Video-based body geometric aware network for 3D human pose estimation*

LI Chaonan, LIU Sheng, YAO Lu, and ZOU Siyu**

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

(Received 3 February 2022; Revised 10 March 2022)

Three-dimensional human pose estimation (3D HPE) has broad application prospects in the fields of trajectory prediction, posture tracking and action analysis. However, the frequent self-occlusions and the substantial depth ambiguity in two-dimensional (2D) representations hinder the further improvement of accuracy. In this paper, we propose a novel video-based human body geometric aware network to mitigate the above problems. Our network can implicitly be aware of the geometric constraints of the human body by capturing spatial and temporal context information from 2D skeleton data. Specifically, a novel skeleton attention (SA) mechanism is proposed to model geometric context dependencies among different body joints, thereby improving the spatial feature representation ability of the network. To enhance the temporal consistency, a novel multilayer perceptron (MLP)-Mixer based structure is exploited to comprehensively learn temporal context information from input sequences. We conduct experiments on publicly available challenging datasets to evaluate the proposed approach. The results outperform the previous best approach by 0.5 mm in the Human3.6m dataset. It also demonstrates significant improvements in HumanEva-I dataset.

Document code: A **Article ID:** 1673-1905(2022)05-0313-8

DOI https://doi.org/10.1007/s11801-022-2015-8

Three-dimensional human pose estimation (3D HPE) provides abundant human 3D structure information, and has become a crucial and hot research topic in recent years. 3D HPE has broad application prospects in the fields of trajectory prediction, posture tracking and action analysis. Despite the tremendous success[1-4] of well-designed deep learning paradigms in recent years, the 3D HPE task still faces challenges from the substantial depth ambiguity and the frequent self-occlusions in the two-dimensional (2D) representations. Previous approaches[5-8] typically decomposed the 3D HPE task into 2D HPE and 2D-to-3D pose lifting. Decoupling the task seems to reduce the difficulty of the problem. However, in some cases with severe self-occlusions, it aggravated the impact of depth ambiguity on the 3D HPE task, since multiple 3D poses could be mapped to the same 2D skeleton.

To alleviate the depth ambiguity, several early methods[9-11] utilized geometric constraints among different body joints to ensure the network can predict a plausible 3D pose. Graph convolutional network (GCN)[12-14] is utilized to capture the implicit kinematic information in 2D keypoints. For graph-structured data, it has a good feature extraction capability. However, CI et al[15] proposed that the inherent weight sharing scheme in GCN limited its feature representation ability, resulting in its poor performance on the 3D HPE task. In this paper, a novel attention mechanism is proposed to effectively improve the spatial feature representation ability of the network and further alleviate the depth ambiguity.

In addition, we find that existing methods for video 3D HPE often yield incoherent and jittery predictions. A major reason behind this is the high variability and nonlinearity of human dynamics that cause the frequent occurrence of self-occlusions. Recently, exploiting the temporal context information among consecutive frames to mitigate the effects of self-occlusions has been demonstrated as an effective technique[7,16,17]. WANG et al[16] proposed a U-shaped GCN network to learn the long-short term motion information among consecutive frames and achieved highly competitive results. Several methods[7,17] rely on dilated temporal convolutions to model long-term temporal context features and achieve significant performance improvement. But these methods inherently have limited temporal connectivity. Recently, the multilayer perceptron (MLP)-Mixer[18] is exploited for image classification due to its high efficiency, excellent scalability and powerful feature modeling capabilities. However, how to take advantage of the capacity of the MLP-Mixer for 3D HPE remains a challenging task.

Furthermore, existing works ignore the fact that spatial configuration constraints and temporal correlations are two types of complementary information. It is obviously a sub-optimal solution to consider only one of them. In

* This work has been supported by the National Key R&D Program of China (No.2018YFB1305200).

** E-mail: edliu@zjut.edu.cn

this work, to take full advantage of spatial-temporal information, we design a novel video-based human body geometric aware network for the 3D HPE task. We introduce a novel skeleton attention (SA) mechanism which is used to adaptively identify the weight of the joints in the pose graph. Next, our network learns the prior knowledge of human structure based on the SA mechanism to improve the spatial representation ability of the network. Moreover, we design a novel temporal MLP-Mixer module to model the long-range temporal context dependencies among each frame in the entire sequence. Our approach exploits the spatial context constraints as well as the temporal context consistency for 3D HPE. The spatial-temporal information can effectively alleviate the depth ambiguity and self-occlusion problems, and significantly improves the accuracy.

