Machine Vision and Applications (2022) 33:82
https://doi.org/10.1007/s00138-022-01334-6

ORIGINAL PAPER

U-shaped spatial–temporal transformer network for 3D human pose estimation

Honghong Yang[1,2] · Longfei Guo[3] · Yumei Zhang[1,3] · Xiaojun Wu[1,3]

Received: 26 April 2022 / Revised: 29 July 2022 / Accepted: 4 August 2022 / Published online: 4 September 2022

Abstract
3D human pose estimation has achieved much progress with the development of convolution neural networks. There still have some challenges to accurately estimate 3D joint locations from single-view images or videos due to depth ambiguity and severe occlusion. Motivated by the effectiveness of introducing vision transformer into computer vision tasks, we present a novel U-shaped spatial–temporal transformer-based network (U-STN) for 3D human pose estimation. The core idea of the proposed method is to process the human joints by designing a multi-scale and multi-level U-shaped transformer model. We construct a multi-scale architecture with three different scales based on the human skeletal topology, in which the local and global features are processed through three different scales with kinematic constraints. Furthermore, a multi-level feature representations is introduced by fusing intermediate features from different depths of the U-shaped network. With a skeletal constrained pooling and unpooling operations devised for U-STN, the network can transform features across different scales and extract meaningful semantic features at all levels. Experiments on two challenging benchmark datasets show that the proposed method achieves a good performance on 2D-to-3D pose estimation. The code is available at https://github.com/l-fay/Pose3D.

Keywords Human pose estimation · Spatial–temporal transformer network · Multi-scale and multi-level feature representations

1 Introduction

3D human pose estimation (HPE) aims to localize the 3D keypoints in an input image or video. With the development of the deep Convolutional Neural Networks, 3D HPE has made great achievements in recent years [1]. However, it is still a challenging task to estimate 3D poses from 2D coordinates due to frequent occlusion, 2D pose prediction errors and depth ambiguity from 2D projection. 3D HPE is a very attractive research field that has a significant influence on many applications such as action recognition, human–robot interaction and athlete motion analysis [2].

The current existing 3D HPE methods can be generally classified into two classes: direct estimation method and 2D-to-3D lifting method [1]. The former one directly regresses the 3D pose joints from 2D images [3, 4], while the latter one first estimates 2D keypoints and then projects them into 3D space [1, 5, 6]. Although 3D HPE achieve promising progress on the basis of the excellent performance of 2D pose estimation [7], some challenging problems still exist when estimating 3D joint locations from monocular images, such as (1) self-occlusions: some human poses cause joints occlusion and may trigger information missing; (2) depth ambiguity: many 3D poses can be projected to the same 2D poses from monocular images; (3) prediction errors: incorrect 2D pose detector may cause inaccurate 3D pose estimation.

To address those issues, several methods incorporate spatial dependencies and temporal consistencies from videos into graph convolutional networks (GCNs) to fit the specific

✉ Yumei Zhang
zym0910@snnu.edu.cn

✉ Xiaojun Wu
xjwu@snnu.edu.cn

[1] Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an 710062, China

[2] Key Laboratory of Intelligent Computing and Service Technology for Folk Song, Ministry of Culture and Tourism, Xi'an, China

[3] School of Computer Science, Shaanxi Normal University, Xi'an 710062, China

needs for 3D pose estimation [8]. Since the spatial dependencies naturally express the correlation between body joints in intra-frame, it can reduce the probability of producing physically impossible 3D structures and is helpful to solve self-occlusions. The temporal information from videos can capture global dependencies from inter-frames, so it is useful for tackling the issue of depth ambiguity. Cai et al. [8] explicitly integrated the specific prior knowledge of human body to construct the spatial–temporal GCNs for 3D pose estimation. Hossain et al. [9] exploited LSTM network with shortcut connections to impose temporal consistency constraint on the predicted 3D poses. Pavllo et al. [10] utilized temporal convolutions to capture the global dependencies from consecutive frames. Though these approaches achieve competitive performance in 3D HPE, there still exist inherent limitations in spatial and temporal correlation. For instance, the CNN-based temporal convolution or temporal correlation windows typically rely on dilation temporal convolution to model long-term dependencies of the nodes. Those are limited in temporal connectivity and are mainly constrained to simply sequential correlation [1]. Additionally, most existing 3D HPE approaches mainly focus on incorporating either spatial constraints or temporal correlations, without considering the complementary characteristics between these two types of information. Furthermore, the topology of the graph convolution in GCNs is a key factor to model the correlations of the input graph nodes. However, once the topology is generated, only single-scale features are extracted and only one transformation exists in each layer of the networks [11]. As a result, the backbones of these methods that incorporate spatial and temporal information into GCNs-based 3D HPE have intrinsic limitations on extracting and synthesizing 3D structure information.

Recently, the vision transformers have been widely introduced into computer vision tasks. Since the transformer architecture embeds the self-attention and position mechanism, it can flexibly model long-range global consistencies information with input sequences. Additionally, as described above, the core factor that influences the performance of 3D HPE is the features extracted by the model. The features with great representation capabilities will boost the performance of 3D HPE. These observations inspire us to devise a U-shaped spatial–temporal transformer network (U-STN) that focus on how to effectively extract spatial and temporal features to improve the performance of 3D HPE. In our work, a multi-scale and multi-level spatial–temporal transformer model is developed for extracting human skeletal features, where the multi-scale Spatial–Temporal Transformer architecture is modeled to learn the intra-frame interactions between different joints and capture global dependencies from inter-frames. Since the multi-scale feature representation can capture information from small to large resolutions of the input data, it can bring rich local to global information. The multi-level feature representation model is devised to fuse different-depth intermediate features from the U-shaped network, it can capture important semantic information at all levels from shallow to deep layers. Additionally, with a skeletal constrained pooling and unpooling operations devised for U-STN, the network can transform features across different scales and extract meaningful semantic features at all levels.

To summarize, the contribution of the proposed method is as follows:

(1) A U-shaped spatial–temporal transformer network is devised for 3D HPE, which incorporates multi-scale and multi-level spatial–temporal transformer feature representations with a prior human skeletal topology to construct the U-shaped network.
(2) The multi-scale Spatial–Temporal Transformer architecture is modeled to learn the intra-frame interactions between different joints and global correlations from inter-frames, where the multi-scale architecture consists of three different scales based on the human skeletal topology.
(3) The multi-level feature representation is introduced to fuse different-depth intermediate features of the U-shaped network, where a skeletal constrained pooling and unpooling operations are performed for transforming features across different scales and extract meaningful semantic features at all levels.

2 Related works

2.1 3D Human pose estimation

3D human pose estimation (HPE) from monocular images has been an attractive research area in computer vision in recent years. The most existing 3D HPE methods can be roughly classified into two classes: direct estimation methods and 2D-to-3D lifting methods. For the former one, some researchers regress 3D pose directly from 2D images without intermediately estimating the 2D pose representation. Li et al. [12] proposed a deep convolution networks with multitask framework to regress the 3D pose. Park et al. [13] designed an end-to-end framework that directly uses a single CNN for 2D joint classification and 3D joint regression. Pavlakos et al. [14] calculated voxel likelihoods for each joint and used them to predict the location of 3D pose. Zeng et al. [15] designed a split-and-Recombine approach for rare and unseen poses prediction, where the human body is split into local groups of joints and then perform local pose configurations in each group. Conversely, the 2D-to-3D lifting methods first exploit 2D pose estimation results from input images and then project them into 3D space. With the intermediate representation from 2D pose detectors, the 2D-to-3D lifting approaches achieve promising highly accurate results for 3D HPE. Our approach falls into this category. Martinez et al. [16] exploited fully connected convolution based-network to directly predict 3D positions from 2D joints. Xu et al. [17] proposed a graph stacked hourglass model to construct an encoder-decoder architecture for 2D-to-3D human pose estimation. Cai et al. [8] embedded the spatial–temporal relationships into graph convolution networks for 3D human pose estimation. The proposed model in the present study is different from the previous works. In our work, a U-shaped spatial–temporal transformer network is designed for feature extraction, where the multi-scale and multi-level architecture is modeled to learn the intra-frame interactions between different joints and global correlations from inter-frames, aiming to enhance the feature representations capability of the proposed U-STN model.

