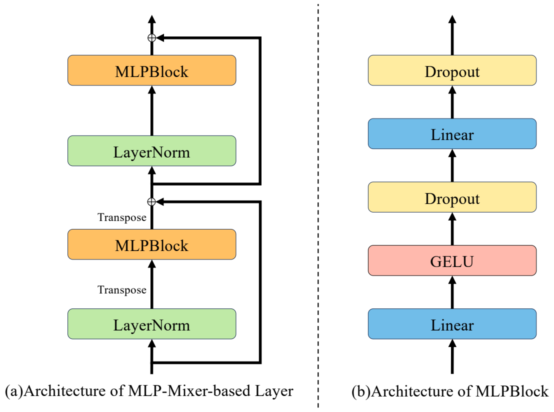
PostoMETRO: Pose Token Enhanced Mesh Transformer for Robust 3D Human Mesh Recovery

Supplementary Material

In Appendix A, we first illustrate the detailed architecture of our MLP-Mixer-based layers, which are used in our pose feature modulator. Then we provide more details about our datasets and experimental settings in Appendix B. More visualization results and occlusion sensitivity analysis results are provided in Appendix C and Appendix D. We also conduct more experiments in Appendix E. Finally, we discuss about societal impact, our limitations and future research directions in Appendix F.

A. Architecture



(a)Architecture of MLP-Mixer-based Layer

(b)Architecture of MLPBlock

Figure S5. Detailed architecture of MLP-Mixer-based layer (a) and MLPBlock (b).

The detailed architecture of our MLP-mixer-based layer is illustrated in Fig. S5. The MLPBlock is a key part, made up of Linear, GELU, and Dropout layers stacked together. Each MLP-Mixer-based layer contains two sets of LayerNorm and MLPBlock. Residual connection is used to ease the training.

B. Datasets

3D/2D Dataset Scale We utilized a large mixed dataset of around 512k samples. Those samples, including 312k from Human3.6M [5], 7k from UP-3D [7], 102k from MuCo-3DHP [14]. and 75k(28k + 47k) from COCO [12], along with 16k from MPII [1]. We adopt pseudo-ground-truth SMPL [13] annotated datasets for part of our COCO(28k) and MPII datasets from open source GitHub repository[3], and the other part of COCO(47k) from EFT[4] after removing items duplicated with the former 28k.

Pretraining Dataset Scale As said in our main paper, we adopt the publicly available[5] pose tokenizer as described in [3]. To train this tokenizer, a large open-source codebase MMPose[6] is used and COCO 2017 dataset(150k instances) and MPII dataset(40K instances) are adopted.

In our main paper, we report the performance of MPT [9] and our proposed method on 3DPW-TEST dataset. Our proposed PostoMETRO outperforms MPT on MPVPE and PA-MPJPE metrics but underperforms on MPJPE. We notice that the scale of the pretraining dataset of MPT is much larger than ours, and that should be brought to concern. The numbers of samples of pretraining datasets are listed in Tab. S8. As can be seen, MPT leverages much more training samples, *i.e.*, 80000k mesh-pose pairs, during pretraining. Due to such a significant gap, we believe that MPT's better performance on the MPJPE metric is partly attributed to the large-scale dataset it utilizes.

Data Preprocessing We directly adopt training data from open source GitHub repository[7] and replace COCO and MPII data as described above. Besides, we download 3DPW-VAL [16] and 3DOH [18] from official websites[8][9] and then parse data. Note that for 3D joint annotations of 3DPW-VAL and 3DOH, we use 3D joints regressed from SMPL [13].

Fine-tune Strategy We first train PostoMETRO on mixed datasets and then fine-tune it on corresponding datasets following existing works. When testing on 3DPW-TEST, we fine-tune our model on 3DPW-TRAIN by setting the learning rate to 2×10^{-5} and training it for 30 epochs. When testing on 3DOH, we fine-tune our model on 3DOH training split by setting the learning rate to 1×10^{-4} and training it for 30 epochs.

