MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation –Supplemental Material–

Wenhao Li[1] Hong Liu[1,*] Hao Tang[2] Pichao Wang[3] Luc Van Gool[2]
[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
[2]Computer Vision Lab, ETH Zurich [3]Alibaba Group
{wenhaoli,hongliu}@pku.edu.cn
{hao.tang,vangool}@vision.ee.ethz.ch pichao.wang@alibaba-inc.com

This supplementary material contains the following details: (1) A brief description of multi-head cross-attention. (2) Additional quantitative results. (3) Additional ablation studies. (4) Additional visualization results.

1. Multi-Head Cross-Attention

In Sec. 3.1 of our main manuscript, we give a brief description of the multi-head self-attention (MSA) block. Given the inputs $x \in \mathbb{R}^{n\times d}$, they are first inearly mapped to queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{n\times d}$, and values $V \in \mathbb{R}^{n\times d}$. Then, the scaled dot-product attention can be computed by:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (1)$$

In this section, we further define the multi-head cross-attention (MCA) among three tensors, x, y, and z. The inputs $x \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^{n\times d}$, and $z \in \mathbb{R}^{n\times d}$ are linearly mapped to queries $Q_x \in \mathbb{R}^{n\times d}$, keys $K_y \in \mathbb{R}^{n\times d}$, and values $V_z \in \mathbb{R}^{n\times d}$, respectively. The scaled dot-product attention in the MCA can be computed by:

$$\text{Attention}_{\text{cross}}(Q_x, K_y, V_z) = \text{Softmax}\left(\frac{Q_x K_y^T}{\sqrt{d}}\right)V_z. \quad (2)$$

The common configuration of MCA uses the same input between keys and values [2, 9, 15], *i.e.*, the inputs $x \neq y = z$. Instead, we adopt a more efficient strategy by using different inputs, *i.e.*, the inputs $x \neq y \neq z$.

2. Additional Quantitative Results

Table 1 shows the results of our proposed MHFormer on Human3.6M under Protocol 2. The input 2D poses are estimated by CPN [4]. Without bells and whistles, our MHFormer achieves promising results that outperform the state-of-the-art approaches.

Several methods [1, 13, 18] adopt a pose refinement module, which is first proposed by ST-GCN [1], to further improve the estimation accuracy. Following [1], we adopt the refine module and the results are shown in Table 2. It can be seen that our method can use the refine module to improve the performance, achieving an error of 42.4 mm in MPJPE which surpasses all other approaches by a large margin.

3. Additional Ablation Studies

Effect of Model Components. Here, we give more details about how to build the different variants of MHFormer in Table 7 of our main manuscript:

- Baseline: The baseline model contains 3 layers for standard Transformer encoder (same architecture as ViT [5]).
- SHR-CHI: We remove the MHG module. SHR-CHI contains L_2=2 SHR and L_3=1 CHI layers.
- MHG-SHR: We replace the CHI layers in MHFormer with SHR layers. MHG-SHR contains L_1=4 MHG and L_3=3 SHR layers.
- MHG-CHI: We replace the SHR layers in MHFormer with CHI layers. SHR-CHI contains L_1=4 MHG and L_3=3 CHI layers.
- MHFormer *: The MHG in MHFormer is simply built upon several parallel Transformer encoders.
- MHFormer: Our proposed method that contains L_1=4 MHG, L_2=2 SHR, and L_3=1 CHI layers. Please refer to Figure 3 in our main manuscript.

Impact of Configurations in MH-CA. As mentioned in Sec. 3.5 of our main manuscript, the common configuration of MCA uses the same input between keys and values [2, 9, 15], which will result in more blocks. We adopt a more efficient configuration by using different inputs among queries, keys, and values. The performance and computational complexity of these two configurations are given in

*Corresponding author: hongliu@pku.edu.cn. This work is supported by National Key R&D Program of China (No. 2020AAA0108904), Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340).

Table 1. Quantitative comparison with the state-of-the-art methods on Human3.6M under Protocol 2. (†) - uses temporal information. **Blod**: best; Underlined: second best.

