ArtTrack: Articulated Multi-person Tracking in the Wild Supplementary Material

Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang,
Evgeny Levinkov, Bjoern Andres, Bernt Schiele

Max Planck Institute for Informatics
Saarland Informatics Campus
Saabrücken, Germany

1. Additional Results on the MPII Multi-Person Dataset

We perform qualitative comparison of the proposed single-frame based *TD/BU* and *BU-full* methods on challenging scenes containing highly articulated and strongly overlapping individuals. Results are shown in Fig. 1 and Figure 2. The *BU-full* works well when persons are sufficiently separated (images 11 and 12). However, it fails on images where people significantly overlap (images 1-3, 5-10) or exhibit high degree of articulation (image 4). This is due to the fact that geometric image-conditioned pairwise may get confused in the presence of multiple overlapping individuals and thus mislead post-CNN bottom-up assembling of body poses. In contrast, *TD/BU* performs explicit modeling of person identity via top-dop bottom-up reasoning while offloading the larger share of the reasoning about body-part association onto feed-forward convolutional architecture, and thus is able to resolve such challenging cases. Interestingly, *TD/BU* is able to correctly predict lower limbs of people in the back through partial occlusion (image 3, 5, 7, 10). *TD/BU* model occasionally incorrectly assembles body parts in kinematically implausible manner (image 12), as it does not explicitly model geometric body part relations. Finally, both models fail in presense of high variations in scale (image 13). We envision that reasoning over multiple scales is likely to improve the results.

2. Results on the We Are Family dataset

We compare our proposed *TD/BU* model to the state-of-the-art methods on the "We Are Family" (WAF) [2] dataset and present results in Table 1. We use evaluation protocol from [3] and report the AP evaluation measure. *TD/BU* model outperforms the best published results [3] across all body parts (87.7 vs 82.0% AP) as well improves on articulated parts such as wrists (+6.4% AP) and elbows (+6.4% AP). We attribute that to the ability of top-down model to better learn part associations compared to explicitly modeling geometric pairwise relations as in [3].

Method	Head	Sho	Elb	Wri	Total
TD/BU	**97.5**	**86.2**	**82.1**	**85.2**	**87.7**
DeeperCut [3]	92.6	81.1	75.7	78.8	82.0
DeepCut [4]	76.6	80.8	73.7	73.6	76.2
Chen&Yuille [1]	83.3	56.1	46.3	35.5	55.3

Table 1: Pose estimation results (AP) on WAF dataset.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP
BU-sparse	84.5	84.0	71.8	59.5	74.4	68.1	59.2	71.6
+ *det-distance*	84.8	84.3	72.9	61.8	74.1	67.4	59.1	72.1
+ *deepmatch*	85.5	83.9	73.0	62.0	74.0	68.0	59.5	72.3
+ *det-distance*	85.1	83.6	72.2	61.5	**74.4**	68.8	62.2	72.5
+ *sift-distance*	**85.6**	**84.5**	**73.4**	**62.1**	73.9	**68.9**	**63.1**	**73.1**

Table 2: Effects of different temporal features on pose estimation performance (AP) (*BU-sparse+temporal* model) on our "MPII Video Pose".

3. Evaluation of temporal features.

We evaluate the importance of combining temporal features introduced in Sec. 3.4 of the paper on our Multi-Person Video dataset. To that end, we consider *BU-sparse+temporal* model and compare results to *BU-sparse* in Tab. 2. Single-frame *BU-sparse* achieves 71.6% AP. It can be seen that using geometry based *det-distance* features slightly improves the results to 72.1% AP, as it enables the propagation of information from neighboring frames. Using *deepmatch* features slightly improves the performance further as it helps to link the same body