When testing on 3DPW-OCC, we directly use PostoMETRO trained from mixed datasets. This differs from the policy in PARE [6], where COCO, Human3.6M, and 3DOH are used. When testing on 3DOH, we further fine-tune our model on 3DOH train set, note that we only train

[3] https://github.com/huawei-noah/noah-research/tree/master/CLIFF

[4] https://github.com/facebookresearch/eft

[5] https://github.com/Gengzigang/PCT

[6] https://github.com/open-mmlab/mmpose

[7] https://github.com/microsoft/MeshTransformer

[8] https://virtualhumans.mpi-inf.mpg.de/3DPW/

[9] https://www.yangangwang.com/papers/ZHANG-OOH-2020-03.html

Methods	Backbone	*Pre* Dataset Scale	3DPW-TEST		
			MPVPE(↓)	MPJPE(↓)	PA-MPJPE(↓)
MPT	HigherHRNet	8000K	79.4	**65.9**	42.8
Ours	HRNet-W48	190K	**76.8**	67.7	**39.8**

Table S8. Comparison between MPT and ours in terms of performance in the 3DPW-TEST dataset and Pretraining dataset scale. Results in bold indicate the best performance. *Pre* is short for Pretraining.

PostoMETRO on 3DOH train set when testing on 3DOH test set.

C. Qualitative Results

Here, we offer more qualitative results in Fig. S7. Note that we highlight the crucial regions with red rectangular boxes. Especially, to validate that PostoMETRO does not collapse when the 2D pose is noisy, we feed noisy pose tokens to our model and test whether it can output plausible results. First, we train our model with pose token without noise and freeze it. Then we add random Gaussian noise to the logits $\hat{L} \in \mathbb{R}^{N\times V}$ (output by the classifier in pose tokenizer) and get noisy logits $\hat{L}_{noisy} \in \mathbb{R}^{N\times V}$. We then use $\hat{L}_{noisy}$ to obtain noisy pose tokens as input for the frozen model and visualize its output. The output and the corresponding noisy 2D poses are shown in Fig. S8, we use red rectangular boxes to denote noisy regions. As can be seen, PostoMETRO can output decent results even when 2D pose is unreliable. These results demonstrate the robustness of PostoMETRO and indicate the complementary role of image tokens to pose tokens, further distinguishing our model from those that rely solely on token-wise pose representation [9].

D. Occlusion Analysis

Following prior works [6, 8, 17], we visualize joint error maps by replacing the classification score with errors between predicted joints and corresponding ground truth. Same as [6, 8], we use MPJPE as our measurement since PA-MPJPE leads to an artificially low error by aligning global orientations. We conduct our experiment on the SOTA non-parametric method, FastMETRO [2], and our proposed PostoMETRO. We provide extensive results in Fig. S9 and Fig. S10, where a warmer color denotes a higher error. It can be seen that PostoMETRO can produce results with lower errors in various scenarios, showing that PostoMETRO is more robust to occlusions, demonstrating its superiority.

E. Extra ablations

Methods	3DPW-Non-OC		
	MPVPE(↓)	MPJPE(↓)	PA-MPJPE(↓)
FastMETRO [2]	99.9	91.1	51.0
Ours	**90.1**	**82.3**	**46.8**

Table S9. Performance on 3DPW-Non-OC. Results in bold indicate the best performance. ResNet-50 is used as backbone. Note that results are obtained ***without*** fine-tuning on 3DPW-TRAIN split.

Backbone	Use GT	3DPW-TEST		
		MPVPE(↓)	MPJPE(↓)	PA-MPJPE(↓)
ResNet-50	✗	78.0	68.4	40.8
ResNet-50	✓	**65.9**	**57.7**	**31.3**

Table S10. Performance on 3DPW-TEST [16] when using or not using ground truth 2D pose tokens. Results in bold indicate the best performance.

Training/Inference Time. We compare PostoMETRO with METRO [10], MeshGraphormer [11] and FastMETRO [2]. We set the training parameters (e.g., epochs) as described in the original works [2, 10, 11]. The training (including pretraining) and inference times are shown in Table S11. PostoMETRO shows competitive efficiency while improving performance.

Results in Non-occlusion Scenarios. We also test our model's performance under non-occlusion scenarios. We construct a non-occlusion subset from 3DPW by removing samples used in 3DPW-OCC and 3DPW-PC. The comparison between FastMETRO and PostoMETRO is listed in Tab. S9. Our method performs significantly better, hence proving its effectiveness even in non-occlusion scenarios.

