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

Boyang Zhang†, Kehua Ma†, Suping Wu*, Zhixiang Yuan
School of information Engineering, Ningxia University, Yinchuan, China
boyangchang@foxmail.com; kehuamm@163.com; pswuu@nxu.edu.cn; yzxnxu@163.com

Figure 1. For complex scenes or extreme lighting conditions, our method uses human discriminative representation to recover accurate human meshes, especially in limb estimation, outperforming the state-of-the-art method TCMR. Best viewed on-screen by zooming in.

Abstract

Recovering 3D human mesh from videos has recently made significant progress. However, most of the existing methods focus on the temporal consistency of videos, while ignoring the spatial representation in complex scenes, thus failing to recover a reasonable and smooth human mesh sequence under extreme illumination and chaotic backgrounds. To alleviate this problem, we propose a two-stage co-segmentation network based on discriminative representation for recovering human body meshes from videos. Specifically, the first stage of the network segments the video spatial domain to spotlight spatially fine-grained information, and then learns and enhances the intra-frame discriminative representation through a dual-excitation mechanism and a frequency domain enhancement module, while suppressing irrelevant information (e.g., background). The second stage focuses on temporal context by segmenting the video temporal domain, and models inter-frame discriminative representation via a dynamic integration strategy. Further, to efficiently generate reasonable human discriminative actions, we carefully elaborate a landmark anchor area loss to constrain the variation of the human motion area. Extensive experimental results on large publicly available datasets indicate superiority in comparison with most state-of-the-art. The Code will be made public.

1. Introduction

3D human mesh recovery from images and videos has been widely concerned in recent years. Existing methods for estimating human pose and shape from a single image are based on parametric human models such as SMPL [17] etc, which takes a set of model parameters as input

*represents corresponding author, † represents the equal contribution.

and finally outputs a human body mesh. These methods capture the statistical information on human body shape and provide human body mesh for various applications. While these methods recover body mesh from a single image [3, 12, 15, 16] can accurately predict human pose, they may be jittery and intermittent when applied to videos. The reason for this problem is that the body pose is inconsistent over successive frames and does not reflect the body's motion in the rapidly changing complex scenes of the video. This thus leads to temporal non-smoothness and spatial non-accuracy. Several approaches [5, 7, 12, 14, 22] have been proposed to efficiently extend single image-based methods to video. They utilize different temporal encoders to learn the temporal representation directly from videos to better capture temporal information. However, these methods only encode spatial features, ignoring the effective utilization of spatial fine-grained features and human motion discriminative features. Therefore, it fails to recover a reasonable and smooth human sequence in chaotic and extreme illumination scenes. For example, TCMR [5] recovers the unsatisfactory motion on the left arm of the actor in Figure 1 in complex scenes.

The background and the human in spatial features have a complex relationship. When spatial features are input to the network, it is difficult for the network to distinguish between the human body and the background. At the same time, this relationship is not conducive to our discovery of fine-grained and discriminable features. Specifically, in extreme illumination and chaotic scenes, messy background severely interferes with human details and movement information, thus the network cannot reason about accurate human detail features in complex scenes and lacks the ability to discriminate reasonable human movements. We consider both intra-frame and inter-frame multi-level spatial representations are ideal cues to efficiently reason about spatial fine-grained information and temporal contextual discriminative information. In addition, learning to represent features at different stages is expected to strengthen the model to strip away the complex background and find human-separable motion features, thereby further improving human-specific discriminative capabilities.

Based on the above perspectives, we propose a two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos. In contrast to previous approaches using common spatial features for encoding temporal features, we attempt to segment spatial features into distinct hierarchical of spatial representations and process them separately in different stages. Specifically, the network learns and models intra-frame and inter-frame multi-level discriminative representations by segmenting spatial features along feature channels and temporal dimensions in two stages. In the first stage of the intra-frame discriminative representation, we design a dual excitation mechanism that combines self-excitation and channel excitation mechanism to simulate and activate human motion while attenuating the interferences of complex backgrounds. In addition, we design a frequency domain enhancement module to capture motion information that can highlight motion features in the frequency domain. In the second stage of inter-frame discriminative representation, we offer a new discriminative representation: the superposition of fragments, which enhances the spatio-temporal representation of past and future frames by a dynamic integration strategy, while modeling the discriminative representation of the temporal context. Furthermore, to ensure the integrity and plausibility of discriminative motion representation in consecutive frames, we also carefully design a new landmark anchor area loss to optimize the network, thereby further helping the model to reconstruct accurate 3D human actions and poses.