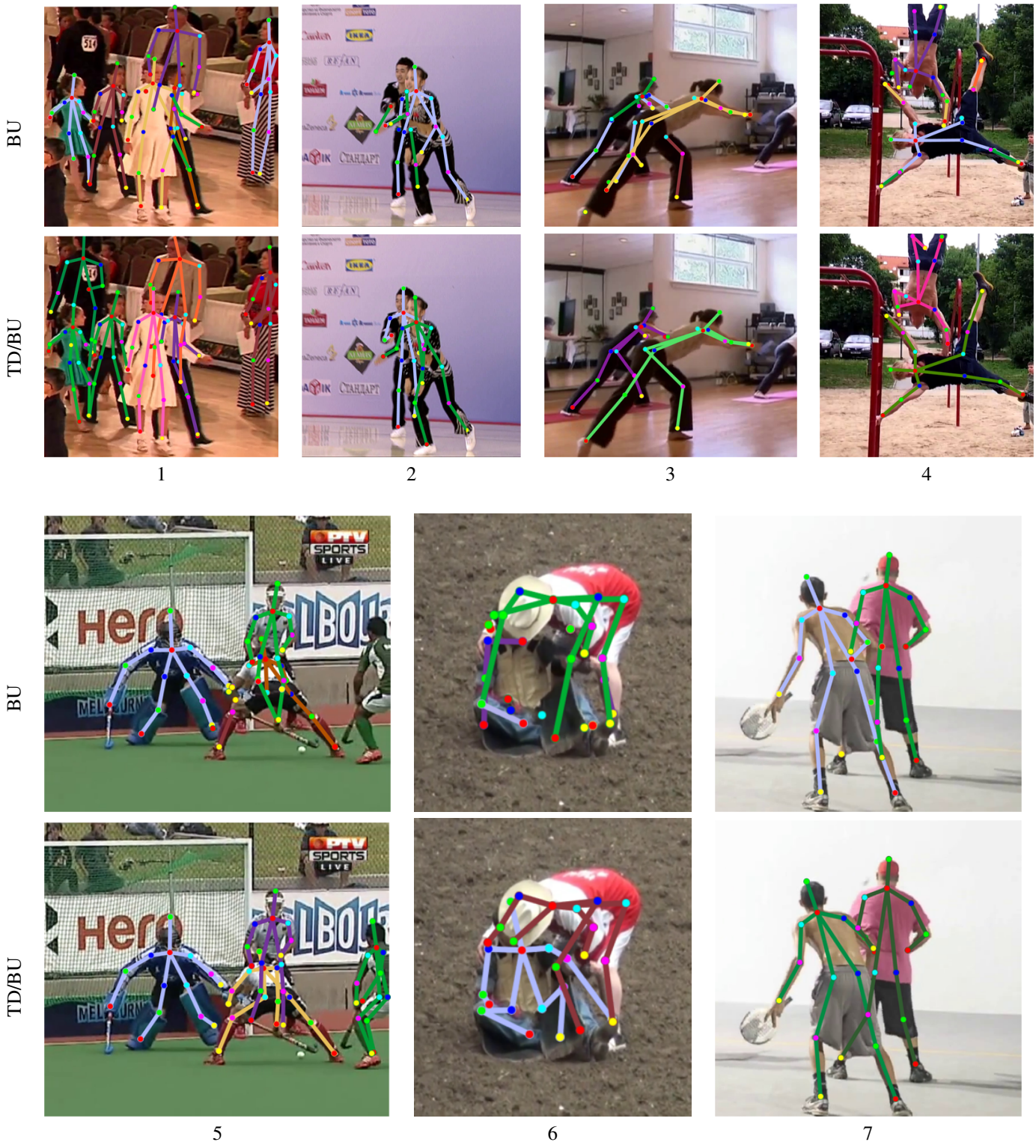

Figure 1: Qualitative comparison of single-frame based *TD/BU* and *BU-full* on MPII Multi-Person dataset.

