3D Human Pose Estimation with Spatial and Temporal Transformers – Supplementary Material

Ce Zheng[1], Sijie Zhu[1], Matias Mendieta[1], Taojiannan Yang[1], Chen Chen[1], Zhengming Ding[2]
[1]Center for Research in Computer Vision, University of Central Florida, USA
[2]Department of Computer Science, Tulane University, USA
{cezheng,sizhu,mendieta,taoyang1122}@knights.ucf.edu;
chen.chen@crcv.ucf.edu;zding1@tulane.edu

In this supplementary material, we provide the following items:

- Comprehensive visualizations of spatial and temporal attention maps.
- Frame-wise comparison to track the average MPJPE of all the joints across frames.
- More qualitative comparison of estimated 3D poses.
- Estimated 3D poses using the proposed PoseFormer on the in-the-wild videos collected from YouTube.

We also include a demo video to showcase the 3D human pose estimation results of our proposed PoseFormer.

1. Attention Visualization

We present more visualization examples of spatial attention maps and temporal attention maps for all 8 heads when evaluating our PoseFormer model on Human3.6M test set S11 with the *SittingDown* action. For the spatial self-attention maps in Fig. 1, the x-axis corresponds to the query of 17 joints and the y-axis indicates the attention output. The attention heads return different attention intensities which represent the various local relations learned among the input joints. For the temporal self-attention maps in Fig. 2, the x-axis corresponds to the query of 81 frames and the y-axis indicates the attention output. Long term global dependencies are captured by different attention heads. The spatial and temporal attention maps have demonstrated that PoseFormer successfully encodes the local relationship between 2D joints as well as models global dependencies cross the arbitrary frames regardless of the distance.

2. Frame-wise Analysis

We perform frame-wise estimation analysis by computing the average MPJPE of all estimated joints in each frame. As shown in Fig. 3, we measure the frame-wise MPJPE through Human3.6M [3] test set S11 with *Eating* and *Photo* actions. Our PoseFormer (red line) yields lower MPJPE in most frames of both actions, compared with our baseline (temporal transformer only) and the state-of-the-art method [1].

3. More Qualitative Results

We provide more visual comparison between the 3D estimated pose and the ground truth. We evaluate PoseFormer on the Human3.6M test set S11 with the *Greeting* and *WalkDog* actions. Compared with the state-of-the-art method [1] and our baseline, PoseFormer achieves more accurate estimations as shown in Fig. 4.

4. Performance on Videos in-the-wild

Our model was trained on the indoor dataset: Human3.6M that the background is static and the camera capture setting is known. Estimating 3D human pose from in-the-wild videos is more challenging due to the dynamic environment and unknown camera setting. There are often high variations in foreground/background objects appearances and severe occlusions in unconstrained environment. We also evaluate the performance of our PoseFormer on some online videos from YouTube as shown in Fig. 5. We first use AlphaPose [2] as 2D pose detector to generate 2D poses from the video frames, then apply PoseFormer for 3D pose estimation. We observe that PoseFormer achieves acceptable performance in most of the frames, but there are still some failure cases (see Fig. 5) due to inaccurate 2D pose detection, occlusion, and fast motion. Since PoseFormer is a 2D-to-3D lifting approach, any incorrect detected 2D poses may lead to inaccurate 3D pose estimation. Occlusion is a key challenge remains in 3D HPE since the information is missing. Moreover, estimation from the extreme fast motion may be affected by the motion blurring of several frames.