Method	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
SimpleBaseline (ICCV'17 [10]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang *et al.* (AAAI'18 [6]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
PoseAug (CVPR'21) [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
SGNN (ICCV'21) [16]	33.9	37.2	36.8	38.1	38.7	43.5	37.8	35.0	47.2	53.8	40.7	38.3	41.8	30.1	31.4	39.0
ST-GCN (ICCV'19) [1] (†)	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
VPose *et al.* (CVPR'19) [12] (†)	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Xu *et al.* (CVPR'20) [14] (†)	**31.0**	**34.8**	34.7	34.4	36.2	43.9	31.6	33.5	**42.3**	49.0	37.1	33.0	39.1	26.9	31.9	36.2
Liu *et al.* (CVPR'20) [8] (†)	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	**32.4**	37.0	25.2	27.2	35.6
UGCN (ECCV'20) [13] (†)	32.9	35.2	35.6	34.4	36.4	42.7	**31.2**	32.5	45.6	50.2	37.3	32.8	36.3	26.0	**23.9**	35.5
Anatomy3D (TCSVT'21) [3] (†)	32.6	35.1	32.8	35.4	36.3	40.4	32.4	32.3	42.7	49.0	36.8	**32.4**	36.0	24.9	26.5	35.0
PoseFormer (ICCV'21) [17] (†)	32.5	**34.8**	**32.6**	34.6	**35.3**	**39.5**	32.1	**32.0**	42.8	**48.5**	**34.8**	**32.4**	35.3	24.5	26.0	34.6
MHFormer (Ours) (†)	31.5	34.9	32.8	**33.6**	**35.3**	39.6	32.0	32.2	43.5	48.7	36.4	32.6	**34.3**	**23.9**	25.1	**34.4**

Table 2. Quantitative comparison on Human3.6M under MPJPE. **Blod**: best; Underlined: second best.

Method	Refine module	MPJPE (mm)
MGCN (ICCV'21) [18]	✓	49.4
ST-GCN (ICCV'19) [1]	✓	48.8
UGCN (ECCV'20) [13]		45.6
UGCN (ECCV'20) [13]	✓	44.5
MHFormer (Ours)		43.0
MHFormer (Ours)	✓	**42.4**

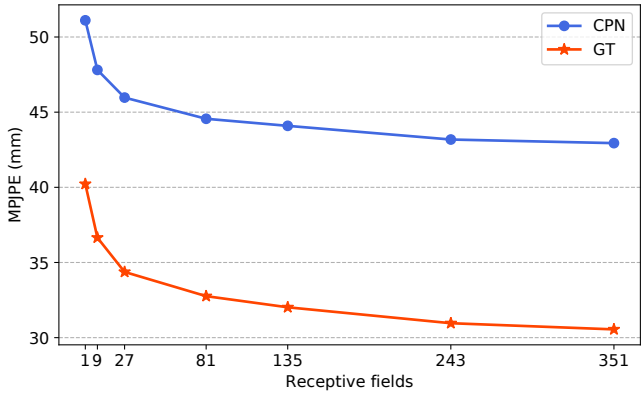
Table 3. Ablation study on different configurations of MH-CA on Human3.6M under MPJPE. Here, * means using the same input between keys and values in MH-CA.

Method	Params (M)	FLOPs (G)	MPJPE (mm)
MH-CA *	22.07	1.21	46.1
MH-CA	18.92	1.03	**45.9**

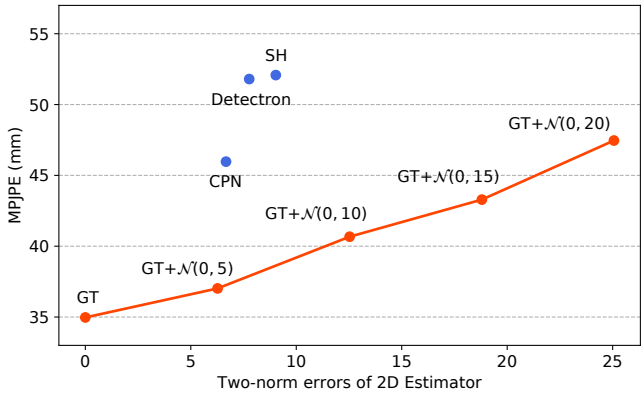
Table 3. We can see that using the same input between keys and values in MH-CA (MH-CA *) requires more parameters and FLOPs but cannot bring further performance gains. It illustrates the effectiveness of our efficient strategy in MCA.