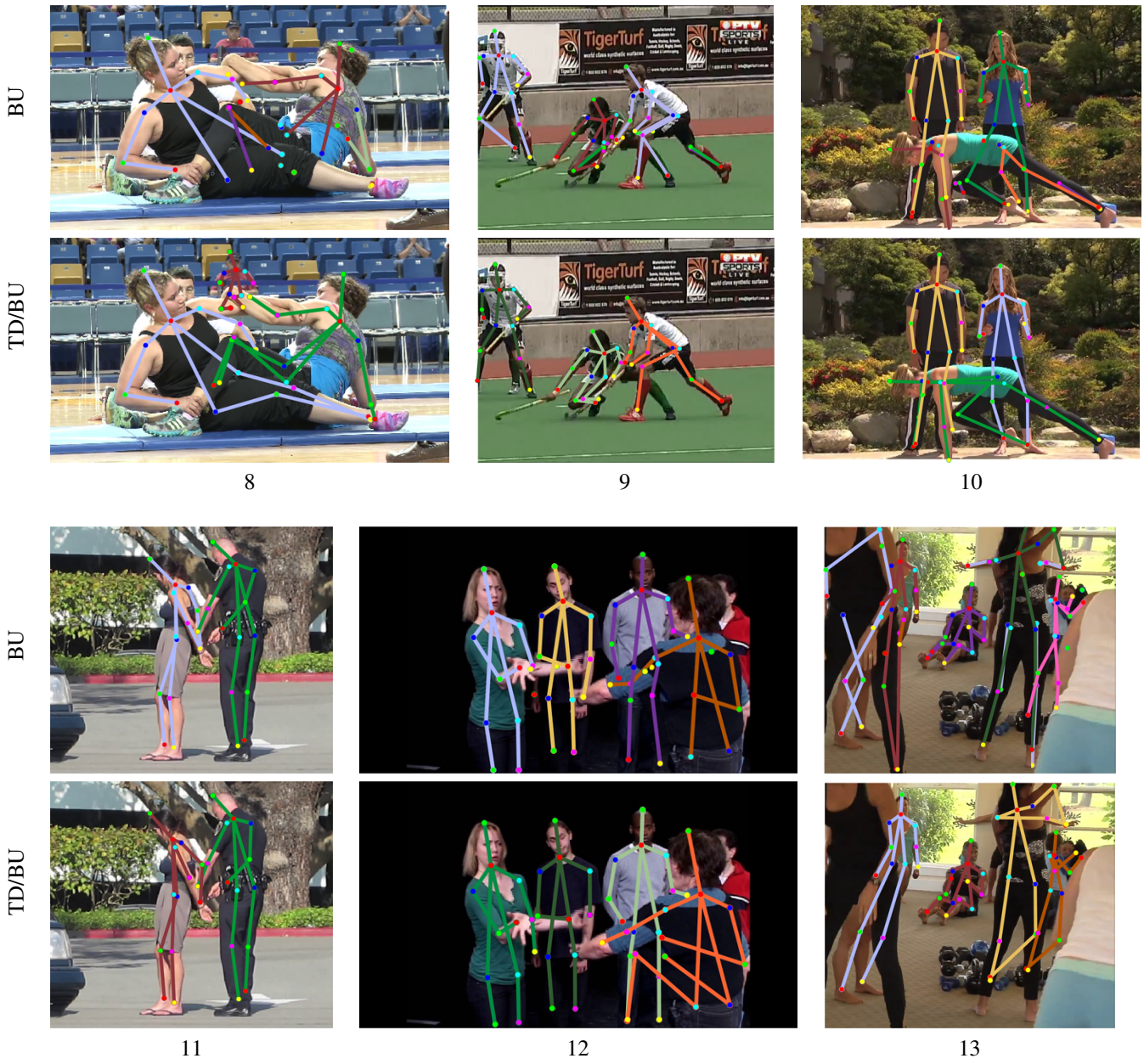

Figure 2: Successfull (8-11) and failure (12-13) pose estimation results by single-frame based *TD/BU* and comparison to *BU-full* on MPII Multi-Person dataset.

part of the same person over time based on the body part appearance. It is especially helpful in the case of fast motion where *det-distance* may fail. The combination of both geometry and appearance based features further improves the performance to 72.5%, which shows their complementarity. Finally, adding the *sift-distance* feature improves the results to 73.1%, since it copes better with the sudden changes in background and body part orientations. Overall, using a combination of temporal features in *BU-sparse+temporal* results in a 1.5% AP improvement over the single-frame *BU-sparse*. Most of the improvement can be seen on the challenging parts such as ankles (+3.9% AP) and wrists (+2.6% AP), whereas This demonstrates the advantages of the proposed approach to improve pose estimation performance using temporal information.

References

[1] X. Chen and A. Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 1

[2] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV'10*. 1

[3] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV'16*. 1

[4] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR'16*. 1