AdaptPose: Cross-Dataset Adaptation for 3D Human Pose Estimation by Learnable Motion Generation

-Supplementary Material-

Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z. Jane Wang
University of British Columbia
{mgholami, rababw, zjanew}@ece.ubc.ca {wandt, rhodin}@cs.ubc.ca

This Appendix provides ablations on the domain discriminator, 2D detections, and 3D discriminator. We also provide further qualitative results that compare AdaptPose against previous methods. Moreover, some failure cases of AdaptPose are visualized.

1. Ablation: Domain Discriminator

In this section, we provide further visualization of the performance of the domain discriminator. Figure 1 shows the distribution of camera viewpoints of the source, target, and generated datasets. Human3.6M and 3DHP are the source and target datasets, respectively. Human3.6M includes four chest-view cameras, while 3DHP includes 14 cameras that cover chest-view, top-view, and bottom-view. Figure 1 shows that the camera viewpoints of the target dataset are more diverse than those of the source dataset. We define the viewpoint by the relative rotation matrix between the subject and the camera. Figure 1 shows that AdaptPose successfully generates camera viewpoints that follow the distribution of the target camera viewpoints.

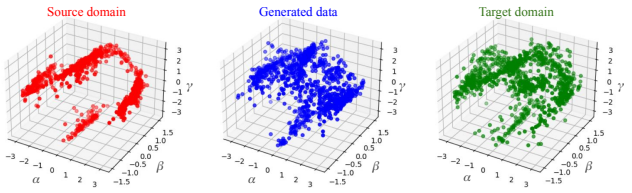

Figure 1. Camera viewpoints of the source (Human3.6M), target (3DHP), and generated data. The generated data follows the diversity and pattern of viewpoints of the target dataset. α, β, and γ are Euler-angles in radiant. The viewpoint is defined by the relative rotation matrix between the person and the camera.

2. Ablation: 2D Detections

In this section, we perform experiments on the influence of 2D detection. Using ground truth 2D for cross-dataset evaluation is the fairest comparison since most of the previous studies use the same data [3, 4]. Therefore, we used ground truth 2D in our evaluations and compared our results with previous work using the same setting. However, ground truth 2D is not always available. In this section, we employ AlphaPose [2] to obtain 2D poses of the target dataset. The model is pre-trained on MPII [1] and is not fine-tuned on the target dataset. To obtain directly comparable numbers we use the same 2D detection to evaluate 3D pose estimators of Pavllo *et al*. [5] and Gong *et al*. [3]. Table 1 provides the cross-dataset evaluation results while using ground truth 2D and detected 2D. In this experiment, the source and target datasets are Human3.6M and 3DHP, respectively. AdaptPose outperforms other methods using both ground truth 2D and detected 2D.

Table 1. Experiment on 2D detection. Source: Human3.6M, target: 3DHP. P2 is mean per joint position error (MPJPA) and P1 is MPJPA after Procrustes alignment of the estimated and ground truth 3D.

	AlphaPose 2D		GT 2D	
Method	P2	P1	P2	P1
Pavllo *et al*. [6]	86.9	127.1	66.5	96.4
PoseAug [3]	87.2	125.7	59.0	92.6
Ours	**83.4**	**120.5**	**53.6**	**77.2**

3. Ablation: 3D Discriminator

Figure 3 shows the structure of the 3D discriminator. A small perturbation ($< 10\,\mathrm{deg}$) is applied to the bone vectors of input 3D and then the perturbed version is fed to the part-wise KCS matrices of right/left arm and right/left leg. The original 3D pose is also fed to a KCS matrix. The perturbation branch enables the model to explore plausible 3D poses out of the source domain. Figure 4 shows the convergence curve of AdaptPose with and without perturbation of the source dataset. Without perturbation, the cross-dataset error of the model decreases to 77.2 in the first 9 epochs and then slightly increases in the following epochs. After