Impact of Receptive Fields. For the video-based 3D human pose estimation task, the number of receptive fields directly influences the estimation results. Figure 1 (a) shows the results of our model with different receptive fields (between 1 and 351) on Human3.6M. Increasing the receptive field can improve the result under both CPN and GT 2D pose inputs, which demonstrates the great power of our method in long-range dependency modeling with a long input sequence.

Impact of 2D Detections. To show the effectiveness of our method on different 2D pose detectors, we carry out experiments with the detections from Stack Hourglass (SH) [11], Detectron [12], and CPN [4]. In addition, to evaluate the robustness of our method to various levels of noise, we also conduct experiments on 2D ground truth plus different levels of additive Gaussian noise. The results are shown in Figure 1 (b). It can be observed that the curve has a nearly linear relationship between MPJPE of 3D poses and two-norm errors of 2D poses. These experiments validate both the effectiveness and robustness of our proposed method.



(a) Different receptive fields under MPJPE.



(b) Different 2D detections under MPJPE.

Figure 1. **(a)** Ablation studies on different receptive fields of our method on Human3.6M under MPJPE metric. **(b)** The effect of 2D detections on Human3.6M under MPJPE. Here, $\mathcal{N}(0, \sigma^2)$ represents the Gaussian noise with mean zero and σ is the standard deviation. (CPN) - Cascaded Pyramid Network; (SH) Stack Hourglass; (GT) - 2D ground truth.

4. Additional Visualization Results

3D Reconstruction Visualization. Figure 2 and Figure 3 show qualitative results of our method on Human3.6M dataset, MPI-INF-3DHP dataset, and challenging in-the-wild videos. Moreover, Figure 4 shows the qualitative comparison

with the baseline method and the previous state-of-the-art method (PoseFormer [17]) on some wild videos. It can be seen that our method can produce more accurate and reasonable 3D poses, especially when the human action is complex and rare.

Hypothesis Visualization. For visualization purposes, we add additional regression layers and finetune our model to output intermediate hypotheses. Figure 5 shows the visualization results of intermediate 3D pose hypotheses generated by our proposed method. We can see that our MHFormer can generate different plausible 3D pose solutions, especially for ambiguous body parts with depth ambiguity, self-occlusion, and 2D detector uncertainty.

Attention Visualization. Visualization results of the multi-head attention maps of the first layers from the Multi-Hypothesis Generation (MHG) module and Self-Hypothesis Refinement (SHR) module (351-frame model with 3 hypotheses) are shown in Figure 6 and Figure 7, respectively. It can be found that the maps of multiple hypotheses contain diverse patterns and semantics. This indicates multiple representations in our method actually learn various modal information of pose hypotheses.

References

[1] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2272–2281, 2019. 1, 2

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 357–366, 2021. 1

[3] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3D human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209, 2021. 2

[4] 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 (CVPR)*, pages 7103–7112, 2018. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[6] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3D pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

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

[8] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5064–5073, 2020. 2

[9] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, Xuesheng Qian, and Xiaoyun Yang. A video is worth three views: Trigeminal transformers for video-based person Re-Identification. *arXiv preprint arXiv:2104.01745*, 2021. 1

[10] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017. 2

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 483–499, 2016. 2

[12] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7753–7762, 2019. 2

[13] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3D pose estimation from videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 764–780, 2020. 1, 2

[14] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3D human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 899–908, 2020. 2

[15] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[16] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3D pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11436–11445, 2021. 2

[17] 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 International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4

