Supplementary Material *for*
Capturing Humans in Motion: Temporal-Attentive 3D Human Pose and Shape Estimation from Monocular Video

7. Datasets

3DPW. 3DPW [37] is mainly captured from outdoors and in-the-wild. It combines a hand-held camera and a set of inertial measurement unit (IMU) sensors attached at the body limbs to calculate the near ground-truth SMPL parameters. It contains a total of 60 videos of different lengths. We use the official split to train and test the model, where the training, validation, and test sets are composed of 24, 12, and 24 videos, respectively. In addition, we report MPVPE on 3DPW because it is the only dataset that contains ground-truth 3D shape annotations among the datasets we used.

MPI-INF-3DHP. MPI-INF-3DHP [27] is a dataset consisting of both constrained indoor and complex outdoor scenes. It is captured using a multi-view camera setting with a markerless motion capture system, and the 3D joint annotation is calculated through the multiview method. Following existing methods [6,20], we use the official split to train and test the model. The training set contains 8 subjects, each of which has 16 videos, and the test set contains 6 subjects, performing 7 actions in both indoor and outdoor environments.

Human3.6M. Human3.6M [16] is one of the largest motion capture datasets, which contains 3.6 million video frames and corresponding 3D joint annotations. This dataset is collected in an indoor and controlled environment. Same as existing methods [6, 20], we train the model on 5 subjects (*i.e.*, S1, S5, S6, S7, and S8) and evaluate it on 2 subjects (*i.e.*, S9 and S11). We subsampled the dataset to 25 frames per second (fps) for training and testing.

AMASS. AMASS [26] is a large-scale human motion sequence database that unifies 15 existing motion capture (mocap) datasets by representing them within a common framework and parameterization. These motion sequences are annotated with Mosh++ to generate SMPL parameters. AMASS has a total of 42 hours of mocap, 346 subjects, and 11, 451 human motions. Following the setting of the existing method [20], we use this database to train our MPS-Net.

PoseTrack. PoseTrack [1] is a 2D benchmark dataset for video-based multi-person pose estimation and articulated tracking. It contains a total of 1, 337 videos, divided into 792, 170, and 375 videos for training, validation, and testing. Each person instance in the video is annotated with 15 keypoints. Same as existing methods [6, 20], we use the training set for model training.

InstaVariety. InstaVariety [18] is a 2D benchmark dataset captured from Instagram using 84 hashtags. There are 28, 272 videos in total, with an average length of 6 seconds, and OpenPose [4] is used to acquire pseudo ground-truth 2D joint annotations. Same as existing methods [6, 20], we adopt this dataset for training.

	3DPW			
Method	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	ACC-ERR ↓
MPS-Net (HAFI, 2 frames/group)	52.6	85.4	101.0	7.8
MPS-Net (HAFI, 3 frames/group)	**52.1**	**84.3**	**99.7**	**7.4**
MPS-Net (HAFI, 4 frames/group)	52.5	85.9	101.2	7.6

Table 5. Effect of the number of frames per group in the HAFI module. The training and evaluation settings are the same as the experiments on the 3DPW dataset [37] in Table 1.

8. Evaluation metrics

Four standard evaluation metrics [6, 20, 25] are considered, including MPJPE, PA-MPJPE, MPVPE, and ACC-ERR. Specifically, MPJPE is calculated as the mean of the Euclidean distance between the ground-truth and the estimated 3D joint positions after aligning the pelvis joint on the ground truth location. PA-MPJPE is calculated similarly to MPJPE, but after the estimated pose is rigidly aligned with the ground-truth pose. MPVPE is calculated as the mean of the Euclidean distance between the ground truth and the estimated 3D mesh vertices (output by the SMPL model). ACC-ERR is measured as the mean difference between the ground-truth and the estimated 3D acceleration for every joint.

9. Ablation study of HAFI module

To analyze the effect of the number of frames per group in the HAFI module, we conduct ablation studies on MPS-Net with different HAFI module settings under the 3DPW dataset [37]. The results in Table 5 indicate that considering the temporal features of three adjacent frames as a group can enable our MPS-Net to achieve the best 3D human pose and shape estimation. Therefore, for all experiments (*i.e.*, Table 1 to Table 4), the HAFI module defaults to three frames as a group.

On the other hand, we further add the results considering only the HAFI module on MPS-Net (called MPS-Net-only HAFI) to supplement the ablation experiments in Table 3, with results of 54.0, 87.6, 103.5, and 7.5, respectively, for PA-MPJPE, MPJPE, MPVPE and ACC-ERR. Compared with MPS-Net-only MoCA (see Table 3), MPS-Net-only HAFI yields better ACC-ERR but worse on PA-MPJPE, MPJPE and MPVPE. However, by coupling MoCA with HAFI, MPS-Net (Ours) achieves the best result.