The core contributions of our work are as follows:

- We present a co-segmentation network based on discriminative representation for recovering human mesh from videos. Our method motivates and learns spatio-temporal discriminative features at different stages.
- In Stage 1, our proposed dual excitation mechanism and frequency domain enhancement effectively enhance human motion features and mitigate background interference. In Stage 2, we develop a dynamic integration strategy to integrate the discriminative representations of distinct stages. We also carefully design a landmark anchor area loss to constrain the generation of the reasonable pose.
- Both the quantitative and qualitative results of our method show the effectiveness of the proposed method on widely evaluated benchmark datasets in comparison with state-of-the-arts.

2. Related Works

3D Human Mesh Recovery from a Single Image. Most of the recent methods for 3D mesh recovery are based on parametric models, such as SMPL [17], SCAPE [2], etc. These methods predicted model parameters from a single image and build statistical human models. The initial work [4,8,20] predicted the 3D human body using keypoints and silhouettes. These methods usually require additional data and cannot effectively handle complex scenes or in-the-wild images. With the rapid development of DCNN performance, some methods regressed SMPL parameters directly from pixels. [12] proposed an end-to-end trainable human mesh recovery system that generated plausible 3D human meshes using adversarial loss constraints. [16] combined both HMR and SMPLify methods into a training loop to derive better results. However, when applied to video, since

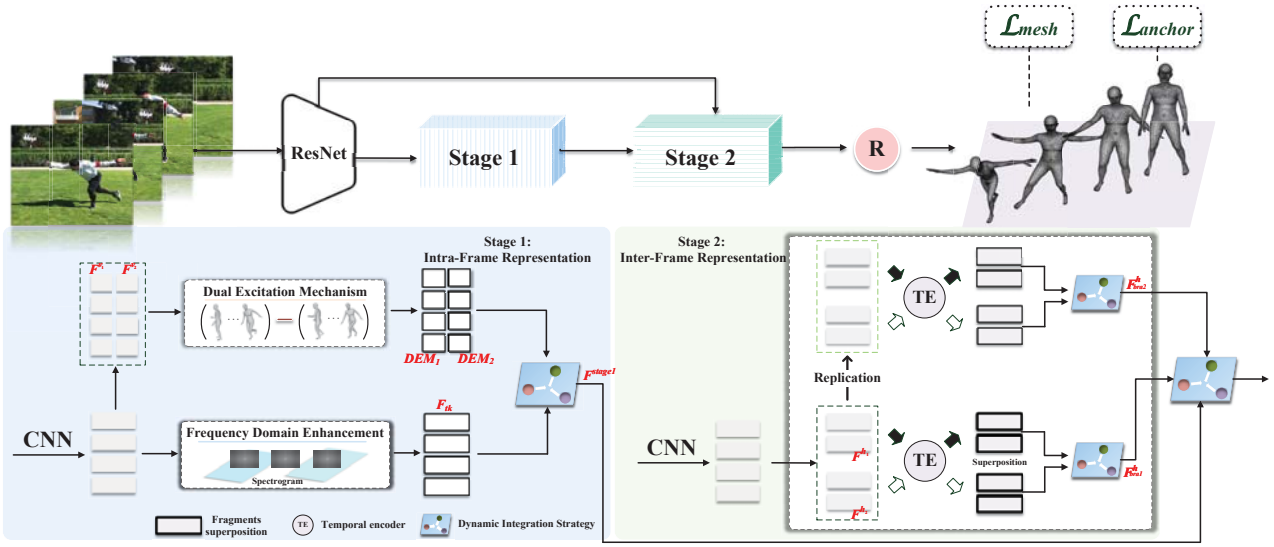

Figure 2. An overview of our framework. Our method divides the spatial features extracted by ResNet into two stages to model multi-level representations. The Stage 1 focuses on intra-frame discriminative features. Stage 2 focuses on inter-frame discriminative features. Finally, the network uses the SMPL regressor to regress the refined features to the body mesh. Best viewed on-screen by zooming in.

human poses tend to change between frames and create motion, the single-image-based approach can predict bodies with different poses in successive frames, resulting in the recovery of jittery and intermittent human mesh sequences and generating temporal inconsistency.

3D Human Mesh Recovery from Videos. Some methods exploited temporal information to estimate the body mesh in videos. [13] employed a one-dimensional fully convolutional temporal encoder to encode spatial features as temporal features, which learned the temporal representation by predicting the 3D pose of past and future frames to reduce temporal inconsistency. [22] proposed a framework for 3D mesh recovery based on skeleton disentanglement. The framework divided the 3D mesh recovery task into different spatial and temporal subproblems in a decoupled manner. Different from these methods, our method divides the spatial features into multiple granularities and simultaneously encodes them, making spatio-temporal features independently uniform. VIBE [14] used a BiGRU [6] to temporally encode spatial features throughout the video. The discriminator was also introduced to make the regressor produce a more plausible human body. MEVA [18] solved this problem from coarse to fine. This method first estimated the coarse 3D human motion using a variational motion estimator and predicted the residual motion using motion residual regression (MRR). TCMR [5] removed the residual connection from spatial features to temporal features as well as reduced temporal inconsistency. MPS-Net [26] calibrated the temporal range in the sequence in an attentional manner and then encodes spatial features. But these methods are easily interfered by complex scenes, resulting in poor human reconstruction accuracy.

