DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation Supplementary Material

Leonid Pishchulin[1], Eldar Insafutdinov[1], Siyu Tang[1], Bjoern Andres[1],
Mykhaylo Andriluka[1,3], Peter Gehler[2], and Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Germany
[2]Max Planck Institute for Intelligent Systems, Germany
[3]Stanford University, USA

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 ($\approx$ 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 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). Funetuning 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.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AlexNet scale 1, lr 0.001, lr step 80k, # iter 140k, IoU pos/neg 0.5	82.2	67.0	49.6	45.4	53.1	52.9	48.2	56.9	35.9
AlexNet **scale 4**, lr 0.001, lr step 80k, # iter 140k, IoU pos/neg 0.5	85.7	74.4	61.3	53.2	64.1	63.1	53.8	65.1	39.0
AlexNet scale 4, **lr 0.003, lr step 160k, # iter 240k**, IoU pos/neg 0.5	87.0	75.1	63.0	56.3	67.0	65.7	58.0	67.4	40.8
AlexNet scale 4, lr 0.003, lr step 160k, # iter 240k, **IoU pos/neg 0.4**	87.5	76.7	64.8	56.0	68.2	68.7	59.6	68.8	40.9
AlexNet scale 4, lr 0.003, lr step 160k, # iter 240k, IoU pos/neg 0.4, **data augment**	87.8	77.8	66.0	58.1	70.9	66.9	59.8	69.6	42.3
AlexNet scale 4, **lr 0.004, lr step 320k, # iter 1M**, IoU pos/neg 0.4, data augment	88.1	79.3	68.9	62.6	73.5	69.3	64.7	72.4	44.6
+ finetune LSP, lr 0.0005, lr step 10k, # iter 40k	92.9	81.0	72.1	66.4	80.6	77.6	75.0	77.9	51.6
VGG scale 4, lr 0.003, lr step 160k, # iter 320k, IoU pos/neg 0.4, data augment	91.0	84.2	74.6	67.7	77.4	77.3	72.8	77.9	50.0
+ finetune LSP lr 0.0005, lr step 10k, # iter 40k	**95.4**	**86.5**	**77.8**	**74.0**	**84.5**	**78.8**	**82.6**	**82.8**	**57.0**

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).

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 evaluation 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 PCP results to the state of the art. The *DeepCut* models outperform all other methods by a large margin. The best known PCP 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.

	Torso	Upper Leg	Lower Leg	Upper Arm	Fore-arm	Head	PCP
AFR-CNN (unary)	93.2	82.7	77.7	75.5	63.5	91.2	78.3
+ *DeepCut SP*	93.3	83.2	77.8	76.3	63.7	91.5	78.7
+ appearance pairwise	93.4	83.5	77.8	76.6	63.8	91.8	78.9
+ *DeepCut MP*	93.6	83.3	77.6	76.3	63.5	91.2	78.6
Dense-CNN (unary)	96.2	87.8	81.8	81.6	**72.3**	95.6	83.9
+ *DeepCut SP*	**97.0**	**88.8**	**82.0**	**82.4**	71.8	**95.8**	**84.3**
+ *DeepCut MP*	96.4	**88.8**	80.9	**82.4**	71.3	94.9	83.8
Tompson et al. [15]	90.3	70.4	61.1	63.0	51.2	83.7	66.6
Chen&Yuille [1]	96.0	77.2	72.2	69.7	58.1	85.6	73.6
Fan et al. [17]*	95.4	77.7	69.8	62.8	49.1	86.6	70.1
Pishchulin et al. [11]	88.7	63.6	58.4	46.0	35.2	85.1	58.0
Wang&Li [16]	87.5	56.0	55.8	43.1	32.1	79.1	54.1

* re-evaluated using the standard protocol, for details see project page of [17]

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

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

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AFR-CNN (unary)	95.3	88.3	78.5	74.2	87.3	84.2	81.2	84.2	58.1
Dense-CNN (unary)	**97.4**	**92.0**	**83.8**	**79.0**	**93.1**	**88.3**	**83.7**	**88.2**	**65.0**
Chen&Yuille [1]	91.5	84.7	70.3	63.2	82.7	78.1	72.0	77.5	44.8
Ouyang et al. [9]	86.5	78.2	61.7	49.3	76.9	70.0	67.6	70.0	43.1
Pishchulin et. [11]	87.5	77.6	61.4	47.6	79.0	75.2	68.4	71.0	45.0
Kiefel&Gehler [7]	83.5	73.7	55.9	36.2	73.7	70.5	66.9	65.8	38.6
Ramakrishna et al. [12]	84.9	77.8	61.4	47.2	73.6	69.1	68.8	69.0	35.2

