

Spatiotemporal Learning Transformer for Video-Based Human Pose Estimation

Di Gai, Runyang Feng, Weidong Min, *Member, IEEE*, Xiaosong Yang, Pengxiang Su, Qi Wang, Qing Han

Abstract—Multi-frame human pose estimation has long been an appealing and fundamental issue in visual perception. Owing to the frequent rapid motion and pose occlusion in videos, this task is extremely challenging. Current state-of-the-art methods seek to model spatiotemporal features by equally fusing each frame in the local sequence, which weakens the target frame information. In addition, existing approaches usually emphasize more on deep features while ignoring the detailed information implied in the shallow feature maps, resulting in the dropping of crucial features. To address the above problems, we propose an effective framework, namely spatiotemporal learning transformer for video-based human pose estimation (SLT-Pose), which consists of a Personalized Feature Extraction Module (PFEM), Self-feature Refinement Module (SRM), Cross-frame Temporal Learning Module (CTLM) and Disentangled Keypoint Detector (DKD). To be specific, we propose PFEM which extracts and modulates the individual frame features to adapt to the varying human shape, and integrates single-frame features to obtain the spatiotemporal features. We further present SRM to establish global correlation spatial cues on the target frame to attain the refinement feature. Then, a CTLM is designed to search for the information most closely related to the target frame from the spatiotemporal features to intensify the interaction between the target frame and the local sequence, using both the shallow detailed and the deep semantic representations. Finally, we employ DKD to extract the disentangled characteristics of each joint and encode the articulated joint pairs in the human body, promoting the model to reasonably and accurately predict the keypoint heatmaps. Extensive experiments on three human motion benchmarks, including PoseTrack2017, PoseTrack2018, and Sub-JHMDB dataset, demonstrate that SLT-Pose plays favorably against state-of-the-art approaches in terms of both objective evaluation and subjective visual performance.

Index Terms—Human pose estimation, transformer, video, spatiotemporal feature learning

I. INTRODUCTION

The goal of human pose estimation is to detect all the instances from images or videos and estimate the location

This work was supported by the National Natural Science Foundation of China under Grant No. 62076117 and 62166026, the Jiangxi Key Laboratory of Smart City under Grant No. 20192BCD40002 and the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011.

(Corresponding author: Weidong Min.)

Di Gai, Weidong Min, Qi Wang and Qing Han are with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang, 330031, China. They are also with the Jiangxi Key Laboratory of Smart City and Institute of Metaverse, Nanchang University, Nanchang, 330031, China. (e-mail: gaidi@ncu.edu.cn; minweidong@ncu.edu.cn; wangqi@ncu.edu.cn; hanqing@ncu.edu.cn).

Runyang Feng is with the School of Artificial Intelligence, Jilin University, Changchun, 130000, China. (e-mail: runyang2019.feng@gmail.com)

Xiaosong Yang is with the National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, U.K. (e-mail: xyang@bournemouth.ac.uk).

Pengxiang Su is with the School of Software, Nanchang University, Nanchang, 330031, China. (e-mail: supengxiang@ncu.edu.cn)

information of kinematic keypoints. It is a fundamental task in the field of computer vision and has a broad spectrum of applications, including motion recognition [1], [2], human body reconstruction [3], [4], human motion prediction [5], and human-computer interaction [6].

Most of the earlier approaches [7], [8] focused on estimating human pose from static images, employing a probabilistic graphical model or tree model to simulate the human structure. Unfortunately, these methods demand manual annotation of a large number of features and have a limited ability to deal with the challenges arising from complex backgrounds. Benefiting from the recent advances in deep learning techniques, researchers have attempted to utilize deep convolutional neural networks to address pose estimation tasks. DeepPose [9] directly obtains the final result by estimating the coordinate information of keypoints via an iterative architecture. Subsequently, fully convolutional networks like [10]–[13] learn the spatial correlation features among body joints by generating heat maps of keypoints and then output the estimation results. Since static human body images may be blurred and occluded, leading to information loss from a single image. To adapt to the video streams, these methods usually work in a frame-by-frame manner. The direct application of these methods to video results in inferior generalization ability, which regularly triggers low accuracy of predicted keypoint locations and exhibits temporal discontinuity.

When applied in practical scenarios, such as surveillance tracking and human behavior understanding, it is necessary to process video data and recognize human posture from the video. Accordingly, it is a tremendous challenge to effectively utilize the rich temporal information between video frames to efficiently promote human pose estimation. Recently, prevalent approaches have been developed to enhance the performance of video-based pose estimation models by fusing extra temporal contextual information from adjacent frames to complement the missing motion features of the target image [14]–[16]. [17] the pose machine is a chain network structure derived from the pose machine, which is a representative model based on the long and short-term memory network (LSTMpose machine, LPM). On the other hand, the convolutional-recurrent neural network approach can extract temporal and spatial features from the video to facilitate the model to generate accurate estimation results. The LPM constructs a simple recruiting structure with the weight-sharing training strategy and recurrent components with strong short-term memory, which improves the learning ability of the model for temporal correlation between video sequences. Nevertheless, the above methods focus more on local features

of the human body, and it is hard to comprehensively interpret the human pose from a global viewpoint. Fortunately, the Transformer-based image recognition model [18] has the ability to adequately construct global relationships and exhibits promising performance, hence researchers have also applied it to the pose estimation problem.

The Transformer-based model shows advantages in modeling long-range correlations by constructing semantic connections between each patch, which consequently provides a global understanding of human pose. However, existing methods tend to aggregate temporal features of neighboring frames directly from short video sequences, which exceedingly concentrates on global features. The similarity with the spatial features of the target frames fails to be modeled sufficiently to obtain the most closely related features, *i.e.*, each frame are processed equally, weakening target information. Moreover, state-of-the-art methods simply adopt the deep semantic feature maps from deep networks for informal learning while ignoring the extraction of detailed information implied in shallow feature maps [2,10], which results in the dropping of some key features.

To satisfy these requirements, we propose a novel framework, namely spatiotemporal learning transformer for video-based human pose estimation (SLT-Pose), which consists of four components. 1) We propose the Personalized Feature Extraction Module (PFEM), where we first use the backbone to process five consecutive frames of character information separately and fuse the multi-scale feature information from different resolution sub-networks to enhance the representation of features in each frame. Subsequently, since traditional convolutional operations are difficult to adapt to the complexity and variability of the human shape, we utilize deformable convolutions (DCN) to increase the feature extraction capability for deformed targets while expanding the perceptual domain. Finally, the spatiotemporal features of the local sequence are obtained by integrating all the single-frame features with a stack of basic blocks. 2) A Self-feature Refinement Module (SRM) is designed to establish global correlation spatial cues on the target frame to attain the refinement feature. 3) We present a Cross-frame Temporal Learning Module (CTLTM) to strengthen the interaction between the target frame and the local sequence by locating the information most intimately associated with the target frame from the spatio-temporal features. 4) We suggest Disentangled Keypoint Detector (DKD), which exploits the disentangled features of each joint to capture the joint pairs of human body and promote the model to predict the keypoint heat map rationally and accurately.

In summary, our main contributions are summarized as follows:

- A PFEM is used to encode personalized features of the human body and effectively aggregate keypoint across frames to obtain the spatiotemporal features of the local sequence. Additionally, an SRM is applied to extract and polish the features of the target frame.
- A CTLTM is conceived to detect the information related to the target frame from the spatiotemporal features of the local sequence and fuse the obtained shallow local features with the deep refinement feature of the target

frame to improve interaction. Moreover, a series of parallel keypoint detection heads extract the corresponding joint features, and then model the relevant joint pairs of features on the human body structure in DKD.

