Supplementary Material *for* Two-stage Co-segmentation Network Based on Discriminative Representation for Recovering Human Mesh from Videos

In this supplement, we will provide additional more experimental details, and results.

1. Datasets

We use batches of mixed 2D and 3D datasets. The 3D dataset contains 3DPW, Human3.6M, MPI-INF-3DHP. The 2D dataset contains InstaVariety, Penn Action, PoseTrack.

3DPW. The 3DPW is a 3D dataset that captures the natural scenes of the SMPL human body, using IMU sensors and handheld cameras to pose. There are also a large number of complex scenes in the 3DPW dataset. Therefore, we use it as one of the main evaluation datasets of our method. We followed the official segmentation protocol to train and test our model. Where the training set, validation set, and test set consist of 24, 12, and 24 videos respectively. In addition, we report MPVPE on 3DPW which only has real 3D shapes to use. We use the 14 joints defined by Human3.6M to evaluate PA-MPJPE and MPVPE.

Human3.6M. Human3.6M is the most standard dataset for 3D pose optimization and evaluation, and it is a large-scale dataset consisting of 3.6 million RGB images of 11 different professional actors performing 15 everyday activities, both 2D and 3D ground truth are available for supervised learning. We subsampled the dataset to 25 fps (initially 50 fps) for training and evaluation of acceleration errors. The 14 joints defined by Human3.6M were used to calculate PA-MPJPE and MPJPE.

MPI-INF-3DHP. The MPI-INF-3DHP is a dataset created using an unmarked motion capture system and multiple cameras. The learning data includes more than 1.3 million frames of videos of 11 people captured by 14 cameras at the same time.

InstaVariety. This dataset is a 2D human dataset curated by HMMR, whose videos are collected from Instagram using 84 motion-related hashtags. There are 28K videos with an average length of 6 seconds.

Penn Action. A benchmark for human pose estimation and tracking. This database contains 2326 video sequences of 15 different actions and human joint annotations for each sequence.

PoseTrack. The PoseTrack is a large-scale benchmark test for human pose estimation and joint tracking in video. Includes more than 1356 video sequences and 46K annotated video frames. We used 792 videos from the official training set, which contains 2D pose annotations in the middle 30 frames of the video.

2. Training Details

During the training phase, we feed F_{all}, F_{bra1}^{stage2}, F_{bra2}^{stage2} to the SMPL regressor, regressing SMPL parameters and camera parameters. In the test phase, we only perform regression parameters on F_{all} to obtain the final human body mesh.

Table 1. The impact of the excitation method on the network.

Method	*PA-MPJPE*↓	*Accel*↓
Channel excitation	61.7	9.0
Single excitation	62.2	8.8
Dual excitation	**61.4**	**8.5**

3. Effectiveness of Dual Excitation Mechanism

We explore the effectiveness of modeling human motion representations. As shown in Table 1 of the supplementary material, to demonstrate that our method is improved by focusing on human motion, we then compare the channel excitation, single excitation, and our dual excitation mechanisms. We observe that the temporal consistency error gradually decreases with channel excitation, single excitation, and our dual excitation mechanism. Channel excitation aims to excite meaningful feature channels and is applied to each frame of the video independently, but this method does not consider temporal information. The single excitation mechanism is simple to apply spatial feature error, but it lacks long-distance time series modeling of overall sequence features. Our dual excitation mechanism considers long-term temporal relationships by introducing self-attention, while we make differences to adjacent spa-

