Spatio-temporal Tendency Reasoning for Human Body Pose and Shape Estimation from Videos

Boyang Zhang
boyangchang@foxmail.com
Suping Wu*
pswuu@nxu.edu.cn
Hu Cao
caohu19980219@gmail.com
Kehua Ma
919056348@qq.com
Pan Li
18235448104@163.com
Lei Lin
2409937088@qq.com

School of information Engineering
Ningxia University
Ningxia, CHN

Abstract

In this paper, we present a spatio-temporal tendency reasoning (STR) network for recovering human body pose and shape from videos. Previous approaches have focused on how to extend 3D human datasets and temporal-based learning to promote accuracy and temporal smoothing. Different from them, our STR aims to learn accurate and natural motion sequences in an unconstrained environment through temporal and spatial tendency and to fully excavate the spatio-temporal features of existing video data. To this end, our STR learns the representation of features in the temporal and spatial dimensions respectively, to concentrate on a more robust representation of spatio-temporal features. More specifically, for efficient temporal modeling, we first propose a temporal tendency reasoning (TTR) module. TTR constructs a time-dimensional hierarchical residual connection representation within a video sequence to effectively reason temporal sequences' tendencies and retain effective dissemination of human information. Meanwhile, for enhancing the spatial representation, we design a spatial tendency enhancing (STE) module to further learns to excite spatially time-frequency domain sensitive features in human motion information representations. Finally, we introduce integration strategies to integrate and refine the spatio-temporal feature representations. Extensive experimental findings on large-scale publically available datasets reveal that our STR remains competitive with the state-of-the-art on three datasets. Our code are available at https://github.com/Changboyang/STR.git.

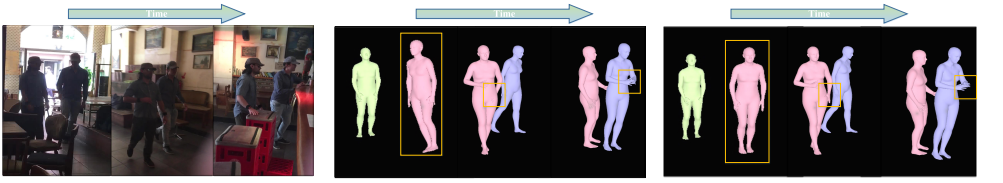

Figure 1: From left to right are the input video sequence, the reconstructed human sequence of TCMR [2], and the reconstructed human sequence of STR. STR demonstrated more realistic and smoother human action in extreme light illumination than the SOTA method TCMR.

1 Introduction and Related Work

The basic goal of 3D human body pose and shape estimation (a.k.a., 3D human motion estimation) in video aims to estimate 3D human pose and shape from motion videos, which have a wide range of computer vision applications. Existing methods are mainly based on parametric models such as SMPL [14] and SCAPE [1] to represent the human body. Depending on the input, we can roughly divide existing methods into two categories: Image-based methods [8, 11, 22] and video-based methods [2, 9, 10, 22]. The latter not only needs to ensure the single-frame image reconstruction effect but also to recover a time-smoothed human video. So it is more complicated than single-frame image reconstruction. Kanazawa et al. [9] encode temporal features via 1D convolution. Although this method obtains smoother human body sequences, the insufficient modeling of spatial features leads to lower accuracy of estimated human poses. Based on this, Sun et al. [20] propose a skeleton decoupling-based paradigm to improve spatial accuracy. Although this approach improves spatial accuracy, it neglects the grasp and reasoning of spatio-temporal feature tendency and fails to balance temporal smoothness and spatial accuracy. Kocabas et al. [10] train an adversarial learning network and use AMASS [16] to discriminate between real human motion and human motion generated by the temporal human body pose and shape regression network. Later, Choi et al. [2] abandon the strong dependence on the current static frame and propose a mesh recovery system for PoseForecast that effectively pays attention to temporal information.

