Pose-Appearance Relational Modeling for Video Action Recognition

Mengmeng Cui, Wei Wang, Kunbo Zhang, *Member, IEEE*,
Zhenan Sun, *Senior Member, IEEE*, Liang Wang, *Fellow, IEEE*

***Abstract*—Recent studies of video action recognition can be classified into two categories: the appearance-based methods and the pose-based methods. The appearance-based methods generally cannot model temporal dynamics of large motion well by virtue of optical flow estimation, while the pose-based methods ignore the visual context information such as typical scenes and objects, which are also important cues for action understanding. In this paper, we tackle these problems by proposing a Pose-Appearance Relational Network (PARNet), which models the correlation between human pose and image appearance, and combines the benefits of these two modalities to improve the robustness towards unconstrained real-world videos. There are three network streams in our model, namely pose stream, appearance stream and relation stream. For the pose stream, a Temporal Multi-Pose RNN module is constructed to obtain the dynamic representations through temporal modeling of 2D poses. For the appearance stream, a Spatial Appearance CNN module is employed to extract the global appearance representation of the video sequence. For the relation stream, a Pose-Aware RNN module is built to connect pose and appearance streams by modelling action-sensitive visual context information. Through jointly optimizing the three modules, PARNet achieves superior performances compared with the state-of-the-arts on both the pose-complete datasets (KTH, Penn-Action, UCF11) and the challenging pose-incomplete datasets (UCF101, HMDB51, JHMDB), demonstrating its robustness towards complex environments and noisy skeletons. Its effectiveness on NTU-RGBD dataset is also validated even compared with 3D skeleton-based methods. Furthermore, an appearance-enhanced PARNet equipped with a RGB-based I3D stream is proposed, which outperforms the Kinetics pre-trained competitors on UCF101 and HMDB51. The better experimental results verify the potentials of our framework by integrating various modules.**

***Index Terms*—Action recognition, 2D pose-appearance, relational modelling, temporal attention LSTM.**

I. Introduction

As an important part of video understanding tasks, human action recognition has been widely used in the applications like human-machine interaction, automatic surveillance, video indexing and retrieval. It is challenging for the complexity of the videos captured in real-life conditions. Apart from the varying actions and environments, camera motion and visual occlusion also increase the difficulty of recognition.

A variety of works have been proposed in recent years. Earlier methods like the Space-Time Interest Points [1] and Dense Trajectories (DT, iDT) [2], [3] are based on handcrafted features. Descriptors extracted from image and optical flow are utilized to represent appearance and motion information. However, these methods are computationally expensive and limited in modelling complex actions. With the development of deep learning in visual classification, more methods have been proposed based on deep neural networks. Mainstream deep methods for video action recognition can be divided into appearance-based methods and pose-based methods.

Convolutional networks are commonly used in the appearance-based methods, e.g., the architectures consisting of 2D CNNs [4], [5], 3D CNNs [6], [5], [7], and the combination of CNNs and RNNs [8], [9], [10]. 2D CNN models mostly work in a two-stream manner with video frames and optical flow as spatial and temporal inputs. Although optical flow brings significant improvements to model performance, it can only represent the motion information between adjacent frames, and lacks the ability to capture temporal relationships of long-range actions [11], [12]. 3D CNNs can learn spatial and temporal features simultaneously by 3D convolution kernels. However, many 3D CNNs-based methods still need to cooperate with optical flow to better exploit motion information. There exists another method to employ RNNs to model the temporal dynamics of spatial features extracted by CNN module. However, the CNN features tend to represent global appearance information rather than local temporal dynamics within video frames, which limites the performance of these CNN-RNN architectures.

The pose-based methods, on the contrary, can directly capture motion information from human poses. The input pose modalities of mainstream methods can be classified into the pose skeletons [13], [64] and the mid-level estimated pose features [14], [15], [16], [17], [18], [19]. 3D pose-based methods achieve great success in the datasets under experimental conditions [20]. However, it is impractical to capture reliable 3D keypoints for videos in the wild, which limits the application in real-world scenarios. Instead of using skeleton keypoints, recent approaches choose to leverage mid-level pose features extracted from specific pose estimators to build recognition models. These methods act as pose encoders to perform spatial or temporal enhancement on different pose modalities (i.e., featuremaps, joint heatmaps), and cooperate with CNN models for feature learning and action classification. 2D poses, as another product of the pose estimator, are easier to obtain and more general than 3D poses and intermediate pose features.