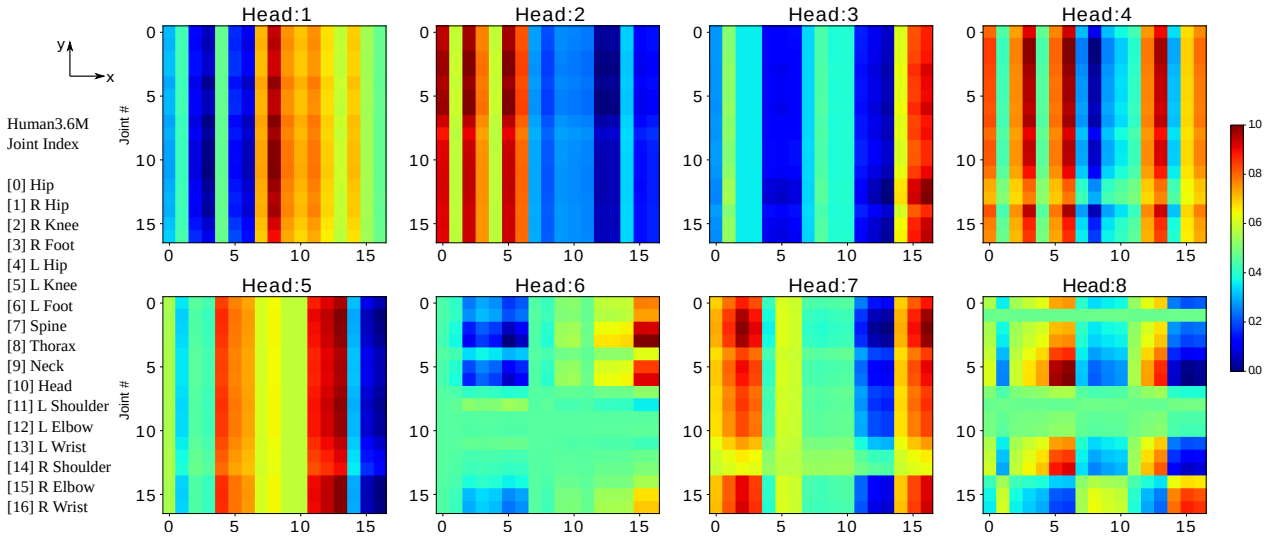

Figure 1. Visualization of self-attentions in the spatial transformer. The x-axis (horizontal) and y-axis (vertical) correspond to the queries and the predicted outputs, respectively. The pixel $w_{i,j}$ (i: row, j: column) denotes the attention weight of the j-th query for the i-th output. Red indicates stronger attention. The attention output is normalized from 0 to 1.