2.2 Spatial–temporal convolution

Since the spatial information from 2D input image naturally expresses the correlation relationship among body joints in each frame and the temporal information from videos can capture global dependencies from adjacent frames. Both of them are useful for tackling the problems of depth ambiguity and self-occlusion in 3D HPE. Liu et al. [18] embedded spatial–temporal information into graph network for 3D HPE, which leveraged human kinematic constraints and dilated temporal convolution to learn spatial–temporal features of the input sequences. Pavllo et al. [10] constructed a fully convolutional model and has introduced the temporal convolution and semi-supervised training for 3D HPE. Wang et al. [6] introduced a motion model into a spatial–temporal graph convolution networks, aiming to better infer the depth information for each frame. Li et al. [19] proposed a Multi-Hypothesis Transformer (MHFormer) that learns spatial–temporal representations of multiple plausible pose hypotheses to solve the depth ambiguity and self-occlusion in 3D HPE. Different from the above methods, our work mainly via constructing a multi-scale and multi-level spatial–temporal transformer model to capture the local and global relationship among graph nodes, where not only the spatial and temporal information were incorporated into the feature extraction, but also the prior human skeletal topology is introduced to construct the U-shaped network to meet the specific demand for 3D HPE.

2.3 Transformer in HPE

As the transformer architecture embeds the attention mechanism, it can flexibly model long-range dependencies in input sequences. Some works resort transformer architecture to improve 3D HPE performance. Li et al. [20] designed a strided transformer encoder network for lifting 2D joint locations to 3D HPE. Zheng et al. [1] embedded the spatial and temporal information into transformer architecture, aiming to comprehensively model the local relationships and the global dependencies information. Lin et al. [21] constructed a multi-layer transformer encoder module to capture the short-and-long-range interactions among body joints and reconstructed 3D human joint coordinates from a single image. With the self-attention and position mechanism, the transformer model has powerful ability to model global dependencies information of the input. While the multi-scale and multi-level features take advantage of the benefits of model depths and scales, they capture the features from small to large resolutions and provide important semantic information at all levels from shallow to deep of the model. Hence, without the multi-scale and multi-level features from different depths and different scales of the model, the extracted spatial–temporal features by transformer model are less generalizable and limit the performance of model for 3D HPE. In our work, we combine transformer model, multi-scale and multi-level features together to let the network inherit the advantages of them, makes the model more expressive.

2.4 Multi-scale and multi-level feature representations

Feature representation capability is a core factor that influences the image-based tasks. Some works concentrate on how to construct a multi-scale and multi-level feature representation module to enhance the expressiveness of the model. Feature Pyramid Network [22] is a typical multi-scale feature module used for object detector. It integrates small to large resolutions features to achieve a better understanding in the spatial domain. Stacked Hourglass network incorporated multi-scale features to learn rich image features from local to global, which enables the model to preserve spatial relationships among human joints for 2D HPE [23]. Sun et al. [24] fused multi-scale and multi-level features from different branches and different depths of the HRNet for keypoint prediction. Zhao et al. [25] embedded multi-level features from shallow to deep layers into pyramid network, aiming to capture better feature representations for object detection. Hua et al. [26] designed a cross-view U-shaped graph convolutional network (CVUGCN) for 3D HPE, which take advantage of spatial configurations and cross-view correlations to accurately refine the coarse 3D poses in a weakly-supervised manner. Xu et al. [17] designed a graph stacked hourglass network to extract multi-scale and multi-level features for human skeletal representations. In our work, a skeletal constrained pooling and unpooling operations is introduced to transform features across different scales and extract semantic feature at all levels of U-shaped network.

3 The proposed method

3.1 Problem formulation

The proposed method follows the 2D-to-3D lifting architecture for 3D HPE in videos. Given a sequence of 2D pose joint locations $X = \left\{x_{t,j} | t = 1, \ldots T;\ j = 1, \ldots J\right\}$ estimated by an off-the-shelf 2D pose detector as input, the goal of 3D HPE is to reconstruct 3D joint coordinates $S = \left\{s_{t,j} | t = 1, \ldots T;\ j = 1, \ldots J\right\}$ for a center frame, where $x_{t,j} \in \mathbb{R}^{J \times 2}$ and $s_{t,j} \in \mathbb{R}^{J \times 3}$ denote the j_th joint location of 2D and 3D at frame $t \in T$, respectively. T and J are the number of video frames and the joints, respectively. Different with the dominant CNN based 3D pose estimation models, we have designed a U-shape spatial–temporal transformer network for 3D HPE. The proposed network first employs the spatial–temporal transformer model to learn the intra-frame interactions between different joints and global correlations from inter-frames. Then, a skeletal constrained pooling and unpooling operations is introduced to construct the U-shape model by transforming features across different scales and extracting semantic features at all levels of network. By combining the spatial–temporal transformer model with multi-scale and multi-level features together in a U-shaped architecture, we construct a U-shaped spatial–temporal transformer network (U-STN), which inherits the advantages of them and makes the model more expressive for 3D HPE.

3.2 Spatial–temporal transformer feature extraction model

With the self-attention and position mechanism, the transformer model has powerful ability to model short-and-long range relations of input sequences. As correlations among nodes in intra-frame and inter-frame are crucial for 3D HPE, we design a spatial–temporal transformer feature extraction model to comprehensively encoder the local and global skeleton features both in space and time dimensions with the spatial transformer model and the temporal transformer model, respectively. The transformer self-attention encoders the relations among surrounding joints, which efficiently captures the local joint correlation in intra-frame and global dependencies of body joint among inter-frames. The framework of the spatial–temporal transformer feature extraction model is shown in Fig. 1.

3.2.1 Spatial transformer model (STM) for local correlation feature extraction

The spatial transformer model (STM) employs self-attention inside each frame to capture location relationship between different joints. With the comprehensive connectivity of 2D joints, the STM can learn stronger feature representations for each frame by employing the spatial self-attention to encode the spatial relations of joint-to-joint in each frame. Let each 2D joint $x_j^t \in \mathbb{R}^{J \times 2}$ at frame t as an input token, the general vision transform architecture in [27] is employed to extract high dimensional features for all input tokens in spatial domain. Firstly, the spatial positional embedding $E_{SPos} \in \mathbb{R}^{J \times C}$ is performed on 2D coordinate of each joint by a linear projection, where the positional embedding is used to retain spatial position information of the joints in each frame as follows:

$$Z_0^t = [x_1 E;\ x_2 E;\ \cdots;\ x_J E] + E_{SPos} \tag{1}$$

where $E \in \mathbb{R}^{(J \cdot 2) \times C}$ is a linear projection matrix to transform each path to a high dimension features, C is the dimension of spatial embedding.

Then, the high dimensional features of joint $Z_0^t \in \mathbb{R}^{J \times C}$ is fed into the self-attention layer of spatial transformer encoder model, which consists of the multi-head self-attention layer (MSA) with multilayer perceptron (MLP) and normalization layer (LN(·)).The MSA uses the multi-head attention to model the relations from different positions of the input with embedded features. After the L layers spatial transformer encoder to process the features Z_0^t, the encoder output $Z_L^t \in \mathbb{R}^{J \times C}$ of the STM can be represented as follows:

$$\begin{aligned} & Z_l' = \mathrm{MSA}(\mathrm{LN}(Z_{l-1}^t)) + Z_{l-1}^t, \quad l = 1, 2, \cdots, L \\ & Z_l = \mathrm{MLP}(\mathrm{LN}(Z')) + Z_l', \quad l = 1, 2, \cdots, L \\ & Z_L^t = \mathrm{LN}(Z_L^t) \end{aligned} \tag{2}$$

where the output of spatial encoder features $Z_L^t \in \mathbb{R}^{J \times C}$ of STM is fed into the temporal transformer model for extraction the global dependencies of input sequences.