Accuracy of Pose Tokens. For the purpose of exploring the upper limits of the token-wise 2D pose's assistance in the 3DHPSE task, we feed the tokens generated from the ground truth 2D pose by the pose encoder into the transformer. When using a classifier to predict pose tokens, pose confidence is generated, but it does not exist when using ground truth pose since pose encoder does not output confidence. Therefore, we use a very high score (fixed at 10 in the experiments) and concatenate it with the ground truth pose tokens for fine-tuning. The experimental results in Tab. S10 demonstrate a significant improvement

Method	Training time	Inference time	3DPW-TEST		
			MPVPE(↓)	MPJPE(↓)	PA-MPJPE(↓)
METRO [10]	∼1400 GPU hrs	15.8FPS	88.2	77.1	47.9
MeshGraphormer [11]	∼1300 GPU hrs	15.0FPS	87.7	74.7	45.6
FastMETRO [2]	**∼330GPU hrs**	17.0FPS	84.1	72.5	44.6
Ours (ResNet-50)	∼340 GPU hrs	**18.7FPS**	78.0	68.4	40.8
Ours (HRNet-W48)	∼390 GPU hrs	11.5FPS	**76.8**	**67.7**	**39.8**

Table S11. Training and Inference time compared with other baselines. Our proposed method shows competitive efficiency while improving performance.

Num. Blocks	3DPW-TEST
	PA-MPJPE(↓)
1	50.1
2	49.6
4	**48.9**
8	49.6

Table S12. Ablation of mixer block number in pose feature modulator. Results are obtained ***without*** fine-tuning on 3DPW-TRAIN split.

in performance when utilizing ground truth pose tokens. When setting ResNet-50 as backbone, by using ground-truth 2D pose tokens, our method scores 65.9mm, 57.7mm and 31.3mm on MPVPE, MPJPE, PA-MPJPE respectively, highlighting the substantial benefit of accurate 2D pose in the process of 3D human mesh recovery, further implying the strong potential of our proposed methods.

Ablation of Numbers of Mixer Layers. We ablate how deep our MLP-Mixer should be. Results in Tab. S12 show that increasing the number of mixer blocks decreases the error, with the optimal performance achieved at 4 blocks. However, further increments deteriorate model performance, possibly due to optimization challenges. Therefore, we use 4 blocks of MLP-Mixer as our pose feature modulator in our work.

F. Discussion

Societal Impact. Our proposed method can be used to detect 3D human body poses, thus applicable in certain scenarios, such as monitoring worker activities in industrial manufacturing environments or positioning patient poses in medical settings. However, in these low fault-tolerant environments, additional model assistance may be necessary when using the model.

Limitations & Future research. Given that PostoMETRO is a data-driven approach, it may fail when there is a significant difference between the test samples and those in our datasets. Here we show some failure cases in Fig. S6. As can be seen, when the persons in the image exhibit extreme poses, *e.g.*, skateboarding, PostoMETRO might not perform well and yield unsatisfactory results(*e.g.*, body part misalignment bounded by red boxes), due to the lack of abundant training samples of such poses in our training sets. A straightforward solution is to use datasets with more diverse human poses. Setting that aside, exploring how to faithfully reconstruct the human mesh with extreme poses within the constraints of existing data is an interesting future work.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2, 5, 1

[2] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, pages 342–359. Springer, 2022. 1, 2, 4, 5, 6, 7, 3, 8

[3] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023. 3, 8, 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1, 5

[6] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2, 3, 5, 6, 8, 1

[7] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer

vision and pattern recognition, pages 6050–6059, 2017. 2, 5, 1

[8] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023. 5, 6, 8, 2

[9] Kevin Lin, Chung-Ching Lin, Lin Liang, Zicheng Liu, and Lijuan Wang. Mpt: Mesh pre-training with transformers for human pose and mesh reconstruction. *arXiv preprint arXiv:2211.13357*, 2022. 3, 5, 6, 1, 2

