A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation

Supplementary Material

1. Overview

The supplementary material is organized into the following sections:

- Section 2: Additional qualitative results on on cross-dataset evaluation.
- Section 3: Additional ablation study on overall framework.
- Section 4: Additional ablation of generation process.
- Section 5: Additional ablation of the number of augmentors.
- Section 6: Hyperparameters' analysis.
- Section 7: Experiments on the multi-frame setting.
- Section 8: Visualizations of distributions between source data and synthesized data.
- Section 9: Visualizations of weak-augmented and strong-augmented poses.
- Section 10: Details of Laplacian weighted similarity.

2. Extra Qualitative Results on Cross-dataset Evaluation

Fig. 1 shows extra qualitative results on cross-dataset evaluation (3DHP on the left side and 3DPW on the right side). HRNet [12] is applied as the 2D pose estimator and VPose [10] is the 2D-to-3D lifting backbone. We use **Source-only**, **PoseAug** [14], **DH-AUG** [6], **Ours**, and **Ground Truth (GT)** for qualitative comparison. It is evident that our method outperforms other baselines significantly.

3. Extra Ablation Study on the Overall Framework

Table 1. Ablation study on the overall framework with varied 2D pose predictions and 2D-to-3D lifting backbones.

Method	2D-to-3D Lifting Backbones	2D Pose Predictions			
		DET	CPN	HR	GT
Ours w/o DiffGen	SemGCN	80.1	78.8	74.0	71.8
Ours w/o DiffDis	SemGCN	79.4	77.2	73.5	70.9
Ours w/o MetaOpt	SemGCN	77.9	75.7	72.4	70.3
Ours	SemGCN	**76.5**	**74.1**	**70.7**	**68.9**
Ours w/o DiffGen	VPose	77.1	76.7	68.6	69.0
Ours w/o DiffDis	VPose	76.7	74.8	67.0	67.3
Ours w/o MetaOpt	VPose	74.5	72.4	64.9	65.5
Ours	VPose	**72.4**	**70.9**	**62.4**	**63.1**
Ours w/o DiffGen	PoseFormer	76.7	75.9	68.8	69.5
Ours w/o DiffDis	PoseFormer	76.1	75.0	67.2	67.7
Ours w/o MetaOpt	PoseFormer	74.5	72.3	65.3	66.1
Ours	PoseFormer	**72.2**	**70.5**	**62.8**	**63.4**
Ours w/o DiffGen	MixSTE	75.4	74.9	71.8	71.2
Ours w/o DiffDis	MixSTE	74.7	73.0	66.6	66.4
Ours w/o MetaOpt	MixSTE	72.8	71.3	63.4	63.5
Ours	MixSTE	**70.5**	**68.2**	**60.4**	**61.0**

Tab. 1 offers valuable insights into the impact of using different 2D pose detectors like **DET** [4], **CPN** [2], HRNet (**HR**) [12] and Ground Truth (**GT**) for 2D poses predictions, and 2D-to-3D backbones **SemGCN** [15], **VPose** [10], **PoseFormer** [16], **MixSTE** [13] on the generalization for the ablation on the overall framework. Here we use MPJPE values on the 3DHP [11] dataset for comparisons. For instance, with GT as 2D predictions and PoseFormer as the backbone, the absence of differential generation (DiffGen) leads to a 6.1mm increase, while excluding differential discrimination (DiffDis) results in a corresponding 4.3mm increase. As for the removal of meta optimization (MetaOpt), it causes a degradation of 2.7mm. The results highlight the integral role of each module in enhancing the overall framework's efficacy across diverse 2D predictions and backbones, affirming our method's robustness and versatility in improving DG for 3D HPE performance.

4. Extra Ablation Study on the Generation Process

Table 2. Ablation study on the generation process with varied 2D pose predictions and 2D-to-3D lifting backbones.