- On three human motion benchmark datasets, including PoseTrack2017, PoseTrack2018, and Sub-JHMDB dataset, our method achieves the state-of-the-art performance in both quantitative and qualitative.

II. RELATED WORK

A. Human pose estimation in images

The image-based human pose estimation method refers to the exact recognition of the location of keypoints from a single RGB image. In the previous stage, researchers have extensively used traditional models based on probability maps to overcome this challenge by representing the connection relationships between joints through a tree diagram structure or a forest model and further generating coordinates of key points in the human body [7], [8], [19]. However, the above methods require a large number of features to be manually labeled in advance and have poor nonlinear characterization capability to perfectly perform the pose estimation task. Recently, numerous deep learning-based models have been proposed and have revealed satisfactory properties. [20] proposes a stacked hourglass network in which the subgraphs that correspond to the human body region are adopted as input through a series of convolution and pooling operations to obtain the base feature maps, and then the feature maps are fed into the corresponding hourglass structure to achieve relatively accurate estimation results. [21] presents a spatial contextual information (SCI) based human pose estimation method, which uses a cascade prediction fusion (CPF) stacked hourglass network as the basic framework. This method learns the underlying keypoint heatmap to introduce the pose graph neural network (PGNN) for pose estimation, which optimizes the key point estimation. [22] exploits a high-resolution feature representation learning network to alleviate the issue of partial information loss caused by the downsampling operation, which consistently maintains a high-resolution feature map as the main line of operation. When downsampling is performed, a novel low-resolution sub-network is generated, and the network is continuously expanded horizontally to avoid the information loss caused by upsampling and downsampling operation via multiple parallel sub-networks that maintain the high-resolution features from beginning to end. [23] designs a high-resolution feature representation learning network for single human pose estimation, which utilizes deconvolution to accomplish upsampling and introduces multiscale supervision. In addition, the method incorporates linked embedded vectors to propose a scale-aware higher-resolution feature representation learning strategy. OPEC-Net [24] applies AlphaPose+ as the pose estimation network and uses the graph neural network to discover the body structure information to deduce the coordinates of the occluded keypoints, which allows the modification of the information obtained from the initial estimation. Static images inevitably suffer from motion blur caused by rapid motion and lack of information due to pose occlusion, which poses a

challenge for the exact estimation of human pose and limits the wide range of applications.

B. Human pose estimation in videos

Currently, image-based human pose estimation methods perform high prediction precision when processing a static human body in a single image. However, when it is used directly to process video tasks, there is significant performance degradation. To address this challenge, researchers have developed a number of algorithms for human pose estimation applied to video scenes, which are centered on exploiting the temporal features implied in the video data to facilitate the connection of information between video frames. [25] constructs a simple baseline for estimating human heatmaps from deep and low resolution feature maps based on a few deconvolutional layers added on a backbone network. It first encodes each frame of human features using a detector and then links its detections in a graph, where the temporal similarity of the features between frames is measured by intensive optical flow to reduce the interference of noisy information. [17] creates a chained network structure model, which improves the learning ability between video sequences by constructing recurrent structures aided by weight-sharing training strategies and recurrent convolution modules with strong short-term memory. A clip tracking network is presented in [15], which first estimates the basic human bounding box in keyframes, and then crops the subgraphs of the region from all frames of the video slice to constitute a regional video slice. The slice is then fed into a 3D high-resolution representation network, which yields the pose of the person positioned in all frames of the regional video slice that belongs to the keyframes. Finally, the estimated human pose is combined with temporal and spatial smoothing to obtain an accurate human pose estimate. [26] proposes a new model of graph neural network (GNN) that explicitly accounts for both spatial temporal and visual information. It first takes the historical pose sequence of the person as input and predicts the position of each key point. Then, the predicted pose and the detected pose features are aggregated to output the final estimated pose, which provides a pose detection process independent of the target frame image information to achieve stronger robust human pose estimation. [27] projects a hierarchical alignment framework and a complementary information extraction strategy to deal with the issue of negative information accumulation due to spatial misalignment between pose features of the target frame and adjacent frames. It progressively updates the adjacent frame information by coarse-to-fine the human features, thus aligning them with the target frame at the feature level. Then, it further utilizes a mutual information loss function for efficient supervision of inter-frame features to assist the model in capturing relevant cues from the input video sequence. Unfortunately, the above approach fails to adequately model the global information reducing the utilization of valid information from adjacent frames, which leads to a decrease in the accuracy of pose estimation.

C. Transformer-based method for human pose estimation

Earlier, Transformer was applied to machine translation, which excels in developing long-range dependencies between modeled sequence elements. CNN-based models have a narrow information-receiving domain, making it difficult to build features with remote dependencies. In contrast, transformer can effectively establish long-range dependencies and has powerful characterization capabilities. Recently, researchers have used transformer to solve a variety of computer vision tasks, including, human pose estimation, motion prediction, image segmentation, etc., which demonstrates remarkable performance on a broad range of benchmark datasets. TransPose [28] implements the attention layer of the transformer to efficiently capture long-range relationships and implicitly reveals the dependencies between keypoints to provide interpretations of the global spatial relationships which exist in the whole video for the model layer-by-layer inference. Its last attention layer acts as an aggregator to fuse the contributions of image cues to form the maximum positions of keypoints. Further, TransPose indicates that the revealed dependencies are image-specific and fine-grained, which can provide the model with valid information for treating special cases (*e.g.*, occlusion and blurring). [16] introduces a new Transformer-based network to address the problem of jitter and inconsistencies arising in skeleton estimation by learning pose and motion distributions in an unsupervised manner to encode the temporal evolution model of the scene and skeleton. The ability of the Skeletor to implicitly model the spatio-temporal context of human motion, applying masking and noise enhancement at both frame- and joint-level enables it to overcome inaccuracies in detection and corrects partial or entire skeleton corruption. [3] presents a unified framework dubbed Multi-view and Temporal Fusing Transformer, which has the ability to tackle diverse viewpoints and arbitrary length video information in uncalibrated scenes. It adaptively measures the relationship between each image pair and reconstructs features from multiple viewpoints by means of a relative-attention mechanism, thus reducing the reliance of the model on annotation information. [29] suggests a Transformer-based fully end-to-end multi-person Pose Estimation framework called PETR, which treats the pose estimation task as a hierarchical set prediction problem and removes a large number of modules that require manual features. PETR first studies the multiple pose features to deduce a set of full-body poses, and then further refines the kinematic relationships between human joints so that the model can adaptively capture the features most relevant to the target key points. [30] takes full advantage of a redundant auxiliary frame sequence to learn representative representations and designs a new architecture called Strided Transformer. It has the ability to model long-range dependencies between human features and capture local contextual information to a single-vector representation in a hierarchical global and local fashion to reduce computational costs. Nevertheless, the mentioned methods equally treat each frame in the sequence, which may weaken the information of the target frame, thus adding redundant information and limiting the model performance.