However, such methods for estimating human pose and shape from videos still have limited performance on some challenging problems. As shown in Figure 1, when capturing motion images in an unconstrained environment (low natural illumination and blurred human motion), it leads to poor model parameter estimation and thus reconstructs an unreasonable human body. Although some approaches [7, 10, 19] attempt to improve performance by adding external data resources, these methods do not take full advantage of the potential information in the underlying data. Meanwhile, some methods [2, 15] are inherently limited to modeling video temporal relationships. While these methods improve the temporal consistency of human pose estimation in video, they lack spatial understanding and reasoning capabilities, leading to biased predictions. In general, during human movement, the current motion depends on the state of the previous motion and influences the subsequent motion sequences. However, when there are problems with extreme illumination or motion blur in the video, the current motion does not effectively obtain the state of the previous motion and can negatively affect future motion. Since human motion has a similar development tendency, we call the above problem tendency reasoning. We find that temporal tendency reasoning helps

to explore long dependencies between frames and to obtain information from frames with a larger temporal range. The enhancement of spatial tendency helps the network to focus more on human-related features in unconstrained scenes and mitigate background effects, thus estimating the pose accurately for each frame.

Based on the above observations and problems, we offer a spatio-temporal tendency reasoning (STR) approach for estimating human body pose and shape from videos. Our STR reason temporal and spatial tendencies separately and use integration strategies to aggregate them with each other. In particular, in the temporal tendency reasoning (TTR) module, in order to preserve the efficient dissemination of temporal tendency in motion sequences, we partition spatial features and the corresponding recurrent layers into hierarchical subsets, in which the network reasons the temporal tendency in each subset in an incremental manner. Subsets of different layers are then concatenated together by a residual structure to reason the tendency of the whole motion sequence. For the modeling of spatial tendency, we employ the means of spatial tendency enhancement to learn human movement. Existing work utilizes VAE [15, 23] and optical flow methods [24] to learn human action representations. Unlike them, our spatial tendency enhancing (STE) module models the human motion representation by calculating the difference in motion between adjacent frames. In STE, we perform time-domain spatial enhancement and frequency-domain spatial enhancement separately. Both employ motion representations to adaptively generate weights. These weights can be used to excite spatial-sensitive features in the time domain and high-frequency motion features in the frequency domain, allowing the network to uniformly learn human body spatial features and motion features. Furthermore, we introduce integration strategies to refine and integrate human features through self-integration strategy and cross-integration strategy.

Our main contributions to this work are outlined below:

- We propose a spatio-temporal tendency reasoning for human body pose and shape estimation from videos, which can alleviate the problem of human reconstruction in unconstrained scenes.
- We design a temporal tendency reasoning module and a spatial tendency enhancement module, respectively, to facilitate the effective propagation of motion information over long-distance frames and to stimulate spatially sensitive features. We also propose integration strategies module to enhance the integrated representation of different features.
- Experimentally, both the quantitative and qualitative results of our method show the effectiveness of the proposed method on widely evaluated benchmark datasets.

2 Approach

Figure 2 shows the overall pipeline of our STR. Given an input video $V = \{I_t\}_{t=1}^{T}$ of length T, we utilize the ResNet-50 [8] to extract feature vectors $F = \{f_i\}_{i=1}^{T} \in \mathbb{R}^{T\times 2048}$ of each frame. Next, F passes through the TTR and STE modules to reason the temporal tendency and enhance the spatial tendency respectively and fuses the results of the two modules through the integration network to output the enhanced tendency features. Meanwhile, F is fed into self-integration and cross-integration to output two enhanced spatio-temporal features. Ultimately, the outputs of these modules are fused through an integration network and the results are used to regress the SMPL parameters.