[10] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 1, 2, 5, 6, 3

[11] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 5, 6, 2, 3

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3, 5, 1

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2

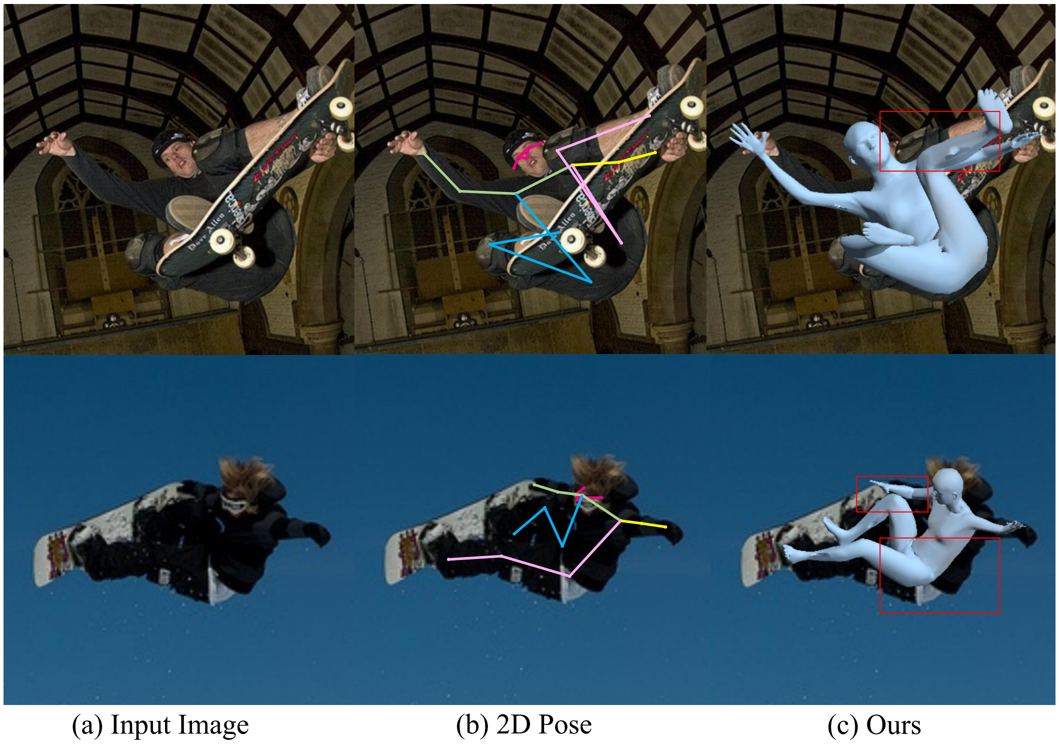
[14] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1, 5