3. Approach

3.1. Problem Formulation

Given a video $V_T = \{I_t\}_{t=1}^{T}$ as input, where I_t denotes the frame t, we feed each frame I_t to the feature extractor, ResNet [9] pretrained by [16], to extract spatial features $F_T = \{f_t\}_{t=1}^{T}$, where $f_t \in \mathbb{R}^{2048}$. Our task is to recover a human mesh sequence $M_T = \{\vec{\theta}_t, \vec{\beta}_t\}_{t=1}^{T}$, where $\vec{\theta}, \vec{\beta}$ denote the human pose and shape parameters at frame t.

$$M_T = \Phi(F_T) \tag{1}$$

where Φ represents our objective function for estimating the human body. To better represent the variability of the human body, we employ the SMPL parametric mesh model. SMPL provides a function $S(\vec{\theta}, \vec{\beta})$, which takes the pose parameters $\vec{\theta} \in \mathbb{R}^{72}$ and shape parameters $\vec{\beta} \in \mathbb{R}^{10}$ as input and outputs the body mesh. The model transforms the mesh M vertices to the body joints J by a mapping, $i.e.$, $J(\Theta) = WM$, where W denotes the linear regressor.

3.2. Stage 1: Intra-Frame Representation

As shown in Figure 2, in the first stage, we segment the extracted features F_T into two patches along the channel dimension, the patches F^{v_1}, F^{v_2} are shaped as $N \times T \times \frac{C}{2}$, where N is the batch size, the T and C represent temporal dimension and feature channels. We aim to obtain an intra-frame spatial multi-level representation to focus on spatial

fine-grained information by segmenting the spatial features at this stage. Intuitively, due to the variability of human actions and the uncertainty of motion scenes, these spatial features are hard to be learned by the network. By segmenting the complete spatial features into two fine-grained features, the network could simplify the complexity and difficulty of human spatial feature learning and focus on localized regions.

With these local spatial feature patches, we need to locate the human in complex scenes, i.e., focus on meaningful human motion features, while mitigating irrelevant background interference information. We consider focusing on human motion for feature representation. Many previous works have utilized optical flow methods [27] or event cameras [31] to represent motion versus capturing the motion of moving objects. They are mostly limited to specific motion capture devices or require separate networks for learning. Unlike them, our dual excitation mechanism models actual human movements by learning discriminative spatio-temporal representations. Both motion features and spatio-temporal features are combined for unified learning, and no additional motion estimation network is required. A dual excitation mechanism is shown in Figure 3. The dimensions of the input spatial features F are $N \times T \times C$. The F is first fed into the self-excitation mechanism, which is composed of dot-product attention [23]. The self-excitation mechanism first self-excites the global spatial features to learn meaningful human features. Then the network seeks motion representation among the available human features. In general, human motion can be expressed as discriminative features that can be approximately represented as the difference between two adjacent frames $F(t)$ and $F(t+1)$. Instead of simply subtracting the raw spatial features, we consider leveraging the self-excitation features.

$$MO(t) = \sigma(\frac{\theta(t) \cdot \phi(t)^{\mathrm{T}}}{\sqrt{a_k}})\psi(t) - F(t) \quad 1 \leq t \leq T \tag{2}$$

where $\theta(t) = W_\theta F(t)$, $\phi(t) = W_\phi F(t)$, $\psi(t) = W_\psi F(t)$, σ represents the softmax function, a_k denotes the reduction factor, $MO(t) \in \mathbb{R}^{N \times 1 \times C}$ is the human motion feature. By subtraction, the interferences of similar complex backgrounds are reduced. We concatenate all the motion features $MO(1)$, ..., $MO(T)$ to construct a discriminative representation MO. We then segment MO along the channel dimension to obtain MO^{v_1} and MO^{v_2}. Then a global average pooling is used to summarize time information. Since we intend to excite the discriminative feature channel, we opt to pool the temporal dimensions.

$$MO_p^{v_i} = GAP(MO^{v_i}) \quad i = 1, 2 \tag{3}$$

Ultimately, the sigmoid function is employed to obtain the distinction weights. Meanwhile, we feed the two patches F^{v_1}, F^{v_2} into the temporal encoder [6] to extract the temporal features $F_G^{v_i}$. We obtain the output of the dual excitation mechanism by channel-wise multiplication between the temporal features F_G and the discriminative weights.

$$DEM_i = F_G^{v_i} \odot \delta(MO_p^{v_i}) \quad i = 1, 2 \tag{4}$$

where $DEM_i \in \mathbb{R}^{T \times N \times \frac{C}{2}}$, δ indicates the sigmoid function. The dual excitation mechanism dynamically generates weights and then utilizes the weights to excite spatio-temporal features related to human movement and discriminative features. As weights change dynamically during training, the network is forced to perceive changes in human motion.