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

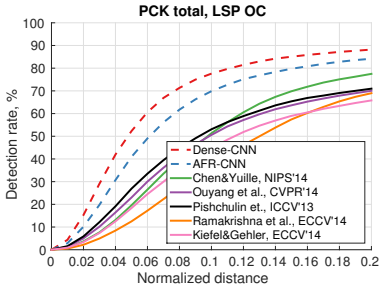

Figure 1: Pose estimation results over all PCK thresholds on LSP (OC) dataset.

	Torso	Upper Leg	Lower Leg	Upper Arm	Fore-arm	Head	PCP
AFR-CNN (unary)	92.9	86.3	79.8	77.0	64.2	91.8	79.9
Dense-CNN (unary)	**96.0**	**91.0**	**83.5**	**82.8**	**71.8**	**96.2**	**85.0**
Chen&Yuille [1]	92.7	82.9	77.0	69.2	55.4	87.8	75.0
Ouyang et al. [9]	88.6	77.8	71.9	61.9	45.4	84.3	68.7
Pishchulin et. [11]	88.7	78.9	73.2	61.8	45.0	85.1	69.2
Kiefel&Gehler [7]	84.3	74.5	67.6	54.1	28.3	78.3	61.2
Ramakrishna et al. [12]	88.1	79.0	73.6	62.8	39.5	80.4	67.8

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

entire PCK curve for smaller distance thresholds (c.f. Fig. 1).

Comparing the performance by *AFR-CNN* and *Dense-CNN* to the state of the art, we observe that both proposed approaches significantly outperform other methods. Both deep learning based approaches of Chen&Yuille [1] and Ouyang et al. [9] are outperformed by $+10.7$ and $+18.2\%$ PCK when compared to the best performing *Dense-CNN*. Analysis of PCK curve for the entire range of PCK distance thresholds reveals even larger performance differences (c.f. Fig. 1). The results using OC annotations confirm our findings from PC evaluation and clearly show the advantages of the proposed part detection models over the state-of-the-art deep learning methods [1, 9], as well as over earlier pose estimation methods based on hand-crafted image features [11, 7, 12].

PCP evaluation measure. Results using OC 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].

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 16, 17, 19). 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, top row) 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). $c \neq c'$ pairwise terms allow to partition the initial set of part detection candidates into valid pose configurations (each image, row 2: person-clusters highlighted by dense colored connections). $c = c'$ pairwise terms facilitate clustering of multiple body part candidates of the same body part of the same person (each image, row 2: markers of the same type and color). In addition, $c = c'$ pairwise terms facilitate a repulsive property that prevents nearby part candidates of the same type to be associated to different people (image 1: detections of the left shoulder are assigned to the front person only). Furthermore, *DeepCut MP Dense-CNN* allows to either merge or deactivate part hypotheses thus effectively performing non-maximum suppression and reasoning about body part occlusions and truncations (image 3, row 2: body part hypotheses on the background are deactivated (black crosses); image 6, row 2: body part hypotheses for the truncated body parts are deactivated (black crosses); image 1-6, 8-9, 13-14, row 3: only visible body parts of the partially occluded people are estimated, while non-visible body parts are correctly predicted to be occluded). These qualitative examples show that *DeepCuts MP* can successfully deal with the unknown number of people per image and the unknown number of

Figure 2: Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (rows 2, 5) to the traditional two-stage approach *Dense-CNN det ROI* (rows 1, 4) and to the approach of Chen&Yuille [2] (rows 3, 6) on WAF dataset. *det ROI* does not reason about occlusion and often predicts inconsistent body part configurations by linking the parts across the nearby 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.

