Supplementary Material of "PoseFix: Model-agnostic General Human Pose Refinement Network"

Gyeongsik Moon
Department of ECE, ASRI
Seoul National University
mks0601@snu.ac.kr

Ju Yong Chang
Department of EI
Kwangwoon University
juyong.chang@gmail.com

Kyoung Mu Lee
Department of ECE, ASRI
Seoul National University
kyoungmu@snu.ac.kr

In this supplementary material, we present more experimental results that could not be included in the main manuscript due to the lack of space.

1. Comparison with conventional end-to-end trainable multi-stage refinement

In Table 1 of the main manuscript, we compared the accuracy of the conventional end-to-end trainable multi-stage refinement model (E2E-refine) and the proposed model-agnostic refinement model (MA-refine). We tried to show the effectiveness of the proposed model-agnostic refinement model by making the number of parameters of the E2E-refine and MA-refine same.

However, as the conventional refinement requires careful model design, simply adding a refinement module which has the same network architecture with the PoseFix can result in sub-optimal performance. Therefore, we compare the accuracy of the refinement module of the state-of-the-art refinement-based method (*i.e.*, CPN [3]) and the PoseFix. The CPN consists of two parts. The first one, GlobalNet, is the baseline of the CPN. The second one, RefineNet, refines the pose estimation results of the GlobalNet. We use the GlobalNet as the pose estimation model and compare the accuracy improvement of the RefineNet and PoseFix. We trained and tested the CPN with GlobalNet only and both of the GlobalNet and RefineNet, using their released code.

Table 1 shows our PoseFix improves AP more than state-of-the-art refinement module (*i.e.*, RefineNet) by a large margin. This comparison demonstrates the benefit of the model-agnostic refinement over conventional end-to-end trainable multi-stage refinement more clearly.

2. Performance improvement of the state-of-the-art methods by PoseFix

We show performance improvement brought by the PoseFix on the PoseTrack 2018 dataset [1] in Table 2. Compared with PoseRefiner [5] which has similar approach with ours, the proposed PoseFix brings significantly large accuracy improvement.

Methods	AP	$AP_{.50}$	$AP_{.75}$	AP_M	AP_L
RefineNet [3]	69.1 (+1.8)	87.9 (+0.4)	76.6 (+2.2)	65.7 (+1.6)	75.5 (+2.2)
PoseFix (Ours)	**71.5 (+4.2)**	**88.0 (+0.5)**	**77.6 (+3.2)**	**68.0 (+3.9)**	**78.1 (+4.8)**

Table 1: AP comparison between state-of-the-art conventional end-to-end trainable multi-stage refinement model (RefineNet [3]) and the proposed model-agnostic refinement model (PoseFix) on the MS COCO [9] validation set. The number in the parenthesis denotes the AP change from the input pose (*i.e.*, GlobalNet of the CPN [3]).

Methods	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
PoseRefiner [5]	74.0 (-0.4)	76.8 (-0.1)	72.2 (+0.0)	65.4 (+0.2)	70.5 (+1.3)	69.7 (-0.3)	63.7 (+0.8)	70.6 (+0.2)
PoseFix (Ours)	**79.0 (+4.6)**	**81.6 (+4.7)**	**76.4 (+4.2)**	**69.7 (+4.5)**	**75.2 (+6.0)**	**74.3 (+4.3)**	**67.0 (+4.1)**	**75.0 (+4.6)**

Table 2: AP comparison between PoseRefiner [5] and PoseFix on the PoseTrack 2018 validation set [1]. The number in the parenthesis denotes the AP change from the input pose (*i.e.*, Simple [13]).

We additionally show the change of each error's frequency of the AE [10] and Mask R-CNN [6] on the MS COCO dataset [9] in Figure 1 and 2, respectively. As the Figures show, our PoseFix improves the performance by fixing all types of pose errors.