In addition, we enhance the spatial features using a non-parametric frequency domain enhancement module. We use the Fourier transform to perform the frequency domain enhancement. Given a sequence f_t with $t \in$ [0, T-1], the Fast Fourier transform $FFT()$ produces a new representation $\widetilde{F_k}$ as the sum of all original input features f_t. After transformation, we get the enhanced motion feature $\widetilde{F_k}$. Since the Fourier transform has a phase shift property, it makes the Fourier transform algorithm sensitive to the overall motion. When the human body and the background are present in the image, the human motion is usually reflected in the frequency spectrum in the video. In the frequency spectrum, the varying human motion can be represented as high-frequency information. Since the spectrum has correspondence with spatial motion as well as discrepancies between observations in the time and frequency domains, we use Fourier inversion $FFT^{-1}()$ to reformulate the high-frequency motion features into spatial features F_k in time domain and supplement the human motion information in the time domain. We then encode F_k to obtain F_{tk} using temporal encoder [6]. Finally, DEM_i, F_{tk} are dynamically integrated to obtain F^{stage1} and sent to the second stage. The dynamic integration strategy will be introduced in the following Section 3.3.

3.3. Stage 2: Inter-Frame Representation

In the second stage, we segment the extracted spatial features along the temporal dimension into two fragments F^{h_1} and F^{h_2}, each fragment is $N \times \frac{T}{2} \times C$. Our goal is to obtain an inter-frame spatio-temporal multi-level discriminative representation by segmenting the spatial features and modeling spatio-temporal context, thus effectively enriching the important information of inter-frame subfragment representations and capturing spatio-temporal contextual cues. We then send two fragments to the temporal encoder [6] for encoding, learning the temporal features in each sub-fragment separately. We replicate subfragments $F_{st}^{h_i}$ to obtain the same spatio-temporal features, which are

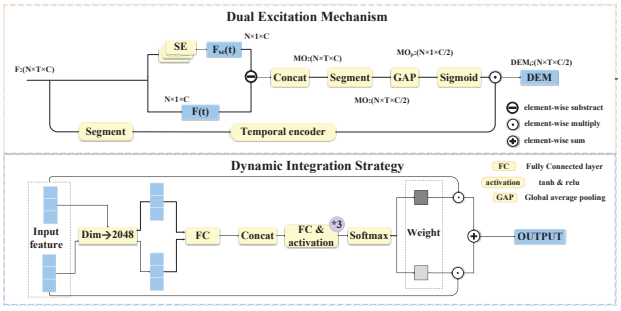

Figure 3. The upper part of the figure shows the implementation of the dual excitation mechanism (DEM), where SE stands for self-excitation. The following is the implementation of the dynamic integration strategy (DIS). Best viewed on screen when zoomed in.

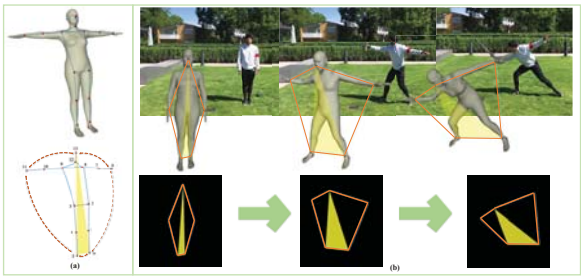

Figure 4. (a) corresponds to part of the 3d joint and SMPL kinematic tree, the numbers represent landmarks. (b) shows the sequence of human motion limb changes. The orange line and the yellow area represent the effective calculation for the landmark area loss.

grouped into two branches. In the first branch, we apply a dynamic integration strategy to the two encoded subfragments, which integrates past subfragments with future subfragments. This aims to allow the network to discover the discriminative information between the subfragments and to identify the more relevant ones. After integration, the first branch outputs F^h_{bra1}.

Human features have a similar tendency when moving. Some potential motion features will fade away with temporal tendency. Therefore, we design a fragment superimposition mechanism in the hope to enhance the spatio-temporal features. The fragment superposition mechanism adopts the same fragment addition to aggregate consistent fragments. The fragment superposition spatio-temporal features have stronger spatio-temporal information, which enables the intrinsic human features in the fragments to be enhanced and the motion features to be retained. This helps to explore the intrinsic feature information of human motion. Hence, in the second branch, we first superimpose the replicated spatio-temporal features from the two sub-fragments to obtain the superimposed, enhanced spatio-temporal features $F^{h_i}_{en}$. Then similar to the first branch, the two enhanced features $F^{h_1}_{en}$, $F^{h_2}_{en}$ are dynamically integrated to produce the second branch result F^h_{bra2}.