[15] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 3

[16] 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. 2, 5, 6, 8, 1

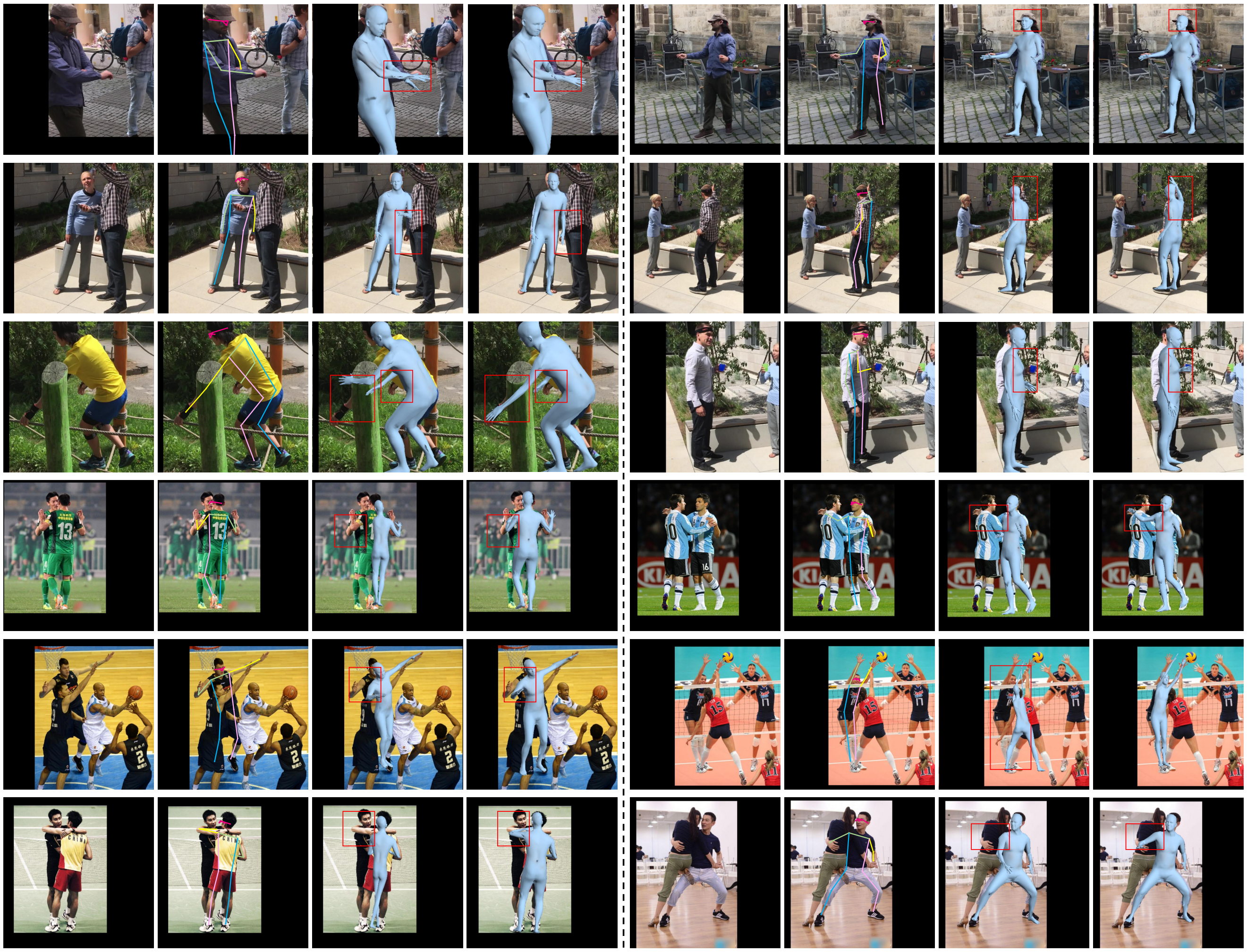
[17] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 2

[18] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 2, 5, 6, 8, 1



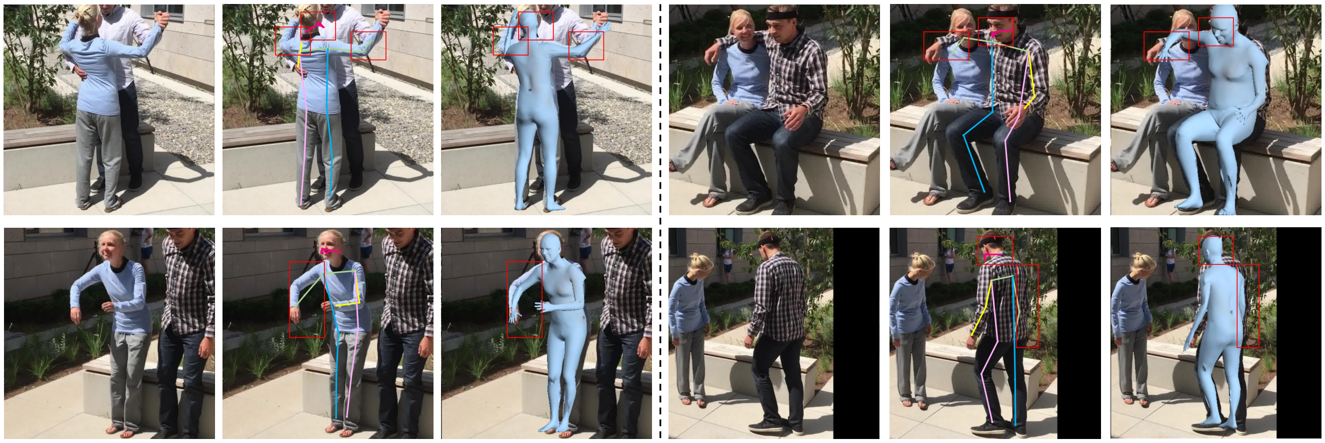
(a) Input Image (b) 2D Pose (c) Ours

Figure S6. **Qualitative results of PostoMETRO in some challenging cases**. From left to right: (a) Input image, (b) 2D pose, (c) Our results.