3. Comparison with state-of-the-art methods

We compare the performance of the PoseFix with state-of-the-art methods, which include PAFs [2], G-RMI [11], AE [10], RMPE [4], Mask R-CNN [6], CFN [7], CPN [3], Integral [12], MultiPoseNet [8], and Simple [13] on the MS COCO [9] test-dev set. All the performance are from their papers. We used Simple [13] as the input pose of the PoseFix. As they did not release the human detec-

Methods	AP	$AP_{.50}$	$AP_{.75}$	AP_M	AP_L	AR	$AR_{.50}$	$AR_{.75}$	AR_M	AR_L
RMPE [4]	61.0	82.9	68.8	57.9	66.5	-	-	-	-	-
PAFs [2]	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
Mask R-CNN [6]	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
AE [10]	65.5	86.8	72.3	60.6	72.6	70.2	89.5	76.0	64.6	78.1
Integral [12]	67.8	88.2	74.8	63.9	74.0	-	-	-	-	-
G-RMI [11]	64.9	85.5	71.3	62.3	70.0	69.7	88.7	75.5	64.4	77.1
G-RMI* [11]	68.5	87.1	75.5	65.8	73.3	73.3	90.1	79.5	68.1	80.4
MultiPoseNet [8]	69.6	86.3	76.6	65.0	76.3	73.5	88.1	79.5	68.6	80.3
CFN [7]	72.6	86.1	69.7	**78.3**	64.1	-	-	-	-	-
CPN [3]	72.1	91.4	80.0	68.7	77.2	78.5	**95.1**	85.3	74.2	84.3
CPN++ [3]	73.0	91.7	80.9	69.5	78.1	79.0	**95.1**	85.9	74.8	84.7
Simple [13]	73.7	**91.9**	81.1	70.3	80.0	79.0	-	-	-	-
Simple [13]	73.3	91.2	80.9	69.8	79.7	78.7	94.8	85.4	74.2	84.8
+ PoseFix (Ours)	**74.9**	91.2	**81.9**	71.1	**81.2**	**79.9**	94.8	**86.3**	**75.5**	**86.0**

Table 3: Comparison of APs with the state-of-the-art methods on the test-dev set. "*" means that the method involves extra data for training. "++" indicates results using ensemble.

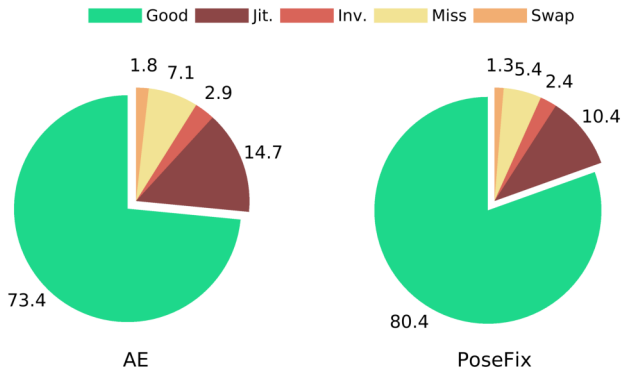

Figure 1: Frequency of each error type change when the PoseFix is applied to the AE. The frequency is calculated on the MS COCO [9] validation set.

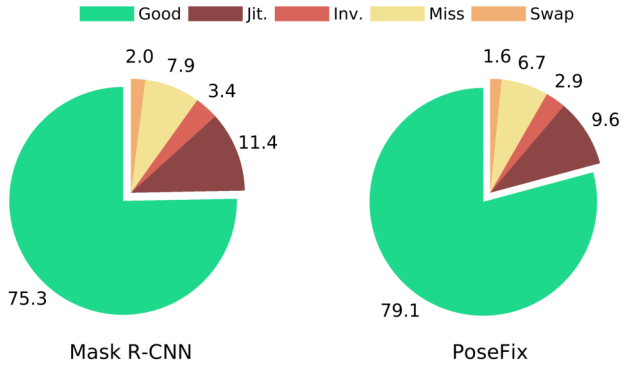

Figure 2: Frequency of each error type change when the PoseFix is applied to the Mask R-CNN. The frequency is calculated on the MS COCO [9] validation set.

tion model and result, we used our human detection model which achieves 57.2 AP for the human category on the test-dev set. The Simple [13] with our human detection model outputs slightly worse performance (73.3 AP) than the original Simple [13] (73.7 AP).

As shown in Table 3, our PoseFix outperforms all existing methods. It is noticeable that our method can achieve better performance when a new state-of-the-art method is proposed by using it as the input pose of our method. We also tried to compare the performance of the PoseFix with Fieraru *et al*. [5] which has a similar approach to ours. As they did not report the performance on the MS COCO [9] dataset, we implemented their system and tested it. However, their model outputs bad result on the COCO dataset.