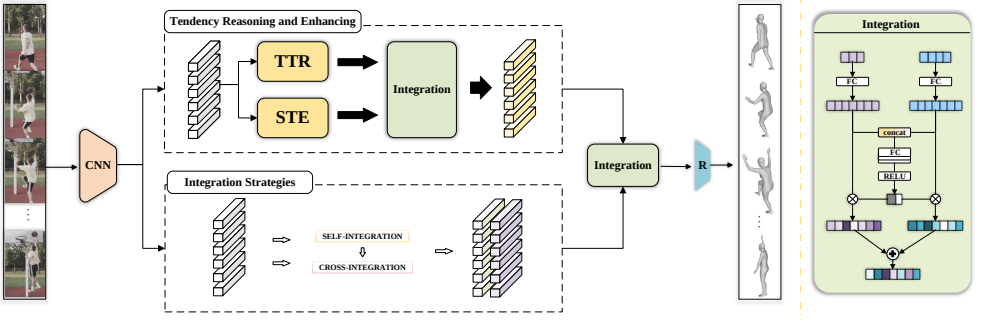

Figure 2: An overview of our framework. Given a video sequence, the aim of our method is to reconstruct the corresponding human sequence. Our method consists of two modules, a tendency reasoning enhancing module and an integration strategy module. The tendency reasoning enhancing module consists of a temporal tendency reasoning and a spatial tendency enhancing module. The integration strategies consist of a self-integration strategy and a cross-integration strategy.

2.1 SMPL Model

We employ the SMPL statistical model to characterize humans. SMPL defines a function $M(\theta,\beta)$, which takes a set of pose parameters $\theta \in \mathbb{R}^{3\times J}$ of the J skeletal joints and shape parameters $\beta \in \mathbb{R}^{10}$ as input, and outputs a full-body triangulated mesh $M \in \mathbb{R}^{N\times 3}$ with $N =$ 6890 vertices. The model transforms the mesh vertices to the body joints J by a mapping, here $J = W \cdot M$, where W is a pre-trained linear regressor.

2.2 Temporal Tendency Reasoning

As shown in Figure 3, we split F evenly into 4 sub-fragments in the temporal dimension to construct 4 sub-branches, where each fragment is shaped as $B \times \frac{T}{4} \times C$. More specifically, the F_1 sub-fragment does not undergo any operation, our TTR takes the other three sub-fragments from F_2 to F_4 as the inputs to three identical GRU [4] to learn temporal representation. To further reason temporal tendency, then for the four branches, our TTR adopts a hierarchical cascade architecture to successively fuse the results of the two adjacent branches and transmits progressively them to the next branch to generate new F_2, F_3 and F_4. We formulate this process as follows,

$$\begin{aligned} F_i^o &= F_i, && i = 1, \\ F_i^o &= GRU(F_i^o) + F_{i-1}^o, && i = 2,3,4 \end{aligned} \tag{1}$$

where $F_i^o \in \mathbb{R}^{B\times \frac{T}{4}\times C}$ is the output of i-th sub-fragment. For F_1, we do not temporally encode it to maximize the preservation of spatial features. And for F_2, we hope to supplement the current features from the relevant spatial features in F_1. Under extreme lighting conditions, fragments interact in this form to facilitate the efficient propagation of invisible information over long-distance frames. In TTR, different sub-fragments focus on different temporal tendencies in a video. TTR can aggregate temporal tendency across multiple fragments to reason temporal tendency across whole motion sequences. This not only explores long-term

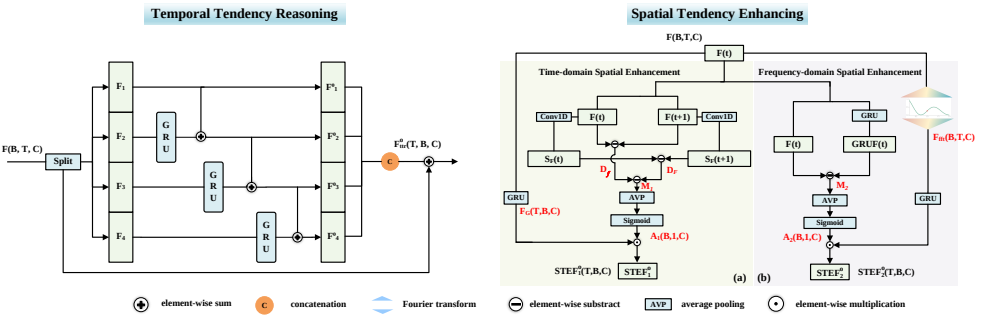

Figure 3: Illustration of the temporal tendency reasoning module (left) and spatial tendency enhancing (right). TTR module inputs feature F and outputs the reasoned spatio-temporal feature F^o_{ttr}. STE module inputs feature F and then passes through time-domain spatial enhancement (a) and frequency-domain spatial enhancement (b) respectively, finally outputting two different enhanced spatio-temporal features $STEF^o_1$, $STEF^o_2$.