(a) Input Image (b) 2D Pose (c) FastMETRO (d) Ours (a) Input Image (b) 2D Pose (c) FastMETRO (d) Ours

Figure S7. **Qualitative results on 3DPW (rows 1-3) and OCHuman (rows 4-6) datasets.** From left to right: (a) Input image, (b) 2D Pose decoded from pose tokens, (c) FastMETRO [2] results, (d) Our results.



(a) Input Image (b) Noisy 2D Pose (c) Ours (a) Input Image (b) Noisy 2D Pose (c) Ours

Figure S8. **Performance of PostoMETRO with noisy pose token**. From left to right: (a) Input image, (b) Noisy 2D pose, (c) Our results.

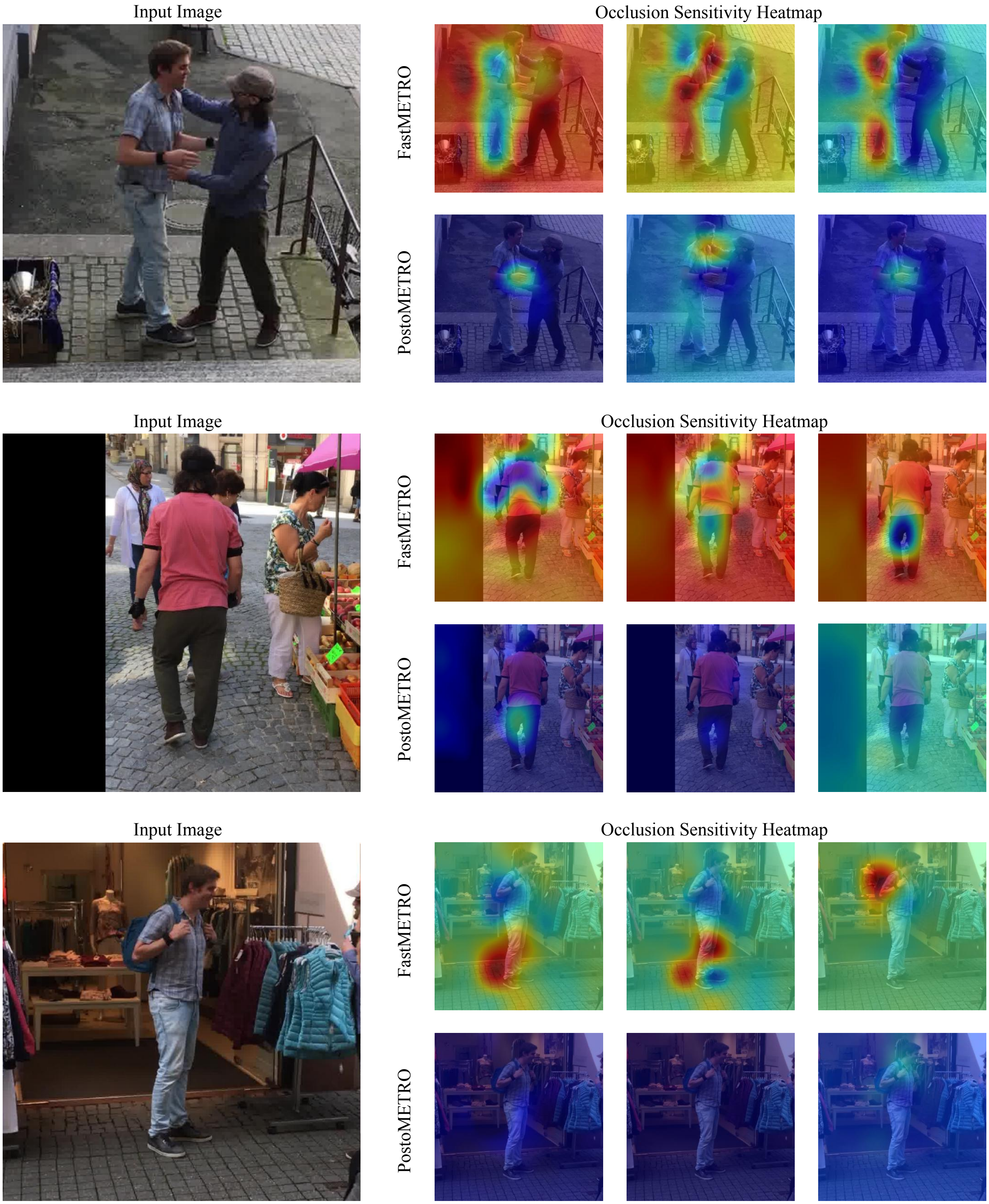


Figure S9. Occlusion Sensitivity Maps of FastMETRO [2] and PostoMETRO.