4. Qualitative results

We show some qualitative results when the PoseFix is applied to Mask R-CNN [6] on the MS COCO [9] test-dev set in Figure 3 and 4, which show the input images, input poses, and refined poses. Figure 5 shows the final output of the PoseFix when the input pose is from Simple [13].

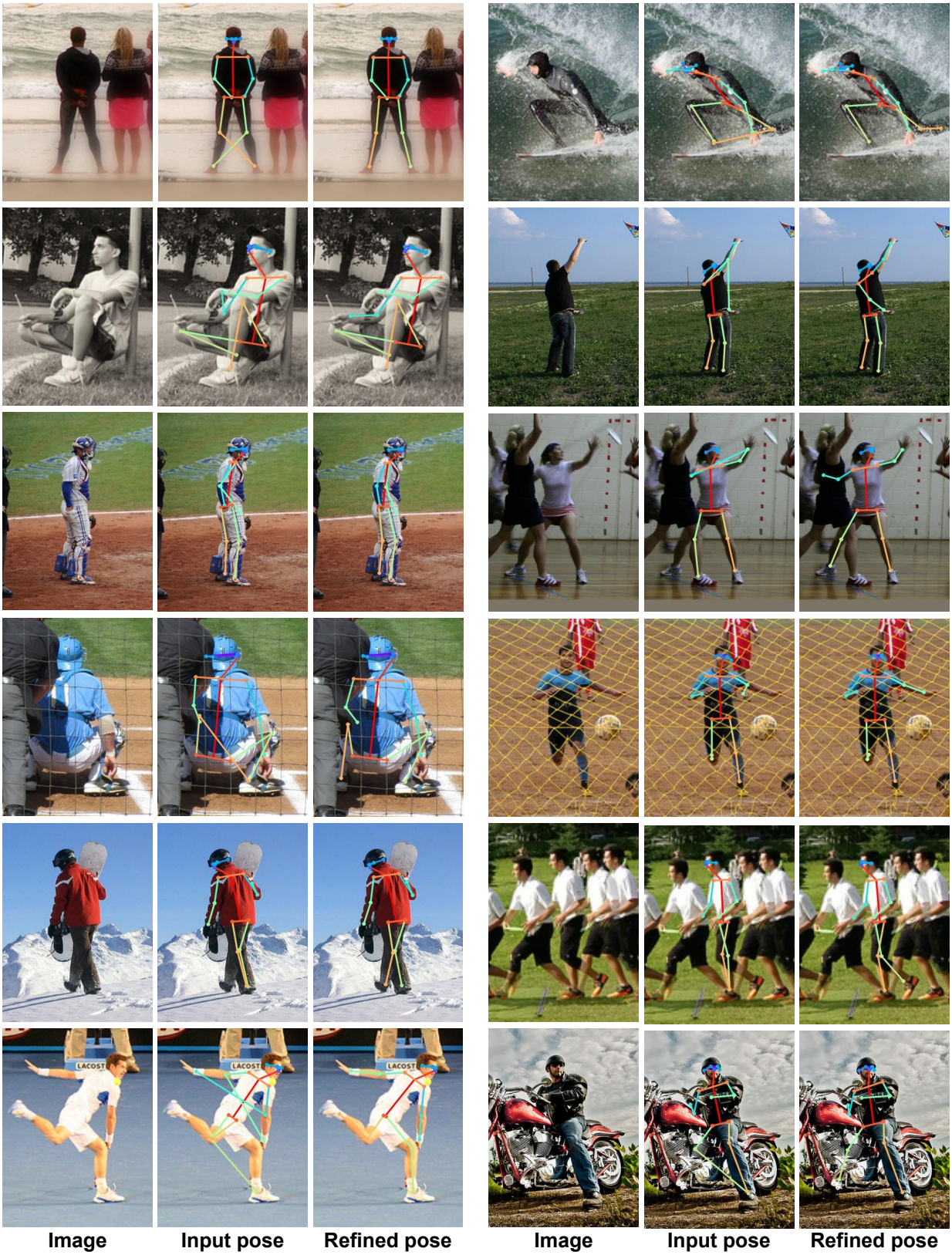

Figure 3: Qualitative results of the PoseFix on the test-dev set.

