Supplement Material: Global Adaptation meets Local Generalization: Unsupervised Domain Adaptation for 3D Human Pose Estimation

Wenhao Chai[1] Zhongyu Jiang[2] Jenq-Neng Hwang[2] Gaoang Wang[1] ✉
[1] Zhejiang University [2] University of Washington

A. Experiments of detected 2D pose

GT 2D pose is used in all the experiments in the paper. Besides, we further evaluate the performance under detected 2D pose as shown in Table A. Since GPA only uses 2D box rather than specific 2D pose, the performance does not drop a lot. We reimplement all the baseline models and PoseAug in H3.6M.

B. Ablation studies on 3DPW

We conduct additional ablation studies on 3DPW dataset in Table B. We believe that the pose diversity is the limitation (*e.g.* the rare poses are still hard to estimate after PoseDA).

C. More discussion on Global Position Alignment (GPA)

We show that global position alignment module actually align the 2D pose distribution in terms of scale and root position on 2D images in Fig. B. While other works did camera view estimation [3, 4] or generation [1, 2] as an auxiliary task to address the global position adaptation problems, our method achieves alignment explicitly and directly.

D. More discussion on Local Pose Augmentation (LPA)

The most counter-intuitive conclusion in this paper is why adaptation methods perform worse than augmentation methods. In the discussion section, we include more detailed ablation studies on LPA. As shown in Fig. A, we sampled two poses from generated poses trained with a 2D discriminator and the target dataset. They have similar 2D poses but different 3D poses, which shows the reason why simply applying local pose adaptation based on a 2D pose discriminator may not have the final adaptation performance.

As Tab. C, compared with our final method, the 2D discriminator trained with 2D poses from the target dataset improves the performance from 66.07 mm to 65.46 mm in MPJPE since the discriminator makes scale and location adaptation better. However, once the 2D discriminator is removed, we can achieve a better result, 61.36 mm. The reason is that the 2D pose discriminator suppresses the diversity of generated 3D poses and makes the generator generates poses with similar 2D poses, but different 3D poses, as Fig. A shows.

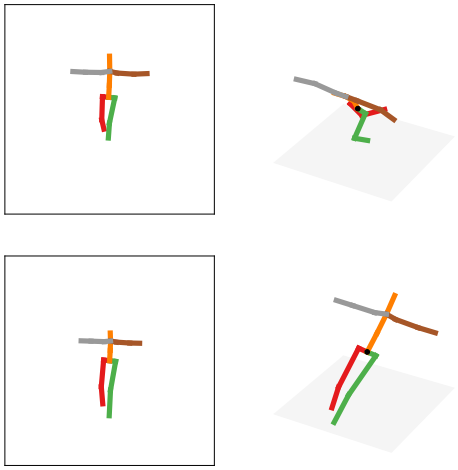

Figure A. These 2 sampled poses are from generated poses and the target dataset. They have similar 2D poses but different 3D poses, indicating that adaptation based on 2D poses may not lead to adaptation on 3D poses.