In order to allow the network to perceive key video fragments and provide a discriminative representation between fragments, we introduce a dynamic integration strategy. As shown in Figure 3, the dynamic integration strategy receives a set of variable, semantic independent features, then concatenates all features along the channel dimension and goes through multiple fully connected and activation layers. Ultimately we apply the softmax function to generate the adaptive discriminative weights $A_h = a^n_{i=1} \in \mathbb{R}^n$, where n is the number of input features. When different features have greater discrimination, the network will adaptively generate discriminative weights. The discriminative weights indicate that the network focuses on fragments of different importance, thus generating inter-frame discriminative representation.

At the end of the network, we obtain the final spatio-temporal representation F_{all} by dynamically integrating the multi-level spatial features F^h_{bra1}, F^h_{bra2}, F^{stage1}, where F^{stage1} is the result of the first stage. F_{all} is sent to the SMPL regressor and return to the final human body mesh.

3.4. Loss Function

To make the generated human discriminative movements closer to the real ones, besides the mesh parameter loss, we elaborate a landmark anchor area loss for the global discriminative representation of human actions. Overall, our loss aims to constrain human motion and pose to make deep features more discriminative, where contains the landmark area loss $L_{\mathcal{A}}$ and the mesh parameter loss $L_{\mathcal{M}}$.

$$L = \lambda_{\mathcal{A}} L_{\mathcal{A}} + L_{\mathcal{M}} \tag{5}$$

Landmark Anchor Area Loss. Human motion is often accompanied by changes in the relative area between joints. When the human motion in a video changes dramatically (e.g., running) or is ambiguous by background interference, it becomes critical for the network to learn this area change in order to maintain a reasonable human pose for recovery. Inspired by the human skeleton, we select five 3D landmark joints (head, left wrist, right wrist, left ankle, right ankle) that determine motion as anchor joints. These distal joints are the five most flexible joints in the human body and are heavily influenced by their parent joints. These joints react more strongly to the rotation of the parent joint when motion occurs, which is semantic and interpretable. We argue that these five joints can help the network to better learn discriminative human action information. As shown in Figure 4, our landmark anchor area loss is divided into two parts which are the anchor perimeter and anchor area. First, we calculate the anchor area, and we pick three 3D

landmark joints (head, left ankle, and right ankle) in these five anchors as the key anchors to calculate the area. Then we calculate the triangular motion area formed by these key anchors. Given the three predicted 3D anchor joints $J_i = \{J_h, J_{la}, J_{ra}\} \in \mathbb{R}^3$, we first calculate the spatial Euclidean distance $D(J_i, J_{i+1})$ between any two 3D anchor joints. Then we calculate the intermediate variable I, which is half of the sum of the three sides in the triangle. Finally, we calculate the final anchor area by the Helen formula.

$$I = \frac{1}{2}\sum_{i=1}^{3} D(J_i, J_{i+1}) \tag{6}$$

$$T_\Delta = \left(I \cdot \prod_{i=1}^{3} (I - D(J_i, J_{i+1})) \right) \tag{7}$$

Similarly, for the anchor perimeter, we calculate and sum the distances T_P formed between the five anchor joints.

$$T_P = \sum_{e=1}^{5} D(J_i, J_{i+1}) \tag{8}$$

e represents the number of edges formed by the five anchors. To this end, $l2$-norm is adopted to calculate the anchor motion area difference between the ground truth $\hat{T}_\Delta$ & $\hat{T}_P$, and predicted values T_Δ^{pred} & T_P^{pred},

$$L_\mathcal{A} = \|T_\Delta^{pred} - \hat{T}_\Delta\|_2^2 + \|T_P^{pred} - \hat{T}_P\|_2^2 \tag{9}$$

It is worth noting that our landmark anchor area loss is based on the same gesture translational invariant in the scene, so it focuses on the overall motion pose reasonably. And it is more insensitive to unalignment between 3D human joints and ground truth joints than 3D joint loss. Compared to previous works, our method not only has a joint loss as the previous method, but also constrains the perimeter and area geometry to regulate the overall human motion pose representation.

Mesh Parameter Loss. Mesh parameter loss mainly improves the learning accuracy of the network, which consists of three L2 losses between the predicted and groud-truth 2D/3D joint positions and SMPL parameters. The mesh parameter loss is derived as

$$\begin{aligned} L_\mathcal{M} = \omega_{3dj} \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2 + \omega_{2dj} \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2 \\ + \omega_{shape}\|\beta - \hat{\beta}\|_2 + \omega_{pose} \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2 \end{aligned} \tag{10}$$

where X_t stands for 3d joints, x_t for 2d joints, θ and β represent the SMPL parameters, and $\omega(.)$ denotes the corresponding loss weights.