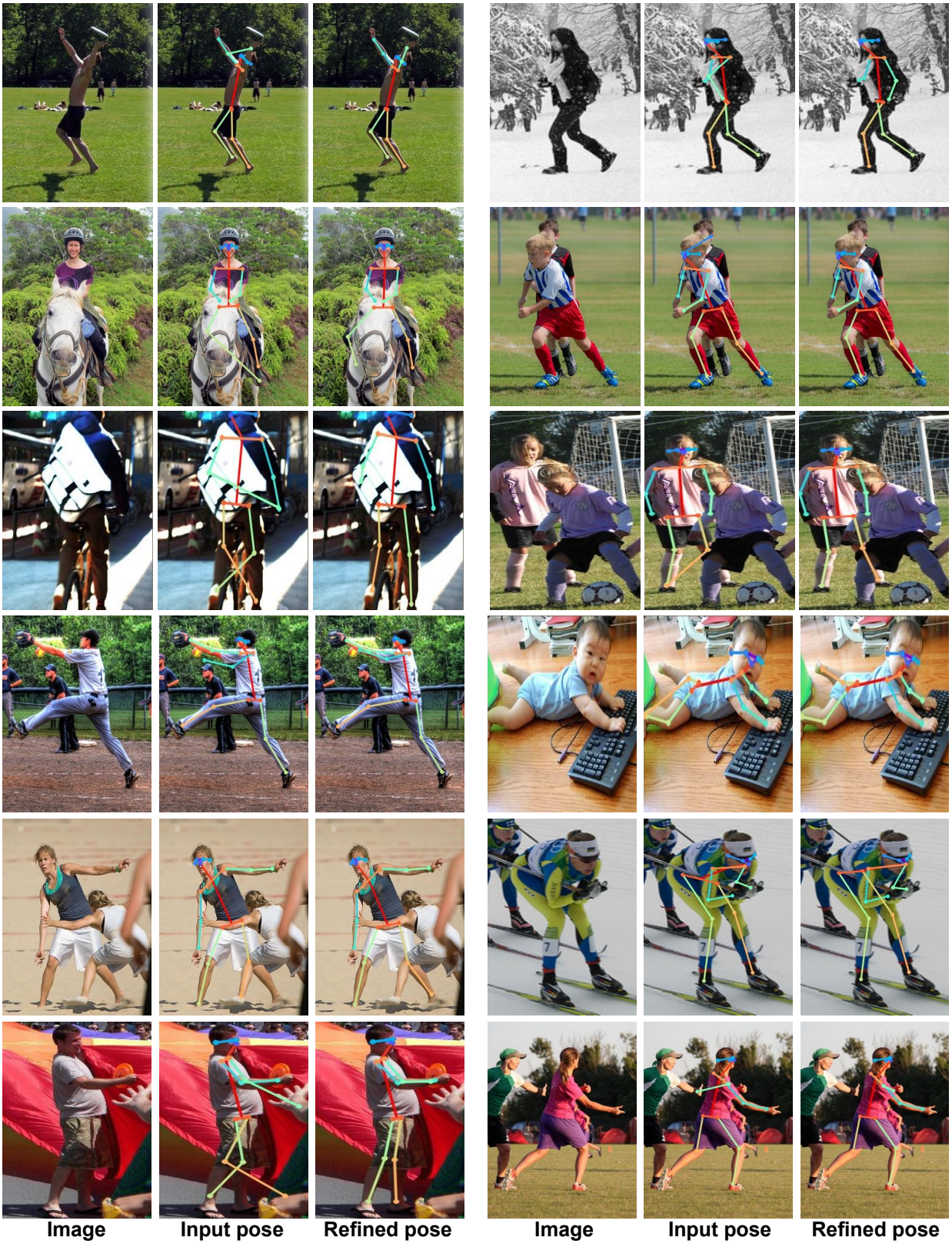

Figure 4: Qualitative results of the PoseFix on the test-dev set.

Figure 5: Qualitative results of the PoseFix on the test-dev set.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2017.

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CVPR*, 2018.

[4] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

[5] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. *CVPRW*, 2018.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[7] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017.

[8] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[10] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017.

[11] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017.

[12] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *ECCV*, 2018.

[13] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *ECCV*, 2018.