3.2.2 Temporal transformer model (TTM) for feature's global dependencies extraction

Since the self-attention in the temporal transformer model (TTM) can effectively learn the correlations of each joint among inter-frames by analyzing the embedding changes in the same body joint along the temporal dimension. The temporal transformer model (TTM) is used to extract global dependencies among spatial feature representations across the input sequences. We first flatten the spatial encoder features of the STM $Z_L^t \in \mathbb{R}^{J \times C}$ at each frame into a vector $\mathbf{Z}^t \in \mathbb{R}^{1 \times (J \times C)}$, and concatenated them to from the input $\mathbf{Z}_0 = \{\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^T$ for TTM, where $\mathbf{Z}_0 \in \mathbb{R}^{T \times (J \times C)}$. Then, the temporal positional embedding $E_{\mathrm{TPos}} \in \mathbb{R}^{T \times (J \times C)}$ is performed for $\mathbf{Z}_0$ to retain the position information of the input frames. The process of temporal feature encoder is same

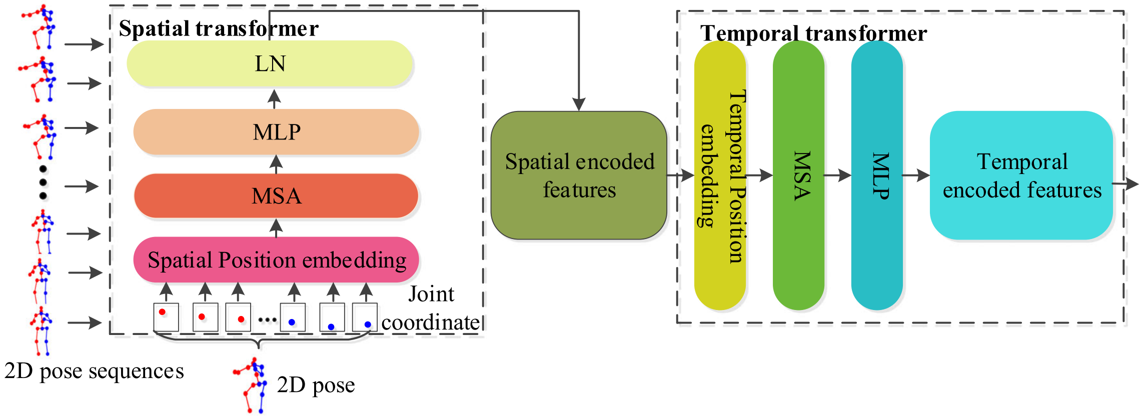


Fig. 1 The spatial–temporal transformer feature extraction model

with STM, which is described in Eqs. (1) and (2). After performing the L identical layers MSA and MLP, the output of the temporal transformer can be represented as the temporal encoded features $Y \in \mathbb{R}^{T\times(J\times C)}$.

3.3 Skeletal constrained pooling and unpooling layer

The most existing 3D HPE methods take the 2D skeleton joints as a whole graph data, and only use a single-scale and single-resolution features to construct the topology relationship of the input data. These methods ignore the fact that the joints of human body have different relative motion space; for instance, the knee and elbow have large motion space than adjacent joints like hip and shoulder. This may limit the performance of model for 3D HPE. As the feature representation capabilities is a core factor to influence the expressiveness of the model, we extract the multi-scale and multi-level features to form U-shaped network, where the pooling and unpooling operation are essential for construct the multi-scale features for our U-shaped network. Thus, we introduce a skeletal constrained pooling and unpooling operations to transform features across different scales and extract semantic feature at all levels, aiming to learn more comprehensive body-joints relationship features and enrich the performance of our model.

3.3.1 Spatial pooling layer

Since the pooling and unpooling required for multi-scale features are mainly defined for image, which ignores the nodes' geography information of the graph and is not suitable for graph-structured data, resulting in information loss for graph representation. Hence, in our work, according to the connection relationship among human body joints, we design a multi-scale skeleton structure with 17, 11 and 7 nodes for scale $s = 1$, 2 and 3 ($s = 1, 2, 3$ involves large-middle-small scales), respectively. As shown in Fig. 2, the large scale $s = 1$ with all 17 keypoints can extract local features of each keypoints within a small receptive field. The small-scale $s = 3$ with 7 nodes can capture global contour features within a large receptive field. We exploit the spatial pooling layer to transform the corresponding features into the lower-scale skeleton structure features, it is important for reducing the size of feature map and enlarging the receptive fields. Given the feature matrix $X_s \in \mathbb{R}^{V\times 2}$ at s-scale, we first construct the pooling matrix $M^s \in \mathbb{R}^{U\times V}$ to reduce v nodes in scale s to U groups in $s+1$ scale, then, 1*1 convolutions is used to adaptively fuse the features as follows:

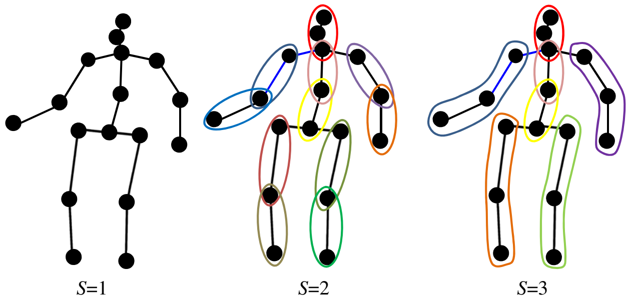


Fig. 2 Multi-scale structure based on human skeletal topology

$$X'_{s+1} = conv_{2D}((W^s \odot M^s) \otimes X_s) \tag{3}$$

where $X'_{s+1} \in \mathbb{R}^{U\times 2}$, $M^s \in \{0, 1\}$ denotes whether the v-th joints in s scale belongs to the $s+1$ scale in u-th pooling group or not. In our work, $M^2 \in \mathbb{R}^{11\times 17}$ and $M^3 \in \mathbb{R}^{7\times 11}$. $W^s \in \mathbb{R}^{U\times V}$ is the trainable weight to measure the importance of joint v in group u. $\odot$ is the element-wise multiplication, $\otimes$ is the matrix multiplication.

3.3.2 Spatial unpooling layer

Since the unpooling operation is essential for restoring lower-scale skeletal information from original resolutions, we have designed a skeleton constrained spatial unpooling layer to pass the lower-scale features and fused them to form higher-scale features. With the U groups nodes proceed by spatial pooling layer in s + 1 scale, the corresponding node features matrix is $X'_{s+1} \in \mathbb{R}^{U \times C}$. Then, a 2D transposed convolution $\mathrm{conv}_T(\cdot)$ is used to recover the higher-scale skeletal representations as follows:

$$X''_s = conv_T(([M^s]^T \odot [W^s]^T) \otimes X'_{s+1}) \tag{4}$$

where $[\cdot]^T$ is the transpose matrix for M^s and W^s. After implementing Eq. (4), we transform features X'_{s+1} from s + 1 scale to the features X''_s in s scale.

3.4 U-shaped spatial–temporal transformer network (U-STN)

As shown in Fig. 3, the proposed U-shaped spatial–temporal transformer network consists of three stages: (1) multi-scale feature extraction with skeletal constrained pooling and unpooling, (2) Multi-level features extraction with spatial–temporal transformer model, and (3) multi-scale and multi-level feature merging.