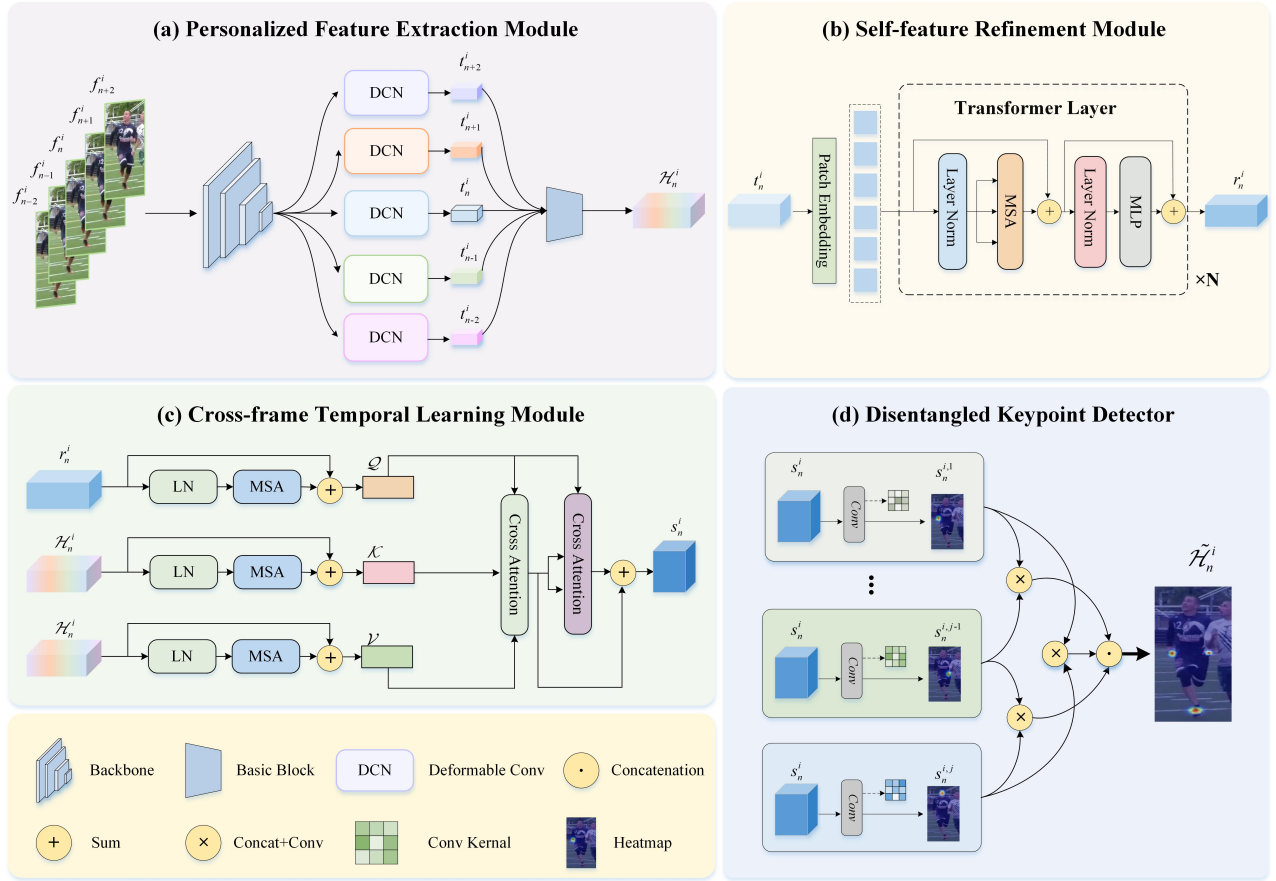


Fig. 1. The overall structure of SLT-Pose contains four modules: (a) Personalized Feature Extraction Module, (b) Self-feature Refinement Module, (c) Cross-frame Temporal Learning Module, and (d) Disentangled Keypoint Detector.

III. OUR METHOD

Our target is to establish a transformer-based framework for executing multi-person pose estimation task in videos. The pipeline of proposed STL-Pose is delineated in Figure 1, which involves four components. 1) we first obtain a person i in a consecutive frame sequence $\mathcal{F}_n^i = [f_{n-2}^i, f_{n-1}^i, f_n^i, f_{n+1}^i, f_{n+2}^i]$ by an off-the-shelf object detector, and feed them to the personalized feature extraction module to derive the spatiotemporal features of the local sequence $\mathcal{H}_n^i$. 2) The target frame feature t_n^i is then prominently polished using a self-feature refinement module to obtain feature r_n^i . 3) The refinement feature r_n^i of the target frame and the spatiotemporal feature $\mathcal{H}_n^i$ of the local sequence are fed together to the cross-frame temporal learning module to yield the comprehensive feature s_n^i . 4) We pass the comprehensive feature s_n^i to the disentangled keypoint decoder to output the heatmap $\mathcal{H}_n^i$ of each corresponding keypoint. In the following, we elaborate each component in detail.

A. Personalized Feature Extraction Module

High-resolution features contain detailed spatial information, which is beneficial for localization tasks, such as locating the positions of nodes. Meanwhile, low-resolution features have rich semantic information, which is conducive to classification tasks, such as the grouping of nodes. Consequently,

it is crucial to effectively fuse the characteristics of different resolutions to improve the accuracy of human pose estimation. High Resolution Net (HRNet) [22] maintains high resolution feature representation by parallelizing high resolution and low resolution sub-networks and progressively adding low resolution features. By repeatedly performing multi-scale feature fusion, information from different resolution sub-networks is repeatedly fused and feature representation is enhanced, which is a strategy to preserve high and low resolution features that can significantly raise the accuracy of human pose estimation. Based on this, we use HRNet as the backbone to obtain coarse features.

$$\langle f_{n-2}^i, \dots, f_{n+2}^i \rangle \xrightarrow{\text{Backbone}} \xrightarrow{\text{DCN}} \langle t_{n-2}^i, \dots, t_{n+2}^i \rangle, \quad (1)$$

where $f_n^i \in \mathbb{R}^{3 \times 384 \times 288}$, $\langle f_{n-2}^i, \dots, f_{n+2}^i \rangle$ denotes the consecutive frame sequence and $\langle t_{n-2}^i, \dots, t_{n+2}^i \rangle$ is the feature sequence after backbone and DCN processing.

$$\langle t_{n-2}^i, \dots, t_{n+2}^i \rangle \xrightarrow[\text{Basic Block}]{\text{stack of}} \mathcal{H}_n^i, \quad (2)$$

where $t_n^i \in \mathbb{R}^{48 \times 96 \times 72}$, $\mathcal{H}_n^i$ is spatiotemporal features of the local sequence. We adjust a residual structure from [31] to design the Basic Blocks in Equation 2.

The human body shape in the video is usually complex and variable, and feature extraction using merely traditional

convolutional kernels is challenging to adapt to this context. Deformable convolution is the addition of an offset to each sample point in a conventional convolution kernel, which permits each sample point to be sampled in the immediate region of the original sample point, instead of restricting it to a rectangular region. This region can be arbitrary in shape, which expands the perceptual field and increases the feature extraction capability for deformable targets. For a given input feature f_n^i , the expression for each pixel p_0 in the output feature t_n^i is as in formula Eq. (3),

$$t_n^i(p_0) = \sum_{p_n \in R} w(p_n) \times f_n^i(p_0 + p_n + \Delta p_n), \quad (3)$$

where w denotes the weight of the convolution kernel on p_n , $R \in \{(-1, 1), (0, 1), \dots, (1, -1)\}$, and Δp_n is the coordinate offset, which is obtained by convolutional learning.

B. Self-feature Refinement Module

The perceptual field of CNN is limited by the convolutional kernel size, and it is extremely restricted to observe the global information of the whole human body, while the transformer module can establish global correlation cues for each patch of the human body. The Self-feature Refinement Module contains multiple identical transformer layers. An MSA module and an MLP module are connected in tandem via Layer Norm (LN) and skip connection operations to form each transformer layer. Before inputting the feature map into the first transformer layer, the Self-feature Refinement Module will use a Patch Embedding module to perform patching and embedding operations on the feature map.