Method	2D-to-3D Lifting Backbones	2D Pose Predictions			
		DET	CPN	HR	GT
Ours w/o W-PP	SemGCN	82.0	80.4	74.7	72.9
Ours w/o S-PP	SemGCN	80.3	79.2	73.9	71.7
Ours w/o W-OG	SemGCN	77.9	75.6	72.5	69.8
Ours w/o S-OG	SemGCN	79.8	78.7	74.3	71.1
Ours	SemGCN	**76.5**	**74.1**	**70.7**	**68.9**
Ours w/o W-PP	VPose	79.5	78.8	73.4	72.6
Ours w/o S-PP	VPose	77.1	76.9	72.2	71.3
Ours w/o W-OG	VPose	73.7	73.1	65.7	65.8
Ours w/o S-OG	VPose	76.2	75.6	68.4	68.2
Ours	VPose	**72.4**	**70.9**	**62.4**	**63.1**
Ours w/o W-PP	PoseFormer	78.9	78.2	73.7	73.2
Ours w/o S-PP	PoseFormer	76.7	76.4	73.0	72.1
Ours w/o W-OG	PoseFormer	73.3	72.9	65.6	66.3
Ours w/o S-OG	PoseFormer	76.0	75.6	68.8	68.5
Ours	PoseFormer	**72.2**	**70.5**	**62.8**	**63.4**
Ours w/o W-PP	MixSTE	78.3	77.1	72.5	72.4
Ours w/o S-PP	MixSTE	75.9	74.2	66.3	66.8
Ours w/o W-OG	MixSTE	72.0	70.6	62.6	63.3
Ours w/o S-OG	MixSTE	74.8	73.4	64.5	64.7
Ours	MixSTE	**70.5**	**68.2**	**60.4**	**61.0**

Tab. 2 provides observations regarding the influence of various 2D predictions and 2D-to-3D backbones on the generalization in the context of the generator ablation study. Here we use MPJPE values on the 3DHP dataset [11] for comparisons. For instance, with GT as 2D predictions and PoseFormer as the backbone, the absence of W-PP leads to a 9.8mm increase, while excluding S-PP results in a corresponding 8.7mm increase. As for the removal of W-OG

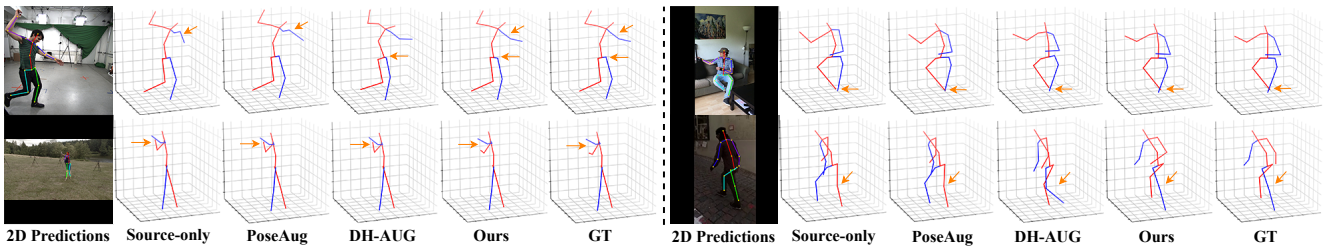

Figure 1. Extra qualitative results on cross-dataset evaluation. Left is the 3DHP dataset, and right is the 3DPW dataset.

and S-OG, they cause a degradation of 2.9mm and 5.1 respectively. The results underscore the crucial contribution of each module in augmenting the effectiveness of the generation process across various 2D predictions and backbone models. This affirms the robustness and versatility of our method in enhancing the performance of DG for 3D HPE.

5. Extra Ablation Study on the Number of Augmentors

Table 3. Ablation study on the number of augmentors with varied 2D pose predictions and 2D-to-3D lifting backbones.

