

Supplementary Materials: APP: Adaptive Pose Pooling for 3D Human Pose Estimation from Videos

Anonymous Authors

1 OVERVIEW

Our supplementary materials contain three parts:

- More implementation details of the 2D pose detectors used in our experiments, including how to obtain multi-level featuremaps and changes in the structure.
- More qualitative results of our proposed method on Human3.6M dataset.

2 MORE IMPLEMENTATION DETAILS

In this section, we discuss how to extract multi-level featuremaps more. Unlike the vanilla CPN [1] version, each featuremap computing branch is upsampled into the same size (height and width) using `nn.Upsample`. We change the upsample operation to the downsample operation across the channel dimension. To get the same size of featuremaps extracted by HRNet, the channel dimension of each level of featuremap originally is 256. Specifically, we downsample it to 32/64/128/256 by averaging operations. However, this downsampling causes information loss. As shown in 1, we observe that when MotionAGFormer-B [3] is selected, there is a 0.2mm increase for our method (from 38.4mm to 38.6mm), which is worse than the vanilla version.

Table 1: In comparison with baselines on Human3.6M, we choose different lifting models, 2D pose detectors and 2D poses.

Method	T	2D Pose Detector	2D Pose	MPJPE↓	P-MPJPE↓
Transformer [5] NeurIPS'17	243	-	SH	39.2	32.8
APP w. Transformer (Ours)	243	CPN	SH	39.0	32.7
APP w. Transformer (Ours)	243	HRNet-w32	SH	39.0	32.7
MixSTE [6] CVPR'22	243	-	CPN	40.9	32.6
APP w. MixSTE (Ours)	243	CPN	CPN	39.8	32.1
APP w. MixSTE (Ours)	243	HRNet-w32	CPN	39.9	32.1
MotionBERT [7] (finetune) ICCV'23	243	-	SH	37.5	-
APP w. MotionBERT (finetune) (Ours)	243	CPN	SH	37.0	31.5
APP w. MotionBERT (finetune) (Ours)	243	HRNet-w32	SH	37.0	31.6
MotionAGFormer-B [3] WACV'24	243	-	CPN	39.2	31.5
APP w. MotionAGFormer-B (Ours)	243	CPN	CPN	39.0	31.3
APP w. MotionAGFormer-B (Ours)	243	HRNet-w32	CPN	38.9	31.3
MotionAGFormer-B [3] WACV'24	243	-	SH	38.4	32.6
APP w. MotionAGFormer-B (Ours)	243	CPN	SH	38.6	32.5
APP w. MotionAGFormer-B (Ours)	243	HRNet-w32	SH	38.4	32.3
HoT [2] CVPR'24	243	-	SH	39.8	33.5
APP w. HoT (Ours)	243	CPN	SH	39.7	33.4
APP w. HoT (Ours)	243	HRNet-w32	SH	39.7	33.4

3 MORE QUALITATIVE RESULTS

Transformer [5], MixSTE [6], MotionBERT [7], MotionAGFormer-B [3], and HoT [2] are selected to generate the pose features P . However, this time, we chose the CPN as a 2D pose detector to extract multi-level featuremaps. The results on the Human3.6M are illustrated in Table 1. The improvement of MPJPE and P-MPJPE is quite similar when using HRNet-w32. As mentioned in Sec 2, there is a slight performance degradation (from 38.4 to 38.6mm) in terms of MPJPE.

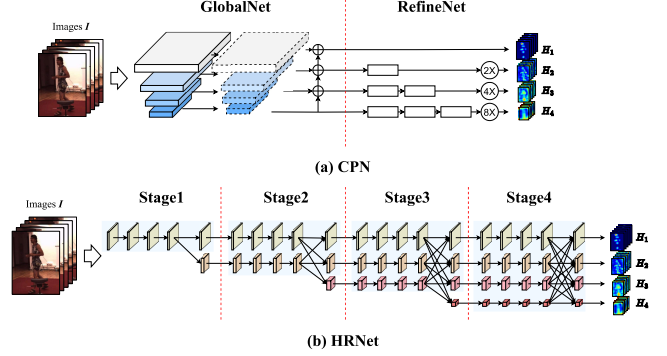


Figure 1: The overview of the CPN [1] and HRNet [4], which are 2D pose detectors used to extract multi-level featuremaps. (a) shows the structure of the CPN, we made some minor changes to ensure the multi-level featuremaps' size was consistent with HRNet-w32. (b) is the framework of HRNet, which consists of four stages.

REFERENCES

[1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7103–7112.

[2] Wenhao Li, Mengyuan Liu, Hong Liu, Pichao Wang, Jialun Cai, and Nicu Sebe. 2024. Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[3] Sorous Mehraban, Vida Adeli, and Babak Taati. 2024. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6920–6930.

[4] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5693–5703.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[6] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. 2022. 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*. 13232–13242.

[7] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15085–15099.