ADVERSARIAL LEARNING ENHANCEMENT FOR 3D HUMAN POSE AND SHAPE ESTIMATION

Yidian Sun , Jiwei Zhang , Wendong Wang

State Key Laboratory of Networking and Switching Technology,School of Computer Science,
Beijing University of Posts and Telecommunications, China

ABSTRACT

Adversarial learning plays an important role in recovering 3D human pose and shape from monocular videos. However, the effectiveness of this process is not often considered. Hence we aim to improve the performance of adversarial learning in 3D human pose and shape estimation. The performance of adversarial learning is mainly influenced by two parts: generator and discriminator. For the generator, we utilize temporal information on a deeper level by adding an attention-based temporal encoder in generator to model the time series of features, which contributes to a more appropriate data representation for pose and shape regression. For the discriminator, we innovatively make use of human skeleton topology information when extracting features from the estimation results. To realize this, we base the discriminator's design on the graph convolution network. In addition, to eliminate the jitter in the estimation results, we design a rotation disentangled smoothing module to process the estimated rotation parameters. We did adequate experiments on public in-the-wild datasets 3DPW and MPI-INF-3DHP. On both datasets, our method achieves higher accuracy and lower acceleration error compared with previous methods.

Index Terms— 3D human pose estimation, adversarial learning, self-attention

1. INTRODUCTION

Current methods of 3D human pose and shape estimation adopt adversarial learning widely. However, there still exists space for improvement. Adversarial learning mainly consists of two parts: generator and discriminator. For the generator, previous methods fail to fully utilize the temporal information to generate pose and shape. And for the discriminator, human skeleton topology information is ignored when distinguishing poses. In order to enhance the performance of adversarial learning, we propose following solutions with regard to these two problems.

For the generator, recent methods use CNN or RNN structure to extract temporal information [1, 2, 3]. RNN only utilizes the information before and at time point t, while the range of information that CNN can extract is limited by receptive field. Therefore we propose an attention-based temporal encoder that can make full use of temporal information. It considers not only the past information but also future information with low latency and high efficiency compared with previous designs. For the discriminator, its job is to distinguish between estimated pose from generator and real pose sampled from AMASS [4] dataset. During the process, we take human skeleton topology information into consideration. By intuition, human skeleton is a graph, so we propose a GCN-based discriminator to realize it. In addition, we propose a method to do post-processing on estimated pose called rotation disentangled smoothing. Our method can reduce the acceleration error without decreasing accuracy of estimation. This is a plug-in method that suits any 3D pose and shpae estimation method with rotation parameter output.

Our contributions can be summed up as follows: Firstly, we enhance the adversarial learning process by utilizing attention-based temporal encoder to boost generator's ability and using our GCN-based discriminator to provide better supervision to generator. Secondly, we proposed a generic smoothing method to process estimation rotation parameters. It can reduce the jitter without accuracy loss. Thirdly, we evaluate our method on two in-the-wild datasets 3DPW [5] and MPI-INF-3DHP [6] to show our improvement.

2. RELATED WORK

3D human pose estimation. Current model-based methods utilize parameterized human model called SMPL [7]. Those methods estimate body shape and pose parameters or body mesh vertex locations from input images. Kanazawa *et al.* [1] proposed an end-to-end network to estimate keypoints rotation from video frames. This method first uses pre-trained CNN to extract image feature and then use MLP to regress pose and shape parameters. They also utilize a sub-network called discriminator to add extra supervision to generator network. In order to utilize temporal information in videos, Kanazawa *et al.* [2] and Kocabas *et al.* [3] proposed temporal encoder after pre-trained CNN to extract temporal information. The method proposed in [3] also modified the design of discriminator by using GRU(Gated Recurrent Unit) and leveraging large dataset AMASS as the real pose input of discriminator. Kolotouros *et al.* [8] proposed a method to combine the optimization method and regression method in a self-improving manner.