Method	2D-to-3D Lifting Backbones	2D Pose Predictions			
		DET	CPN	HR	GT
WA	SemGCN	88.4	87.7	87.3	85.3
SA	SemGCN	81.7	80.5	77.2	76.9
Ours	SemGCN	**76.5**	**74.1**	**70.7**	**68.9**
WA	VPose	79.5	77.6	75.0	74.5
SA	VPose	76.1	75.9	72.7	71.0
Ours	VPose	**72.4**	**70.9**	**62.4**	**63.1**
WA	PoseFormer	80.8	78.2	75.4	75.1
SA	PoseFormer	75.0	74.6	70.5	70.7
Ours	PoseFormer	**72.2**	**70.5**	**62.8**	**63.4**
WA	MixSTE	77.3	78.0	73.5	73.2
SA	MixSTE	74.1	74.4	70.2	70.3
Ours	MixSTE	**70.5**	**68.2**	**60.4**	**61.0**

In Tab. 3, we compare our dual-augmentor framework with single-augmentor frameworks across different 2D predictions and 2D-to-3D backbones on the generalization for 3D HPE. Here we use MPJPE values on 3DHP [11] for comparisons. For instance, with GT as 2D predictions and PoseFormer as the backbone, Ours outperforms SA by 7.3mm, and it surpasses WA by 11.7mm. The results highlight the importance of dual-augmentor frameworks across diverse 2D predictions and backbones, affirming our method's robustness and versatility in improving DG for 3D HPE.

6. Hyperparameter Analysis

In Tabs. 4 through 8, we perform a parameter analysis on α_1, α_2, β, γ, and k via the two cross-dataset evaluation tasks on 3DHP and 3DPW respectively. Notably, considering the analogous functions of β_1 and β_2 in serving as trade-off parameters in the differential discrimination, we consolidate them into a single parameter denoted as β in Tab. 6. These results not only validate the appropriateness of our parameter choices but also demonstrate the stability

Table 4. Parameter analysis of α_1 on 3DHP and 3DPW

	3DHP			3DPW	
α_1	PCK ↑	AUC ↑	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓
0.35	92.2	60.0	63.8	74.5	107.9
0.40	92.7	60.4	63.5	74.1	107.5
0.45	92.9	60.7	63.1	73.5	106.9
0.50	92.9	60.7	63.1	73.2	106.6
0.55	92.9	60.5	63.6	73.8	107.0
0.60	92.5	60.4	63.8	74.0	107.3
0.65	92.1	60.1	63.9	74.4	107.7

Table 5. Parameter analysis of α_2 on 3DHP and 3DPW

	3DHP			3DPW	
α_2	PCK ↑	AUC ↑	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓
0.20	92.5	60.2	63.8	74.1	107.5
0.25	92.7	60.5	63.5	73.5	106.9
0.30	92.9	60.7	63.1	73.2	106.6
0.35	92.9	60.7	63.1	73.2	106.6
0.40	92.9	60.7	63.1	73.2	106.6
0.45	92.7	60.4	63.4	73.4	106.8
0.50	92.4	60.0	63.6	73.8	107.0

Table 6. Parameter analysis of β on 3DHP and 3DPW

	3DHP			3DPW	
β	PCK ↑	AUC ↑	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓
1	92.2	60.0	63.6	73.8	107.0
2	92.4	60.2	63.5	73.5	106.9
3	92.7	60.7	63.4	73.4	106.8
4	92.9	60.7	63.1	73.2	106.6
5	92.7	60.4	63.5	73.5	106.9
6	92.5	60.2	63.8	74.1	107.5
7	92.1	60.0	63.9	74.5	107.9

Table 7. Parameter analysis of γ on 3DHP and 3DPW

	3DHP			3DPW	
γ	PCK ↑	AUC ↑	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓
0.7	92.4	60.0	63.8	74.1	107.5
0.8	92.7	60.2	63.8	73.8	107.0
0.9	92.9	60.5	63.3	73.5	106.9
1	92.9	60.7	63.1	73.2	106.6
1.1	92.7	60.7	63.6	73.8	107.0
1.2	92.4	60.5	63.8	74.1	107.5
1.3	92.2	60.0	63.9	74.4	107.7

Table 8. Parameter analysis of k on 3DHP and 3DPW

	3DHP			3DPW	
k	PCK ↑	AUC ↑	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓
1	92.9	60.7	63.1	73.2	106.6
4	92.9	60.7	63.1	73.5	107.0
7	92.5	60.4	63.4	73.8	107.3
10	92.2	60.1	63.8	74.3	107.5

of our proposed framework across varying parameter settings.