[18] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11477–11487, 2021. 1, 2

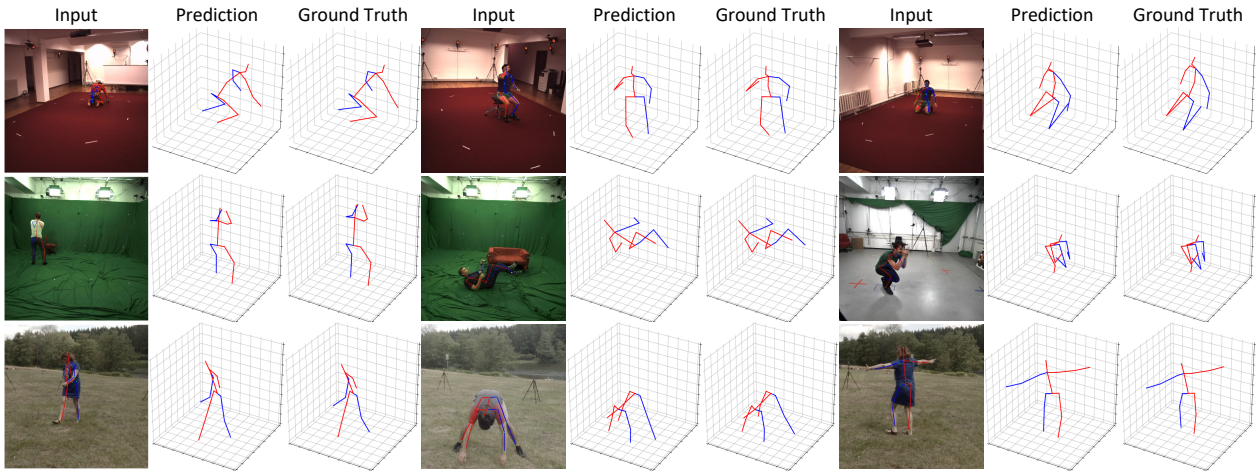


Figure 2. Qualitative results of our proposed method on Human3.6M dataset (first 1 row) and MPI-INF-3DHP dataset (last 2 rows).

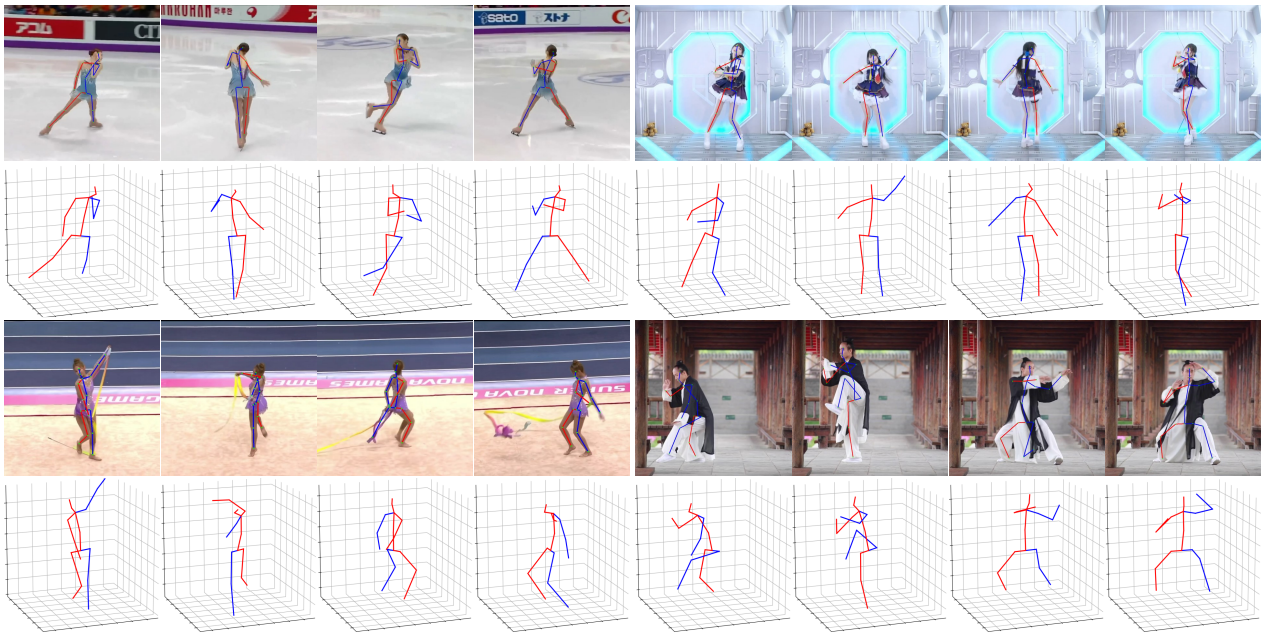


Figure 3. Qualitative results of our proposed method on challenging in-the-wild videos.