Graph convolution network. There are extensive articles talking about generative adversarial network. Previous work [1, 2] utilize a CNN-based motion discriminator to provide supervision but it can not model motion sequences. The discriminator in [3] solved this problem but none of them considered the human skeleton topology. Recent work of graph convolutional network focus on knowledge graph embedding [9], action recognition [10] and traffic forecasting [11]. We propose a GCN-based discriminator to provide adversarial learning in our network.

3. TECHNICAL APPROACH

The whole pipeline of our proposed method is shown in Fig. 1. The input of our neural network is video, we first dump it into frames represented as $V = \{I_t, 1 \leq t \leq T\}$ where T is sequence number which is the length of input sliding window size. After we get the input frames, we use pre-trained CNN to extract image features f_t. And then, our temporal encoder models the time series of image features. Finally, we use a multi-layer perceptron to regress rotation in 6D representation [12] of key points.

The supervision of our network include 4 parts: 2D keypoints loss L_{2D}, 3D keypoints loss L_{3D}, SMPL parameters loss L_s and adversarial loss L_a. Total loss L is the weighted sum of those 4 parts.

SMPL parameters loss definition is:

$$L_s = ||\beta - \tilde{\beta}||_2 + \sum_{t=0}^{T} ||\theta_t - \tilde{\theta}_t||_2, \tag{1}$$

where β is shape parameter and θ_t is rotation parameter of body joints.

In order to support the dataset without SMPL annotation, we can get 3d joints location J_t by using SMPL body model. So the 3D keypoints loss L_{3D} is:

$$L_{3D} = \sum_{t=0}^{T} ||J_t - \tilde{J}_t||_2, \tag{2}$$

Due to the lack of 3D dataset, we make use of dataset with only 2D annotation. So we have to estimate the camera parameter to project 3D coordinate of key points to 2D space. After we got 2D pixel coordinate of key points, the 2D key points loss L_{2D} is:

$$L_{2D} = \sum_{t=0}^{T} ||x_t - \tilde{x}_t||_2, \tag{3}$$

For the adversarial loss, we choose the loss function of LS-GAN [13].

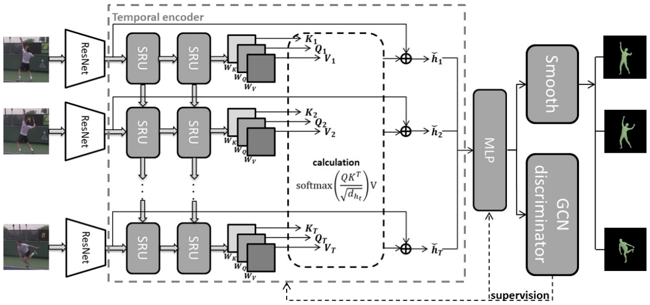

Fig. 1. The whole pipeline of our proposed method. First, we choose ResNet50 as backbone. And then is attention-based temporal encoder. The GCN-based discriminator provide extra supervision to network.

3.1. Attention-based temporal encoder

This module is after the feature extractor $\mathcal{F}(I_t) = f_t \in \mathbb{R}^{2048}$ where $t \in [1, T]$ and I_t is the frame at time t. Since the feature extractor extracts image feature individually for each time slot, we need to consider the temporal information between consecutive frames. A natural interpretation of this idea is that human pose at time t will be influenced by the pose before time t. Therefore we use simple recurrent units(SRU) [14] to extract the relationship between frames, so the feature vector we use to regress human pose at time t will contain the information before time t.