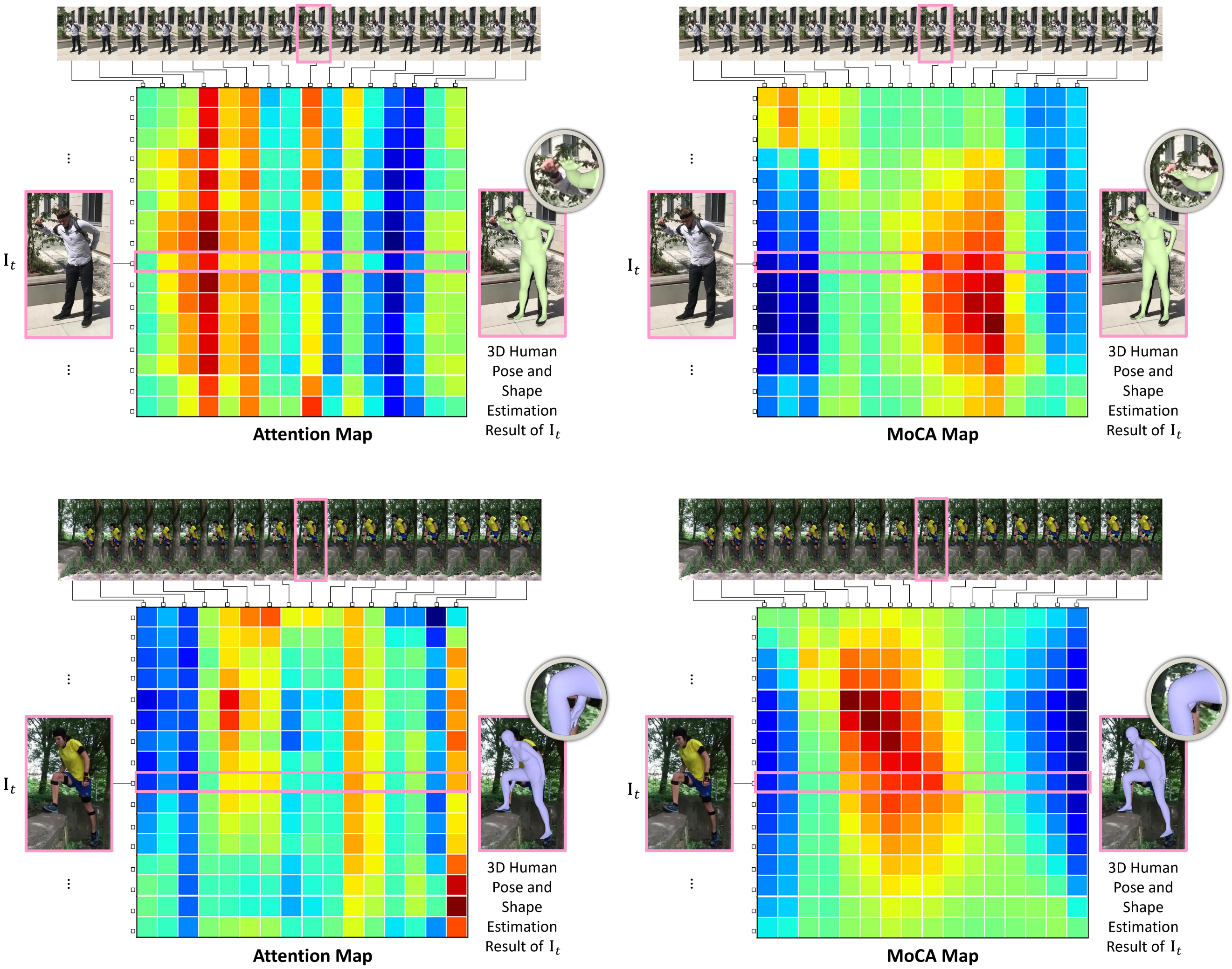

Figure 9. Visual comparison of 3D human pose and shape estimation of MoCA module and non-local block [38] on the 3DPW dataset [37]. Where the attention map is generated from the non-local operation, and the MoCA map is generated from the MoCA operation. In the attention and MoCA maps, red indicates a higher attention value, and blue indicates a lower one. The results demonstrate that the MoCA map generated by our MoCA operation can indeed allow the MPS-Net to focus attention on a more appropriate range of action sequence to improve the estimation results.

10. Qualitative comparison between MoCA module and non-local block

To further verify whether the proposed MoCA operation can improve the 3D human pose and shape estimation by introducing NSSM to recalibrate the attention map generated by the non-local operation [38], we conduct the following qualitative experiments. Specifically, we visualize the 3D human pose and shape estimation resulted from the methods of MPS-Net-only Non-local and MPS-Net-only MoCA in Table 3, respectively. The results can be seen from the two examples in Figure 9 that the MoCA map generated by our MoCA operation can indeed allow the MPS-Net to focus attention on a more appropriate range of action sequence, thereby improving the accuracy of 3D human pose and shape estimation. On the contrary, the attention map generated by non-local operation is often unstable, and it is easy to focus attention on less correlated frames and ignore the continuity of human motion in the action sequence, which reduces the accuracy of estimation. Such a result is quantitatively demonstrated by the improvement of MPS-Net-only MoCA in the MPJPE, PA-MPJPE, and MPVPE errors (see Table 3).

On the other hand, we also add the results considering only the setting of NSSM on MPS-Net (called MPS-Net-only NSSM) to supplement the ablation experiments in Table 3, with results of 53.3, 88.0, 104.1, and 26.0, respectively, for PA-MPJPE, MPJPE, MPVPE and ACC-ERR. These (cf. Table 3) indicate that MPS-Net-only NSSM and MPS-Net-only Non-local are complementary, and their fusion, *i.e*., MPS-Net-only MoCA, can further achieve better results.

11. Effect of motion speed and input fps

To demonstrate how motion speed and frame rate influence MPS-Net performance, we conducted experiments with various demo videos (*e.g*., fast-moving dancing or general walking video) at different frame rates, from 24fps to 60fps, and found that it has little impact on MPS-Net. Frame rate info for each demo video is available on our demo website.