4. Experiments

4.1. Implementation Details

For data processing and network initialization, we set the length of the input video sequence to 16 and the input frame rate to 25-30 frames per second and initialize the backbone and regression using pre-trained SPIN [16]. To identify the appropriate human regions, following [12], we use the groundtruth box for cropping in both the training and testing. The cropped image is resized to 224×224. Meanwhile, to reduce time and memory, we utilize ResNet [9] to precompute spatial features from cropped images. We train the network for 30 epochs using an NVIDIA RTX 2080Ti GPU. Our network is implemented on the PyTorch.

4.2. Evaluation Datasets and Metrics

Evaluation Datasets. We use 3DPW [24], Human3.6M [10], MPI-INF-3DHP [19], InstaVariety [13], Penn Action [29], and PoseTrack [1] for training and evaluation. 3DPW is the only in-the-wild dataset that contains accurate groundtruth SMPL parameters. Specific dataset details and implementation details are in the supplementary material.

Evaluation Metrics. For the human mesh recovery accuracy metrics per frame in the video, we calculated the mean error per joint position (MPJPE) and Procrustes-aligned MPJPE (PA-MPJPE) as the main metrics of accuracy. And we measured the Euclidean distance (MPVPE) between the ground truth vertex and the predicted vertex. In addition to this, we calculated the mean of the difference between the predicted 3D coordinates and the ground truth acceleration (Accel) for the temporal evaluation.

4.3. Comparison with State-of-the-Art Methods

Quantitative results. First, Table 1 tabulates the comparison of our approach with previous state-of-the-art (SOTA) video-based and image-based methods. All methods except HMMR are trained on a training set including 3DPW. Overall our method outperforms the previous video-based methods with respect to per-frame 3D pose accuracy. Our 3D pose errors MPJPE, and MPVPE in 3DPW (the wild dataset) are substantially reduced compared to TCMR [5]. This demonstrates that our method can effectively reconstruct reasonable and accurate human bodies in complex outdoor scenes. In the MPI-INF-3DHP dataset, our reconstruction error PA-MPJPE can still be optimal. Even in Human3.6m dataset for simple scenes, our method still outperforms TCMR.

To verify the robustness of our method, we also compare the results of the above methods without training on 3DPW. As shown in Table 2, our method outperforms the previous image-based and video-based methods in both pose accuracy PA-MPJPE and acceleration error on MPI-INF-3DHP dataset. With no in-the-wild dataset involved in the train-

Table 1. Evaluation of state-of-the-art methods on 3DPW, MPI-INF-3DHP, Human3.6m. All video methods except HMMR do not use Human3.6M ground truth SMPL parameters from Mosh, but use 3DPW train set for training. Red is the best and blue is the second best.

Method	3DPW				MPI-INF-3DHP			Human3.6M			Input Type or Frame Number
	MPJPE↓	PA-MPJPE↓	MPVPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	
EFT [11] *(2020)*	-	52.2	-	-	-	67.0	-	-	43.8	-	image
Zanfir et. al [28] *(2020)*	90.0	57.1	-	-	-	-	-	-	-	-	image
STRAPS [21] *(2020)*	-	66.8	-	-	-	-	-	-	55.4	-	image
HMMR [13] *(2019)*	116.5	72.6	139.3	15.2	-	-	-	-	56.9	-	20
VIBE [14] *(2020)*	91.9	57.6	-	25.4	103.9	68.9	27.3	78.0	53.3	27.3	16
MEVA [18] *(2020)*	86.9	54.7	-	11.6	96.4	65.4	11.1	76.0	53.2	15.3	90
TCMR [5] *(2021)*	86.5	52.7	103.2	6.8	97.6	63.5	8.5	73.6	52.0	3.9	16
MPS-Net [26] *(2022)*	84.3	52.1	99.7	7.4	96.7	62.8	9.6	69.4	47.4	3.6	16
Ours	83.4	51.7	98.9	7.2	98.2	62.5	8.6	73.2	51.0	3.6	16

Table 2. Evaluation of state-of-the-art methods on MPI-INF-3DHP. All methods do not use 3DPW in training.

	Method	MPI-INF-3DHP		
		MPJPE	PA-MPJPE	Accel
image	HMR [12] *(2018)*	124.2	89.8	-
image	SPIN [16] *(2019)*	105.2	67.5	-
image	DC-GNet [30] *(2021)*	97.2	62.5	-
video	VIBE [14] *(2020)*	97.7	63.4	29
video	TCMR [5] *(2021)*	96.5	62.8	9.5
video	TePose [25] *(2022)*	99.5	62.9	17.2
video	Ours	95.2	61.4	8.5

ing, our method can still recover accurate human mesh in complex scenes. The continuous improvement in the performance of our approach emphasizes the importance of multi-level spatio-temporal features. It is worth noting that our method improves the pose accuracy while maintaining similar acceleration errors as TCMR. More experimental results are available in the additional material.