We use multi-layer SRU with skip connection to encode image features to hidden state h_t. We choose SRU as our encoder because its processing time is much short than other common architecture such as LSTM [15] and GRU [16]. Because SRU simplified the gate design making its parameters fewer than LSTM and GRU. Although we use SRU to extract temporal information between each time slot, for hidden state h_t, it will only contain the information before t if we do not use bi-directional RNN architecture. But if we use bi-directional structure the latency will be much higher than before. So we propose a self-attention mechanism to solve this problem. First, we need three projection layers W_Q, W_K, W_V to calculate the projection of h_t. After projection is the activation function. We choose leaky-relu for best performance.

$$Q = lrelu(W_Q h_t), \tag{4}$$

$$K = lrelu(W_K h_t), \tag{5}$$

$$V = lrelu(W_V h_t), \tag{6}$$

After we obtain three representations Q, K, V, we follow the [17] method, using the dot-product attention to calculate attention weight:

$$\hat{h}_t = softmax(\frac{QK^T}{\sqrt{d_{h_t}}})V, \tag{7}$$

where d_{h_t} is the dimension of SRU hidden state and $\hat{h}_t$ is the final representation to regress human pose parameters.

3.2. GCN-based discriminator

In order to adopt adversarial learning, we need a discriminator to provide extra supervision for our generator network. The input of discriminator contains two parts: the estimated pose sequence from our generator called fake pose, and ground-truth pose sequence from our dataset called real pose. The outputs of discriminator are the scores of those inputs. The score near zero means it is fake pose and the score near one means it is real pose. So the discriminator's work is to distinguish between the real pose and fake pose.

When distinguishing human poses, there are two important aspects: temporal information and human skeleton topology. So we propose a GCN-based discriminator to extract both temporal and skeleton information. We use GLU(Gated Linear Units) to extract temporal information and a graph layer [11] to extract skeleton information. Input of our discriminator is 3D joints $x_{fake}, x_{real} \in \mathbb{R}^{B\times T\times J\times 3}$ where B is batch size, T is sequence length and J is number of joints. For SMPL model J equals to 24. The reason why we use 3D joints location as the inputs is that the graph relationship of joints location outperforms joints rotation.

3.3. Rotation disentangled smoothing

Although current methods have high accuracy, the estimation results are often accompanied by random jitter. Acceleration error is the metric to measure the degree of jitter. By analyzing the cause of jitter, we find it can be divided into rotation axis jitter and rotation angle jitter. So we propose dealing with them separately, using the method called rotation disentangled smoothing. We need to convert estimation result to axis-angle representation $\vec{\theta}_t = \{\vec{\theta}_t^j, j \in [0, 24)\}$ and then we divide it into two parts: rotation axis $\vec{\varphi}_t = \{\vec{\varphi}_t^j, j \in [0, 24)\}$ and rotation angle $a_t = \{a_t^j, j \in [0, 24)\}$. Rotation axis $\vec{\varphi}_t^j$ is a unit vector and rotation angle is a scalar of joint j in time t. The relationship between these two parts is:

$$\vec{\theta}_t^j = a_t^j \vec{\varphi}_t^j = [x, y, z], \tag{8}$$

$$a_t^j = \sqrt{x^2 + y^2 + z^2}, \tag{9}$$

$$\vec{\varphi}_t^j = [\frac{x}{a_t^j}, \frac{y}{a_t^j}, \frac{z}{a_t^j}], \tag{10}$$

After we get these two parts, we utilize 1 euro filter [18] to process rotation axis and rotation angle respectively:

$$\hat{a}_t^j = filter(a_t^j), \tag{11}$$

$$\hat{\vec{\varphi}}_t^j = filter(\vec{\varphi}_t^j), \tag{12}$$

Using the smoothed rotation axis and angle, we can apply equation (8) to recover axis-angle representation.

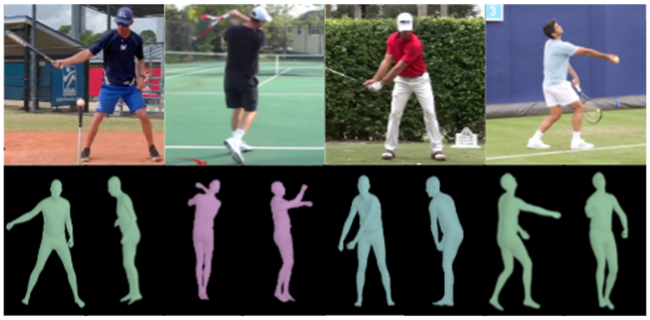

Fig. 2. Qualitative evaluation results of some out-door scenes: baseball, tennis, golf and serve. Left-side is original view, right-side is view rotated by 90 degrees.