Patch Embedding: For the target frame feature $t_n^i \in \mathbb{R}^{H \times W \times C}$, it first needs to be reshaped into a sequence of flattened patches $tp_n^i \in \mathbb{R}^{N \times (P^2 \times C)}$. Where (H, W, C) is the shape of the feature map, (P, P, C) is the shape of each patch, and $N = HW/P^2$ is the number of patches. With a constant latent vector size D used in all transformer layers, we need to perform a linear projection to compress the sequence of flattened patches into an $N \times D$ feature map.

Multihed Self-Attention (MSA): The formula for the self-attention mechanism is expressed as in Eq. (4),

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V, \quad (4)$$

where Q is the query matrix, K is the key matrix, V denote value matrix, and d_k represent the dimension of the Q and K .

The MSA module is a combination of several self-attention mechanism modules, which is given by the Eq. (5),

$$\begin{aligned} \text{MSA}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O, \\ \text{head} &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \end{aligned} \quad (5)$$

where the W^O , W_i^Q , W_i^K and W_i^V is the projections parameter matrices.

Multilayer Perceptron (MLP): The MLP is mainly composed of two fully connected layers and a linear activation layer.

$$\text{MLP} = \max(0, XW_1 + b_1)W_2 + b_2, \quad (6)$$

where W_1 , b_1 , and W_2 , b_2 represent the weights and biases of the two fully connected layers, respectively.

In addition, an LN layer is applied before each module, and a skip connection is employed after each module. Thus, each transformer layer can be expressed as in Eq. (7).

$$\begin{aligned} tp_n^{i'} &= \text{MSA}(\text{LN}(tp_n^i)) \oplus tp_n^i, \\ r_n^i &= \text{MLP}(\text{LN}(tp_n^{i'})) \oplus tp_n^{i'}, \end{aligned} \quad (7)$$

where $r_n^i \in \mathbb{R}^{48 \times 96 \times 72}$ is the final feature processed by Self-feature Refinement Module.

C. Cross-frame Temporal Learning Module

Due to the possibility of motion blur and pose masking in the target frame resulting in some absent information, the Self-feature Refinement Module only manipulates the target frame features which do not take into account the successive frames and are unable to leverage the affluent spatiotemporal information. Therefore, we have to aggregate the local sequence features to compensate for the missing information. Nevertheless, the existing methods merely stack the features of distinct frames directly by channel and then process them with CNN, which can weaken the information of the target frame. In addition, the existing transformer obtains the deep feature map from the deep network and ignores the effective information implied in the shallow feature map, which causes the loss of some pivotal features. To address above issues, we propose the Cross-frame Temporal Learning Module, which adopts the features of the current frame as the search features and the spatiotemporal features as the auxiliary information to explore the relevant cues to the current frame. Specifically, the spatio-temporal features of the local sequence are represented as key matrix $\mathcal{K}$ and value matrix $\mathcal{V}$, and the refined target frame feature r_n^i is used as a query matrix $\mathcal{Q}$, and then the cross-attention calculation is performed,

$$\begin{aligned} \langle \text{MSA}(\text{LN}(r_n^i)) \oplus r_n^i \rangle W_q &\rightarrow \mathcal{Q}, \\ \langle \text{MSA}(\text{LN}(\mathcal{H}_k^i)) \oplus \mathcal{H}_k^i \rangle W_k &\rightarrow \mathcal{K}, \\ \langle \text{MSA}(\text{LN}(\mathcal{H}_v^i)) \oplus \mathcal{H}_v^i \rangle W_v &\rightarrow \mathcal{V}, \end{aligned} \quad (8)$$

where $\text{MSA}(\cdot)$ denotes multi-headed self-attention, $\text{LN}(\cdot)$ is layer normalization, and $\oplus$ denotes an add operation. W_q , W_k , and W_v denote the linear transformation matrices respectively, which are multiplied with the input vectors to obtain the projections on different spaces to enhance the representational capability of the model. The specific formulation is expressed as in formula Eq. (9),

$$\text{CrossAttention}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)\mathcal{V}, \quad (9)$$

where $\text{softmax}(\cdot)$ is the activation function, and d denotes the dimensions of the matrix.

D. Disentangled Keypoint Decoder

Each joint has different motion patterns, and there are significant differences between them. For example, in the walking state, the left wrist and the right wrist typically have opposite motion trends, which leads to a considerable discrepancy in the spatial characteristics of the two joints. Most existing methods simply use a detector to couple the features of each joint, but the non-articulated joint features introduce interference to the precise understanding of human body information, which can contribute to a dramatic decrease in the accuracy of pose estimation.

To tackle this challenge, we propose a Disentangled Keypoint Decoder, which employs a series of parallel 3×3 convolutions to model the salient features of different joints. Formally,

$$\begin{aligned} s_n^i &\xrightarrow[Conv]{3 \times 3} s_{n,1}^i, \\ s_n^i &\xrightarrow[Conv]{3 \times 3} s_{n,2}^i, \\ &\vdots \\ s_n^i &\xrightarrow[Conv]{3 \times 3} s_{n,J}^i, \end{aligned} \quad (10)$$

where s_n^i is the comprehensive feature, $s_{n,j}^i$ denotes the feature of the joint j of the person i after the corresponding convolution process, j is the current joint, and J is the total number of joints.

Further, joint spatial dependencies in the human body are sufficiently considered to couple the tight features of articulated joint pairs (e.g., left wrist and left elbow) so as to raise the structural stability of pose estimation. The formula is expressed as,

$$\bigodot_{n,i}^{j \in [1:J]} \left(s_n^{i,j} \otimes s_n^{i,\tilde{j}} \right) \rightarrow \tilde{\mathcal{H}}_n^i, \quad (11)$$

where $\bigodot$ denotes the concatenation operation, $s_n^{i,\tilde{j}}$ is a feature with skeletal connectivity to $s_n^{i,j}$, $\otimes$ is a concatenation operation followed by a convolution operation and $\tilde{\mathcal{H}}_n^i$ is the target heat map. n is the number of frames, i is the person i in this frame, and J is the total number of joints.

E. Loss Function

Most of the existing work uses standard human pose estimation loss function [4], [27]. However, different keypoints have different estimation difficulties. For example, the wrist and ankle have faster movement and regularly show motion blur, while the hip joint has less movement and the corresponding image is clearer. It is extremely unreasonable to assign the same weight to all keypoints. We assign higher weights to joints with larger movement velocities. Empirically, it forces the model to focus on these keypoints that are difficult to estimate. Formally,

$$\mathcal{L}_{\mathcal{H}} = \frac{1}{I \times J} \sum_{i=1}^I \sum_{j=1}^J \varphi_n^{i,j} \times \left(y_n^{i,j} \times \left\| \tilde{\mathcal{H}}_n^{i,j} - \tilde{\mathcal{H}}_n^{i,j} \right\|_2 \right), \quad (12)$$

where I is the total number of characters in frame n , $\varphi_n^{i,j}$ is the weight of keypoints, $y_n^{i,j}$ is visibility, and $\mathcal{H}_n^{i,j}$ is Ground Truth.

IV. EXPERIMENT

A. Experimental Settings

Datasets: PoseTrack is a dataset about human bone keypoints and multi-person human keypoint tracking dataset, which contains single-frame keypoint detection, multi-frame keypoint detection, and multi-person keypoint tracking. **The PoseTrack2017** dataset contains a total of 514 video clips and 16,219 pose annotations, of which 250 video clips are divided as the training set, 50 video clips as the validation set, and 214 video clips as the test set [32]. **The PoseTrack2018** dataset comprises 1138 video clips and 153,615 pose annotations, including 513 clips in the training set, 170 clips in the validation set, and 375 clips in the test set [33]. **The Sub-JHMDB** dataset is a fully annotated dataset for human motion recognition and human pose estimation, which contains 316 video clips with a total of 11,200 frames providing 15 joint annotations per person [34].

