PSVT: End-to-End Multi-person 3D Pose and Shape Estimation with Progressive Video Transformers Supplementary Material

Zhongwei Qiu[1,3,4], Qiansheng Yang[2], Jian Wang[2], Haocheng Feng[2], Junyu Han[2],
Errui Ding[2], Chang Xu[3], Dongmei Fu[1,4], Jingdong Wang[2]
[1]School of Automation and Electrical Engineering, University of Science and Technology Beijing
[2]Baidu, [3]University of Sydney, [4]Beijing Engineering Research Center of Industrial Spectrum Imaging

In this supplementary material, we first give the algorithm details of PSVT in Section 1. Then, we will provide more experimental results and analysis of PSVT on the crowded scenarios in Section 2. Then, some failure cases and the limitations of PSVT are discussed in Section 3. Finally, more visualization results on in-the-wild images or videos are shown in Section 4.

1. Algorithm Details

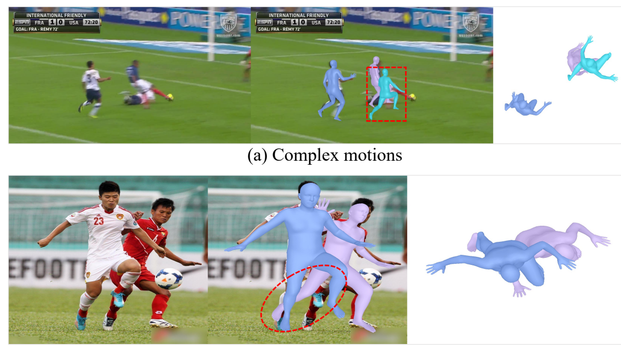
Algorithm 1 PSVT with progressive decoding mechanism and pose-guided attention.

Input: Video $\mathbf{V}$: $\{I^t, t \in [1,T]\}$; Backbone network of HRNet-32: $\phi(\cdot)$; Spatio-Temporal Encoder: STE$(\cdot)$; Spatio-Temporal Pose Decoder: STPD$(\cdot)$; Spatio-Temporal Shape Decoder: STSD$(\cdot)$; Token Aligning: TA$(\cdot)$; Joints weights of projecting mesh to 3D joints: $\mathcal{W}$.

Output: Human meshes $\mathcal{M} = \{\mathcal{M}_i^t | t \in [1,T], i \in [1,N]\}$; 3D joints $J = \{J_i^t | t \in [1,T], i \in [1,N]\}$;

- Initializing $\mathcal{Q}_{pose}$,$\mathcal{Q}_{shape}$;
- $F = \{F^t | t \in [1,T]\} = \{\phi(I^t) | t \in [1,T]\}$;
- $\tau_e = \{\tau_e^t | t \in [1,T]\} = \{\text{STE}(F^t) | t \in [1,T]\}$;
- **for** $t = 1; t <= T; t++$ **do**
    - Updating pose queries $\hat{\mathcal{Q}}_{pose} = \psi(\mathcal{Q}_{pose}^t, \tau_{pose}^{t-1})$
    - $\tau_{pose}^t = \text{STPD}(\hat{\mathcal{Q}}_{pose}, \tau_e^t)$;
    - Updating shape queries $\hat{\mathcal{Q}}_{shape} = \psi(\mathcal{Q}_{shape}^t, \tau_{shape}^{t-1})$
    - Token aligning $\hat{\mathcal{Q}}_{shape} = \text{TA}(\hat{\mathcal{Q}}_{shape}, \tau_{pose}^t)$;
    - $\tau_{shape}^t = \text{STSD}(\hat{\mathcal{Q}}_{shape}, \tau_e^t)$;
    - Regressing joints maps: M_{2D}, M_o, and M_d from τ_{pose}^t;
    - Localizing Top-N center points $P = \{(x_i, y_i, d_i) | i \in [1,N]\}$ from joints maps;
    - Regressing shape maps: M_s from τ_{shape}^t;
    - Decoding mesh $\mathcal{M}^t = \{\mathcal{M}_i^t(\theta, \beta, \alpha) | i \in [1,N]\}$ with the center points of P;
    - Projecting 3D joints $J^t = \{\mathcal{W}\mathcal{M}_i^t | i \in [1,N]\}$;
- **end for**

The algorithms details of PSVT with progressive decoding mechanism and pose-guided attention are shown in Algorithm 1. The backbone network is HRNet-32 [6]. The progressive decoding mechanism is a bidirectional propagation scheme, which includes forward propagation and backward propagation. For clarity, only forward propagation is shown in Algorithm 1.



(a) Complex motions

(b) Complex occlusions

Figure 1. The visualization results of some failure cases on in-the-wild images.

2. Evaluation on the Crowded Scenarios

To better evaluate the effectiveness of Pose-Guided Attention (PGA) in PSVT, we test PSVT on the crowded dataset. Following [3–5, 7, 9], we test PSVT on the occluded subset of 3DPW dataset [8]. As shown in Table 1, we test PSVT without using temporal information for a fair comparison with BEV [7]. PSVT achieves 49.67, 79.80, and 92.04 in PA-MPJPE, MPJPE, and MPVE, respectively. Compared with SOTA multi-person method (BEV), PSVT achieves relative gains of 7.2%, 12.0%, and 12.0%, respectively. These results show that PSVT has a stronger ability to handle images with occluded persons since the PGA.

3. Failure Cases and Limitations

Although PSVT achieves state-of-the-art results on multiple widely-used 3D pose and shape estimation datasets,

Figure 2. More visualization results on the in-the-wild images from CrowdPose [2] dataset. More visualization results on in-the-wild videos are in the attached documents.

Methods	Frame	PA-MPJPE	MPJPE	MPVE
BEV [7]	1	53.55	90.64	104.55
PSVT (Ours)	1	**49.67**	**79.80**	**92.04**

Table 1. The comparison between BEV [7] and PSVT on the 3DPW-OC [9], the crowded subset of 3DPW [8] dataset.

there still are some limitations and failure cases. As shown in Figure 1 (a), for some human instances with complex motions, it's difficult for PSVT to predict its pose and shape accurately. As shown in Figure 1 (b), for some human instances with complex occlusions, it's also difficult for PSVT to predict their poses and shapes accurately.

4. More Visualization Results

To evaluate the generalization ability of PSVT, we test PSVT on the in-the-wild images from CrowePose [2] dataset and videos from PoseTrack [1] dataset. As shown in Figure 2, PSVT performs well on these images with crowded or strange poses, which shows the stronger generalization ability of PSVT.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, pages 5167–5176, 2018. 2

[2] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019. 2

[3] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Learning recurrent structure-guided attention network for multi-person pose estimation. In *ICME*, pages 418–423. IEEE, 2019. 1

[4] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Weakly-supervised pre-training for 3d human pose estimation via perspective knowledge. *PR*, page 109497, 2023. 1

[5] Zhongwei Qiu, Qiansheng Yang, Jian Wang, and Dongmei Fu. Ivt: An end-to-end instance-guided video transformer for 3d pose estimation. In *ACM MM*, pages 6174–6182, 2022. 1

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1

[7] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of

3d people in depth. In *CVPR*, pages 13243–13252, 2022. 1, 2

[8] 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 *ECCV*, pages 601–617, 2018. 1, 2

[9] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 1, 2