Mengmeng Cui, Wei Wang, Kunbo Zhang, Zhenan Sun, Liang Wang are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. Kunbo Zhang, Zhenan Sun, Liang Wang are also with the University of Chinese Academy of Sciences, Beijing 101408, China. E-mail: mengmeng.cui@cripac.ia.ac.cn, wangwei@nlpr.ia.ac.cn, kunbo.zhang@ia.ac.cn, znsun@nlpr.ia.ac.cn, wangliang@nlpr.ia.ac.cn. *(Corresponding author: Wei Wang.)*

Manuscript received October 26, 2021.

Since recent 2D pose estimators [21], [22] achieve excellent performance in speed and accuracy, 2D pose-based methods can be a proper choice for action recognition. However, body joints occlusion and truncation existed in the unconstrained real-world videos lead to great degradation to the recognition performance. Therefore, how to model discriminative pose and motion cues to form an explicit understanding of on-going action, and ensure model robustness towards defective pose skeletons and complex environments, is a problem worthy of further investigation.

Motivated by the problems mentioned above, a **P**ose-**A**ppearance **R**elational **N**etwork (PARNet) is proposed for robust action recognition based on 2D poses and video frames. As shown in Figure 1, the overall architecture of PARNet consists of a Temporal Multi-Pose (TMP) RNN Module, a Spatial Appearance (SA) CNN Module, and a Pose-Aware (PA) RNN Module. These three modules are respectively built for the temporal modelling of 2D poses, the spatial modeling of video frames, and the relational modeling of these two modalities. Considering tremendous multi-person action scenarios in real life, such as the confrontational or cooperative sports (e.g., boxing and dancing), and the activities with irrelevant people in the background (e.g., high-jump crowed with audiences), attentional selection is applied on the multiple poses detected by OpenPose [21]. Instead of individually processing each person in the video [23], or directly aggregating them together [15], our approach is able to simultaneously attend to multiple informative targets while ignoring irrelevant characters. Through the relational modelling of pose and appearance features in the PA Module, the action-sensitive appearance information is captured at each iteration step, and the generated pose-aware representation can provide context supplement to the dynamic representations of TMP Module. Extensive experiments are conducted on the pose-complete datasets (KTH, Penn Action and UCF11), the challenging pose-incomplete datasets (UCF101, HMDB51 and JHMDB), and the NTU-RGBD dataset with depth information. The proposed PARNet achieves much better performances compared with the state-of-the-arts and demonstrates its robustness towards unconstrained real-world scenarios. When integrating PARNet with the RGB stream of I3D model to compare with the Kinetics pre-trained methods, the appearance-enhanced PARNet still outperforms several competitors on UCF101 and HMDB51, which further verifies the advantage of our method.

Our main contributions are summarized as follows:

- We introduce a robust action recognition architecture which integrates both 2D pose and visual appearance through a relational modeling strategy. The dynamic representations from the TMP Module, the global appearance representation from the SA Module, and the pose-aware representation from the PA Module are combined to generate a comprehensive representation for action recognition, ensuring robust performances for unconstrained indoor/outdoor videos.
- Attentional selections are performed iteratively to select target persons in the TMP Module and action-sensitive information in the PA Module, i.e., the persons performing action and the interactive scenes/objects, which improve the performance as well as the interpretability of the model.
- Different from previous studies which use LSTM for modeling temporal dynamics in videos, we deploy a memory enhanced Temporal Attention LSTM, which is able to capture stronger contextual dependency for long-term actions.

The source code will be released for future works[1].

II. Related Works

In this section, we concentrate on the deep learning approaches for action recognition and categorize them into 1) the appearance-based methods and 2) the pose-based methods.

A. Appearance-based Action Recognition

Visual appearance is widely used for action recognition in videos with its intuitive and informative properties. Different from image classification, video classification needs to capture both spatial and temporal information within the frame sequence. TSN [4] is based on 2D CNN framework which operates on the snippets sampled from several evenly-segmented clips. It works in a two-stream manner, feeding with video frames and optical flow to learn the appearance and motion features. RSTAN [12] conducts random sampling for multiple times and takes an average of the scores of the segmented groups to produce the final classification. Proposals like [9], [24] place RNN layers on top of CNN models to learn the temporal evolution of dynamic actions, where attention mechanism is mostly adopted to capture the salient area of featuremaps at each iteration step. Compared with 2D CNNs, 3D CNNs are considered as a more suitable solution for their spatio-temporal modelling ability. Thanks to the presence of large-scale action datasets such as Kinetics [25], 3D CNNs can be initialized with weights pre-trained on these datasets and get rid of the overfitting problem. Among 3D solutions, I3D [6] achieves an impressive performance with the optionally combined two-stream frameworks. With the purpose of improving model efficiency, ECO [5] is developed with 2D CNN to learn the appearance feature of each independent frame and 3D CNN to capture the temporal relationships between adjacent frames. To select the discriminative frames that are relevant to actions, SAST [26] adds a temporal attention model between 2D and 3D CNNs. SMART [63] builds relationship between frame-level local features and video-level global features through attention module. Although state-of-the-art performances are obtained on several challenging datasets, the appearance-based CNN approaches can not give an explicit clues of their classification choice. Philippe et al. [23] analyse the recognition bias of 3D CNNs, proving that the classification is often made by static context such as objects and scenes instead of actual action in the video. Some other approaches have also found such scene bias of CNNs and proposed various solutions. Jinwoo et al. [68] introduce novel losses in pre-training to prevent model from making predictions from