4. EXPERIMENTAL EVALUATION

For the training part, our dataset contains 3 parts: 3D dataset, 2D dataset and 3D un-pair dataset. Frames in 3D dataset and 2D dataset are the inputs of our generator. For 3D datasets, we use MPI-INF-3DHP for training. 3DPW is not used for fair comparison. For 2D datasets, we use Insta-variety for training. For the 3D un-pair datasets, we only use the annotations in AMASS as the discriminator's input, i.e. x_{real}.

For the evaluation part, we test on MPI-INF-3DHP and 3DPW test sets. We focus on 3 metrics: (1)MPJPE: mean per joints position error. (2)PA-MPJPE: mean per joints position error after Procrustes alignment. (3)accel error: acceleration error which measures the jitter of the estimated pose.

4.1. Comparison with previous methods

As shown in Table 1, Temporal-HMR [2] method trained without 3DPW dataset achieved low acceleration error but large PA-MPJPE in 3DPW test set, while VIBE [3] method of high accuracy trained under the same condition results in high acceleration error. So the estimated pose will contain much jitter. Our method can achieve lower acceleration error and higher accuracy than VIBE. When testing on the MPI-INF-3DHP dataset, our method can still achieve high accuracy with lower acceleration error.

4.2. Ablation study

Baseline in Table 2 and 3 is the original VIBE implementation training without 3DPW dataset. The rest of the experiments in this chart all use the same training datasets.

As shown in Table 2 and 3, for SRU+atten(attention) configuration, it is about how our temporal encoder affects the estimation result. Fig. 3 shows that for each time slot, the weight is different, which suggests the importance of different time slots varies. 3D visualization shows that no slot has zero weight, thus all the information across the timescale is utilized.

	3DPW				MPI		
Models	MPJPE	PA-MPJPE	PVE	Accel	MPJPE	PA-MPJPE	Accel
HMR [1]	130.0	76.7	-	37.4	89.8	72.9	-
Kanazawa *et al.* [2]	116.5	72.6	139.3	**15.2**	-	-	-
Kolotouros *et al.* [8]	96.9	59.2	116.4	29.8	105.2	67.5	-
VIBE [3]	93.5	56.5	113.4	27.1	97.7	63.4	-
Ours	**92.6**	**55.2**	**111.9**	22.0	**96.8**	**62.2**	**28.6**

Table 1. Comparison with previous methods on the 3DPW and MPI-INF-3DHP test set.

Models	MPJPE	PA-MPJPE	Accel
Baseline(VIBE)	93.5	56.5	27.1
SRU+atten	94.1	55.8	27.8
SRU+atten+GCN	92.8	55.3	27.7
Ours	**92.6**	**55.2**	**22.0**

Table 2. Impacts of different components on 3DPW. "Ours" stands for SRU+atten+GCN+smooth.

Models	MPJPE	PA-MPJPE	Accel
Baseline(VIBE)	97.7	63.4	-
SRU+atten	97.8	63.0	32.1
SRU+atten+GCN	97.0	62.7	32.0
Ours	**96.8**	**62.2**	**28.6**

Table 3. Impacts of different components on MPI-INF-3DHP.