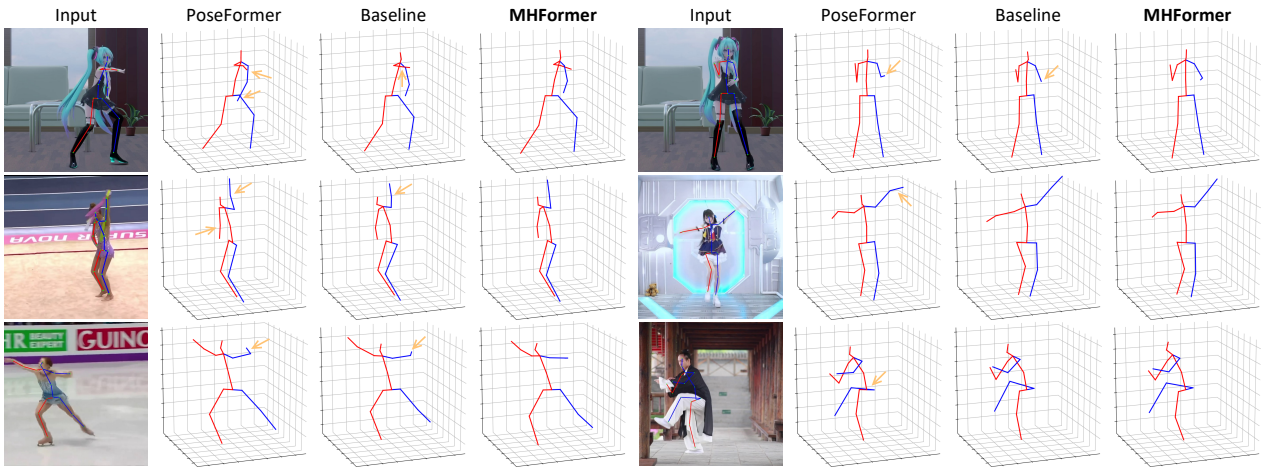


Figure 4. Qualitative comparison among the proposed method (MHFormer), the baseline method, and the previous state-of-the-art method (PoseFormer) [17] on challenging wild videos. Wrong estimations are highlighted by yellow arrows.