dependencies between fragments but also captures information from long-distance frames. Eventually, each sub-branch is concatenated and then added to the original feature F to integrate the spatio-temporal feature representation.

$$F^o_{ttr} = F + concat(F^o_1, F^o_2, F^o_3, F^o_4), \tag{2}$$

Where $F^o_{ttr} \in \mathbb{R}^{T\times B\times C}$ is the output of the TTR module. In this way, the temporal modeling of the entire video sequence is transformed into temporal tendency reasoning, i.e., the temporal tendency of different sub-branches are combined hierarchically to form a complete temporal tendency. This temporal tendency reasoning is more conducive to the network's learning of long time sequences.

2.3 Spatial Tendency Enhancing

When encountering extreme illumination, the network cannot express human-related features well. Motion information is an important clue for understanding human behavior in videos. Spatial tendency enhancing(STE) aims to enhance human representation and distinguish human-related features by focusing on motion. We observe that the pixel values of human motion regions change over time larger than background regions. So STE exploits the temporal differences of adjacent frame-level features to focus on motion features while suppressing irrelevant information in the background. We elaborated on two parts of STE, which are time-domain spatial enhancement and frequency-domain spatial enhancement. As shown on the right of Figure 3(a), in time-domain spatial enhancement, we first use 1D convolution on feature F of shape $B\times T\times C$ to learn its time-domain spatial representation S_F.

Then our STE iteratively calculates the difference between the features of two adjacent frames to construct a difference sequence D_f. For S_F, our STE models the spatial difference sequence D_F in the same way. Finally, the two difference sequences are subtracted to

calculate the time-domain spatial representation offset M_1.

$$\begin{aligned} &M_1(t) = D_F(t) - D_f(t), && 0 < t < T \\ &D_F(t) = S_F(t+1) - S_F(t), D_f(t) = F(t+1) - F(t) && 0 < t < T \end{aligned} \tag{3}$$

where t represents a frame in a sequence of T frames. Then we leverage the global average pooling layer to aggregate temporal information and employ a sigmoid layer to learn the spatial offset weight map A_1. While learning the offset weight map, we also send the original features to the GRU [4] layer to learn temporal representation and obtain spatio-temporal features F_G. F_G is multiplied by the spatial motion weight map A_1 to obtain time-domain spatio-temporal features $STEF_1^o$.

$$\begin{aligned} &STEF_1^o(t) = F_G(t) * A_1, \\ &A_1 = sigmoid(AVP(M_1)), F_G = GRU(F) \end{aligned} \tag{4}$$

$*$ denotes the channel-wise multiplication.

The Fourier transform is sensitive to high-frequency features of human motion. For Figure 3(b), in frequency-domain spatial enhancement, we first perform the Fast Fourier Transform on F to obtain the frequency domain feature representation and then use the inverse Fast Fourier Transform to convert the frequency domain feature back to the temporal domain feature F_{fft}. We consider that the FFT-IFFT operation can preserve human information and highlight the high-frequency motion features to compensate for the lack of time-domain representation. Next, we send F and F_{fft} to the GRU [4]. The Fourier transform is sensitive to overall spatial motion, the $GRU(F)$ and F are overall subtracted to obtain the spatio-temporal offset map M_2 to model the information of the spatio-temporal difference. Finally, we apply formulas 4, M_2, F_{fft} as input, and output spatio-temporal feature $STEF_2^o$ with enhanced spatial tendency.

$$\begin{aligned} &STEF_2^o(t) = GRU(F_{fft}(t)) * A_2, \\ &A_2 = sigmoid(AVP(M_2)), \\ &M_2(t) = GRU(F(t)) - F(t), \qquad 1 \le t \le T \end{aligned} \tag{5}$$

STE enhances spatial human motion tendency by focusing on continuous frame differences and overall sequence differences. STE fully considers the properties of the time and frequency domain to model the actual motion features. Meanwhile different from using motion estimation network to learn human motion, STE uniformly learns motion features and spatio-temporal features, which can effectively enhance spatial tendency.