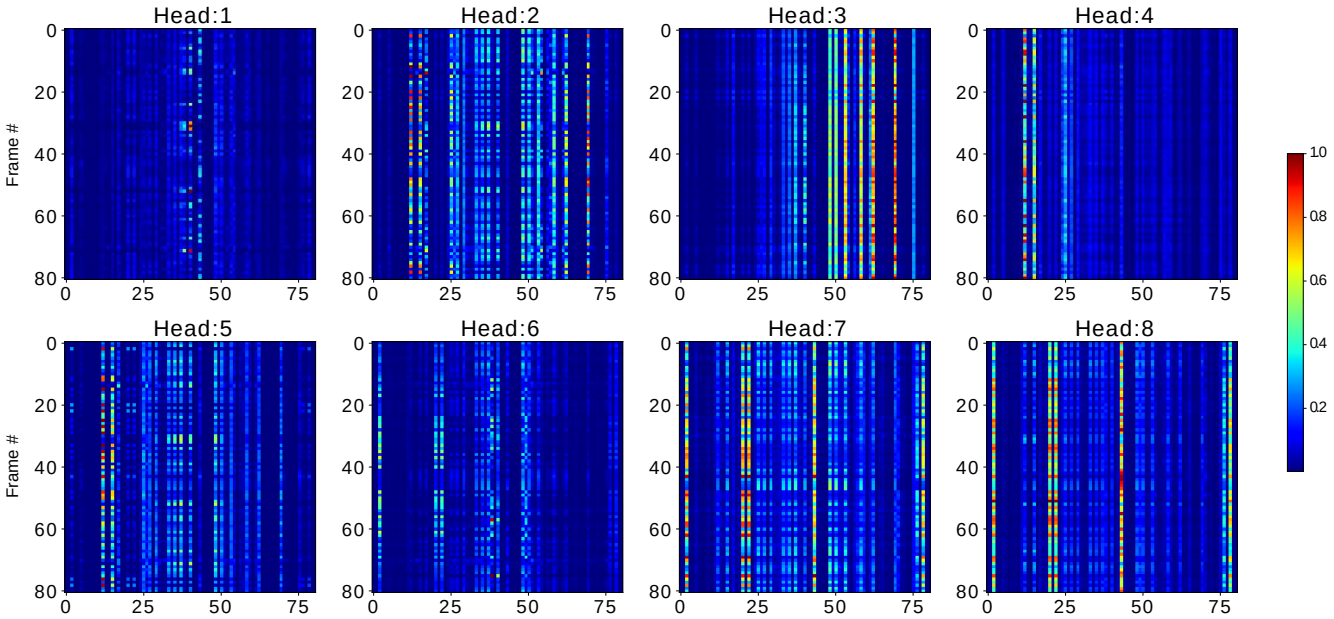

Figure 2. Visualization of self-attentions in temporal transformer. The x-axis (horizontal) and y-axis (vertical) correspond to the queries and the predicted outputs, respectively. The pixel $w_{i,j}$ (i: row, j: column) denotes the attention weight of the j-th query for the i-th output. Red indicates stronger attention. The attention output is normalized from 0 to 1.

References

[1] 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*, 2021. 1, 3, 4

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 1

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014. 1

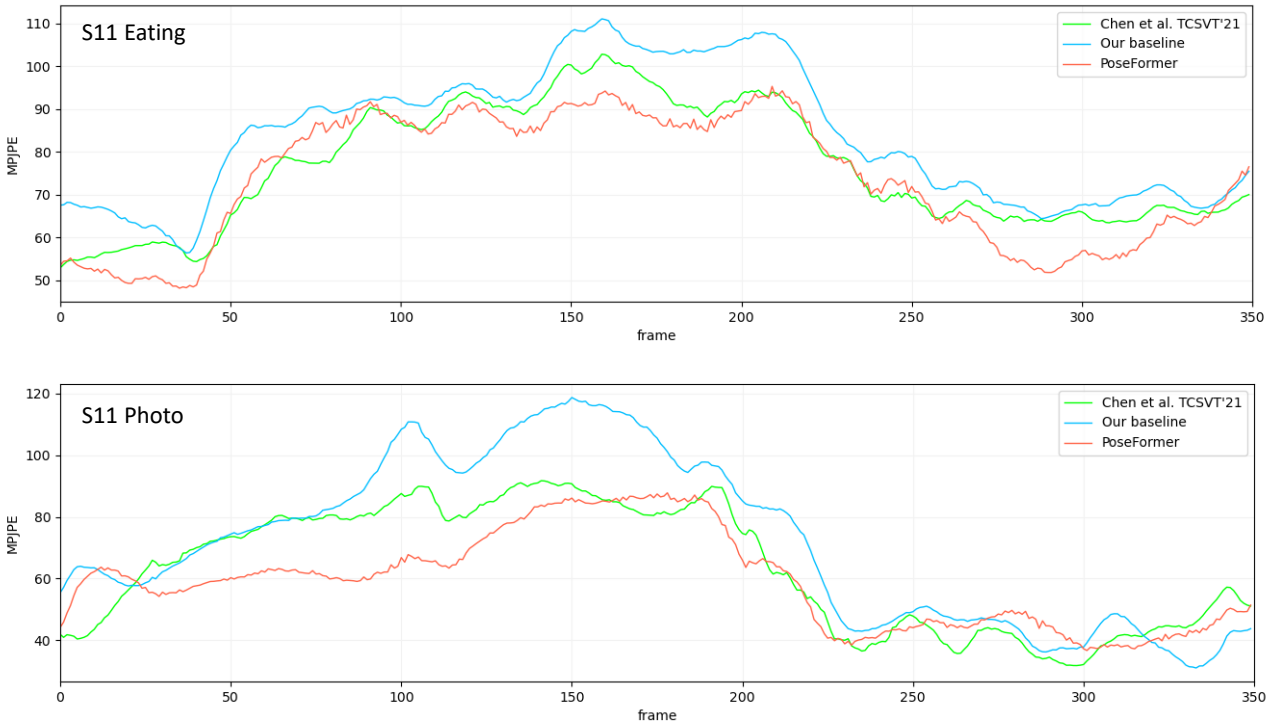

Figure 3. Frame-wise comparison between our method (PoseFormer), our baseline, and the SOTA approach Chen *et al*. [1] on Human3.6M test set. Top-figure: S11 with the *Eating* action. Bottom-figure: S11 with the *Photo* action.