	H3.6M				3DHP			
Method	DET	CPN	HRNet	GT	DET	CPN	HRNet	GT
SemGCN	77.3	73.8	67.2	58.9	101.9	98.7	95.6	95.6
+ PoseAug	75.5 (-1.8)	73.5 (-0.3)	66.1 (-1.2)	58.0 (-0.9)	89.9 (-11.9)	89.3 (-9.4)	89.1 (-6.5)	86.1 (-9.5)
+ PoseDA	**75.0** (-2.3)	**71.9** (-1.9)	**63.8** (-3.4)	**53.9** (-5.0)	**80.3** (-21.6)	**80.3** (-18.4)	**80.9** (-14.7)	**78.3** (-17.3)
SimpleBaseline	69.2	65.1	60.3	53.6	91.1	88.8	86.4	81.2
+ PoseAug	68.4 (-0.8)	64.5 (-0.6)	59.7 (-0.6)	51.8 (-1.8)	78.7 (-12.4)	78.7 (-10.1)	76.4 (-10.1)	76.2 (-5.0)
+ PoseDA	**67.9** (-1.3)	**63.0** (-2.1)	**56.8** (-3.5)	**50.2** (-3.4)	**67.3** (-23.8)	**67.3** (-21.5)	**67.3** (-19.1)	**64.7** (-16.5)
ST-GCN	73.8	76.8	62.9	57.2	95.5	91.3	87.9	81.2
+ PoseAug	73.8 (-0.0)	72.9 (-3.9)	61.3 (-1.6)	51.2 (-6.0)	83.5 (-12.1)	**77.7** (-13.6)	76.6 (-11.3)	74.9 (-6.3)
+ PoseDA	**73.0** (-0.8)	**68.6** (-8.2)	**61.2** (-1.7)	**48.4** (-8.8)	**78.8** (-16.7)	77.8 (-13.5)	**75.1** (-12.8)	**69.5** (-11.7)
VideoPose3D	70.4	79.2	70.7	64.7	92.6	89.8	85.6	82.3
+ PoseAug	**67.1** (-3.3)	70.4 (-8.8)	63.6 (-7.1)	56.7 (-8.0)	78.3 (-14.4)	78.4 (-11.4)	73.2 (-12.4)	73.0 (-9.3)
+ PoseDA	67.4 (-3.0)	**62.2** (-17.0)	**55.7** (-15.0)	**49.9** (-14.8)	**64.6** (-28.0)	**65.4** (-24.4)	**64.5** (-21.1)	**61.4** (-20.9)

Table A. Performance comparison in MPJPE (↓) for various pose estimators on H3.6M and 3DHP datasets. DET, CPN, HRNet and GT denote 3D pose estimation model trained on several widely used different 2D pose sources, respectively. For H3.6M Exp., source: H3.6m-S1; target: H3.6m-S5, S6, S7, S8. For 3DHP Exp., source: H3.6M; target: 3DHP.

Method	MPJPE (↓)	PA-MPJPE (↓)	PCK (↑)	AUC (↑)
SimpleBaseline	153.44	100.95	59.79	28.59
+ LPA	136.64 (-16.8)	79.07 (-21.88)	63.07 (+3.28)	28.99 (+0.40)
+ GPA	131.41 (-22.03)	90.10 (-10.85)	67.53 (+7.74)	28.94 (+0.35)
+ PoseDA	**121.93** (-31.51)	**78.39** (-22.56)	**69.23** (+6.16)	**29.72** (+0.73)
VideoPose3D	101.46	61.49	80.50	41.17
+ LPA	96.72 (-4.74)	58.96 (-2.53)	83.42 (+2.92)	43.17 (+2.00)
+ GPA	92.44 (-9.02)	58.59 (-2.90)	83.94 (+3.44)	45.05 (+3.88)
+ PoseDA	**87.70** (-13.76)	**55.30** (-6.19)	**84.98** (+4.48)	**46.10** (+4.93)

Table B. Ablation study on components and pose lifting network of our method. Source: H3.6M. Target: 3DPW.

$\mathcal{G}_{pose}$	$\mathcal{D}_{3D}$	$\mathcal{D}_{2D}$	MPJPE (↓)	PCK (↑)	AUC (↑)
-	-	-	66.07	90.87	60.07
$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$	73.55	88.96	56.41
$\mathcal{S}$	$\mathcal{S}$	$\mathcal{T}$	65.46	91.27	60.03
$\mathcal{S}$	$\mathcal{S}$	-	**61.36**	**92.05**	**62.52**

Table C. The input of the pose generator $\mathcal{G}_{pose}$, the 3D pose discriminator $\mathcal{D}_{3D}$, and the 2D pose discriminator $\mathcal{D}_{2D}$ in Local Pose Augmentation (LPA) module. S, T denote poses from the source or target domain. Source: H3.6M. Target: 3DHP.