3.4.1 Multi-scale features extraction with skeletal constrained pooling and unpooling

The U-STN starts from a high-resolution branch features with 27*17*2 human body joints. A new branch is formed by performing downsample with skeletal constrained spatial pooling as described in Sect. 3.3. We have designed a 3 branch (s = 1,2,3 involves large-middle-small scales) based on skeleton structures with three different nodes of 17, 11 and 7, respectively. With the spatial pooling and unpooling in multi-scale features extraction as calculated in Eqs. (3) and (4), the multi-scale features learn to integrate features from different resolutions at various scales, where more channels are introduced at relatively low scales of the U-STN for skeleton representation. Thus, with the consecutive spatial pooling performed in multi-scale, more channels are introduced and gradual enlarged the receptive field for feature extraction. This is useful for capturing information from small to large resolutions of the input skeleton and reduces information loss due to scale changes.

3.4.2 Multi-level features extraction with spatial–temporal transformer model

Since the multi-level features from different depths of the model can capture important semantic information at all levels from shallow to deep, we design a U-shaped spatial–temporal transformer feature extraction model to capture the multi-level intermediate features from different scales. With the multi-scale features from three different skeletal structures, we first performer spatial transformer model as described in Sect. 3.2 for each resolution feature. Let X_s denotes the feature matrix at s-scale and is input to the STM. Then, after performing Eqs. (1) and (2), the output of STM $[Z^t_L]_s$ is fed into the temporal transformer model. The corresponding output Y_s of the TTM from s-scale is the features of level s. By performing the spatial–temporal transformer model for the multi-scale features from three different skeletal structures, we achieve three different level features with different shapes and channels. Since the spatial–temporal transformer model learns the features in intra-frame interactions between different joints and inter-frame correlations from adjacent frames, it brings valuable semantic information for 3D HPE.

3.4.3 Feature merge model for multi-scale and multi-level features

As multi-scale features present information within the spatial domain of the graph-structured data and the multi-level features provide semantic information at all levels from shallow to deep, we have designed a feature merge model for concatenating them to obtain the final overall features for 3D HPE.

For the three different resolution features processed by spatial–temporal transformer model from three scales, we first perform unsample operation by employing the skeleton constrained unpooling layer designed in Sect. 3.3 to embed the lower-scale skeleton features into higher-scales. Let $Y_s \in \mathbb{R}^{J \times (C \times T)}$ denotes the features processed by the spatial–temporal transformer model in scale s, $s = 2$, the corresponding node parts in three scale $s = 2$ are $s = 2$. The features from the lower-scale are processed by the skeleton constrained unpooling layer to achieve the higher-scale features as follows:

$$X''_{s-1} = \mathrm{conv}_T(([M^s]^T \odot [W^s]^T) \otimes Y'_s) \tag{5}$$

Then, to better embed multi-level intermediate features from different depths of the U-STN network into the multi-scale features for obtaining the final features for 3D HPE, we design the feature merging model as shown in Fig. 3. With the features X''_s recovered from Eq. (5) and Y_s processed by the spatial–temporal transformer model at each scale s, the

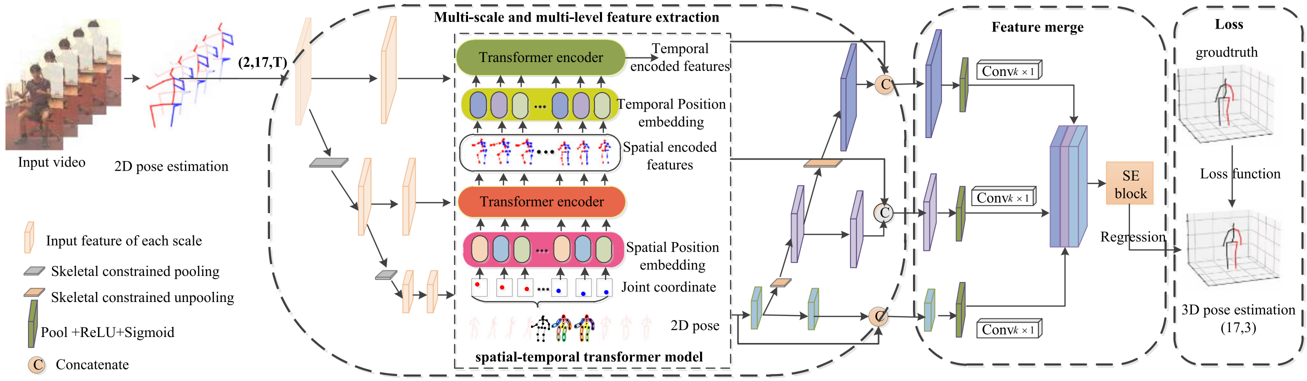

Input video
2D pose estimation
(2,17,T)
Multi-scale and multi-level feature extraction
Transformer encoder
Temporal encoded features
Temporal Position embedding
Spatial encoded features
Transformer encoder
Spatial Position embedding
Joint coordinate
2D pose
spatial-temporal transformer model
Input feature of each scale
Skeletal constrained pooling
Skeletal constrained unpooling
Pool +ReLU+Sigmoid
C Concatenate
Feature merge
Conv$k\times1$
Conv$k\times1$
Conv$k\times1$
SE block
Regression
Loss
groudtruth
Loss function
3D pose estimation (17,3)

Fig. 3 The framework of the proposed U-STN

final features in each scale are achieved by fusing the same scale features from different levels of the network as follows:

$$Y_s' = Y_s \odot \mathrm{Sigmoid}(\mathrm{ReLU}(\mathrm{pool}(Y_s)W_1)W_2) \oplus X_s'' \tag{6}$$

where $pool(\cdot)$ is the average pooling operation, which is performed for all input sequences as well as skeleton nodes in each channel at each scale, aiming to capture channel-wise statistics information for Y_s. $W_1 \in \mathbb{R}^{C\times(C/r)}$ and $W_2 \in \mathbb{R}^{(C/r)\times(C\times U)}$ represent the weights for two fully connected layers, r is the reduction ratio.

Finally, the features from three scales with embedded different depths intermediate features from different levels are used for 3D HPE. Those features are first transformed into the same shape with the 1*1 convolution, which is used to reduce the channels of lower-scale features. Then, concatenate those three scale features from U-STN for overall feature representations as follows:

$$Y_{cat} = \mathrm{Concat}(Y_1', Y_2', Y_3') \in \mathbb{R}^{J\times C} \in \mathbb{R}^{J\times(C\times T)} \tag{7}$$

where J and C represent the number of nodes and feature dimensions in each scale. Since the overall features are concatenated by the multi-scale and multi-level features across different scales and different depths of the U-shaped network, it enables the model to capture the features from small-to-large resolutions and provide rich semantic feature representations among the intermediate features. Then, Squeeze and-Excitation block (SE block) in [28] is used to measure the channel-wise weight for all concatenated features Y_{cat}. Finally, the output of SE block is fed into one linear layer for 3D regression, the corresponding output $s_{t,j} \in \mathbb{R}^{J\times3}|t = 1, \ldots T;\ j = 1, \ldots J$ is the estimated 3D pose for the center frame.

3.5 Loss function

With 2D pose joints $P = \{p_{t,j} \in \mathbb{R}^{J\times2}|t = 1, \ldots T;\ j = 1, \ldots J\}$ in a sequence, our model learns the mapping function $F^* : \mathbb{R}^{J\times2} \rightarrow \mathbb{R}^{J\times3}$ to estimate the 3D joint location $S = \{s_{t,j} \in \mathbb{R}^{J\times3}|t = 1, \ldots T;\ j = 1, \ldots J\}$. The proposed model is trained with the Mean Squared Error (MSE) loss, which is employed to minimize the errors between the estimated and ground truth pose in T frames as follows:

$$\begin{aligned} F^* &= \arg\min_{F} \frac{1}{T}\sum_{t=1}^{T} \ell(F(p_t), \mathrm{y}_t), \\ \ell &= \frac{1}{J}\sum_{j=1}^{J} \left\| s_{t,j} - \mathrm{y}_{t,j} \right\|_2 \\ s_t &= F^*(p_t) \end{aligned} \tag{8}$$

where $\ell(\cdot)$ is the MPJEP (Mean per Joint Position Error) loss function, $s_{t,j}$ and $\mathrm{y}_{t,j}$ are the estimated and ground truth 3D joint location of the j_th joint in t frame.

