</td>
<td>78.1</td>
<td>72.0</td>
<td>77.5</td>
</tr>
<tr>
<td>Ouyang et al. [9]</td>
<td>86.5</td>
<td>78.2</td>
<td>61.7</td>
<td>49.3</td>
<td>76.9</td>
<td>70.0</td>
<td>67.6</td>
<td>70.0</td>
</tr>
<tr>
<td>Pishchulin et. [11]</td>
<td>87.5</td>
<td>77.6</td>
<td>61.4</td>
<td>47.6</td>
<td>79.0</td>
<td>75.2</td>
<td>68.4</td>
<td>71.0</td>
</tr>
<tr>
<td>Kiefel&amp;Gehler [7]</td>
<td>83.5</td>
<td>73.7</td>
<td>59.5</td>
<td>36.2</td>
<td>73.7</td>
<td>70.5</td>
<td>66.9</td>
<td>65.8</td>
</tr>
<tr>
<td>Ramakrishna et al. [12]</td>
<td>84.9</td>
<td>77.8</td>
<td>61.4</td>
<td>47.2</td>
<td>73.6</td>
<td>69.1</td>
<td>68.8</td>
<td>69.9</td>
</tr>
</tbody>
</table>

Figure 1: Pose estimation results over all PCK thresholds on LSP (OC) dataset.

Table 4: Pose estimation results (PCP) on LSP (OC) dataset.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>PCP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR-CNN (unary)</td>
<td>92.9</td>
<td>86.3</td>
<td>79.8</td>
<td>77.0</td>
<td>64.2</td>
<td>91.8</td>
<td>79.9</td>
<td></td>
</tr>
<tr>
<td>Dense-CNN (unary)</td>
<td>96.0</td>
<td>91.0</td>
<td>83.5</td>
<td>82.8</td>
<td>71.8</td>
<td>96.2</td>
<td>85.0</td>
<td></td>
</tr>
<tr>
<td>Chen&amp;Yuille [1]</td>
<td>92.7</td>
<td>82.9</td>
<td>77.0</td>
<td>69.2</td>
<td>55.4</td>
<td>87.8</td>
<td>75.0</td>
<td></td>
</tr>
<tr>
<td>Ouyang et al. [9]</td>
<td>88.6</td>
<td>77.8</td>
<td>71.9</td>
<td>61.9</td>
<td>45.4</td>
<td>84.3</td>
<td>68.7</td>
<td></td>
</tr>
<tr>
<td>Pishchulin et. [11]</td>
<td>88.7</td>
<td>78.9</td>
<td>73.2</td>
<td>61.8</td>
<td>45.0</td>
<td>85.1</td>
<td>69.2</td>
<td></td>
</tr>
<tr>
<td>Kiefel&amp;Gehler [7]</td>
<td>84.3</td>
<td>74.5</td>
<td>67.6</td>
<td>54.1</td>
<td>28.3</td>
<td>78.3</td>
<td>61.2</td>
<td></td>
</tr>
<tr>
<td>Ramakrishna et al. [12]</td>
<td>88.1</td>
<td>79.0</td>
<td>73.6</td>
<td>62.8</td>
<td>39.5</td>
<td>80.4</td>
<td>67.8</td>
<td></td>
</tr>
</tbody>
</table>

2. Additional Results on WAF dataset

Qualitative comparison of our joint formulation DeepCut MP Dense-CNN to the traditional two-stage approach Dense-CNN det ROI relying on person detector, and to the approach of Chen&Yuille [2] on WAF dataset is shown in Fig. 2. See figure caption for visual performance analysis.

3. Additional Results on MPII Multi-Person

Qualitative comparison of our joint formulation DeepCut MP Dense-CNN to the traditional two-stage approach Dense-CNN det ROI on MPII Multi-Person dataset is shown in Fig. 3 and 4. Dense-CNN det ROI works well when multiple fully visible individuals are sufficiently separated and thus their body parts can be partitioned based on the person detection bounding box. In this case the strong Dense-CNN body part detection model can correctly estimate most of the visible body parts (image 1, 6, 7). However, Dense-CNN det ROI cannot tell apart the body parts of multiple individuals located next to each other and possibly occluding each other, and often links the body parts across the individuals (images 1-16, 19-20). In addition, Dense-CNN det ROI cannot reason about occlusions and truncations always providing a prediction for each body part (image 4, 6, 10). In contrast, DeepCut MP Dense-CNN is able to correctly partition and label an initial pool of body part candidates (each image, row 2) into subsets that correspond to sets of mutually consistent body part candidates and abide to mutual consistency and exclusion constraints (each image, row 2), thereby outputting consistent body pose predictions (each image, row 3).}

PCP evaluation measure. Results using PC annotations and PCP evaluation measure are shown in Tab. 4. Overall, the trend is similar to PC evaluation: both proposed approaches significantly outperform the state-of-the-art methods with Dense-CNN achieving the best result of 85.0% PCP thereby improving by $+10\%$ PCP over the best published result [1].
Typical of person 4 are not estimated due to missing part detection candidates. Staying people (image 4, right shoulder and wrist of person 2 are linked to the right elbow of person 3; image 5, left elbow of person 4 is linked to the left wrist of person 3). In contrast, DeepCut MP predicts body part occlusions, disambiguates multiple and potentially overlapping people and correctly assembles independent detections into plausible body part configurations (image 4, left arms of people 1-3 are correctly predicted to be occluded; image 5, linking of body parts across people 3 and 4 is corrected; image 7, occlusion of body parts is correctly predicted and visible parts are accurately estimated). In contrast to Chen&Yuille [2], DeepCut MP better predicts occlusions of person's body parts by the nearby staying people (images 1, 3-9), but also by other objects (image 2, left arm of person 1 is occluded by the chair). Furthermore, DeepCut MP is able to better cope with strong articulations and foreshortenings (image 1, person 6; image 3, person 2; image 5, person 4; image 7, person 4; image 8, person 1). Typical DeepCut MP failure case is shown in image 10: the right upper arm of person 3 and both arms of person 4 are not estimated due to missing part detection candidates.
Figure 3: Qualitative comparison of our joint formulation DeepCut MP Dense-CNN (rows 1-3, 5-7) to the traditional two-stage approach Dense-CNN det ROI (rows 4, 8) on MPII Multi-Person dataset. See Fig. 1 in paper for the color-coding explanation.
Figure 4: Qualitative comparison (contd.) of our joint formulation DeepCut MP Dense-CNN (rows 1-3, 5-7) to the traditional two-stage approach Dense-CNN det ROI (rows 4, 8) on MPII Multi-Person dataset. See Fig. 1 in paper for the color-coding.
visible body parts per person.

References