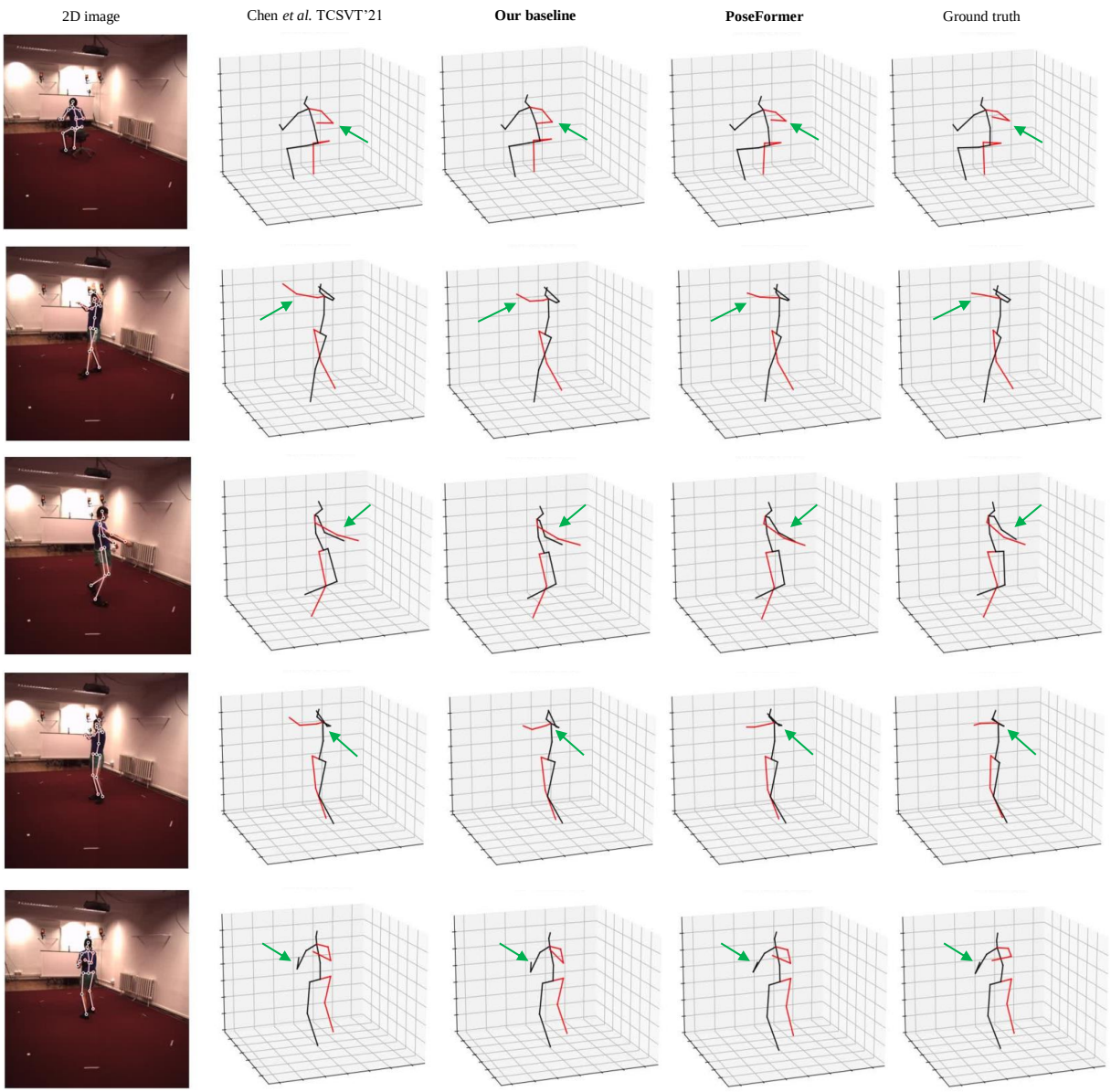

Figure 4. Qualitative comparison between our method (PoseFormer), our baseline, and the SOTA approach Chen *et al*. [1] on Human3.6M test set S11 with the *Greeting* and *WalkDog* actions. The green arrows highlight locations where PoseFormer clearly has better results.