4 Experiments

4.1 Datasets and evaluation metrics

4.1.1 Datasets

To evaluate the efficiency of the proposed U-STN model, we conduct experiments on two widely used 3D HPE datasets: Human3.6 M [29] and HumanEva-I [30]. Human3.6 M is the most widely used 3D HPE dataset in indoor environment under 4 viewpoints. It contains 3.6 million pose images with 11 professional actors performing 17 actions, such as discussion, smoking, taking photograph. Following the setting in [1, 4, 8], the proposed model is trained on five subjects (S1, S5, S6, S7, S8) and is tested on two subjects (S9 and S11). HumanEva-I contains 7 calibrated video sequences that are obtained from a motion capture system. It contains four

Table 1 Ablation study the influence of different scales for 3D HPE

Scales	Node numbers *J*					MPJPE(mm)
	17	11	7	5	2	
1	√					47.7
1,2	√	√				46.34
1,2,3	√	√	√			45.92
1,3,5	√		√		√	46.9
1,2,4	√	√		√		47.4
1,2,3,4	√	√	√	√		47.28
1,2,3,5	√	√	√		√	48.8
1,2,3,4,5	√				√	47.95

subjects performing six common actions (such as walking, jogging, and gestures).

4.1.2 Evaluation metrics

Two common evaluation metrics (MPJPE and P-MPJPE) [31] are used to evaluate the performance of our method. The Mean Per Joint Position Error (MPJPE) is the mean Euclidean distance between the estimated joints and the ground truth over all joints in millimeters, which is referred as Protocol 1. P-MPJPE is used to compute the mean Euclidean distance after alignment the estimated 3D pose and the ground truth by rotations, translations and scale, which is referred as Protocol 2.

4.2 Experimental setup

4.2.1 Implementation details

The experiments are conducted by Python 3.8.2 with Pytorch framework on one NVIDIA RTX 2080 GPU. The proposed model is trained using the Adam optimizer [32] for 200 epochs with weight decay of 0.1. The initial learning rate is 0.00004 and the shrink factor is 0.99. The dropout [33] is 0.2. The batch size is 512 for Human 3.6 M and 64 for HumanEva-I. We employ stochastic depth [34] with a rate of 0.1 for transformer encoder layers. The 2D pose is achieved by the cascaded pyramid network (CPN) for Human 3.6 M dataset and the Mask R-CNN is adopted for HumanEva-I dataset for a fair comparison.

4.3 Ablation study

To verify the effectiveness of each crucial component for the proposed method on 3D HPE, we perform ablation experiments on the Human3.6 M test dataset under protocol 1.

4.3.1 Effects of multiple scale feature representations

To better study how the multiple scale features affect the 3D HPE performance, we have conducted ablation studies by removing multi-level features while reserving multi-scale features in U-STN model. The multi-scale features are achieved by pooling and unpooling in the U-STN model that transforms features across different scales. We remove all pooling and unpooling layers in our architecture, which means the features processed at the highest scale, it denotes as Scale 1. By gradually adding pooling and unpooling on the basis of Scale 1 to form Scale 2 and Scale 3. Besides, apart from the three scales are presented in our method, we introduce two additional scales: S4 (Left arm, right arm, left limb, left limb and torso) and S5 (upper body and lower body). Then our model concatenates each scale's features to achieve the final features for 3D HPE. As shown in Table 1, by using three scales, the model can achieve the lowest error, with 45.92 mm MPJEP. It is obviously that combing two scales has better performance than the case when only one scale features are adopted. However, when we fuse S4 and S5 features together, the MPJPE is increased. This is caused by that the redundancy features are introduced with multi-scale from scale S1 to S5, which hurts the model's performance.

4.3.2 Effects of multiple level feature representations

To validate the effects of features from different depths of the U-STN model, we employ intermediate features from different depths of U-STN for 3D HPE. Since it has validated in Table 1 that our model achieves the best result by fusing multi-scale features from S1, S2 and S3, we only validate the effects of features from three levels derived from those three scales as shown in Table 2. We denote the intermediate features from the Scale 1, Scale 2 and Scale 3 are Level 1, Level 2 and Level 3, respectively. Table 2 shows the results that (1) only intermediate feature of Level 1 is used; (2) combining the intermediate features of Level 1 and Level

Table 2 Ablation study the influence of multiple level features for 3D HPE

	Scale 1	Scale 1,2	Scale 1,2,3	MPJPE
Level 1	√			47.00
Level 1,2	√	√		45.97
Level 1,2,3	√	√	√	45.51

Table 3 Ablation study the influence of different pool/unpool methods for 3D HPE

Method	MPJPE (mm)
Average pooling/unpooling	46.25
Maxpooling/unSampling	46.13
Skeletal constrained pooling/unpool	45.51

Table 4 Ablation study of attention part

Method	Spatial transformer	Temporal transformer	MPJPE (mm)
a	√	×	46.49
b	×	√	46.23
c	×	×	46.88
d	√	√	45.51

2; (3) all three intermediate features of Level 1, Level 2 and Level 3 are used. As shown in Table 2, we can see the lowest MPJPE achieved by combing three level intermediate features. It is obvious that using two level intermediate features is much better than using one intermediate features. This further proves that the multi-level features from different depths of U-STN boost the feature representation capabilities of the proposed model, which is useful for improving the performance of 3D HPE.

4.3.3 Effects of skeletal constrained pooling and unpooling

By comparing the proposed skeletal constrained pooling/unpooling with the traditional average pooling/unpooling and maximum pooling/unSampling, we have analyzed the influence of the skeletal constrained pooling and unpooling for the performance of 3D HPE. As shown in Table 3, the maximum and average pooling achieve inferior results compared with the proposed method. This is mainly caused by that the maximum and average pooling are designed for image, which only compute the node features by maximum or average values of the nodes and ignore the nodes' geography information. Thus, they are not suitable for graph-structured data. While the proposed skeletal constrained pooling /unpooling method considers the nodes structure features when downsample them into lower-scales and pass them into higher-scales, learning valuable features for graph representation, thus, the proposed skeletal constrained pooling/unpooling method achieves the best performance with the lowest MPJEP compared with other pooling and unpooling methods. Further validates that the designed skeletal constrained pooling/unpooling improves the feature representations capability of the proposed model.

4.3.4 Effects of the spatial transformer and temporal transformer

We analyze the impact of the spatial transformer and temporal transformer for 3D HPE by conducting four possible combinations of them: (a) performer spatial transformer model only in U-STN; (b) perform temporal transformer model only in U-STN; (c) performance none of temporal transformer model and spatial transformer model in U-STN and (d) perform both of the temporal transformer model and spatial transformer model in U-STN. Experimental results in Table 4 show that the best result is achieved by applying both of the spatial and temporal transformer model. As the spatial transformer module is designed to encoder local relationships between human body joints from a single frame and the temporal transformer module captures the global dependencies among frames of the input sequence, the performance of only using the spatial transformer model or the temporal transformer model is inferior than applying both of them. This is also consistent with the results in Table 4.

4.4 Comparison with state-of-the-art methods

4.4.1 Results on Human3.6 M

The comparison between results of the proposed method and the SOTA methods on Human3.6M dataset are shown in Tables 5 and 6. In Tables 5 and 6, we report the performance of our model with receptive field $T = 27$ and $T = 243$ on protocol 1 and protocol 2, respectively. The last column is the average performance for all test sequences. Our method achieves average performance of 45.5mm under protocol 1 and 34.8 mm under protocol 2 with receptive field $T = 27$. With the same receptive field, the proposed method outperforms the SOTA methods on Human3.6m dataset,

Table 5 Quantitative comparisons with the state-of-the art methods on Human3.6 M under Protocol #1. (T denotes the number of input frames used in each method.)