7. Multi-frame Domain Generalization for 3D Human Pose Estimation

While our proposed framework is specifically tailored for single-frame tasks in 3D Human Pose Estimation (HPE), such as PoseAug [5], CEE-Net [8], and PoseDA [1], there are alternative approaches that extend their considerations to multi-frame settings through temporal-based techniques, exemplified by PoseAug-V [14] and DH-AUG [6]. In this section, we present a comparative analysis of our method against these temporal-based multi-frame approaches within the 27-frame setting. Notably, our method remains consistent with the single-frame setting, except for variations in input size that contain 27 frames.

Table 9. Cross-dataset evaluation with MPJPE (↓) and PA-MPJPE (↓) on 3DHP and 3DPW (27-frame).

		3DHP		3DPW	
Method	DG	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
VPose [10]	×	96.4	66.5	103.3	63.6
VPose + PoseAug-V [14]	✓	86.5	61.0	91.1	54.3
VPose + DH-AUG [6]	✓	80.9	58.6	87.3	52.5
VPose + Ours	✓	**79.7**	**57.9**	**85.2**	**51.6**
PoseFormer [16]	×	93.3	66.7	118.5	73.4
PoseFormer + PoseAug-V [14]	✓	82.9	63.1	108.3	64.8
PoseFormer + DH-AUG [6]	✓	75.4	61.8	104.4	62.1
PoseFormer + Ours	✓	**74.1**	**61.0**	**102.7**	**61.3**

In Tab. 9, our single-frame approach demonstrates a significant performance advantage over temporal-based methods like PoseAug-V and DH-AUG, despite being a single-frame method without relying on temporal-based techniques. This result underscores the effectiveness and superiority of our proposed method.

8. Visualizations of Distributions between Source Data and Synthesized Data

In Fig. 2 and Fig. 3, we present a qualitative visualization of our augmented data. We utilize S6 from Human3.6M [7] as the source domain (depicted by red dots in both figures). Subsequently, we generate 3,000 data points using both the weak augmentor (depicted by green dots on the left side) and the strong augmentor (depicted by blue dots on the right side).

In these two figures, it is evident that data points generated by the weak augmentor closely resemble the distributions of the source, whereas those generated by the strong augmentor exhibit a notable deviation from the source distributions, thereby demonstrating the effectiveness of these two augmentors.

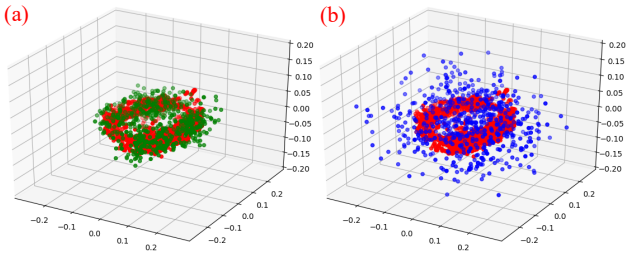

Figure 2. Visualizations of Distributions. Here we choose the right hip joint in S6 set of Human3.6M for illustration. Red dots in both figures are the source data's distributions. Green dots on the left side are 3K points generated by the weak-augmentor, and blue dots on the right side are 3K points generated by the strong-augmentor.