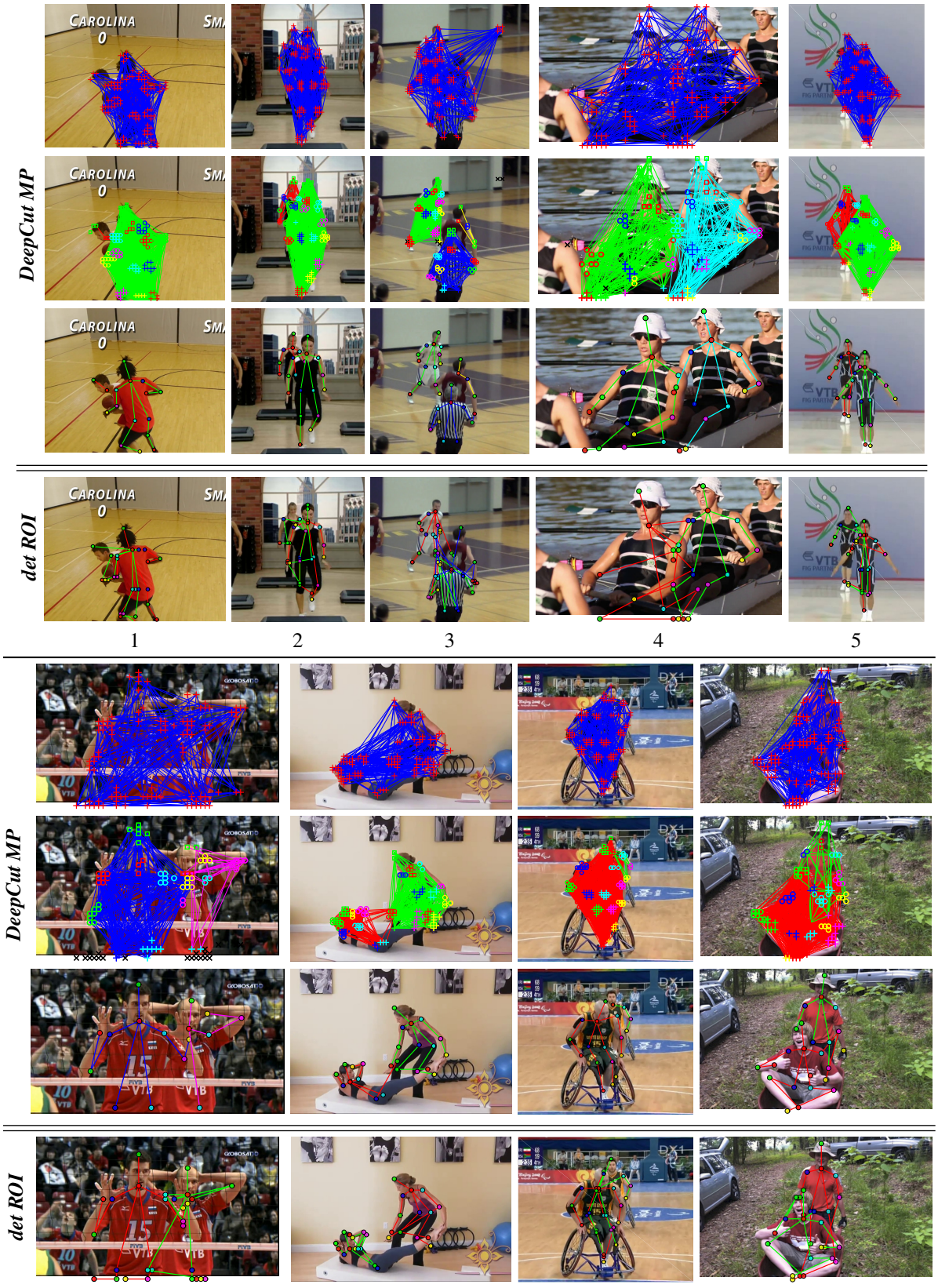

Figure 3: Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (rows 1-3, 5-7) to the traditional two-stage approach *Dense-CNN det ROI* (rows 4, 8) on MPII Multi-Person dataset. See Fig. 1 in paper for the color-coding explanation.