Implementation Details: Experiments are conducted on PoseTrack2017, PoseTrack2018, and Sub-JHMDB datasets using the same configuration, and training, validation and testing are performed independently. During the training period, we employ popular data enhancement strategies including random rotation $[-45^\circ, 45^\circ]$, random scaling of size $[0.65, 1.35]$, image truncation, and horizontal flipping. The size of the input image is fixed uniformly at 384×288 . The parameters of the network are initialized from a Gaussian distribution with mean 0 and variance 0.001, and all bias parameters are initialized to 0. We apply the HRNet-W48 pretrained on COCO dataset as our backbone network. This paper uses Adam as the optimizer, and the initial learning rate is set to $1e^{-4}$, which decreases to $1e^{-5}$, $1e^{-6}$, and $1e^{-7}$ at epochs 8, 15, and 20, respectively. We use 4 NVIDIA RTX 2080Ti GPUs for training, with Batch Size set to 32, for a total of 20 epochs.

Evaluation Metric: Following the method used in the literature [14], [15], [22], [27], we adopt the average accuracy (AP) as an evaluation metric to test the model. In addition, the average accuracy is calculated for each joint separately and the average of the accuracy for all joints is taken as the final mean AP (mAP).

B. Comparison with State-of-the-art Approaches

Evaluation on PoseTrack2017 Dataset: The results of the estimated accuracy of each pose obtained by measuring the model on the PoseTrack2017 validation set are summarized in Table I. From the analysis of the results, it is evident shows that the detection accuracy of the proposed method achieves an excellent result in the relatively complicated to estimate poses such as elbow, wrist, hip, knee and ankle. Surprisingly, compared with PoseTracker [35], PoseFlow [36], JointFlow [37], FastPose [38], TML++ [39], simple baseline (ResNet50) [25], simple baseline (ResNet152) [25], STEembedding [40], HRNet [22], MDPN [41], Dynamic-GNN [26], PoseWarper [42], DCPose [14], IMAPose [43], and GLPose [44], the

TABLE I
QUANTITATIVE RESULTS ON THE POSETRACK2017 VALIDATION SET.

Method	Backbone	Additional Training Data	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
PoseTracker [35]	ResNet-3D	COCO	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow [36]	-	MPII Pose + COCO	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow [38]	-	-	-	-	-	-	-	-	-	69.3
FastPose [38]	-	-	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
TML++ [39]	-	-	-	-	-	-	-	-	-	71.5
Simple baseline [25]	ResNet-50	-	79.1	80.5	75.5	66.0	70.8	70.0	61.7	72.4
Simple baseline [25]	ResNet-152	-	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
STEmbedding [40]	4-stage Stacked Hourglass	-	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet [22]	HRNet-W48	COCO	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN [41]	SimpleBaseline	MPII Pose + COCO	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
Dynamic-GNN [26]	HRNet-W48	COCO	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
PoseWarper [42]	HRNet-W48	COCO	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose [14]	HRNet-W48	COCO	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
IMAPose [43]	HRNet-W48	COCO	88.2	88.7	84.3	78.4	82.9	81.7	73.6	82.9
GLPose [44]	HRNet-W48	COCO	88.1	88.9	84.1	78.1	83.5	81.5	74.2	83.1
SLT-Pose	HRNet-W48	COCO	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2



Fig. 2. Visualization of the results of our method on PoseTrack2017 dataset.

estimation accuracy and the average accuracy of this approach are significantly higher than those of the previously mentioned methods in the seven types of keypoint. The prediction accuracy of small-scale nodes such as wrist, knee and ankle are 1.8%, 3.3% and 2.2% higher than the GLpose accuracy, respectively. The superior estimation results are facilitated by the Cross-frame Temporal Learning Module in this paper, which greatly improves the extraction of useful feature information for those occluded parts and small-scale joints. We further present the visualization results of the proposed SLT-Pose on the PoseTrack 2017 dataset.

As shown in Figure 2, our method yields accurate estimation for both single and multi-person scenes. Even if the human images appear blurred due to rapid motion, our method still assists the model to perform correct estimation by capturing intimately related features from the auxiliary frames.

Evaluation on PoseTrack2018 Dataset: A total of 10 methods including STAF [45], AlphaPose [46], TML++ [39],

MDPN [41], PGPT [47], Dynamic-GNN [26], PoseWarper [42], DCPose [14], IMAPose [43], and GLPose [44] are evaluated in this paper. From the experimental results in the Table II, the human pose estimation accuracy reaches 81.5 mAP, which surpasses the current state-of-the-art methods. It is worth mentioning that for those challenging joints that are more difficult to estimate (e.g., wrist and ankle), the proposed method also obtains an encouraging improvement: 78.5 AP for the wrist is 0.5 AP higher than the second-ranked method, and 74.4 AP for the ankle is 0.6 AP higher than the previous state-of-the-art. Moreover, in terms of performance comparison, those methods that utilize video adjacent frame information for human pose estimation (e.g., PoseWarper and DCPose reported in the Table II) significantly outperform human pose estimation methods that employ only a single video frame (e.g., MDPN in the Table II). Evidently, it is a significant improvement over the methods using only single video frames (e.g., MDPN in the Table II).

TABLE II
QUANTITATIVE RESULTS ON THE POSETRACK2018 VALIDATION SET.

Method	Backbone	Additional Training Data	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
STAF [45]	COCO	-	-	-	-	64.7	-	-	62.0	70.4
AlphaPose [46]	8-stage Stacked Hourglass	COCO	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
TML++ [39]	OpenPose	COCO	-	-	-	-	-	-	-	74.6
MDPN [41]	SimpleBaseline	MPII Pose + COCO	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
PGPT [47]	-	-	-	-	-	60.2	-	-	56.9	76.8
Dynamic-GNN [26]	HRNet-W48	COCO	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
PoseWarper [42]	HRNet-W48	COCO	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
DCPose [14]	HRNet-W48	COCO	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
IMAPose [43]	HRNet-W48	COCO	84.2	86.8	82.7	77.9	80.4	79.4	73.2	80.9
GLPose [44]	HRNet-W48	COCO	84.0	86.9	82.4	77.6	80.4	79.3	73.8	80.8
SLT-Pose	HRNet-W48	COCO	84.3	87.5	83.5	78.5	80.9	80.2	74.4	81.5



Fig. 3. Visualization of the results of our method on PoseTrack2018 dataset.

We visualize the estimation results on the PoseTrack 2018 dataset. As illustrated in Figure 3, our model can demonstrate satisfactory performance for human features that are in a fast-moving or occluded state of the limb, which also proves the validity of SLT-Pose.

Evaluation on Sub-JHMDB Dataset: To confirm the effectiveness of the proposed method, we conduct validation using 11 different methods consisting of Part Models [48], Joint Action [49], Pose-Action [50], CPM [11], Thin-slicing [51], LSTM PM [17], DK(ResNet-50) [52], K-FPN(ResNet-18) [53], K-FPN(ResNet-50) [53], MotionAdaptive [54], and GLPose [44]. The results of the comparison in Sub-JHMDB dataset are tabulated in Table III. The proposed method arrives at an average localization accuracy of 96 mAP for all keypoints, which is 0.9 mAP higher than the GLPose method and significantly outperforms the current state-of-the-art methods. Meanwhile, for those challenging human keypoints, the proposed method achieves 98.6 AP, 91.4 AP and 95.3 AP for the shoulder, wrist and ankle, respectively. Despite the fact that the elbow did not reach the highest value, the proposed method also ranked second.

As described in Figure 4, SLT-Pose is visualized and an-

alyzed on the Sub-JHMDB dataset. In challenging scenarios such as variable number of human and loss of visual information, our model can express accurate estimation performance, which fully demonstrates the robustness of the proposed method.