Recent state-of-the-art approaches[5,6,19] typically rely on the off-the-shelf 2D human keypoints detectors to first detect the 2D keypoints from the image and then lift the 3D pose from the predicted 2D joints. MARTINEZ et al[5] directly predicted the 3D human pose based on 2D keypoints via a simple but effective fully connected residual neural network. CHEN et al[19] treated the 3D HPE task as a data-driven matching problem, and exploited the nearest-neighbor algorithm to retrieve the optimal 3D human pose from the skeleton pool which is generated from 2D keypoints. Since these approaches benefit from intermediate supervision, they significantly outperform those methods that directly estimate 3D pose from images. Therefore, we adopt the two-stage pose estimation paradigm for the 3D HPE task. Due to the excellent feature modeling performance of GCN on non-European data, several GCN-based works[12,13] model 2D keypoints to extract global and local body geometric constraint features of each joint for the 3D HPE task. However, CI et al[15] first proposed that the internal weight sharing scheme in GCN would hinder the feature representation capability of GCN. Similar to GCN, our method relies on the topology of the pose graph to effectively capture human body geometric features. But our work has the following three distinct features. Instead of constructing an adjacency matrix to present the structure of human body topology, our approach builds a more generalized matrix to represent the correlation between 2D joints. The proposed SA mechanism is employed to guide the fully connected neural network to predict the 3D human pose from the 2D keypoints. The pose encoder module essentially forms a human pose graph, where the weight of each joint is dynamically adjusted using the SA mechanism.

Utilizing temporal information[20-23] from videos to mitigate the effects of self-occlusions is an effective technique. HOSSAIN et al[2] adopted long short-term memory (LSTM)[24] cells to construct a recurrent neural network. It was exploited to capture the temporal consistency over a sequence. PAVLLO et al[7] adopted dilated temporal convolution to achieve a larger temporal receptive field, which significantly improves the accuracy. LIU et al[17] proposed a novel temporal attention mechanism to adaptively discriminate significant frames for efficiently extracting temporal consistency among frames. MLP-Mixer[18] was designed for the image recognition tasks. Compared with state-of-the-art convolutional networks, MLP-Mixer treats an image as a sequence of patches and achieves remarkable results. In this work, unlike existing temporal-based methods[7,17], which relied on dilated temporal convolutions to capture temporal dependencies, we utilize MLP-Mixer for learning temporal context information across multiple frames.

To further mitigate the impact of depth ambiguity and self-occlusions on the 3D HPE task. Some methods[20,23] utilized the spatial-temporal information in the 2D skeleton sequence to improve prediction accuracy. ZHENG et al[20] proposed a novel spatial-temporal network architecture that utilized graph convolution and temporal convolution to alternately extract geometric constraint information and temporal context information implicit in 2D skeleton sequences. LIU et al[23] designed a transformer-based spatial-temporal network architecture, which can effectively extract the spatial correlation from 2D keypoints and the temporal context information from all input frames. All of the above methods achieved significant performance improvements. Inspired by this, our approach effectively integrates both temporal and spatial correlations into a neural network and utilizes it for the 3D HPE task.

The overall network framework is illustrated in Fig.1. In the body geometric aware module, the coordinates of each joint in the 2D skeleton are first projected to the high-dimensional embedding space through linear projection operation, and then N stacked pose encoders are utilized to extract the geometric dependencies of each joint in the 2D skeleton. Temporal MLP-Mixer module is used for learning global temporal context dependencies from the entire 2D skeleton sequence. In the end, a regression head is utilized to regress the target 3D pose. Our approach takes 2D skeleton sequences $\boldsymbol{X} \in \mathbb{R}^{f \times J \times 2}$ as input, where f is the input sequence length, J is the number of joints in a 2D skeleton, and it predicts the 3D joint positions in the target frame.

The goal of the body geometric aware module (BGAM) is to learn the geometric dependencies among different body joints from a single 2D skeleton. Following prior works[18,25], the coordinates of each joint are viewed as an individual patch and projected to the high-dimensional embedding space. Specifically, given the 2D pose data $\boldsymbol{x}^i \in \mathbb{R}^{J \times 2}$ of the ith frame in $\boldsymbol{X}$, a linear projection operation is first utilized to transform the 2D coordinates of each joint in $\boldsymbol{x}^i$ into high-dimensional space. After that, we get $\boldsymbol{\theta}_0^i \in \mathbb{R}^{J \times C_s}$, where C_s is the channel dimension, $\boldsymbol{\theta}_0^i$ is then sent to N stacked pose encoders to further extract the spatial configuration constraints.