2.4 Integration Strategies

Integration strategies are classified into self-integration and cross-integration strategies according to the type of input. The aim is to aggregate the outputs of each component via an integration network. The integrated network is shown in Figure 2. First, the network accepts a set of spatio-temporal features as input, then these features are cascaded and passed through multiple RELU activation functions and FC layers, followed by a Softmax activation function to produce a set of weights. This set of weights is then multiplied by the corresponding features and summed to produce the integration features.

We first integrate the output of TTR, and STE to obtain the spatio-temporal tendency reasoning feature F_{STR}.

$$F_{STR} = Integration(F_{ttr}^o, STEF_1^o, STEF_2^o) \tag{6}$$

The implementation procedure of the integration strategies module is described in Algorithm 1. The integration strategy first divides the input features according to the temporal dimen-

Algorithm 1 Integration strategies

Input: All frame features F, number of integration selectable N.
Output: Enhancing features $\hat{F}$

```
/* Splitting F into c parts */
Get F^{c1}, F^{c2} = GRU(SPLIT(F))
<PHASE 1: SELF-INTEGRATION PHASE>
for i < N do
    Get F_i^{c1} = Integration(F^{c1})
    Get F_i^{c2} = Integration(F^{c2})
end for
<PHASE 2: CROSS-INTEGRATION PHASE>
Get F̂_SF = Integration(F_i^{c1}, F_i^{c2})
Get F̂_CF = Integration(F^{c1}, F^{c2})
return F̂_SF, F̂_CF
```

sion to focus on the temporal context. The divided features are then temporally encoded and passed through the self-integration phase, the self-integration process can enhance the expression of the original human features. Finally, the enhanced human features $F_i^{c_1}$ and $F_i^{c_2}$ and the original features F^{c_1}, F^{c_2} are fed into the cross-integration phase that has focused on human information at different times. The integration strategies module outputs $\hat{F}_{SF}$, $\hat{F}_{CF}$. We finally integrate the $\hat{F}_{SF}, F_{STR}$ and $\hat{F}_{CF}$ to obtain the final spatio-temporal features Z. Meanwhile, we feed Z into the SMPL regressor to regress the SMPL parameters.

$$Z = Integration(\hat{F}_{SF}, F_{STR}, \hat{F}_{CF}) \tag{7}$$

2.5 Loss Function

L2 loss was applied to 2D and 3D joint coordinates and SMPL parameters during training.

$$L_{\mathcal{G}} = \omega_{3d} \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2 + \omega_{2d} \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2 + \omega_{shape} \|\beta - \hat{\beta}\|_2 + \omega_{pose} \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2$$

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

3 Experiments

3.1 Implementation Details

Following the [2] parameters, we initialize the backbone and regressor with the pre-trained SPIN by setting the length of the input sequence T to 16. With a mini-batch size of 32, the weights are modified via the Adam optimizer. In order to save training time and memory, we pre-computed the spatial features from the images through ResNet. The initial learning rate is set at 5×10^{-5} and is reduced by a factor of 10 if the accuracy of the 3D pose does not improve after 5 epochs. With a Quadro RTX 6000 GPU, we trained the network for 30 epochs. PyTorch was used to implement the code.

3.2 Evaluation Datasets and Metrics

Evaluation Datasets. The 3DPW[21] is a 3D dataset capturing the SMPL human body in a natural scene. The MPI-INF-3DHP[17] consists of over 1.3 million frames of video of 11 people captured by 14 cameras simultaneously. Human3.6M[5] is a massive dataset of 3.6 million RGB images of 15 daily activities performed by 11 different professional actors.

Evaluation Metrics. 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. We calculated the mean of the difference between the predicted 3D coordinates and the ground truth acceleration (Accel) for the temporal evaluation.

3.3 Comparison Result and Ablation Study

Quantitative Comparison. As shown in Table 1, we first compared our proposed approach with the state-of-the-art video-based and image-based approaches on three datasets. We compare the results with or without the 3DPW dataset (in-the-wild) during training separately. When 3DPW was involved in the training, our method performs admirably on all three test datasets, with the performance on the challenging dataset (3DPW [21] and MPI-INF-3DHP [17]) being particularly impressive. From the quantitative results, we can see that our method provides spatially more accurate 3D human sequence results. In terms of temporal consistency, our approach maintains almost the same acceleration error as TCMR [2], with only a 0.1% increase. Our approach improves spatial accuracy while maintaining a similar temporal acceleration error. Our approach focuses on unconstrained scene (extreme lighting, etc.) problems. We also validate on the Human3.6m dataset (indoor), and quantitative results show that under constrained scenarios, our method still outperforms previous methods in spatial accuracy and temporal consistency.