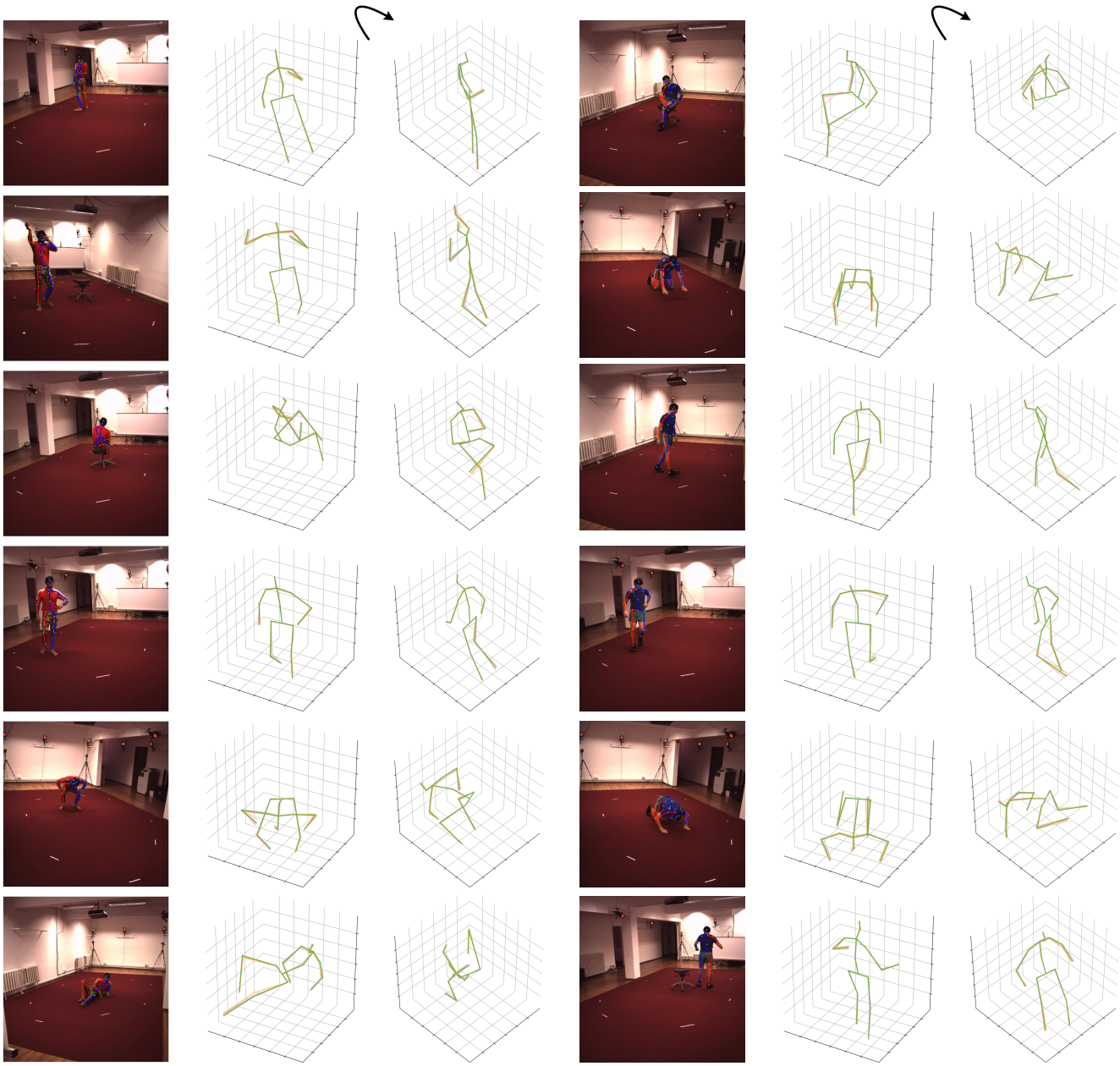
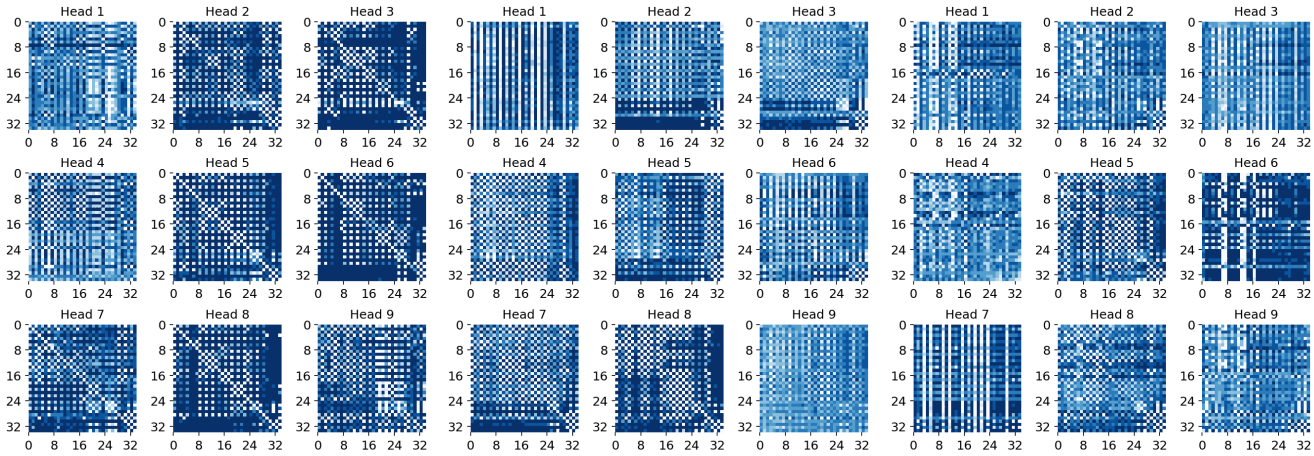
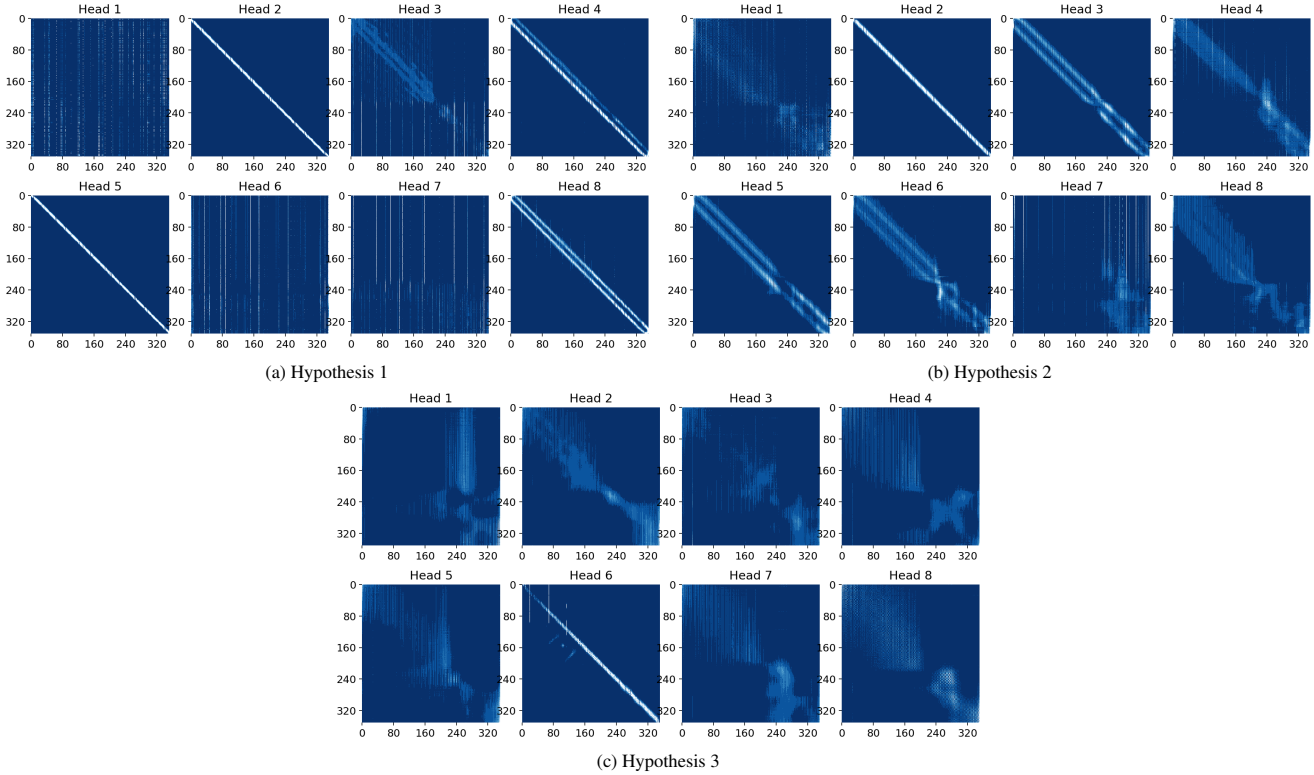


Figure 5. Diverse 3D pose hypotheses generated by MHFormer. For easy illustration, we color-code the hypotheses to show the difference among them, and the hypotheses are shown from two perspectives. Green colored 3D pose corresponds to the final synthesized estimation of our method.



(a) Hypothesis 1 (b) Hypothesis 2 (c) Hypothesis 3

Figure 6. Multi-head attention maps (9 heads) from the Multi-Hypothesis Generation (MHG) module of our 351-frame model with 3 different hypotheses. The brighter color indicates a stronger attention value.



(a) Hypothesis 1 (b) Hypothesis 2

(c) Hypothesis 3

Figure 7. Multi-head attention maps (8 heads) from the Self-Hypothesis Refinement (SHR) module of our 351-frame model with 3 different hypotheses. The brighter color indicates a stronger attention value.