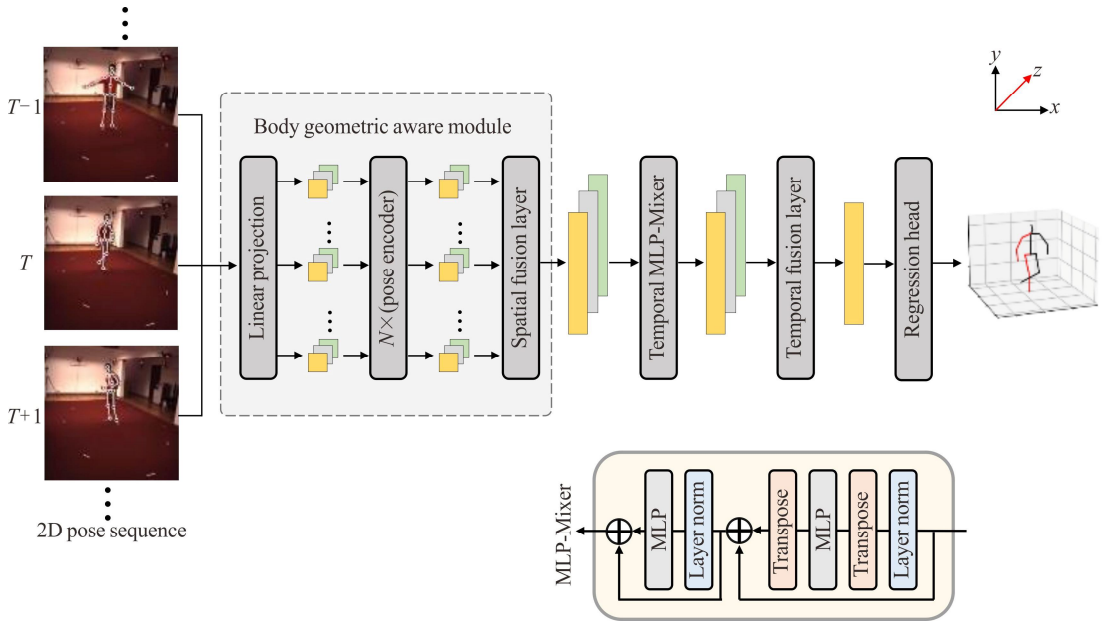


Fig.1 Overall network framework of the proposed scheme

The overall pose encoder architecture is illustrated in the Fig.2(a). It integrates the features of multiple neighbor joints to further enhance the spatial representation ability of the network via the SA mechanism and prevents an implausible pose prediction. To take advantage of the prior knowledge of human structure, we construct a structure matrix $\boldsymbol{S} \in \mathbb{R}^{J\times J}$ with the following formula:

$$S_{(q,p)} = \begin{cases} \dfrac{1}{2^{(k-1)}}, & MD(q,p) \le K \\ 0, & MD(q,p) > K \end{cases}, \tag{1}$$

where $S_{(q,p)}$ is the element in the qth row and pth column in the structure matrix, $MD(q,p)$ is the manifold distance between the joint q and p predefined in the pose graph as shown in Fig.2(b), K is the predefined hyperparameter. For instance, according to the human pose graph illustrated in Fig.2(b), the manifold distance of the directly connected joints left-hip and hip is 1, and the manifold distance of the right-hip and left-hip is 2, because they are separated by hip. The structure matrix is utilized to generate an attention matrix which is used to weigh the importance of different joints in the pose graph as

$$\boldsymbol{W}_{\text{Att}} = \text{sig}(\boldsymbol{W}_s \text{Fla}(S)), \tag{2}$$

where $\boldsymbol{W}_{\text{Att}} \in \mathbb{R}^{J^2\times 1}$ is an attention matrix, $\boldsymbol{W}_s \in \mathbb{R}^{J^2\times J^2}$ is a learnable matrix, sig(·) is a sigmoid activation function, and Fla(·) is a flatten function. Given a feature matrix $\boldsymbol{\delta} \in \mathbb{R}^{J\times C_s}$, the core function SA(·) in the pose encoder can be formulated as

$$\text{SA}(\boldsymbol{\delta}) = \boldsymbol{W}_2(\boldsymbol{W}_{\text{Att}} \otimes \sigma(\boldsymbol{W}_1\boldsymbol{\delta})), \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{J^2\times J}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{J\times J^2}$ are learnable matrixes, σ is an element-wise nonlinearity (Gaussian error linear units (GELU)[26]), and $\otimes$ is an element-wise multiplication operation. Thus the output of N stacked pose encoders can be expressed as

$$\text{MLP}_S(\boldsymbol{\zeta}) = \boldsymbol{W}_4\sigma(\boldsymbol{W}_3\boldsymbol{\zeta}^{\text{T}}), \tag{4}$$

$$\boldsymbol{\varphi}^i_{\ell_s} = \boldsymbol{\theta}^i_{\ell_s-1} + \text{SA}(\text{LN}(\boldsymbol{\theta}^i_{\ell_s-1})), \quad \ell_s = 1, 2, ..., N, \tag{5}$$