Protocol#1	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Martinez [16] ($T = 1$)	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang [35] ($T = 1$)	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Xu [17] ($T = 64$)	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Hossain [9] ($T = 1$)	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Pavllo [10] ($T = 243$)	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
Zou [36] ($T = 1$)	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4
Cai et al [8] ($T = 7$)	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
Liu [18] ($T = 243$)	45.5	48.4	43.9	48.3	49.3	57.6	45.0	45.8	57.3	61.4	49.3	45.3	49.6	33.7	33.4	47.7
Zhao. [37] ($T =1$)	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
Yeh [38] ($T = 243$)	44.8	46.1	43.3	46.4	49.0	55.2	44.6	44.0	58.3	62.7	47.1	43.9	48.6	32.7	33.3	46.7
Lin [39] ($T = 50$)	42.5	44.8	42.6	44.2	48.5	57.1	52.6	41.4	56.5	64.5	47.4	43.0	48.1	33.0	35.1	46.6
Wang [6] ($T = 96$)	41.4	43.9	44.0	42.2	48.0	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
Zheng [1] ($T = 27$)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	47.0
Zheng [1] ($T = 81$)	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
Lin [21] ($T = 1$)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	54.0
Li [20] ($T = 27$)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	46.9
Li [20] ($T = 351$)	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	44.2	43.0	44.2	30.0	30.2	43.7
Ours ($T = 27$)	41.8	45.2	43.5	43.4	46.9	53.0	43.0	41.4	56.2	65.7	46.2	42.9	46.5	33.0	34.1	45.5
Ours ($T = 243$)	40.9	43.8	41.2	42.5	44.5	51.6	41.3	40.8	55.2	62.1	44.8	41.9	44.8	31.5	32.8	43.9

Table 6 Quantitative comparisons with the state-of-the art methods on Human3.6 M under Protocol #2. (T denotes the number of input frames used in each method.)

Protocol#2	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Martinez [16] ($T = 1$)	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang [35] ($T = 1$)	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Gong [40] ($T = 1$)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	39.1
Hossain [9] ($T = 1$)	35.7	39.3	44.0	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Pavllo [10] ($T = 243$)	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Zou [36] ($T = 1$)	35.7	38.6	36.3	40.5	39.2	44.5	37.0	35.4	46.4	51.2	40.5	35.6	41.7	30.7	33.9	39.1
Xu [17] ($T = 64$)	33.9	37.2	36.8	38.1	43.5	43.5	37.8	35.0	47.2	53.8	40.7	38.3	41.8	30.1	31.4	39.0
Cai et al [8] ($T = 7$)	35.7	37.8	36.9	40.7	45.2	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
Liu [18] ($T = 243$)	34.9	37.5	34.9	38.3	37.4	44.0	34.4	34.6	45.1	48.0	49.3	34.8	37.7	26.2	27.1	36.9
Lin [39] ($T = 50$)	32.5	35.3	34.3	36.2	43.0	43.0	33.0	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Xu [41] ($T = 9$)	31.0	34.8	34.7	34.4	43.9	31.6	52.6	33.5	42.3	49.0	37.1	33.0	39.1	26.9	31.9	36.2
Liu [42] ($T = 243$)	32.3	35.2	33.3	35.8	41.5	33.2	44.6	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
Wang [6] ($T = 96$)	32.9	35.2	35.6	34.4	42.7	31.2	42.2	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
Chen [5] ($T = 243$)	32.6	35.1	32.8	35.4	40.4	32.4	35.7	32.3	42.7	49.0	36.8	32.4	36.0	24.9	26.5	35.0
Zheng [1] ($T = 81$)	32.5	34.8	32.6	34.6	39.5	32.1	42.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
Li [20] ($T = 351$)	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
Ours ($T = 27$)	31.1	35.0	32.1	34.1	39.2	39.1	31.5	32.2	45.1	49.8	36.2	31.7	36.1	23.2	26.8	34.8
Ours ($T = 243$)	30.0	33.6	31.0	32.3	36.9	36.4	30.6	31.2	44.9	48.1	35.8	30.5	35.5	22.8	24.5	33.6

achieving better performance of 3D HPE in most evaluation metrics. Compared with the temporal transformed-based method, such as PoseFormer [1], strided Transformer [20] and METRO [21], the proposed method has a better performance with smaller MPJPE and P-MPJPE. For example, the average MPJPE of the proposed method with receptive field $T = 27$ on protocol 1 is 45.5 mm, which is 1.5 mm, 1.4 mm and 8.5 mm smaller than that of method in PoseFormer in [1] (47.0 mm with T =27), Strided Transformer [20] (46.9 mm with $T = 27$) and METRO in [21] (54.0 mm with $T = 1$). The average MPJPE of the proposed method with receptive field $T = 27$ on protocol 2 is 34.8 mm and is 1.3 mm less than that in Strided Transformer [20] (36.1 mm with $T = 243$). Besides, the proposed method also has better performance than many GCN based 3D HPE, such as Graph stacked hourglass network in [17] (51.9 mm on protocol 1 and 35.8 mm on protocol 2 with $T = 64$), Cai et al. method in [8] (48.8 mm on protocol 1 and 39.0 mm on protocol 2 with $T = 7$). Additionally, the proposed method outperforms the U-net based 3D HPE methods, for example, the MPJEP and P-MPJEP of our method are all smaller than UGCN proposed in [6]. The above comparisons clearly demonstrate that the proposed method has a good performance. This is mainly attributed to the fact that the proposed method has encoded the complementary characteristic of local and global skeleton features in intra-frame and inter-frames by the U-shaped multi-scale and multi-level features extraction model, which not only bring rich information from small to large resolutions of the input, but also take important semantic information at all levels from shallow to deep of the model. Moreover, a skeletal constrained pooling and unpooling layer is designed to transform the features from various scales and different depths of network, which is beneficial for the proposed model to effectively integrate the global and local features from full skeleton to local part via shallow to deep layers of the network. This further boosts the feature representation capabilities of the proposed model and enables the model to achieve good performance for 3DHPE.

To further demonstrate the effectiveness of the proposed method, we have compared the MPJPE metric for individual joints in some difficult actions, such as Photo, WalkDog and Smoke in Human3.6 M test set S11. Figure 4 shows the average joint error of action Photo on S11, which is a challenging sequence with serious self-occlusion and rapid movement. The action that moves quickly always needs long frames to capture the correlations. It can be seen from Fig. 4 that our method has smaller errors than compared methods, such as Pavllo et al. [10] and Chen et al. [5]. For body joints, our method achieves significant improvement, e.g., right wrist (109.0 mm), left wrist (95.4 mm), and right elbow (86.3 mm). This further proves that the proposed method can effectively encoder the global dependencies and local information for 3D HPE. It is particularly beneficial for our method to estimation these difficult joints.

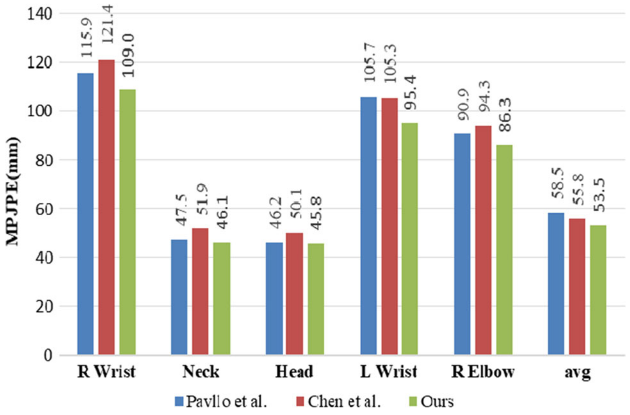


Fig. 4 comparation the average joint error of Photo action on S11

4.4.2 Results on HumanEva-I

To further evaluate the generalization performance of the proposed method, we employ the model trained on the Human 3.6 M to the HumanEva-I dataset. The comparison results of our method with SOTA method on HumanEva-I dataset are shown in Table 7. Although the proposed model only trained on Human 3.6 M, our method achieves promising results, demonstrating that the proposed method has a good generalization capability on unseen dataset.