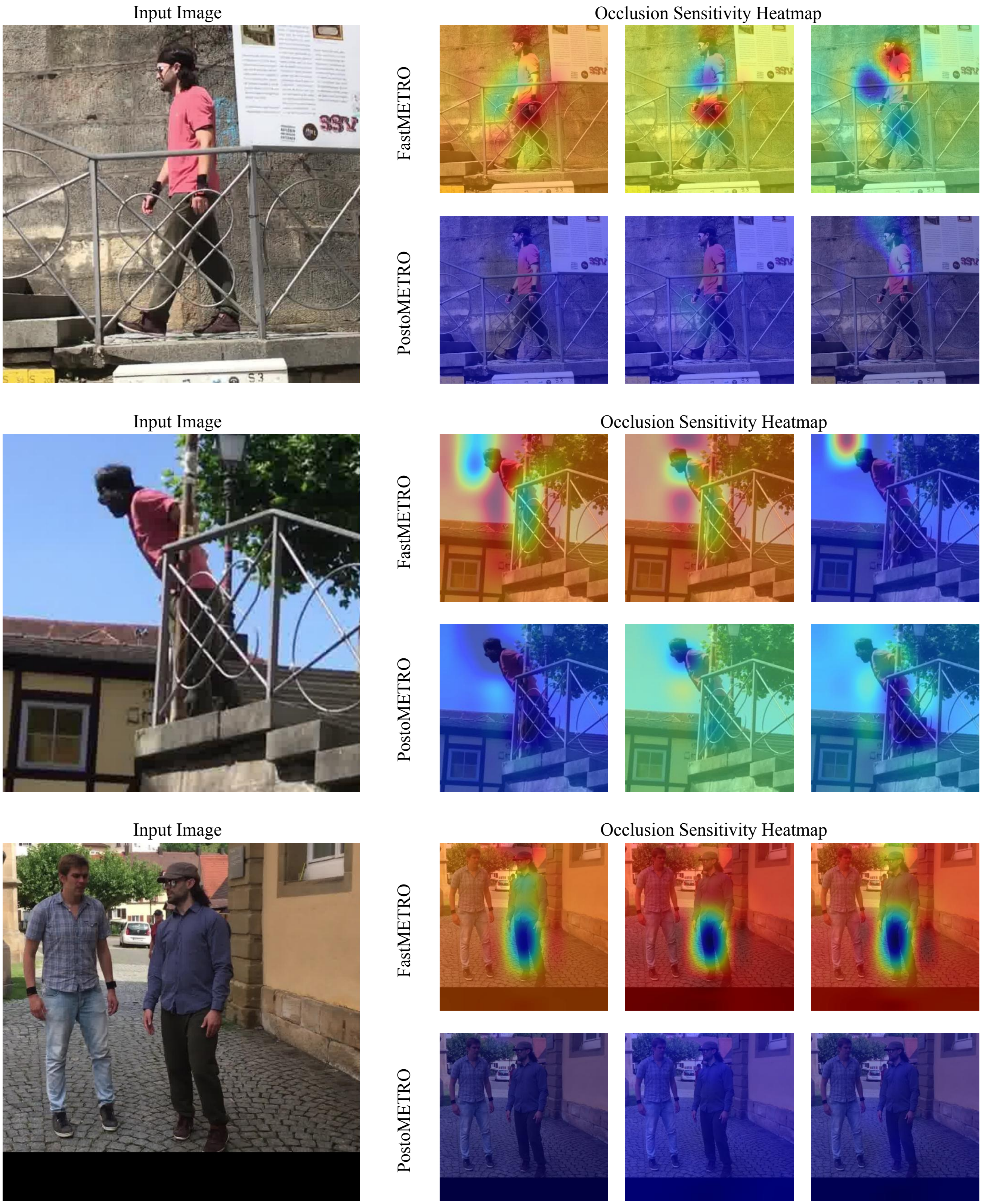


Figure S10. Additional results for Occlusion Sensitivity Maps of FastMETRO [2] and PostoMETRO.