$$\boldsymbol{\theta}^i_{\ell_s} = \boldsymbol{\varphi}^i_{\ell_s} + \text{MLP}_S(\text{LN}(\boldsymbol{\varphi}^i_{\ell_s}))^{\text{T}}, \quad \ell_s = 1, 2, ..., N, \tag{6}$$

where LN(·) represents the layer normalization operator, $\boldsymbol{W}_3 \in \mathbb{R}^{C_s\times C_s}$ and $\boldsymbol{W}_4 \in \mathbb{R}^{C_s\times C_s}$ are learnable matrixes, and MLP_S represents the MLP block operator in the pose encoder. The feature $\boldsymbol{\theta}^i_N \in \mathbb{R}^{J\times C_s}$ is generated after passing through each pose encoder where the network can continuously extract the spatial dependencies of each joint in a 2D skeleton.

As the final layer in BGAM, the spatial fusion layer first normalizes $\boldsymbol{\theta}^i_N$ by the LN(·) operator and then concatenates the features of each joint in $\boldsymbol{\theta}^i_N$ to form a new feature vector $\boldsymbol{\gamma}^i \in \mathbb{R}^{1\times(J\times C_s)}$. After that, $\boldsymbol{\gamma}^i$ will be forwarded to the temporal MLP-Mixer module which learns long-range temporal relationships across frames.

Utilizing the temporal information in the video can significantly reduce jittery and incoherent predictions. Since the BGAM encodes rich spatial features for every frame, the temporal MLP-Mixer module is design for modeling temporal consistency among each frame in the entire input sequence. The BGAM processes all input frames in parallel and produces a feature set $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}^1, \boldsymbol{\gamma}^2, ..., \boldsymbol{\gamma}^f\}$, and then we concatenate each vector in $\boldsymbol{\gamma}$ to form a new feature $\boldsymbol{\beta}_0 \in \mathbb{R}^{f\times(J\times C_s)}$.

As the core modules of the MLP-Mixer, channel-mixing MLP and token-mixing MLP are designed to enable communication between different channels and allow interaction between different spatial locations, respectively. Specifically, the channel-mixing MLP acts on rows of $\boldsymbol{\beta}_0$ and shares information across all rows. Moreover, it enables cross-channel interaction which is important to efficiently learn channel attention. The token-mixing MLP acts on

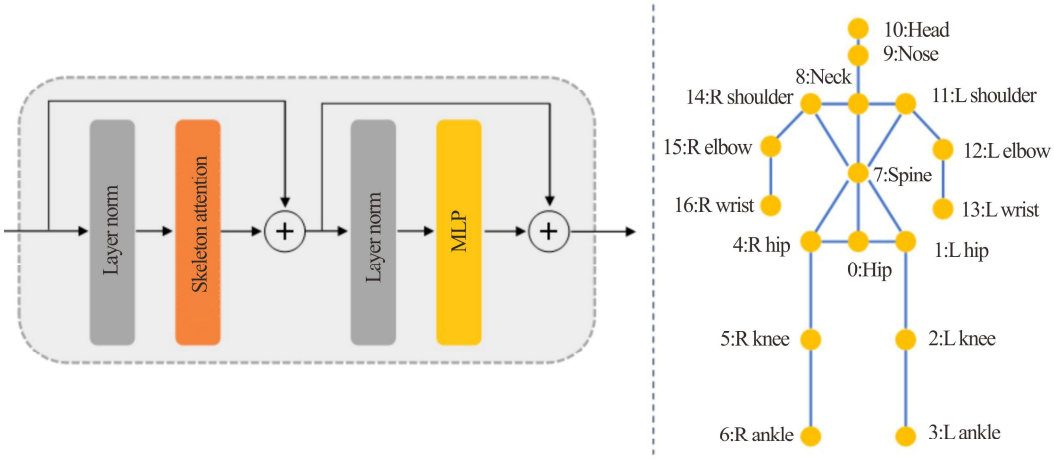
columns of $\boldsymbol{\beta}_0$ to share spatial information across all columns. In our case, it interleaves features of each frame to enable interaction between input sequences. Each MLP block contains two fully-connected layers and a GELU activation function applied independently to each row of its input tensor. Thus the operation of the channel-mixing MLP can be expressed as

$$C_M(\boldsymbol{X}) = \boldsymbol{W}_4\sigma(\boldsymbol{W}_3\mathrm{LN}(\boldsymbol{X}^{\mathrm{T}})), \tag{7}$$

and the token-mixing MLP can be expressed as

$$T_M(\boldsymbol{X}) = \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\mathrm{LN}(\boldsymbol{X})), \tag{8}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{D_s\times f}$, $\boldsymbol{W}_2 \in \mathbb{R}^{f\times D_s}$, $\boldsymbol{W}_3 \in \mathbb{R}^{D_c\times(J\times C_s)}$ and $\boldsymbol{W}_4 \in \mathbb{R}^{(J\times C_s)\times D_c}$ are learnable matrixes, D_S and D_C are hidden dimensions, $T_M(\boldsymbol{X})$ and $C_M(\boldsymbol{X})$ are token-mixing and channel-mixing operations, respectively.