	Method	3DPW				MPI-INF-3DHP			Human3.6M		
		MPJPE↓	PA-MPJPE↓	MPVPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓
single image	HMR [8]	130.0	76.7	-	37.4	124.2	89.8	-	88.0	56.8	-
	GraphCMR [15]	-	70.2	-	-	-	-	-	-	50.1	-
	SPIN [16]	96.9	59.2	116.4	29.8	105.2	67.5	-	-	41.1	18.3
	I2L-MeshNet [23]	93.2	57.7	110.1	30.9	-	-	-	55.7	41.1	13.4
	PyMAF [9]	92.8	58.9	110.1	-	-	-	-	**57.7**	40.5	-
video	HMMR [6]	116.5	72.6	139.3	15.2	-	-	-	-	56.9	-
	Sun *et al.* [20]	-	69.5	-	-	-	-	-	59.1	42.4	-
	VIBE (*w/o 3DPW*)[14]	93.5	56.5	113.4	27.1	97.7	63.4	29.0	65.9	41.5	18.3
	TCMR (*w/ 3DPW*) [2]	86.5	52.7	103.2	6.8	97.6	63.5	8.5	73.6	52.0	3.9
	TCMR (*w/o 3DPW*) [2]	95.0	55.8	111.3 .	**6.7**	96.5	62.8	9.5	62.3	41.1	5.3
	Lee et al. (*w/o 3DPW*) [13]	92.8	**52.2**	106.1	6.8	**93.5**	**59.4**	9.4	58.4	**38.4**	6.1
	Ours (*w/ 3DPW*)	**85.2**	52.4	**101.2**	6.9	96.3	63.1	8.6	73.3	51.9	**3.6**
	Ours(*w/o 3DPW*)	91.5	55.2	108.7	**6.7**	95.3	61.6	**8.4**	67.8	46.6	**3.6**

Table 1: Comparisons of our approach with state-of-the-art methods on 3DPW(in-the-wild), MPI-INF-3DHP(outdoor), Human3.6M(indoor) testing set. We denote whether 3DPW is involved in the training process as *w/ 3DPW*, *w/o 3DPW* respectively.

To verify the effectiveness of our method for unconstrained scenes, our approach is also compared to previous 3D pose and shape estimation algorithms that have not been trained by 3DPW [21]. In the in-the-wild dataset 3DPW, our method outperforms TCMR by about

Figure 4: Qualitative visualization of STR. The top row shows the original image samples, the middle row shows the TCMR [2] results, and the bottom row shows our results.

3.5% and 3.4% on MPJPE and MPVPE, respectively. It also continues to perform well on MPI-INF-3DHP. When our method is trained without 3DPW, the spatial accuracy and temporal smoothness are still optimal, and the relative improvement is more for ours($w/$ 3DPW training). Our method ensures the plausibility of the human pose by reasoning about the spatio-temporal tendency of human motion. When no wilderness dataset is involved in the training and the constraints become less, our method can still reason and enhance the spatio-temporal tendency in the unconstrained environments and has more robustness to outdoor scenes. Notably, Lee et al.'s method [13] generally outperforms our method in accuracy, but weaker than our method in temporal consistency. But since the method of Lee et al. [13] has no published code, we cannot make a qualitative comparison. Furthermore, the reduction of acceleration errors demonstrates the effectiveness of the proposed spatio-temporal tendency reasoning module. In particular, our method can recover smoother human action sequences compared to single image-based methods, i.e., the temporal consistency is greatly improved. In the indoor dataset, compared with the TCMR(w/o 3DPW) [2], the reason why the MPJPE and PA-MPJPE of Human3.6M [6] in Table 1 is not good in that we have not obtained the SMPL annotations of Human3.6M [6].

