Supplementary Material for Human Pose as Compositional Tokens

Zigang Geng[1,3] , Chunyu Wang[3*], Yixuan Wei[2,3], Ze Liu[1,3], Houqiang Li[1], Han Hu[3*]
[1]University of Science and Technology of China [2]Tsinghua University [3]Microsoft Research Asia
https://sites.google.com/view/pctpose

Table 1. Results on the MPII [1] test set (PCKh@0.5). '†' means using extra training datasets. '‡' means using larger image size.

Method	Hea.	Sho.	Elb.	Wri.	Hip.	Kne.	Ank.	Mean
Xiao *et al.* [9]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
Tang *et al.* [7]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Sun *et al.* [6, 8]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Cai *et al.* [4]	98.5	97.3	93.9	89.9	92.0	90.6	86.8	93.0
Bulat *et al.* [3]†	98.8	97.5	94.4	91.2	93.2	92.2	89.3	94.1
Bin *et al.* [2]‡	98.9	97.6	94.6	91.2	93.1	92.7	89.1	94.1
Our (Swin-Base)	98.7	97.5	94.2	90.6	92.9	92.1	88.7	93.8
Our (Swin-Large)	98.9	97.8	94.8	91.1	93.6	93.0	89.7	94.3

1. Results on the MPII Test Set

We provide the results on the MPII [1] test set. Table 1 shows the results on the MPII test set. Our approach outperforms the other methods, even those that utilize extra training datasets or larger image sizes.

2. Results on the H36M under occlusion

To evaluate the performance of PCT under different occlusion conditions, we artificially occlude the images in the h36m test set by either cropping or masking them. Table 2 reports the results of the models with and without PCT. It reveals that the advantages of PCT become more apparent as the level of occlusion increases.

3. More visual illustrations for the sub-structures.

Figure 1 provides more examples of sub-structures represented by our compositional tokens. We use 34 tokens to represent a human pose. We statistically find that almost two tokens are responsible for a sub-structure consisting of a body joint and its related joints, one is for major changes, and the other is for minor jitters. We select some of them to show.

Table 2. Results on the H36M [5] test set (MPJPE mm) under different occlusion conditions.

Mask Ratio	0.0	0.2	0.4	0.6	0.8
w/o PCT	53.9	66.6	94.4	157.6	268.7
PCT	50.8	63.4	88.2	145.5	287.9
Crop Ratio	0.0	0.1	0.2	0.3	0.4
w/o PCT	53.9	53.9	54.8	60.0	84.1
PCT	50.8	50.9	51.2	55.0	74.8

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 1

[2] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. In *ECCV*, pages 606–622, 2020. 1

[3] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. In *International Conference on Automatic Face and Gesture Recognition, FG 2020*, pages 8–15. IEEE, 2020. 1

[4] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning delicate local representations for multi-person pose estimation. In *ECCV*, pages 455–472, 2020. 1

[5] 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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 1

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1

[7] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *ECCV*, pages 197–214, 2018. 1

[8] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep

[*] Equal Advising

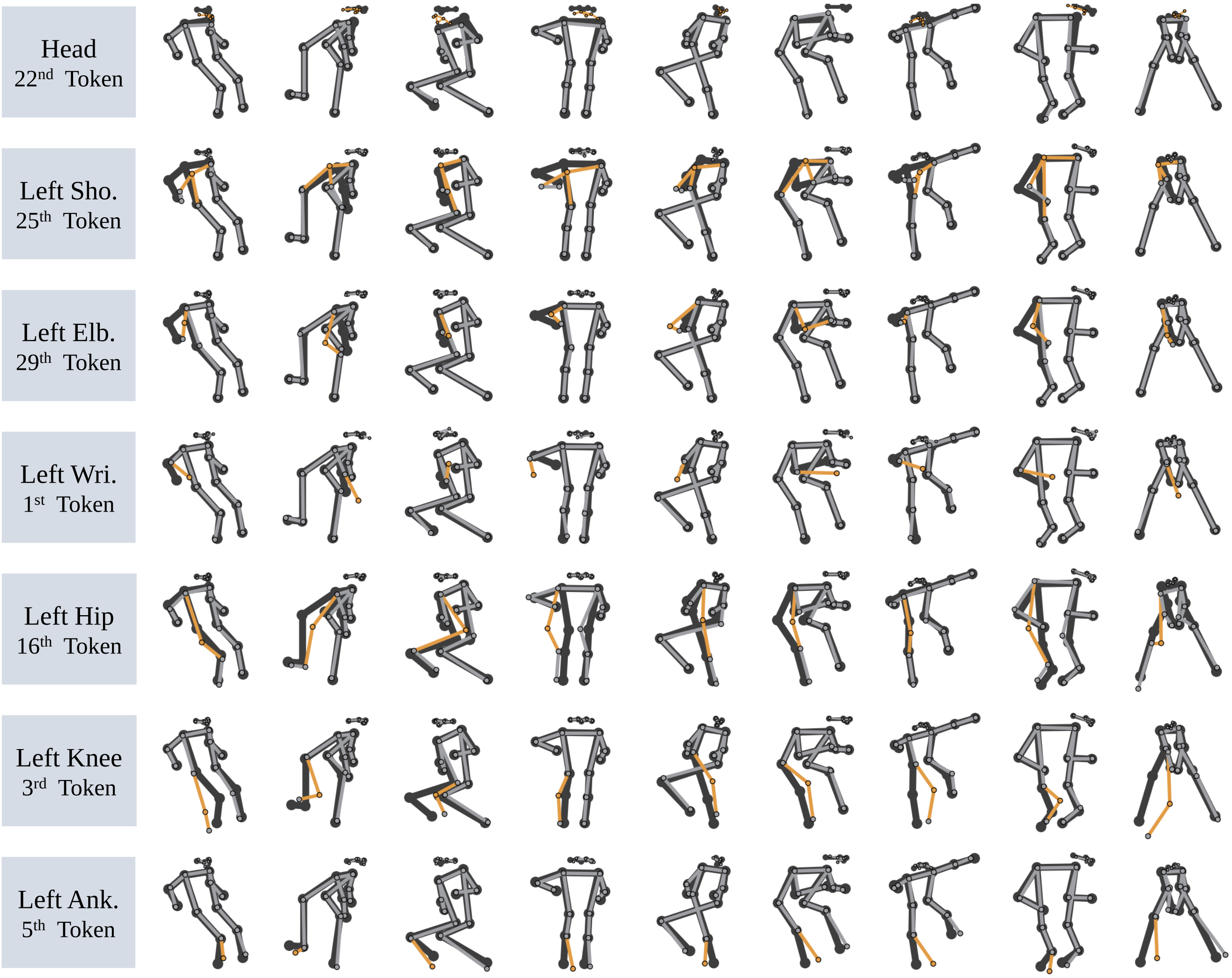

Figure 1. Each token is learned to represent a sub-structure. In each row, we show that if we change the stage of one token to different values, it consistently changes the same sub-structure highlighted by orange. The black poses are before changing.

high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[9] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 472–487, 2018. 1