(a) Pose encoder (b) Pose graph

Fig.2 (a) Architecture of the pose encoder; (b) Pose graph used in the pose encoder

Our temporal MLP-Mixer module consists of L identical MLP-Mixer layers. Given an embedded feature $\boldsymbol{\beta}_0$, the processing procedure of temporal MLP-Mixer module can be formulated as

$$\boldsymbol{U}_\ell = T_M(\boldsymbol{\beta}_{\ell-1}) + \boldsymbol{\beta}_{\ell-1}, \quad \ell = 1,\ 2,\ \ldots,\ L, \tag{9}$$

$$\boldsymbol{\beta}_\ell = C_M(\boldsymbol{U}_\ell)^{\mathrm{T}} + \boldsymbol{U}_\ell, \quad \ell = 1,\ 2,\ \ldots,\ L, \tag{10}$$

where T represents a transpose operation on tensors.

Temporal fusion layer aims to fuse the extracted temporal features $\boldsymbol{\beta}_L \in \mathbb{R}^{f\times(J\times C_s)}$. To this end, it takes the average of the $\boldsymbol{\beta}_L$ along the frame dimension to get a fusion vector $\boldsymbol{\gamma}_{\mathrm{out}} \in \mathbb{R}^{1\times(J\times C_s)}$.

The goal of the regression head is to regress the body joint locations in 3D space from feature vector $\boldsymbol{\gamma}_{\mathrm{out}}$. To this end, we adopt a layer norm followed by a linear layer to regress the 3D pose of the target frame $\tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{J\times 3}$.

A standard L2 loss is adopted to evaluate the error between the predicted 3D human body joints and ground truth 3D human body joints. It can be formulated as

$$L_{3\mathrm{D}} = \frac{1}{J}\sum_{i=1}^{J}\left\|\boldsymbol{P}_i - \tilde{\boldsymbol{P}}_i\right\|_2, \tag{11}$$

where $\tilde{\boldsymbol{P}}_i$ and $\boldsymbol{P}_i$ are the ith 3D joint positions in the estimated and ground truth, respectively.

Both training and testing of our network are performed on two NVIDIA TITAN RTX graphics processing units (GPUs). For Human3.6M[27] dataset, following previous works[7,17], we adopt the cascaded pyramid network (CPN)[28] as 2D keypoints detector. Note that the HumanEva-I[29] dataset uses a 15-joints skeleton produced by Detectron as inputs following[7]. Following Refs.[5,7], the horizontal flip data augmentation strategy is applied for training and testing. We adopt the ADAMW[30] optimizer for training our network with 80 epochs. For hyperparameters, we set N=3, L=8, K=3, D_S=256, and D_C=512. We also set the initial learning rate to 10^{-4}, and adopt cosine annealing[31] learning rate decay strategy to decrease the initial learning rate to 10^{-5}.

HumanEva-I and Human3.6M are utilized to evaluate our approach. Currently, Human3.6M is the largest publicly available dataset for 3D human analysis, containing 3.6 million video frames. It is captured from different viewpoints by 4 synchronized human motion capture cameras. The motions cover 15 daily activities (e.g. posing and eating) performed by 11 professional actors. Following Refs.[7,17], using the same training and testing policy as previous works, we adopt subjects (S1, S5, S6, S7, S8) for training, and two subjects (S9, S11) are applied for testing in our experiments. HumanEva-I is a much smaller dataset, which contains 4 subjects performing 6 common actions (e.g. jogging and walking). Following the previous method[7], we conduct training/testing on three commonly used actions (Box, Jog, Walk) performed by three subjects (S1, S2, S3). Two standard protocols are involved in our experiment for both datasets. Protocol #1 measures the mean per-joint position error (*MPJPE*) in millimeters between the estimated and ground-truth 3D joint locations

without any transformation. Protocol #2 applies a similarity transformation (Procrustes analysis) to the predicted 3D pose before calculating the *MPJPE*, referred to *PA-MPJPE*.