Qualitative Results for Extreme Illumination Scenes. Figure 1 shows the qualitative results of our method in extreme illumination scenes. It can be seen that the reconstruction results of TCMR are interfered by extreme illumination scenes. Although TCMR can reconstruct the overall human mesh, it lacks the motion discrimination features of some joints and is unable to discriminate between the background and the human, resulting in insufficient integrity of the same view. Our method has better performance than TCMR in the reconstruction of moving joints such as hands and legs. This shows that our method makes reasonable use of motion discriminative representations.

Qualitative Results in Complex Scenes and Generalization. In complex scenes with indoor chaotic backgrounds, the estimated pose of our method is more reasonable than the previous SOTA methods VIBE [14] and TCMR [5] as shown in Figure 5. In complex scenes with outdoor chaotic backgrounds, our method still recovers the accurate pose from a different perspective in Figure 6. The results show that our method estimates the correct global

Figure 5. In complex scenes with indoor chaotic backgrounds, our method can recover accurate human mesh compared to [5, 14].

Figure 6. Qualitative results of our approach in complex outdoor chaotic scenes. By observing from different perspectives, our method can recover an accurate human mesh.

Figure 7. Qualitative results. For videos on the web, our method can still recover smooth and accurate human motion sequences.

body rotation. To demonstrate the generalization of our method, as shown in Figure 7, we randomly downloaded a video with motion poses in complex scenes from the Internet. We recover smooth and accurate human sequences for complex scenes out of domain (from the web). Also, the sequences have better continuity and motion discrimination. These qualitative results demonstrate that our method can reasonably utilize discriminative information to recover accurate human sequences.

4.4. Ablation Study

We conduct ablation experiments on MPI-INF-3DHP.

Table 3. The impact of the segmentation method and numbers on the network. Red is the best.

Segmentation	Number of segments	*PA-MPJPE*↓	*Accel*↓
Stage 1	Not segmented	62.6	8.5
Stage 1	4 patches	61.8	8.6
Stage 1	8 patches	62.8	8.8
Stage 2	Not segmented	62.3	8.7
Stage 2	4 fragments	61.8	10.5
Stage 2	8 fragments	63.1	11.6
Ours	2 patches & fragments	61.4	8.5

Effectiveness of Multi-level Spatial Representation. We analyze the number of patches and fragments for segmenting features. First, we do not segment the spatial representation. As shown in Table 3, the reconstruction accuracy decreases substantially when we remove either stage. Besides, in Stage 1, we segment the spatial features into 4 and 8 patches respectively. We noticed too many patches can destroy the information and structure of the original spatial features, resulting in the network's inability to learn discriminative human motion and detail representations. Again, we maintain similar settings in Stage 2. As the number of segments increases, the accuracy and temporal smoothness drop. Too short fragments cover less human sequence information, which makes it harder for the network to reason about sensible human information from the available partial fragments.

Table 4. Effects of the network designs on the performance.

Model	*PA-MPJPE*↓	*Accel*↓
Ours w/o DEM	62.2	8.7
Ours w/o Fragments Superposition	62.5	8.6
Ours w/o FDE	62.3	8.6
Ours	61.4	8.5

Effectiveness of Dual Excitation Mechanism, Fragments Superposition, Frequency Domain Enhancement. As shown in Table 4, we removed the dual excitation mechanism (DEM), and the reconstruction accuracy decreased. We can see that, given the motion-discriminative representation, the discovery of motion-sensitive features will force the network to focus on dynamic information that reflects actual human actions. Meanwhile, we remove the fragment superposition and FDE, respectively, and observe a significant decrease in accuracy. This is because the human discriminative representations gradually weaken over time. Whereas the removal of the FDE causes the network to be less sensitive to the overall motion discriminative features.

Effectiveness of Landmark Anchor Area Loss. Table 5 shows our experiments on the landmark anchor area loss. First, for the anchor area, we select the head and both wrists, five anchor joints, and randomly selected joints to calculate the anchor area. As shown in Table 5, the accuracy is reduced. We notice that anchor selection is particularly important for pose estimation. In contrast, our anchor point selection ensures maximum coverage of the human motion area. We remove the anchor area loss and anchor perimeter loss, respectively. Experiments show that the coexistence of anchor area and anchor perimeter is necessary to achieve the best performance. Finally, we removed the total landmark anchor area loss, which resulted in a decrease in accuracy. Figure 8 shows the qualitative image with and without $L_{\mathcal{A}}$. We can observe that with the constraint of $L_{\mathcal{A}}$, the distance between the wrist joint and the head is more reasonable, and the fitting is also more accurate. This shows that the change in the relative area between joints is crucial for the network to learn human motion and maintain a reasonable pose.