In Fig. 4, the loss curves show the improvement of our method compared with baseline. For the baseline, the adversarial loss of generator increases to 0.5 in only a few epochs and the loss of discriminator also decreases to 0. According to the baseline experiments settings, the maximum value of generator adversarial loss is 0.5, and the minimum value of discriminator loss is 0. Intuitively, this phenomenon means discriminator becomes optimal in few epochs. So it can distinguish estimated pose sequence and real pose sequence. The distribution of estimation results deviates from ground truth till the end of training. There is no confrontation between generator and discriminator in the most time. The generator is too weak to generate pose with high accuracy.

Theoretically, when the discriminator becomes optimal, the loss function of generator is equivalent to Pearson chi-squared divergence. When the distance between the distribution of estimated pose and real pose enlarges, the gradient of divergence tends to zero. It means there is no supervision to generator during the most time of training process. As shown in the right part of Fig. 4, the loss curve of our method becomes reasonable. The performance of adversarial learning has been improved. Generator can make better estimation than before.

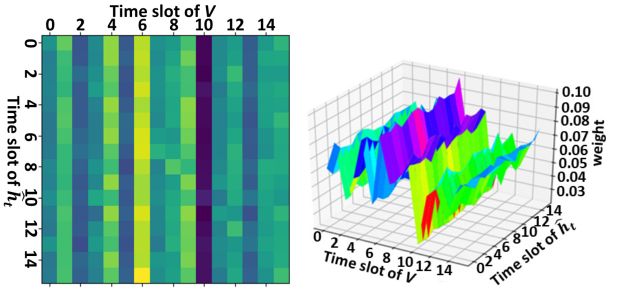

Fig. 3. Visualization of the attention weights of self-attention.

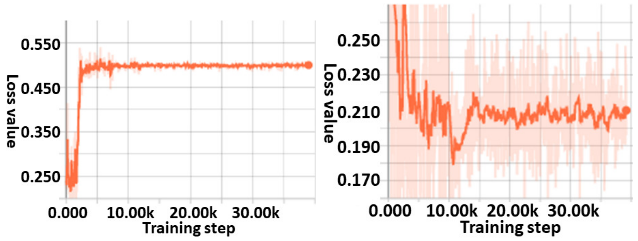

Fig. 4. The generator's adversarial learning loss curves of VIBE(left) and our(right) method.

5. CONCLUSION

We introduce an attention-based temporal encoder to generator part and devise a GCN-based discriminator which provides better supervision. Those two parts enhance the adversarial learning process. We design a rotation disentangled smoothing module to smooth estimated rotation parameters without degradation of estimation accuracy. Finally, we evaluate our method on two in-the-wild datasets 3DPW and MPI-INF-3DHP, which shows improvement in estimation accuracy and acceleration error.

6. ACKNOWLEDGMENTS

This work is supported by The Action Plan Project for Basic Scientific Research Business Fee of Central University to Improve Scientific and Technological Innovation Ability (No. 2019XD-A12) and the Beijing Municipal Natural Science Foundation-Haidian Original Innovation Joint Fund (No.L182034).

7. REFERENCES

[1] 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*, 2018, pp. 7122–7131.

[2] 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*, 2019, pp. 5614–5623.

[3] 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*, 2020, pp. 5253–5263.

[4] 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*, 2019, pp. 5442–5451.

[5] 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)*, 2018, pp. 601–617.

[6] 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)*, 2017, pp. 506–516.

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

[8] 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*, 2019, pp. 2252–2261.

[9] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo, "Knowledge graph embedding: A survey of approaches and applications," in *IEEE Transactions on Knowledge and Data Engineering*, 2017, pp. 2724–2743.

[10] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018, pp. 7444–7452.

[11] Bing Yu, Haoteng Yin, and Zhanxing Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 3634–3640.

[12] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[13] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[14] Tao Lei, Yu Zhang, Sida I Wang, Hui Dai, and Yoav Artzi, "Simple recurrent units for highly parallelizable recurrence," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4470–4481.

[15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," in *Neural computation*, 1997, pp. 1735–1780.

[16] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] Géry Casiez, Nicolas Roussel, and Daniel Vogel, "1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 2527–2530.