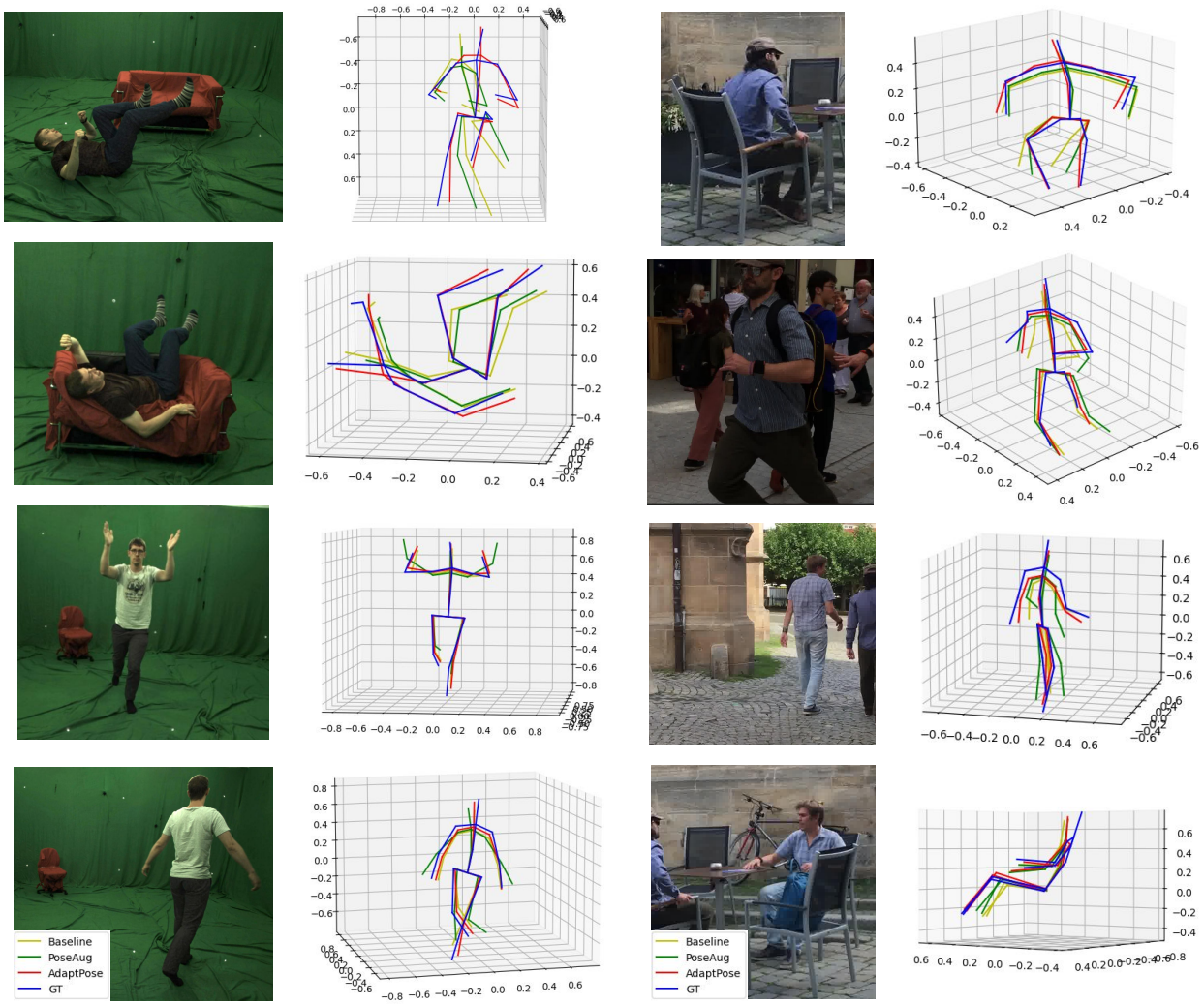

Figure 2. Further qualitative examples from 3DPW (right) and 3DHP (left) datasets. Yellow is Pavllo *et al*. [6], green is PoseAug [3], red is AdaptPose, and blue is the ground truth.

applying the perturbation, the error decreases slower and convergence of the model is more stable. Excluding the 3D discriminator increases the MPJPE by 10 mm (87 mm vs 77 mm from original Adaptpose).

4. Further Ablations

In this section we perform further ablations on 1) source conditioning over standard GAN 2) feedback losses. For the 2 experiments, the corresponding results are 78mm and 84mm, respectively, while the original AdaptPose still obtains the best performance with an MPJPE of 77.2mm (Table 5)

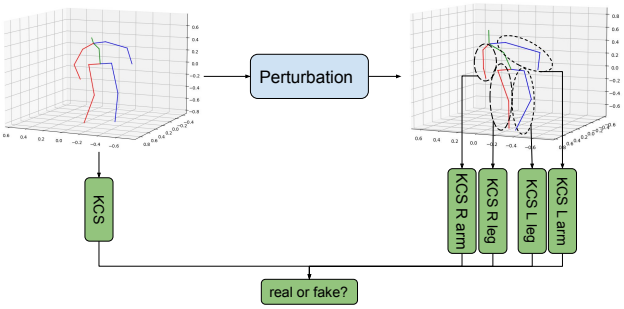

Figure 3. The 3D discriminator of AdaptPose.

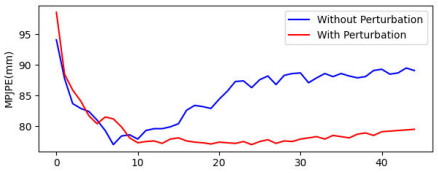

Figure 4. The evaluation error of AdaptPose while training with and without adding perturbation to the 3D discriminator.

5. Further Qualitative Results

Figure 2 provides further qualitative comparisons between AdaptPose, VideoPose3D [6], PoseAug [3], and ground truth 3D. AdaptPose significantly outperforms the previous methods. Figure 2 shows that in the case of body occlusions, AdaptPose is more accurate than other methods. One of the main limitations of our method is scale error. AdaptPose under-performs if there is a large difference between source and target body scales. Such scale ambiguity is inevitable when using only monocular views and no 3D supervision of the target domain is available. Figure 5 provides some examples of scale error for cross-dataset evaluation on 3DPW dataset. Code will be publicly available upon publication.

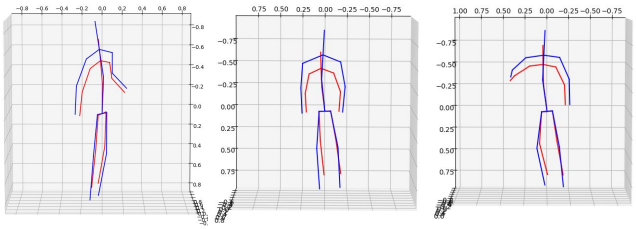

Figure 5. Scale error while performing cross-dataset evaluation on 3DPW dataset. Source: Human3.6M, target: 3DPW.

6. Run-time Comparison

AdaptPose provides fast inference when compared with competitors that conduct online adaptation. AdaptPose takes only 1 second to perform inference on the test-set of 3DPW (35k samples) while BOA [4] needs 12 hours via online adaptation.

7. Same-domain Evaluation

In this section we evaluate the performance of AdaptPose while training and testing on the same dataset (H3.6M). AdatPose improves over the pre-trained baseline model with an MPJPE of 37.6mm compared to 41.3mm for the baseline

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[3] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8575–8584, June 2021. 1, 2, 3

[4] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10472–10481, June 2021. 1, 3

[5] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 1

[6] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3