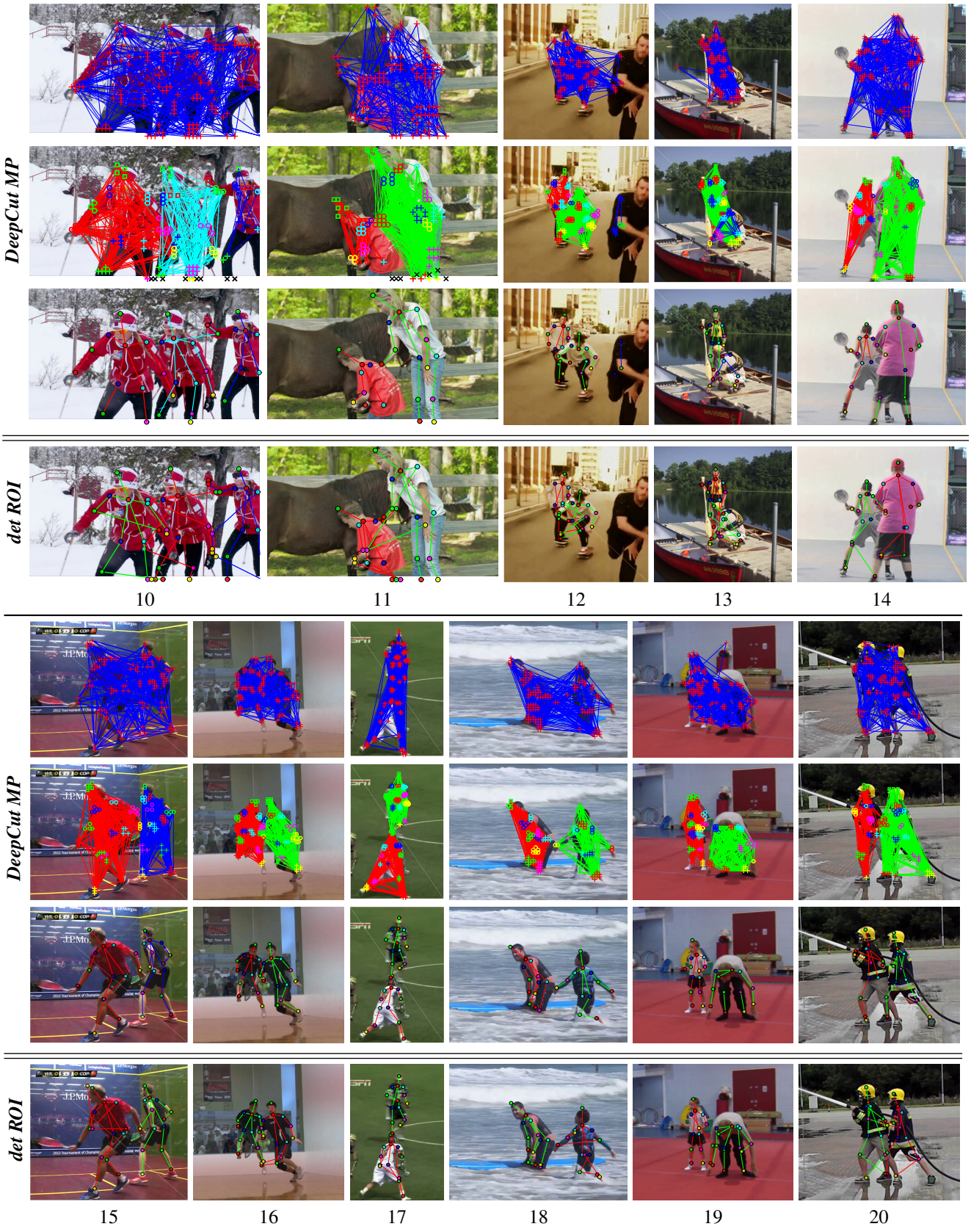

Figure 4: Qualitative comparison (contd.) of our joint formulation *DeepCut MP Dense-CNN* (rows 1-3, 5-7) to the traditional two-stage approach *Dense-CNN det ROI* (rows 4, 8) on MPII Multi-Person dataset. See Fig. 1 in paper for the color-coding.

visible body parts per person.

References

[1] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS'14*. 1, 2, 3

[2] X. Chen and A. Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 3, 4

[3] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV'12*. 2

[4] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR'08*. 1, 2

[5] R. Girshick. Fast r-cnn. In *ICCV'15*. 1

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR'14*. 2

[7] M. Kiefel and P. Gehler. Human pose estimation with fields of parts. In *ECCV'14*. 3

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'12*. 1

[9] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR'14*. 3

[10] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR'13*. 2

[11] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV'13*. 2, 3

[12] V. Ramakrishna, D. Munoz, M. Hebert, A. J. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV'14*. 3

[13] B. Sapp and B. Taskar. Multimodal decomposable models for human pose estimation. In *CVPR'13*. 1

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR,14*. 1

[15] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS'14*. 1, 2

[16] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR'13*. 2

[17] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR'15*. 2