References

[1] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022. 1

[2] 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*, pages 8575–8584, 2021. 1

[3] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019. 1

[4] Zhe Wang, Daeyun Shin, and Charless C Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 1

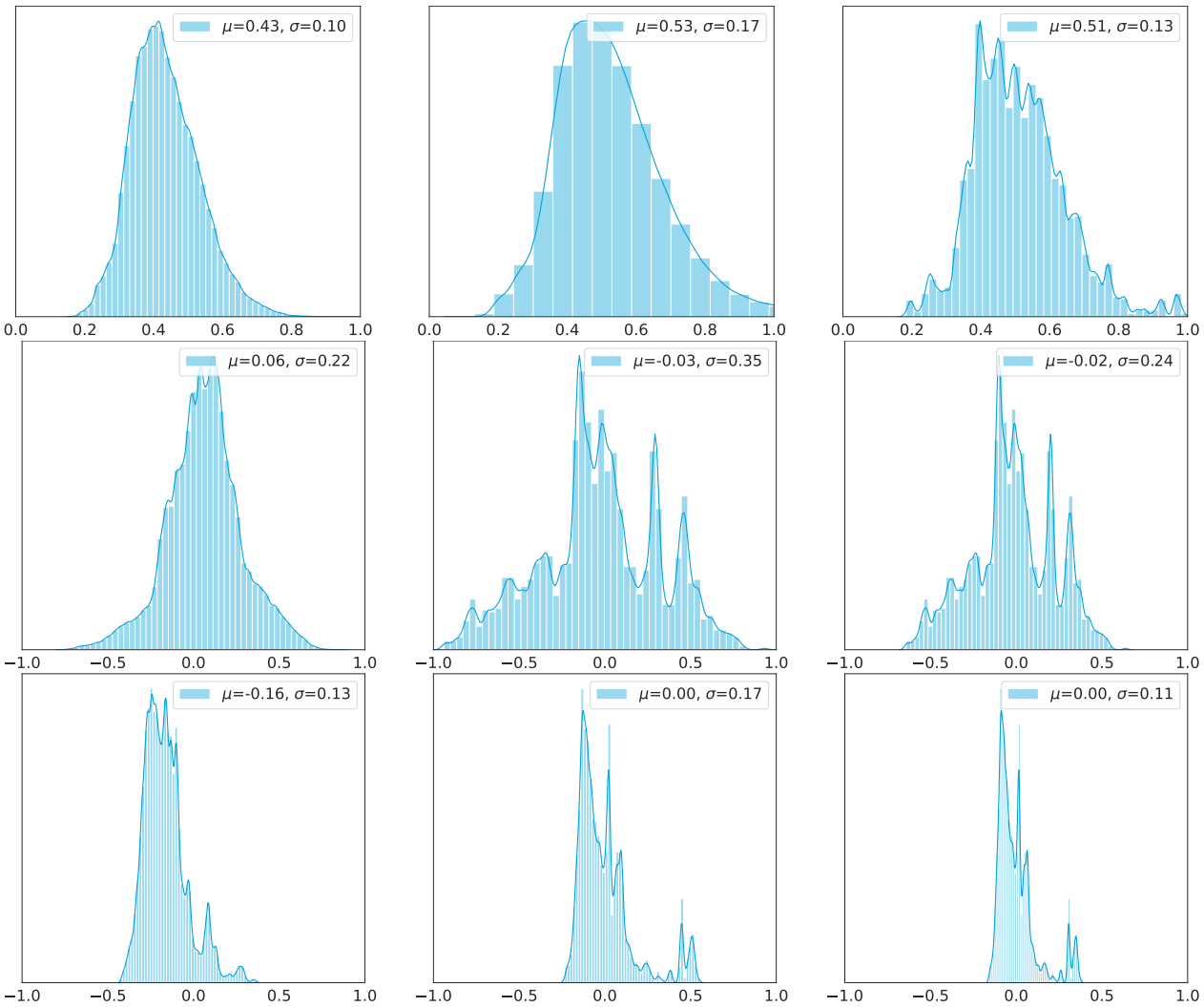

Figure B. Comparison of 2D scale (first row), root position of x-axis (second row), y-axis (third row) in source domain (left), source domain after GPA (middle), target domain (right). Source: H3.6M. Target: 3DHP.