Qualitative comparison. In qualitative experiments, as shown in Figures 1 and 4, our method pays attention to spatial accuracy as well as temporal consistency. TCMR [2] is unable to reason spatial information from more distant frames in the extremely weaker illumination scenes. In addition, because TCMR [2] focuses too much on temporal smoothing enhancement, the human pose variation between frames is relatively small, which also leads to the bias of prediction. In contrast, our method predicts reasonable human action sequences by reasoning the human body information in the current weak illumination from the more distant visible frames. Moreover, our method has a better prediction ability for human movements, especially for limb movements (e.g., walking, arm-waving, etc.).

Discussion. Figure 5(a) shows our reconstructed out-of-dataset video sequence. Our method can predict human actions with a continuous tendency in consecutive frames and has promising generalization capabilities. The transitional properties of the actions show that our method captures the past spatio-temporal tendency and predicts the future spatio-

temporal tendency. As shown in Figure 5(b), compared to TCMR [2], our acceleration error curves are generally flat and have similar trends. As the time step increases, our acceleration error approaches the GT acceleration error. At most time steps, our acceleration error is even lower than the GT acceleration error, which indicates that our method reasons for the correct spatio-temporal tendency of human motion which validates the effectiveness of our method.

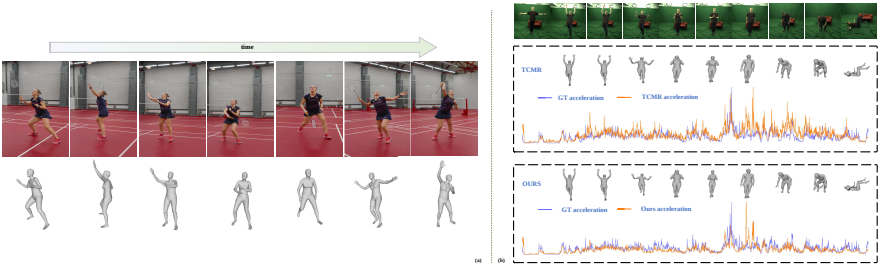

Figure 5: Subfigure (a) is our reconstructed video sequence from the web. Subfigure (b) is the comparison among TCMR, Ours, and GT acceleration errors.

Ablation Study. Table 2 shows that the acceleration error rises by 0.3 and the accuracy decreases by 0.1 after removing the TTR module. When we remove the STE module and the time-frequency domain enhancement module in STE, respectively, the PA-MPJPE increases. This indicates that the model cannot perceive the time-domain sensitive or frequency-domain high-frequency human spatial motion tendency, resulting in a decrease in accuracy. We removed the self- and cross-integration strategy from the integration strategies module and the PA-MPJPE and acceleration errors rise. This shows that the inter-frame features need to complement each other to refine the current spatio-temporal features to recover reasonable human poses.

Model	*PA-MPJPE*↓	*Accel*↓
STR w/o TTR	61.9	8.7
STR w/o STE	62.1	**8.4**
STR w/o STE (time-domin)	61.9	**8.4**
STR w/o STE (frequency-domin)	61.7	8.5
STR w/o Integration strategies(self-)	62.3	9.1
STR w/o Integration strategies(cross-)	62.7	9.0
STR	**61.6**	**8.4**

Table 2: Effects of the network designs on the performance on the MPI-INF-3DHP dataset.

4 Conclusion

We offer a spatio-temporal tendency reasoning method for human pose and shape estimation from videos. STR fully exploits human feature representations in video sequences and enhances spatio-temporal feature representations by reasoning about temporal information representations and exciting sensitive spatial features in human motion sequences. Spatio-temporal features are also refined through integration strategies. We demonstrate that our method provides smooth and accurate human motion through extensive testing.

5 Acknowledgements

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] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

[2] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021.

[3] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020.

[4] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.

[7] 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.

[8] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[9] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[10] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

[11] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

[12] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[13] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12355–12364, 2021.

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[15] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[16] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019.

[17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[18] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020.

[19] 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.

[20] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019.

[21] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[22] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human pose, shape and texture from low-resolution images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[23] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, pages 265–281, 2018.

[24] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10996–11005, 2021.