We report the results under 15 action categories in Tab.1, and the last column represents the average results for all actions. Our approach yields the lowest average error of 43.9 mm under Protocol #1, and it outperforms all previous methods. In addition, our approach achieves the second lower average error of 34.7 mm under Protocol #2. In particular, compared with Ref.[10] which ignores the temporal consistency among frames, our model reduces the *MPJPE* by approximately 22%. This clearly demonstrates the advantage of using temporal context information to improve performance. In addition, the estimation accuracy outperforms that of the GCN-based Ref.[12] by 10% (4.9 mm) under Protocol #1. It is clearly demonstrated that our BGAM can more effectively capture human-structure information from graph-structured data. To further investigate the impact of the noise data introduced by CPN on the 3D HPE accuracy, we use the ground truth 2D keypoints as input. It can be seen from Tab.2 that the *MPJPE* is remarkably reduced from 43.9 mm to 30.7 mm, with an error reduction of approximately 30.1%. More importantly, under Protocol #1, our approach improves the performance reported in Ref.[20] by 0.6 mm, approximately 2% improvement. It indicates that as the noise in the 2D keypoints data decreases, our model can further boost the accuracy.

Tab.1 Quantitative comparison between the proposed method and other methods in Human3.6M dataset (Abbreviations such as Dir. and Disc. in the table represent different action categories (e.g. Direction and Discussion) in Human3.6M[27] dataset)

Protocol #1	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
PAVLAKOS et al[10]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
LUVIZON et al[4]	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
CAI et al[12]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
PAVLLO et al[7]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
XU et al[11]	**37.4**	**43.5**	42.7	42.7	46.6	59.7	41.3	45.1	**52.7**	**60.2**	45.8	43.1	47.7	33.7	37.1	45.6
LIU et al[17]	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
ZHENG et al[20]	41.5	44.8	39.8	42.5	46.5	**51.6**	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Ours	41.6	43.8	**39.5**	**42.4**	**45.5**	53.7	**40.7**	**41.0**	56.0	62.4	**44.3**	**42.9**	**44.1**	**29.8**	**30.2**	**43.9**
Protocol #2	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
PAVLAKOS et al[10]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
CAI et al[12]	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
PAVLLO et al[7]	35.2	40.2	32.7	35.7	38.2	45.5	40.6	36.1	48.8	**47.3**	37.8	39.7	38.7	27.8	29.5	37.8
XU et al[11]	**31.0**	**34.8**	34.7	**34.4**	36.2	43.9	31.6	33.5	**42.3**	49.0	37.1	33.0	39.1	26.9	31.9	36.2
LIU et al[17]	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
ZHENG et al[20]	32.5	34.8	32.6	34.6	35.3	**39.5**	32.1	32.0	42.8	48.5	**34.8**	**32.4**	35.3	24.5	26.0	**34.6**
Ours	32.0	34.9	**32.2**	34.6	**35.0**	41.4	**31.5**	**31.5**	44.7	50.6	36.0	32.8	**34.2**	**23.8**	**24.5**	34.7

The quantitative results on HumanEva-I are also reported in Tab.3, where (-) represents that the corresponding value is not provided in the original paper. Our method again achieves the lowest prediction error under multiple action categories and outperforms all the previous state-of-the-art approaches. It is demonstrated that the proposed method has excellent generalization ability on the small dataset.

We also provide a visual comparison between ours and the previous state of the arts (SOTA) approach, as shown in Fig.3. The two most challenging actions performed by subjects S9, and S11 on the Human3.6M test set (e.g. Walking dog and Directions) are adopted for testing. Severe self-occlusion commonly exists in these actions. However, in most cases, our approach obtains more accurate predictions than the state-of-the-art method[7].

To further investigate the impact of different 2D keypoints input sequence lengths on performance, we select three different input sequence lengths f to conduct experiments on Human3.6M under Protocol #1. All test results are listed in Tab.4. It is indicated that a larger sequence length can reduce the prediction error more effectively. More importantly, our network has a smaller amount of parameters compared to previous methods.

Tab.2 Comparison results with other methods on Human3.6M dataset with 2D ground truth keypoints as input

Methods	*MPJPE*	*PA-MPJPE*
MARTINEZ et al[5]	45.5	37.1
HOSSAIN et al[2]	41.6	31.7
LEE et al[32]	38.4	-
PAVLLO et al[7]	37.2	27.2
LIU et al[17]	34.7	-
ZHENG et al[20]	31.3	-
Ours	**30.7**	**22.7**

Tab.3 Comparison results with other methods on HumanEva-I dataset

Protocol #2	Walk			Jog			Box		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
PAVLAKOS et al[10]	22.3	19.5	**29.7**	28.9	21.9	23.8	-	-	-
MARTINEZ et al[5]	19.7	17.4	46.8	26.9	18.2	18.6	-	-	-
LEE et al[32]	18.6	19.9	30.5	25.7	16.8	17.7	42.8	48.1	53.4
PAVLLO et al[7]	13.9	10.2	46.6	**20.9**	13.1	13.8	23.8	33.7	32
ZHENG et al[20]	14.4	10.2	46.6	22.7	13.4	13.4	-	-	-
Ours	**13.9**	**10.1**	45.3	21.6	**12.7**	**13.1**	**23.1**	**32.3**	**30.7**