C. Ablation Studies

To evaluate the impact of each module in the SLT-Pose, we conduct ablation experiments on the PoseTrack2017 validation set. 1) Verify the contribution of each module to the overall model by eliminating them individually. 2) Investigate the effect of inputting different local sequence lengths on the accuracy of the model. 3) Study the impact of different DCN layers on the overall performance. 4) Analyze the role of transformer layers on the performance of pose estimation. 5) Evaluate the influence of different attention layers on the model.

Verify the contribution of different components: To wholly prove the performance of each module, the performance of individual modules and combinations of multiple modules are evaluated. The ablation experiments for each module are presented in Table IV. The model performance

TABLE III
PERFORMANCE COMPARISONS ON THE SUB-JHMDB DATASET.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Part Models [48]	79.0	60.3	28.7	16.0	74.8	59.2	49.3	52.5
Joint Action [49]	83.3	63.5	33.8	21.6	76.3	62.7	53.1	55.7
Pose-Action [50]	90.3	76.9	59.3	55.0	85.9	76.4	73.0	73.8
CPM [11]	98.4	94.7	85.5	81.7	97.9	94.9	90.3	91.9
Thin-slicing Net [51]	97.1	95.7	87.5	81.6	98.0	92.7	89.8	92.1
LSTM PM [17]	98.2	96.5	89.6	86.0	98.7	95.6	90.0	93.6
DK(ResNet-50) [52]	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0
K-FPN(ResNet-18) [53]	94.7	96.3	95.2	90.2	96.4	95.5	93.2	94.5
K-FPN(ResNet-50) [53]	95.1	96.4	95.3	91.3	96.3	95.6	92.6	94.7
MotionAdaptive [54]	98.2	97.4	91.7	85.2	99.2	96.7	92.2	94.7
GLPose [44]	98.9	98.1	93.6	90.8	98.7	89.8	93.9	95.1
SLT-Pose	99.3	98.6	94.3	91.4	99.2	91.9	95.3	96.0

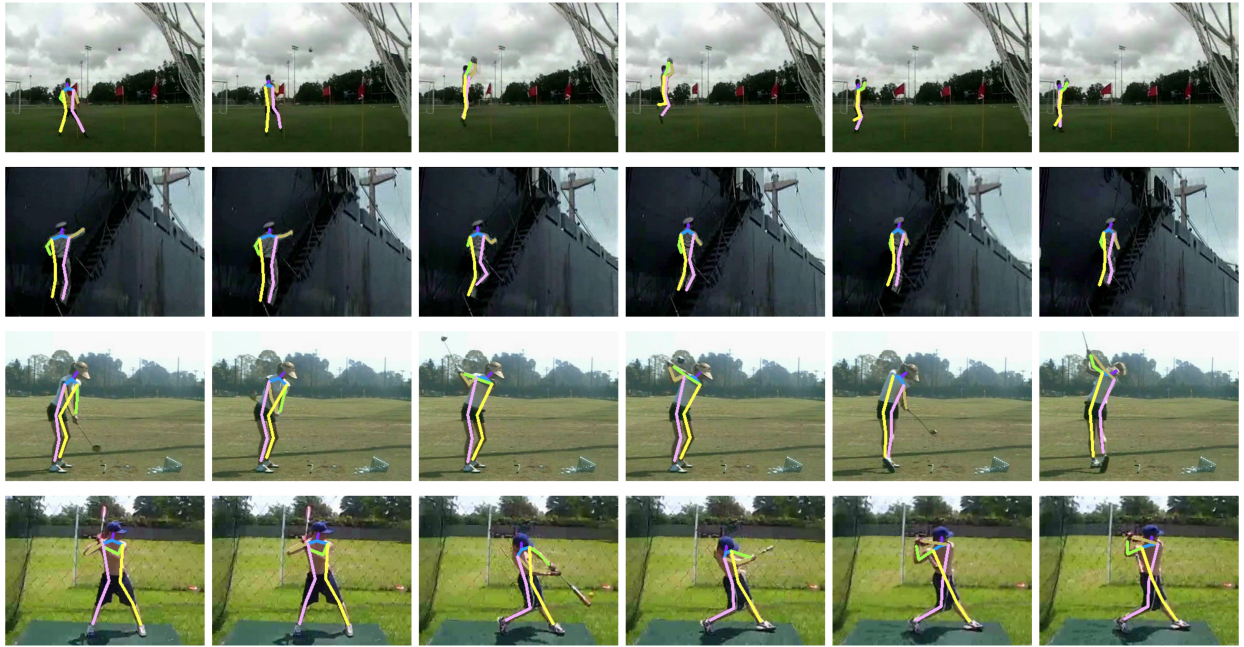


Fig. 4. Visualization of the results of our method on Sub-JHMDB dataset.

is severely negative and degraded by 1.6 mAP when we just employ the local sequence spatiotemporal features obtained by Personalized Feature Extraction Module. When only Self-feature Refinement Module and Disentangled Keypoint Detector are incorporated in the proposed method, the lack of coding of adjacent frame features leads to the neglect of some pivotal supplementary information, reducing the estimation accuracy of the model from 84.2 mAP to 83.6 mAP. When we remove Disentangled Keypoint Detector directly from the proposed method, which loses the ability to understand joint personality and body structure, the robustness of the model is negatively affected and the estimation accuracy is reduced by 0.5 mAP.

Investigate the effect of inputting different local sequence lengths: Empirically, the auxiliary frames contained in the local sequences can provide the model with abundant spatiotemporal correlation information as a supplement to the target frame features. Nonetheless, an ill-conceived setting of

the number of auxiliary frames might struggle to generate sufficient relevant information or introduce noisy features. In order to identify the most suitable length of the input sequence, we analyze the local sequences. Specifically, we utilize short sequences of total lengths 1, 3, 5, and 7 frames as input. As shown in Table V, it may be hard to sufficiently provide the model with the spatiotemporal correlation features when the number of frames is 1 and 3 compared to 5 frames.

Study the impact of the DCN method: To verify the DCN that is the core component of the proposed Personality feature Extraction Module, we replace the five parallel DCNs in SLTPose-DCN with CNN to construct a new model SLTPose-CNN to extract personalized features of the human body in each frame. It is obvious from Table VI that the model accuracy decreases dramatically when CNN is applied. The reason may be that it is difficult to capture plausible features using traditional convolutional kernels due to the constantly

TABLE IV
ABLATION STUDY FOR EACH COMPONENT FOR POSETRACK2017 DATASET.

PFEM	SRM	CTLM	DKD	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
✓				87.2	87.1	84.3	77.9	82.9	81.3	73.2	82.6
	✓			87.5	88.0	84.1	78.1	83.4	81.5	73.8	82.2
			✓	87.3	87.9	84.0	78.0	83.1	81.8	74.2	82.4
✓	✓			88.0	88.7	84.9	78.3	83.9	82.2	74.8	82.7
✓			✓	88.2	89.1	84.6	78.4	83.5	82.7	75.1	83.4
	✓		✓	87.9	88.9	84.8	78.8	83.4	82.6	75.4	83.6
✓	✓	✓		88.3	89.5	85.1	79.7	83.8	83.1	75.7	83.7
✓	✓		✓	88.7	89.4	85.3	79.0	84.0	82.9	75.6	83.9
✓	✓	✓	✓	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2

TABLE V
PERFORMANCE COMPARISON OF USING DIFFERENT COMBINATIONS OF LOCAL SEQUENCE LENGTHS ON POSETRACK2017 DATASET.