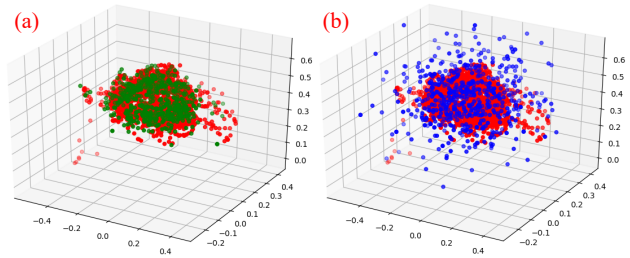

Figure 3. Visualizations of Distributions. Here we choose the right shoulder joint in S6 set of Human3.6M for illustration. Red dots in both figures are the source data's distributions. Green dots on the left side are 3K points generated by the weak-augmentor, and blue dots on the right side are 3K points generated by the strong-augmentor.

9. Visualizations of Weak-augmented and Strong-augmented Poses

In Fig. 4, we provide several examples to visualize the effectiveness of our dual-augmentor system.

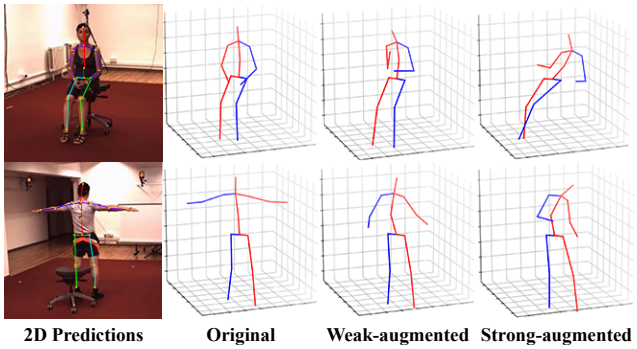

Figure 4. Visualizations of Augmentations.

As depicted in this figure, the poses generated by the

weak augmentor exhibit similarity to the original source poses. In contrast, poses generated by the strong augmentor differ significantly from the source poses. Nevertheless, these strong-augmented poses remain reasonable in their respective scenarios and align with the human model.

10. Details of Laplacian Weighted Similarity

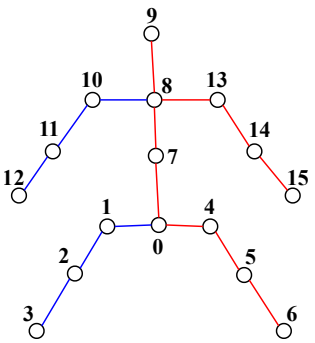

Figure 5. Visualizations of 16-joint Human Body Model.

In Fig. 5, the 16-joint human body model is provided, which follows the settings of previous works [1, 3, 6, 8, 9, 14]. Based on this model, it is straightforward to obtain the two matrices, the adjacency matrix A:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & \dots \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{16\times16}, \quad (1)$$

and the degree matrix D:

$$D = \begin{bmatrix} 3 & & & & & \\ & 2 & & & & \\ & & 2 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & \end{bmatrix}_{16\times16} \quad (2)$$

References

[1] Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14655–14665, 2023. 3, 4

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 1

[3] 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. 4

[4] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 1

[5] 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. 3

[6] Linzhi Huang, Jiahao Liang, and Weihong Deng. Dh-aug: Dh forward kinematics model driven augmentation for 3d human pose estimation. In *European Conference on Computer Vision*, pages 436–453. Springer, 2022. 1, 3, 4

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 3

[8] Haolun Li and Chi-Man Pun. Cee-net: Complementary end-to-end network for 3d human pose generation and estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1305–1313, 2023. 3, 4

[9] Hanbing Liu, Jun-Yan He, Zhi-Qi Cheng, Wangmeng Xiang, Qize Yang, Wenhao Chai, Gaoang Wang, Xu Bao, Bin Luo, Yifeng Geng, et al. Posynda: Multi-hypothesis pose synthesis domain adaptation for robust 3d human pose estimation. In *Proceedings of the ACM International Conference on Multimedia*, 2023. 4

[10] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 1, 3

[11] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 1, 2

[12] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1

[13] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022. 1

[14] Jianfeng Zhang, Kehong Gong, Xinchao Wang, and Jiashi Feng. Learning to augment poses for 3d human pose estimation in images and videos. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence, 45(8):10012–10026, 2023. 1, 3, 4

[15] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019. 1

[16] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 1, 3