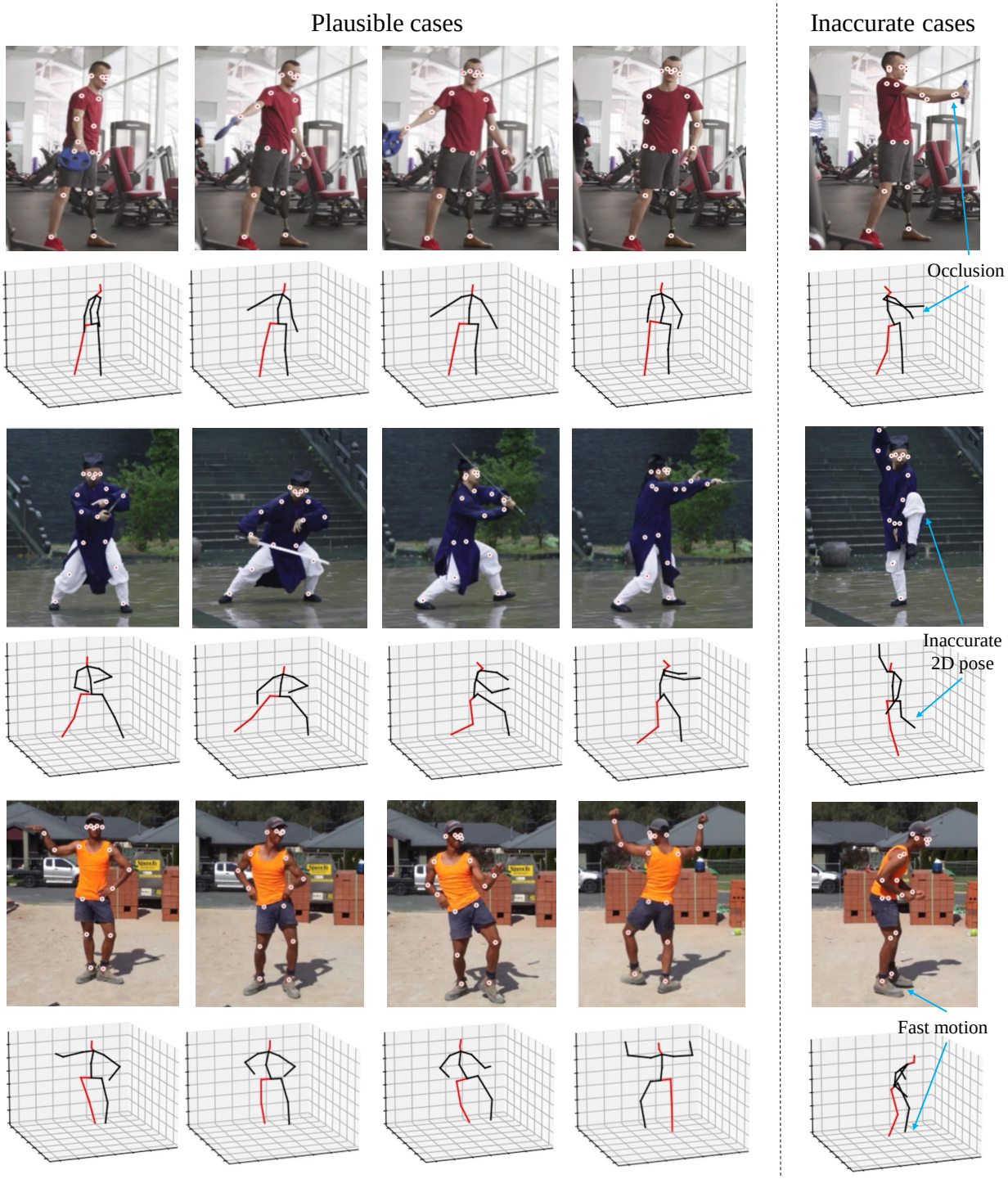

Figure 5. Qualitative results on in-the-wild videos: original frame sequence with detected 2D joints and the recovered 3D poses using PoseFormer.