Input Frames	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
1	87.5	88.6	84.4	78.3	83.4	81.6	74.9	83.3
3	88.4	88.8	85.2	78.9	83.9	82.8	75.0	83.8
5	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
7	88.1	88.3	84.7	79.7	83.5	82.9	75.5	84.1

changing human shape.

TABLE VI
COMPARISON OF THE EFFECTS OF DIFFERENT FEATURE EXTRACTORS ON POSETRACK2017 DATASET.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
SLTPose-CNN	88.5	88.7	85.3	79.1	83.8	82.7	75.3	83.7
SLTPose-DCN	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2

Study the impact of different DCN layers: As indicated in Table VII, we utilize 1, 3 and 5 DCNs to capture personalized human features, respectively. When using one DCN, the model merely considers the commonality of the human body ignoring the inter-frame variability resulting in the model being unable to derive quite reasonable information. According to the temporal distance difference from the target frame, we set 3 DCNs, one of which is used to extract the target frame features, and the remaining DCNs are used for features with a temporal spacing of one frame and two frames, respectively. The estimation accuracy decreases by 0.6 mAP compared to using 5 DCNs, and it indicates that it is irrational to set the number by temporal distance.

TABLE VII
COMPARISON OF THE NUMBER OF DCNs FOR POSETRACK2017 DATASET.

DCN	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
1	88.1	88.4	84.6	78.4	82.7	82.5	74.5	83.3
3	88.8	88.9	85.0	79.0	83.3	83.0	74.8	83.6
5	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2

Analysis on transformer layer: When the number of stacked network layers is low, shallow features are extracted and it is difficult to capture detailed information. When the number of network layers is large, deep information is obtained, which can provide sufficient local features. However,

it is difficult to capture global information when the model is excessively deep. Therefore, it is essential to establish a proper network depth for the performance. As shown in Table VIII, we explore the number of layers of transformer layer included in Self-feature Refinement Module, which are 1, 2, 3, 6, 8, 10, and 12 layers. When the network is shallow, the human pose features captured by the model are inadequate, and its performance improves as the network deepens.

TABLE VIII
QUANTITATIVE COMPARISON OF DIFFERENT TRANSFORMER LAYERS FOR POSETRACK2017 DATASET.

Transformer Layer	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
1	87.9	88.6	84.2	78.9	83.4	82.4	74.4	83.8
2	87.7	88.2	84.8	78.8	83.2	82.8	75.1	83.7
4	87.4	89.3	84.9	78.5	83.0	81.9	76.1	83.8
6	88.5	88.5	85.4	79.2	84.1	82.3	74.6	84.1
8	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
10	88.2	89.0	85.8	78.6	83.4	82.5	75.7	83.6
12	87.8	89.5	85.2	79.0	84.2	82.7	74.5	83.4

Evaluate the influence of different attention mechanisms: We verify the effect of different attention mechanisms on the model. Cross Attention in the prototype model is replaced with Self-attention and named SLTPose-SA, and spatiotemporal features of the local sequence are used uniformly as input to Cross-frame Temporal Learning Module. As shown in Table IX, SLTPose-SA is lower than SLTPose-CA by 0.7 mAP because treating each frame feature in the local sequence equally may cause the most important target frame features to be under-considered which results in a decrease in accuracy. We utilize Cross Attention to highlight the significance of target frame information and to assist the model in obtaining spatio-temporal features from local motion sequences that are closely related to the target frame.

TABLE IX
QUANTITATIVE COMPARISON OF DIFFERENT ATTENTION MECHANISMS FOR POSETRACK2017 DATASET.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
SLTPose-SA	87.7	89.3	85.5	79.1	83.9	82.9	75.6	83.5
SLTPose-CA	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2

Evaluate the influence of different attention layers: In addition, setting a reasonable number of cross attention has an important impact on the accuracy of the model due to the fact

that when the number is too small, it is difficult to completely evaluate the degree of correlation between features, resulting in the target frame features being weakened. When the number is too large, it causes the model to focus excessively on closely related features and ignores other complementary information, which makes it difficult to deal with the absence of features in the target frame caused by blurring or occlusion.

In Table X, we evaluate the model performance when including 1, 2, 3 and 4 layers of cross attention. We can find that the estimation accuracy reaches up to 84.2 mAP when applying 2 layers.

TABLE X
QUANTITATIVE COMPARISON OF DIFFERENT CROSS ATTENTION LAYERS
FOR POSETRACK2017 DATASET.

Cross Attention	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
1	87.8	89.4	85.1	79.0	83.4	82.5	75.4	83.9
2	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
3	88.1	88.6	84.6	78.6	84.0	82.0	75.1	83.5
4	87.7	89.2	84.9	78.3	84.1	81.7	74.6	83.8

Explore the effect of Disentangled Keypoint Detector:

As illustrated in Table XI, when the skeletal connections between joint pairs are not considered (SLT-Pose r/m SC), which results in a decrease in the estimation accuracy of 0.2 mAP. We argue that modeling skeletal connected joint pairs can facilitate an accurate understanding of the human body and enhance performance. The reason is that the pose is a structure consisting of a series of joints following skeletal connections rather than discrete points, and there is vital complementary information between the related joints.

TABLE XI
EXPLORE THE EFFECT OF DISENTANGLED KEYPOINT DETECTOR ON
POSETRACK2017 DATASET.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
SLT-Pose r/m SC	88.2	89.4	85.1	79.2	84.0	82.7	75.3	84.0
SLT-Pose	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2

V. CONCLUSION

In this paper, an effective framework of spatiotemporal learning transformers is presented for human pose estimation. We propose a Personalized Feature Extraction Module to utilize the individual features of each frame to adapt to the complexity and variability of the human shape. Our Self-feature Refinement Module encodes global correlation cues on the human spatial information of the target frame to attain the refinement feature. To address the issue that advanced methods capturing spatiotemporal features from local sequences would equally fuse the information of each frame which leads to the weakening of the target frame information, we design Cross-frame Temporal Learning Module to emphasize the target frame features by excavating the closely related cues from the spatiotemporal features of local sequences. Moreover, the Disentangled Keypoint Detector extracts the characteristics of each joint and models the articulated joint pairs. Extensive experiments confirm that our method significantly surpasses existing work on four different benchmark datasets.

REFERENCES

- [1] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, 2018.
- [2] D. C. Luvizon, D. Picard, and H. Tabia, "Multi-task deep learning for real-time 3d human pose estimation and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2752–2764, 2021.
- [3] H. Shuai, L. Wu, and Q. Liu, "Adaptive multi-view and temporal fusing transformer for 3d human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 1–14, 2022.
- [4] Z. Liu, R. Feng, H. Chen, S. Wu, Y. Gao, Y. Gao, and X. Wang, "Temporal feature alignment and mutual information maximization for video-based human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10996–11006.
- [5] Z. Liu, P. Su, S. Wu, X. Shen, H. Chen, Y. Hao, and M. Wang, "Motion prediction using trajectory cues," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 13 279–13 288.
- [6] T. Tao, X. Yang, J. Xu, W. Wang, S. Zhang, M. Li, and G. Xu, "Trajectory planning of upper limb rehabilitation robot based on human pose estimation," in *17th International Conference on Ubiquitous Robots, UR 2020, Kyoto, Japan, June 22-26, 2020*. IEEE, 2020, pp. 333–338.
- [7] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6312. Springer, 2010, pp. 406–420.
- [8] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 596–603.
- [9] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1653–1660.
- [10] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1302–1310.
- [11] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4724–4732.
- [12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7103–7112.
- [13] G. Papandreou, T. Zhu, L. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11218. Springer, 2018, pp. 282–299.
- [14] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, and X. Wang, "Deep dual consecutive network for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 525–534.
- [15] M. Wang, J. Tighe, and D. Modolo, "Combining detection and tracking for human pose estimation in videos," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 11 085–11 093.
- [16] T. Jiang, N. C. Camgöz, and R. Bowden, "Skeletor: Skeletal transformers for robust body-pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 3394–3402.