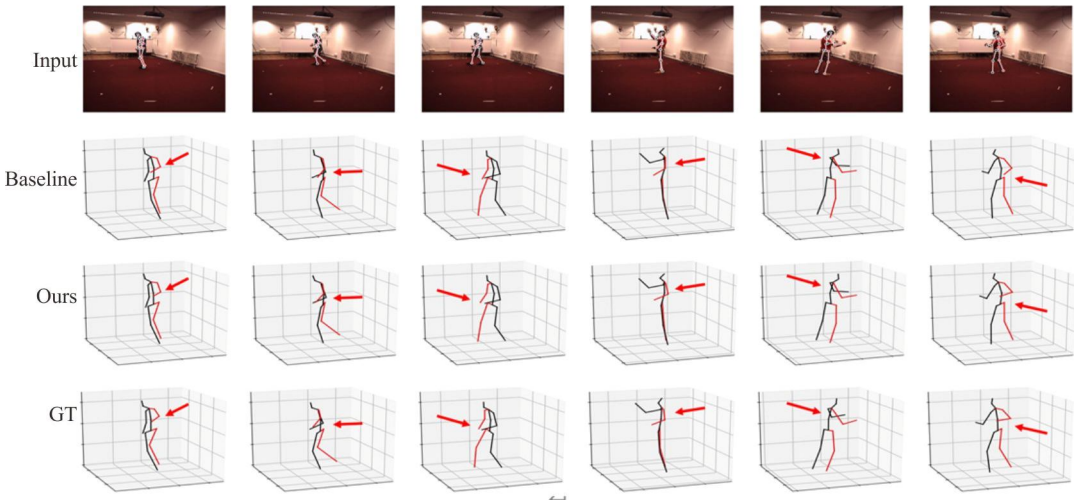


Fig.3 Visual comparison between our approach and the baseline method[7] on Human3.6M test set

Tab.4 Impact of different input sequence lengths in Human3.6M dataset

Methods	f	Parameters (M)	*MPJPE* (mm)
PAVLLO et al[7]	27	8.56	48.8
	81	12.75	47.7
	243	16.95	46.8
LIU et al[17]	27	5.69	48.5
	81	8.46	46.3
	243	11.25	45.1
Ours	27	4.6	48.8
	81	4.9	45.9
	243	**5.5**	**43.9**

To assess the effectiveness of our proposed SA for 3D HPE, we conduct experiments on the Human3.6M datasets using two variants of the network, namely the model without SA and the model with SA (K-NN), where K is the hyperparameter of the structure matrix. The results are presented in Tab.5. Obviously, the model with SA has a consistent improvement in performance. In addition, choosing an appropriate value of K to construct a structure matrix is important for the network to learn geometry priors about the human body. It clearly indicates that our proposed SA can model geometric context dependencies among different body joints.

Tab.5 Impact of SA mechanism

Model	*MPJPE*	Δ
Without SA	45.6	+1.7
With SA (1-NN)	44.5	+0.6
With SA (2-NN)	44.2	+0.3
With SA (3-NN)	43.9	-
With SA (4-NN)	44.3	+0.4

In this paper, we have presented a novel video-based human body geometric aware network for 3D HPE from videos. To improve the spatial feature representation ability of the network, we have introduced a novel SA mechanism to model geometric context dependencies among different body joints. In addition, the proposed temporal MLP-Mixer structure is able to comprehensively learn temporal context information from input sequences. Furthermore, a novel spatial-temporal network framework has also been proposed to effectively integrate spatial-temporal information. Extensive experiments demonstrate that our approach can effectively reduce the error of the 3D HPE task and achieve SOTA performance on two challenging datasets.

Statements and Declarations

The authors declare that there are no conflicts of interest

related to this article.

References

[1] MEHTA D, RHODIN H, CASAS D, et al. Monocular 3D human pose estimation in the wild using improved CNN supervision[C]//2017 International Conference on 3D Vision (3DV), October 10-12, 2017, Qingdao, China. New York：IEEE, 2017：506-516.

[2] HOSSAIN M R I, LITTLE J J. Exploiting temporal information for 3D human pose estimation[C]//Proceedings of the European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin：Springer, 2018：68-84.

[3] LIN J, LEE G H. Trajectory space factorization for deep video-based 3D human pose estimation[C]//2019 British Machine Vision Conference (BMVC), September 9-12, 2019, Cardiff, UK. BMVA, 2019.