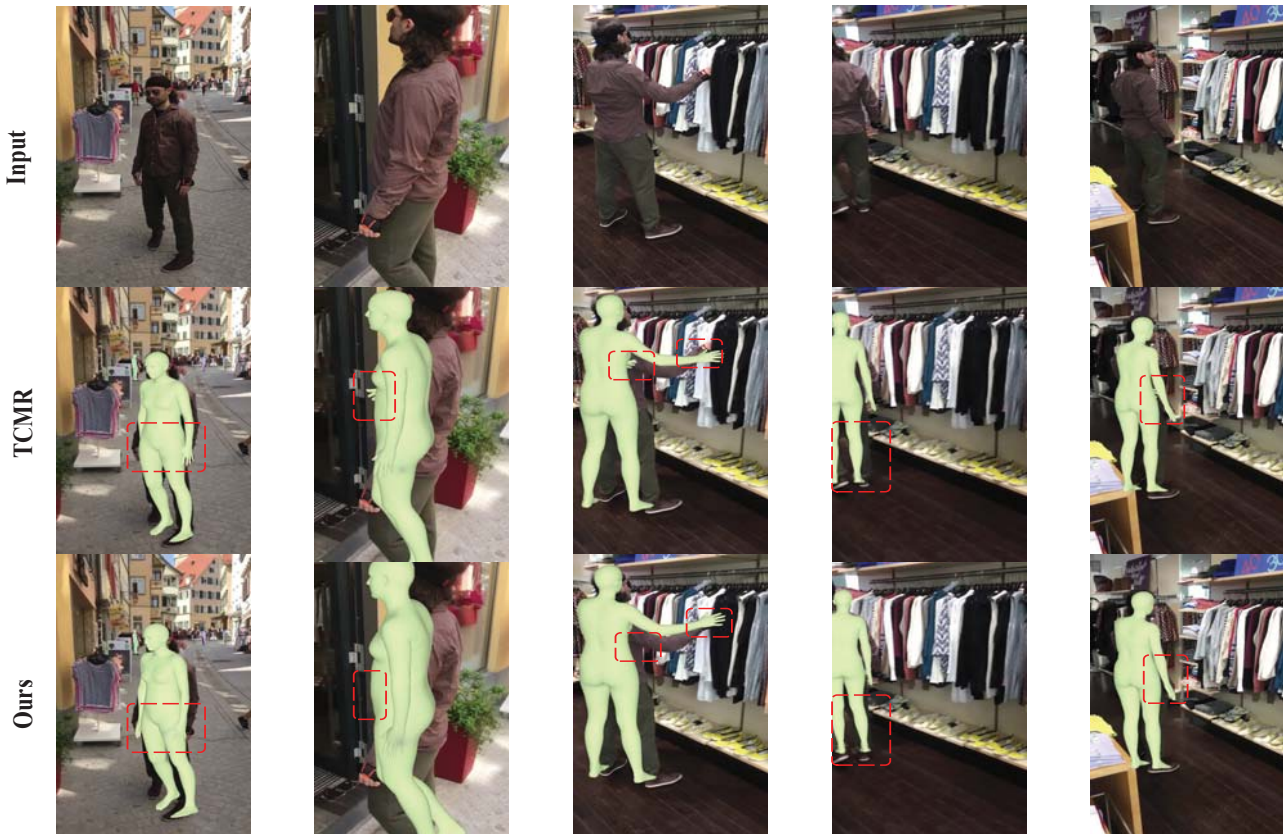

Figure 1. Qualitative results of our approach in complex outdoor scenes.

tial features, thereby motivating motion features.

4. Effectiveness of the Frame Superposition

In Stage 2, spatio-temporal features are self-added to obtain enhanced features. In Table 2 of the supplementary material, we show three variants using frame superposition. First, in our method, frame superposition is only used in one branch of Stage 2. Then we try the other two methods, including both branches in Stage 2 with and without frame superposition. The experimental results show that our method achieves the best results. These illustrates that the spatio-temporal feature self-adding operation helps to reconstruct a reasonable human body, but when both branches use frame superposition or neither, it causes the network to treat both branches equally. Our method learns based on the discriminative representation, so the network is more sensitive to the discriminative representation. Therefore, the integration of the two branches with discriminative spatio-temporal features will result in better performance.

Table 2. The effect of the frame superposition.

Method	*PA-MPJPE*↓	*Accel*↓
Both use frame superposition	62.0	**8.5**
Neither use frame superposition	62.5	8.6
Ours	**61.4**	**8.5**

5. More Qualitative Results

In this section, we show more qualitative results of our approach. These include comparisons with state-of-the-art methods and robustness and generalization.

5.1. Compared to the State-of-the-Art Method

Compared to the state-of-the-art method TCMR, the human body recovered by our method is more reasonable and realistic. As shown in Figure 1 of the supplementary material, our method is more sensitive to minute details within the overall complex surroundings. This includes hand reconstruction and partial pose occlusion. Our method can reason out and reconstruct a reasonable human pose. Although TCMR can recover a reasonable human posture, it

Figure 2. Generalization experiments. In videos on the web, our method shows good results for human body reconstruction in extreme illumination.

is less effective in reconstructing some details (e.g., joints of the hands and legs). Our method handles the details well while maintaining a reasonable posture.

5.2. Robustness and Generalization

As shown in Figure 2 of the supplementary material, we conduct qualitative experiments on the robustness and generalization of our method. We conduct qualitative experiments on out-of-domain (from the Internet) in extreme illumination. It can be seen that our method can reconstruct the human body in motion very well. Meanwhile, our method can still maintain high accuracy in some large pose motions.

5.3. Limitation.

For the landmark anchor loss, we constrained by the body geometry. It can therefore be used in body pose-related tasks, but cannot generalize to non-human reconstruction tasks.