- [17] Y. Luo, J. S. J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, and L. Lin, "LSTM pose machines," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5207–5215.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [19] M. Dantone, J. Gall, C. Leistner, and L. V. Gool, "Human pose estimation using body parts dependent joint regressors," in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 3041–3048.
- [20] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9912. Springer, 2016, pp. 483–499.
- [21] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," *CoRR*, vol. abs/1901.01760, 2019.
- [22] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 5693–5703.
- [23] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 5385–5394.
- [24] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui, "Peeking into occluded joints: A novel framework for crowd pose estimation," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12364. Springer, 2020, pp. 488–504.
- [25] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11210. Springer, 2018, pp. 472–487.
- [26] Y. Yang, Z. Ren, H. Li, C. Zhou, X. Wang, and G. Hua, "Learning dynamics via graph neural networks for human pose estimation and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 8074–8084.
- [27] Z. Liu, R. Feng, H. Chen, S. Wu, Y. Gao, Y. Gao, and X. Wang, "Temporal feature alignment and mutual information maximization for video-based human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10996–11006.
- [28] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11782–11792.
- [29] D. Shi, X. Wei, L. Li, Y. Ren, and W. Tan, "End-to-end multi-person pose estimation with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 11059–11068.
- [30] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, and W. Yang, "Exploiting temporal contexts with strided transformer for 3d human pose estimation," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.
- [31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.90>
- [32] U. Iqbal, A. Milan, and J. Gall, "Posetrack: Joint multi-person pose estimation and tracking," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 4654–4663.
- [33] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5167–5176.
- [34] Z. Fan, J. Liu, and Y. Wang, "Motion adaptive pose estimation from compressed videos," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11699–11708.
- [35] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 350–359.
- [36] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 53.
- [37] A. Doering, U. Iqbal, and J. Gall, "Joint flow: Temporal flow fields for multi person tracking," *CoRR*, vol. abs/1805.04596, 2018.
- [38] J. Zhang, Z. Zhu, W. Zou, P. Li, Y. Li, H. Su, and G. Huang, "Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks," *CoRR*, vol. abs/1908.05593, 2019.
- [39] J. Hwang, J. Lee, S. Park, and N. Kwak, "Pose estimator and tracker using temporal flow maps for limbs," in *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 2019, pp. 1–8.
- [40] S. Jin, W. Liu, W. Ouyang, and C. Qian, "Multi-person articulated tracking with spatial and temporal embeddings," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 5664–5673.
- [41] H. Guo, T. Tang, G. Luo, R. Chen, Y. Lu, and L. Wen, "Multi-domain pose network for multi-person pose estimation and tracking," in *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part II*, ser. Lecture Notes in Computer Science, L. Leal-Taixé and S. Roth, Eds., vol. 11130. Springer, 2018, pp. 209–216.
- [42] G. Bertasius, C. Feichtenhofer, D. Tran, J. Shi, and L. Torresani, "Learning temporal pose estimation from sparsely-labeled videos," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 3021–3032.
- [43] X. Wang, R. Feng, H. Chen, R. Zimmermann, Z. Liu, and H. Liu, "Personalized motion kernel learning for human pose estimation," *Int. J. Intell. Syst.*, vol. 37, no. 9, pp. 5859–5879, 2022.
- [44] Y. Jiao, H. Chen, R. Feng, H. Chen, S. Wu, Y. Yin, and Z. Liu, "Glpose: Global-local representation learning for human pose estimation," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.
- [45] Y. Raaj, H. Idrees, G. Hidalgo, and Y. Sheikh, "Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4620–4628.
- [46] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: regional multi-person pose estimation," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2353–2362.
- [47] Q. Bao, W. Liu, Y. Cheng, B. Zhou, and T. Mei, "Pose-guided tracking-by-detection: Robust multi-person pose tracking," *IEEE Trans. Multim.*, vol. 23, pp. 161–175, 2021.
- [48] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, D. N. Metaxas, L. Quan, A. Sanfeliu, and L. V. Gool, Eds. IEEE Computer Society, 2011, pp. 2627–2634.
- [49] B. X. Nie, C. Xiong, and S. Zhu, "Joint action recognition and pose estimation from video," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1293–1301.
- [50] U. Iqbal, M. Garbade, and J. Gall, "Pose for action - action for pose," in *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*. IEEE Computer Society, 2017, pp. 438–445.

- [51] J. Song, L. Wang, L. V. Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5563–5572.
- [52] X. Nie, Y. Li, L. Luo, N. Zhang, and J. Feng, "Dynamic kernel distillation for efficient pose estimation in videos," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 6941–6949.
- [53] Y. Zhang, Y. Wang, O. I. Camps, and M. Szaier, "Key frame proposal network for efficient pose estimation in videos," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12362. Springer, 2020, pp. 609–625.
- [54] Z. Fan, J. Liu, and Y. Wang, "Motion adaptive pose estimation from compressed videos," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11 699–11 708.



Di Gai Gai Di received the M.E. and Ph.D. degrees in College of Computer Science and Technology from Jilin University, China, in 2018 and 2021, respectively. He is currently a lecturer, School of Software, Nanchang University, China. He also is an assistant researcher in Jiangxi Key Laboratory of Smart City, China. His research interests include medical image processing and pattern recognition, especially on image fusion.



Runyang Feng is a Ph.D. student in School of Artificial Intelligence from Jilin University. His current research focuses on computer vision, 2D human pose estimation, and video representation learning.



Weidong Min received the B.E., M.E. and Ph.D. degrees in computer application from Tsinghua University, China in 1989, 1991 and 1995, respectively. He is currently a Professor and the Dean, School of Metaverse, Nanchang University, China. He is an Executive Director of China Society of Image and Graphics. His current research interests include image and video processing, artificial intelligence, big data, distributed system and smart city information technology.



Xiaosong Yang received the bachelor's and master's degrees in computer science from Zhejiang University, Hangzhou, China, in 1993 and 1996, respectively and the Ph.D. degree in computing mechanics from the Dalian University of Technology, Dalian, China, in 2000. He worked as a Post-Doc. from 2000 to 2002 with the Department of Computer Science and Technology, Tsinghua University for two years and as a Research Assistant from 2001 to 2002 with the Chinese University of Hong Kong. He is currently a Professor with the National Centre for Computer Animation, Bournemouth University, Poole, U.K. He has authored more than 70 papers in journals and refereed conferences. His research interests include deep learning, computer vision, computer animation, motion capture and synthesis, VR, AR, special effects and game development, digital health, data mining, and medical visualization.



Pengxiang Su received the Ph.D. degree in College of Computer Science and Technology from Jilin University, China in 2022. He is a lecturer at School of Software, Nanchang University, China. His research interests include computer vision, human motion analysis, and image recognition.



Qi Wang received the M.E. and Ph.D. degrees in School of Information Engineering from Nanchang University, China in 2018 and 2021, respectively. He is currently a lecturer, School of Software, Nanchang University, China. He also is an assistant researcher in Jiangxi Key Laboratory of Smart City, China. His current research focuses on computer vision and deep learning, particularly object re-identification.



Qing Han obtained the B.E. and M.E. degrees of computer application at Tianjin Polytechnic University in China in 1997 and 2006, respectively. She is now an associate professor at School of Mathematics and Computer Science, Nanchang University, China. Her research interests include image and video processing, network management.