[4] LUVIZON D C, PICARD D, TABIA H. 2D/3D pose estimation and action recognition using multitask deep learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake, UT, USA. New York：IEEE, 2018：5137-5146.

[5] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3D human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York：IEEE, 2017：2640-2649.

[6] PARK S, HWANG J, KWAK N. 3D human pose estimation using convolutional neural networks with 2D pose information[C]//Proceedings of the European Conference on Computer Vision, October 11-14, 2016, Amsterdam, The Netherlands. Berlin：Springer, 2016：156-169.

[7] PAVLLO D, FEICHTENHOFER C, GRANGIER D, et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York：IEEE, 2019：7753-7762.

[8] CHEN X, LIN K Y, LIU W, et al. Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York：IEEE, 2019：7753-7762.

[9] FANG H S, XU Y, WANG W, et al. Learning pose grammar to encode human body configuration for 3D pose estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 2-7, 2018, New Orleans, Louisiana, USA. Cambridge：AAAI Press, 2018：6821-6828.

[10] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York：IEEE, 2017：7025-7034.

[11] XU J, YU Z, NI B, et al. Deep kinematics analysis for monocular 3D human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York：IEEE, 2020：899-908.

[12] CAI Y, GE L, LIU J, et al. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York：IEEE, 2019：2272-2281.

[13] ZHAO L, PENG X, TIAN Y, et al. Semantic graph convolutional networks for 3D human pose regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York：IEEE, 2019：3425-3435.

[14] LIU K, DING R, ZOU Z, et al. A comprehensive study of weight sharing in graph networks for 3D human pose estimation[C]//Proceedings of the European Conference on Computer Vision, August 23-28, 2020, Glasgow, UK. Berlin：Springer, 2020：318-334.

[15] CI H, WANG C, MA X, et al. Optimizing network structure for 3D human pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York：IEEE, 2019：2262-2271.

[16] WANG J, YAN S, XIONG Y, et al. Motion guided 3D pose estimation from videos[C]//Proceedings of the European Conference on Computer Vision, August 23-28, 2020, Glasgow, UK. Berlin：Springer, 2020：764-780.

[17] LIU R, SHEN J, WANG H, et al. Attention mechanism exploits temporal contexts：real-time 3D human pose reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York：IEEE, 2020：5064-5073.

[18] TOLSTIKHIN I, HOULSBY N, KOLESNIKOV A, et al. MLP-mixer： an all-MLP architecture for vision[C]//Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), December 6-12, 2021, Virtual Event. New York：Curran Associates, 2021：24261-24272.

[19] CHEN C H, RAMANAN D. 3D human pose estimation= 2D pose estimation + matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York：IEEE, 2017：7035-7043.

[20] ZHENG C, ZHU S, MENDIETA M, et al. 3D human pose estimation with spatial and temporal transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, QC, Canada. New York：IEEE, 2021：11656-11665.

[21] DABRAL R, MUNDHADA A, KUSUPATI U, et al. Learning 3D human pose from structure and motion[C]// Proceedings of the European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin：Springer, 2018：668-683.

[22] CHENG Y, YANG B, WANG B, et al. Occlusion-aware networks for 3D human pose estimation in video[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York：IEEE, 2019：723-732.

[23] LIU J, ROJAS J, LI Y, et al. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video[C]//2021 IEEE International Conference on Robotics and Automation (ICRA), May 30-June 5, 2021, Xi'an, China. New York：IEEE, 2021：3374-3380.

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8)：1735-1780.

[25] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words：transformers for image recognition at scale[C]//9th International Conference on Learning Representations (ICLR), May 3-7, 2021, Virtual Event, Austria. 2021.

[26] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs)[EB/OL]. (2016-06-27) [2021-12-26]. https：//arxiv.org/abs/1606.08415v1.

[27] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6m：large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(7)：1325-1339.

[28] CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake, UT, USA. New York：IEEE, 2018：7103-7112.

[29] SIGAL L, BALAN A O, BLACK M J. Humaneva：synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International journal of computer vision, 2010, 87(1-2)：4.

[30] KINGMA D P, BA J. Adam：a method for stochastic optimization[EB/OL]. (2014-12-22) [2021-12-26]. https：//arxiv.org/abs/1412.6980v1.

[31] LOSHCHILOV I, HUTTER F. SGDR：stochastic gradient descent with warm restarts[EB/OL]. (2016-08-13) [2021-12-26]. https：//arxiv.org/abs/1608.03983v1.

[32] LEE K, LEE I, LEE S. Propagating LSTM：3D pose estimation based on joint interdependency[C]//Proceedings of the European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin：Springer, 2018：119-135.