4.5 Computational complexity analysis

Table 8 compares the total number of parameters, floating-point operations (FLOPs) and the frame per second (FPS) with SOTA methods in different receptive fields on Human3.6 M under Protocol 1 with MPJPE. Compared with SOTA methods, our model achieves competitive performance for 3D HPE with small receptive field and relatively fewer parameters. The total number of parameters does not increase much when the receptive field is increased. This is caused by the fact that the length of the receptive fields mainly affects the temporal positional embedding in temporal transformer layer, which does not require many parameters. As shown in Table 8, the FPS of our model is lower than the compared methods, it is still acceptable for real-time inference since our model follows the 2D-to-3D lifting method, where the 2D pose detector provides the 2D pose coordinates is usually below 80 FPS.

4.6 Visualization results

The qualitative results of our method on Human 3.6 m are shown in Fig. 5. We only present some challenging examples on S9 and S11 to show the effectiveness of the proposed method. Figure 5 shows the estimated 3D pose by the proposed method and the corresponding ground truth 3D pose.

Table 7 Comparison on HumanEva-I under protocol #2

Protocol #2	Walk			Jog			Box		
	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$
Martinez [16]	19.7	17.4	46.8	26.9	18.2	18.6	–	–	–
Pavlakos [14]	18.8	12.7	29.2	23.5	15.4	14.5	–	–	–
Lee [43]	18.6	19.9	30.5	25.7	16.8	17.7	42.8	48.1	53.4
Pavllo [10]	13.9	10.2	46.6	20.9	13.1	13.8	23.8	33.7	32
GAST [18]	13.7	9.2	46.2	20.1	12.5	12.7	21.8	27.8	27
Ours	12.6	10.1	45.7	20.9	12.0	12.8	21.2	20.2	27.1

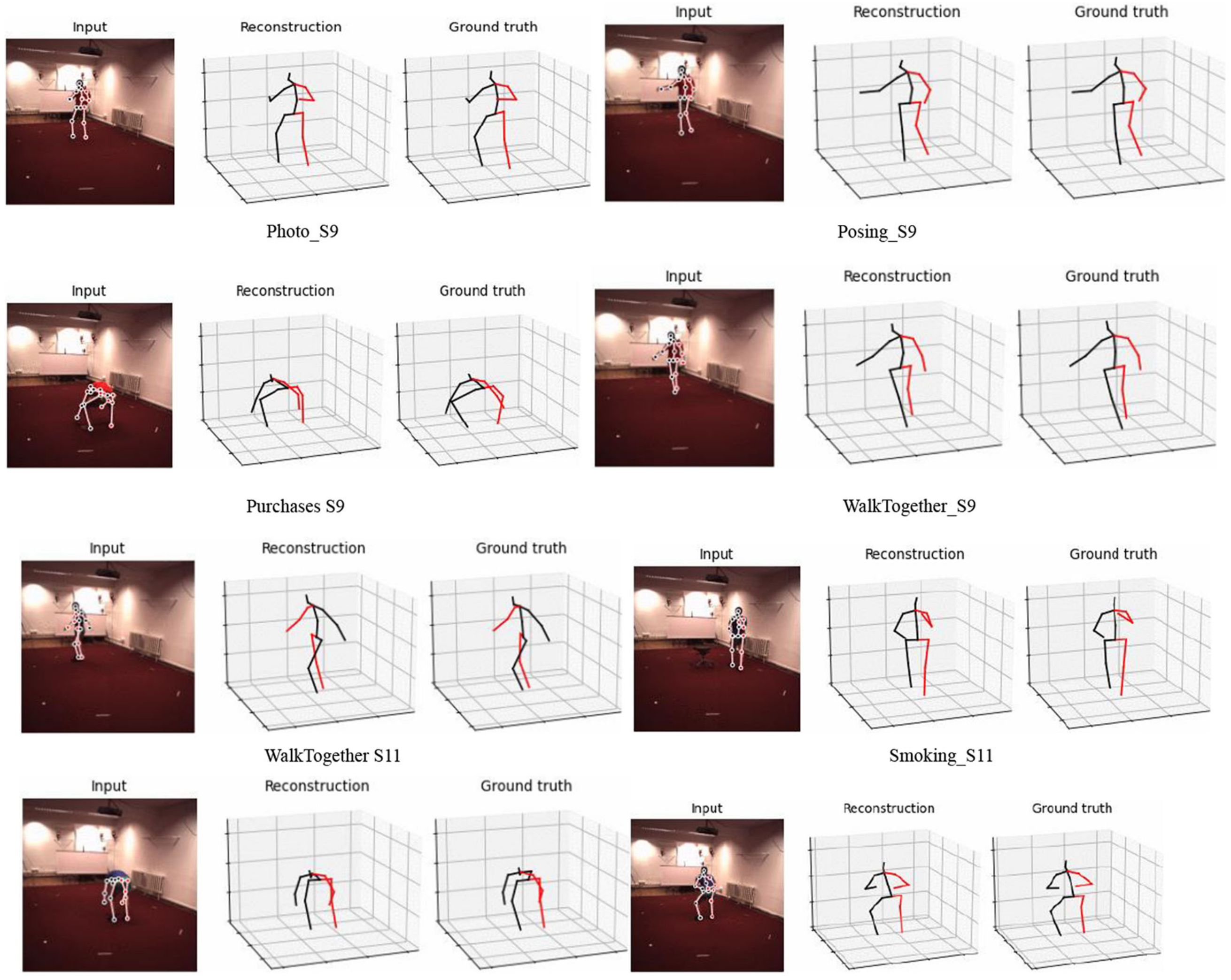
Table 8 Computational Complexity analysis on Human3.6 M under Protocol 1 with different receptive fields

	receptive field T	Parameters (M)	FLOPs (M)	MPJPE(mm)	FPS
Hossain [9]	–	16.95	33.88	58.3	
Pavllo et al. [10]	27	8.56	17.09	48.8	1492
Pavllo et al. [10]	81	12.79	25.48	47.7	1121
Pavllo et al. [10]	243	16.95	33.87	46.8	863
Chen et al. [5]	27	31.88	61.7	45.3	410
Chen et al. [5]	81	45.53	88.9	44.6	315
Chen et al. [5]	243	59.18	116	44.1	264
Zheng [1]	9	9.58	11.2	49.9	320
Zheng [1]	27	9.59	33.9	47.0	297
Zheng [1]	81	9.60	101	44.5	269
Li et al. [20]	27	4.01	128	46.9	118
Li et al. [20]	81	4.06	392	45.4	112
Li et al. [20]	243	4.23	1372	44.0	108
Ours	27	15.18	61.2	45.5	165
Ours	81	15.24	183	44.4	150
Ours	243	15.42	550.8	43.9	145

It can be seen that the proposed method can successfully estimate the 3D pose.

5 Conclusion

In this paper, we have developed a U-shaped spatial–temporal transformer network for 3D HPE from monocular images. To better encoder the complementary characteristic of local and global skeleton features in intra-frame and inter-frames, we design a U-shaped multi-scale and multi-level features extraction model based on spatial–temporal transformer architecture. With the skeletal constrained pooling and unpooling layer to transform the features from various scales and different depths of network, the proposed model can effectively integrate the global and local features from full skeleton to local part, which is useful for boosting the feature representations of the proposed model. The experimental results show that the proposed model achieves the state-of-the-art performance on two benchmark 2D-to-3D pose estimation datasets.