Table 5. The impact of landmark anchor selection on the network. $L_{\mathcal{A}}$ represents the landmark anchor area loss.

Method	Anchor Selection	*PA-MPJPE*↓	*Accel*↓
Anchor area loss	Head-double wrist	62.6	8.7
	Five anchors	62.9	8.7
	Random Anchor	62.3	8.8
	Ours (only area)	61.8	9.0
Anchor perimeter loss	Random Anchor	61.6	8.7
	Ours (only perimeter)	61.7	8.8
$L_{\mathcal{A}}$	*without*	61.9	8.6
	Ours (area + perimeter)	61.4	8.5

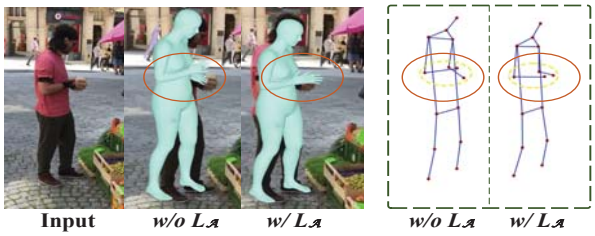

Figure 8. Mesh comparison and corresponding joints comparison. Our $L_{\mathcal{A}}$ is designed to recover a reasonable pose while being insensitive to joint alignment.

5. Conclusion

We present a two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos. We segment spatial features to obtain multi-level spatial representation with dual excitation and dynamic integration strategy to model the spatio-temporal context, and we design a landmark anchor area loss to enhance the discriminative representation. Our approach improves the accuracy of multiple challenge datasets.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant 62062056, in part by the Ningxia Graduate Education and Teaching Reform Research and Practice Project 2021, and in part by the National Natural Science Foundation of China under Grant 61662059.

References

[1] Andriluka et al. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on CVPR*, pages 5167–5176, 2018. 6

[2] Anguelov et al. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[3] Arnab et al. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 3395–3404, 2019. 2

[4] Alexandru O. Balan, Leonid Sigal, Michael J. Black, James Davis, and Horst W. Haussecker. Detailed human shape and pose from images. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2

[5] Choi et al. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 1964–1973, 2021. 2, 3, 6, 7

[6] Dey et al. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international MWSCAS*, pages 1597–1600. IEEE, 2017. 3, 4

[7] Doersch et al. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in NeurIPS*, 32:12949–12961, 2019. 2

[8] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 641–647 vol.1, 2003. 2

[9] Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on CVPR*, pages 770–778, 2016. 3, 6

[10] Ionescu et al. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on PAMI*, 36(7):1325–1339, 2013. 6

[11] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. *2021 International Conference on 3D Vision (3DV)*, pages 42–52, 2021. 7

[12] Kanazawa et al. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on CVPR*, pages 7122–7131, 2018. 2, 6, 7

[13] Kanazawa et al. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 5614–5623, 2019. 3, 6, 7

[14] Kocabas et al. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 5253–5263, 2020. 2, 3, 7

[15] Kolotouros et al. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 4501–4510, 2019. 2

[16] Kolotouros et al. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF ICCV*, pages 2252–2261, 2019. 2, 3, 6, 7

[17] Loper et al. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2

[18] Luo et al. 3d human motion estimation via motion compression and refinement. In *Proceedings of the ACCV*, 2020. 3, 7

[19] Mehta et al. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 3DV*, pages 506–516. IEEE, 2017. 6

[20] I. Sárándi, Timm Linder, Kai Oliver Arras, and B. Leibe. How robust is 3d human pose estimation to occlusion? *ArXiv*, abs/1808.09316, 2018. 2

[21] A. Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *ArXiv*, abs/2009.10013, 2020. 7

[22] Sun et al. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF ICCV*, pages 5349–5358, 2019. 2, 3

[23] Vaswani et al. Attention is all you need. In *Advances in NeurIPS*, pages 5998–6008, 2017. 4

[24] von Marcard et al. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018. 6

[25] Zhouping Wang and Sarah Ostadabbas. Live stream temporally embedded 3d human body pose and shape estimation. *ArXiv*, abs/2207.12537, 2022. 7

[26] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13201–13210, 2022. 3, 7

[27] Zach et al. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 4

[28] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *ArXiv*, abs/2003.10350, 2020. 7

[29] Zhang et al. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE ICCV*, pages 2248–2255, 2013. 6

[30] Shihao Zhou, Mengxi Jiang, Shanshan Cai, and Yunqi Lei. Dc-gnet: Deep mesh relation capturing graph convolution network for 3d human shape reconstruction. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 7

[31] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10976–10985, 2021. 4