Photo_S9

Posing_S9

Purchases S9

WalkTogether_S9

WalkTogether S11

Smoking_S11

Purchases S11

SittingDown_S11

Fig. 5 Visualization results of our method on Human3.6 M test set S9 and S11

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (No.61907028, No.11872036), the Young science and technology stars in Shaanxi Province (2021KJXX-91), the Young Talent fund of University Association for Science and Technology in Shaanxi (No. 20200105) and the Fundamental Research Funds for the Central Universities (No. GK202103114, No. GK2021011004, No. 2022TD-26).

References

1. Zheng, C., Zhu, S., Mendieta, M., et al: 3D human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 11656–11665 (2021)
2. Malik, Z., Shapiai, M.: Human action interpretation using convolutional neural network: a survey. Mach. Vis. Appl. **33**(3), 1–23 (2022)
3. Moon, G., Lee, K.M.: I2l-meshnet: Image to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 752–768 (2020)
4. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7307–7316 (2018)
5. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3D human pose estimation with bone-based pose decomposition. IEEE Trans. Circuits Syst. Video Technol. **32**(1), 198–209 (2022). https://doi.org/10.1109/TCSVT.2021.3057267
6. Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3D pose estimation from videos. In: Proceedings of the European Conference on Computer Vision 2020 (ECCV), pp. 764–780. Springer, (2020)
7. Wang, R., Tong, J., Wang, X.: Enhancing feature fusion for human pose estimation. Mach. Vis. Appl. **31**, 60 (2020). https://doi.org/10.1007/s00138-020-01104-2
8. Cai, Y., Ge, L., Liu, J., et al.: exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In:

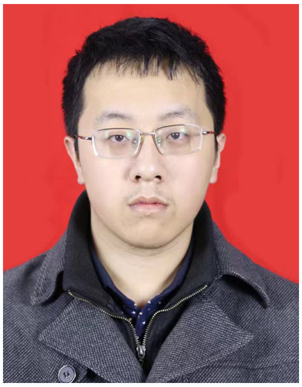
Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2272–2281 (2019)

9. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3D human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 6869–8486, Springer, (2018)
10. Pavllo, D., Feichtenhofer, C., Grangier, D., et al.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7745–7754 (2019)
11. Huang, Z., Shen, X., Tian, X., et al.: Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In: ACM Deep Learning of Multimedia, Seattle, WA, USA, pp. 2122–2130 (2020). https://doi.org/10.1145/3394171.3413666
12. Li, S., Chan, A.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision, pp. 332–347 (2014)
13. Park, S., Hwang, J., Kwak, N.: 3D human pose estimation using convolutional neural networks with 2d pose information. In: European Conference on Computer Vision (ECCV), pp. 156–169, Springer, (2016)
14. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7025–7034(2017)
15. Zeng, A., Sun, X., Huang, F., et al.: SRNet: improving generalization in 3D human pose estimation with a split-and-recombine approach. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 507–523 (2020)
16. Martinez, J., Hossain, R., Romero, J., Little, J.J: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2659–2668 (2017) https://doi.org/10.1109/ICCV.2017.288.
17. Xu, T., Takano, W.: Graph stacked hourglass networks for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16105–16114 (2021)
18. Liu, J., Guang, Y., Rojas, J.: A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 3374–3380 (2021)
19. Li, W., Liu, H., Tang, H., et al.: MHFormer: multi-hypothesis transformer for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13147–13156 (2022)
20. Li, W., Liu, H., Ding, R., et al.: Exploiting temporal contexts with strided transformer for 3D human pose estimation. IEEE Trans. Multimed. (2022). https://doi.org/10.1109/TMM.2022.3141231
21. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1954–1963, (2021) https://doi.org/10.1109/CVPR46437.2021.00199
22. Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, H., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944 (2017) https://doi.org/10.1109/CVPR.2017.106
23. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision 2020 (ECCV), pp. 483–499 (2020)
24. Sun, K., Xiao, B., Liu, D., et al.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5686–5696 (2019)
25. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: a single-shot object detector based on multi-level feature pyramid network. In: The Thirty-Third AAAI Conference on Artificial Intellilgence (AAAI), pp. 9259–9266, (2019) https://doi.org/10.1609/aaai.v33i01.33019259
26. Hua, G., Li, W., Zhang, Q., et al.: Weakly-supervised 3D human pose estimation with cross-view U-shaped graph convolutional network. In: IEEE Transactions on Multimedia, arXiv preprint http://arxiv.org/abs/2105.10882, (2022) https://doi.org/10.48550/arXiv.2105.10882
27. Dosovitskiy, A., Beyer, L., Kolesnikov., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint http://arxiv.org/abs/2010.11929 (2021) https://doi.org/10.48550/arXiv.2010.11929
28. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Trans. Patt. Anal. Mach. Intell. **42**(8), 2011–2023 (2020). https://doi.org/10.1109/TPAMI.2019.2913372
29. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Patt. Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
30. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. **87**(1), 4–27 (2010)
31. Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: a http://arxiv.org/abs/2012.13392v4, https://doi.org/10.48550/arXiv.2012.13392
32. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR), pp. 1–15 (2015), https://doi.org/10.48550/arXiv.1412.6980.
33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: 1Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014)
34. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Proceedings of the European conference on computer vision (ECCV), pp. 646–661 (2016)
35. Fang, H., Xu, Y., Wang, W., Liu, X., Zhu, S.: Learning pose grammar to encode human body configuration for 3D pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, pp. 6821–6828 (2018)
36. Zou, Z., Tang, W.: Modulated graph convolutional network for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 11477–11487 (2021)
37. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N..: Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 3425–3435 (2019)
38. Yeh, R.A., Hu, Y., Schwing, A.G.: Chirality nets for human pose regression. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS), pp. 8163–8173 (2019) https://doi.org/10.48550/arXiv.1911.00029
39. Lin, J., Lee, G.H.: Trajectory space factorization for deep video-based 3d human pose estimation. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 1–13(2019) https://doi.org/10.48550/arXiv.1908.08289

40. Gong, K., Zhang, J., Poseaug, J.F.: A differentiable pose augmentation framework for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8575–8584(2021) https://doi.org/10.48550/arXiv.2105.02465
41. Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W.: Deep kinematics analysis for monocular 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 896–905, (2020) https://doi.org/10.1109/CVPR42600.2020.00098
42. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S., Asari, V.: Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5063–5072 (2020) https://doi.org/10.1109/CVPR42600.2020.00511.
43. Lee, K., Lee, I., Lee, S.: Propagating lstm: 3D pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 123–141 (2018) https://doi.org/10.1007/978-3-030-01234-2_8

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Honghong Yang received the M.Sc. and Ph.D. degrees from the Department of Automation, Northwestern Polytechnical University, Xi'an, China, in 2014 and 2018, respectively. She is an associate professor at the Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an. Her research interests include computer vision, pose estimation and detection. (Email: yanghonghong0615@163.com).



Longfei Guo received the B.S. degree, in 2018. He is currently pursuing the M.S. degree with Shaanxi Normal University. His research interests include artificial intelligence, pose estimation and object tracking. (Email: lfyuanwuqu@163..com).

Yumei Zhang received the Ph.D. degree in control engineering from Northwestern Polytechnical University, Xi'an, China, in 2009. Currently, she is a professor at Shaanxi Normal University. Her research interests include signal processing and artificial intelligence. (Email: zym0910@snnu.edu.cn).

Xiaojun Wu received the Ph.D. degree in system engineering from Northwestern Polytechnical University, Xi'an, China, in 2005. He is a professor at Shaanxi Normal University. His research interests include pattern recognition, intelligent system and system complexity. (Email: xjwu@snnu.edu.cn).