[1] https://github.com/Mona9955/PARNet

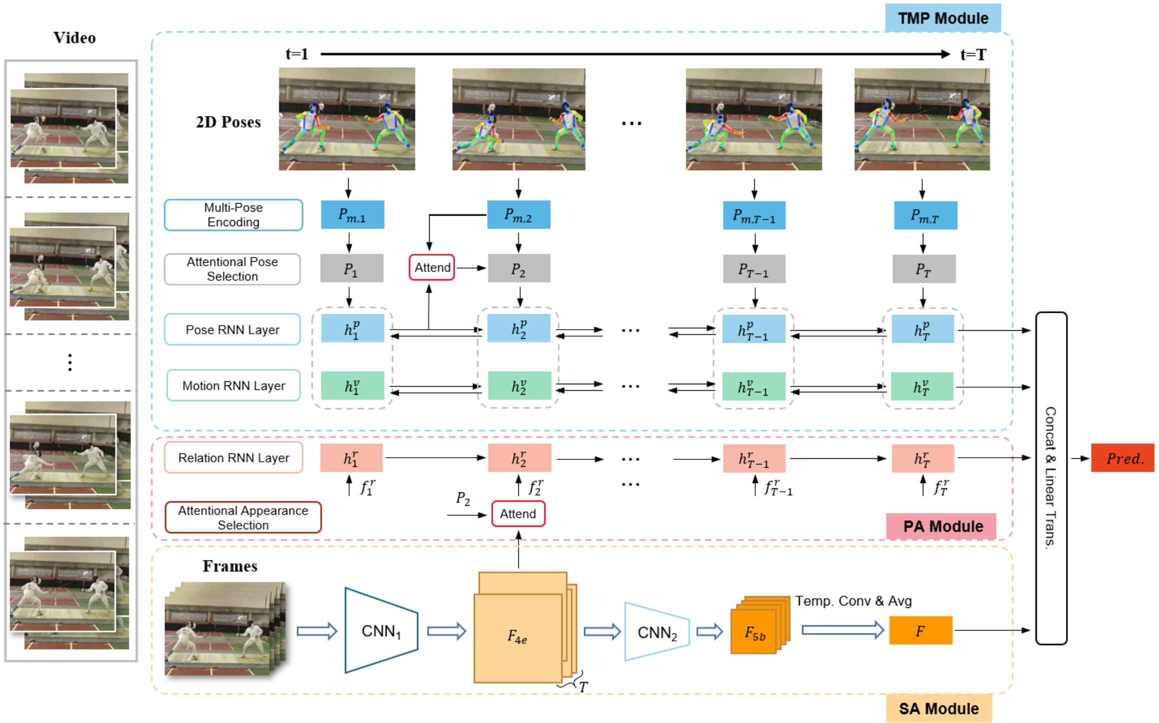


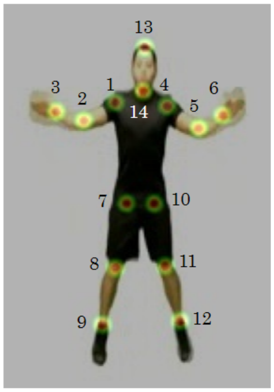
Fig. 1: Architecture of the proposed PARNet. The sampled T=16 video frames and the corresponding 2D pose skeletons are fed into the SA Module and TMP Module, respectively. In the **TMP Module**, the encoded pose-fusion vector P_t is exported to the temporal iteration of pose representation $\tilde{h}_T^p$ and motion representation $\tilde{h}_T^v$. In the **SA Module**, 2D CNN is utilized to generate the global appearance representation F. The mid-level featuremaps of SA are modulated by P_t through the Attentional Appearance Selection in the **PA Module**. This process is performed at each time step along with RNN iteration until the pose-aware representation $\tilde{h}_T^r$ is obtained. Finally, $[\tilde{h}_T^p; \tilde{h}_T^v; \tilde{h}_T^r; F]$ are concatenated to form the comprehensive representation of video actions.

the scenes. Yingwei et al. [69] propose a procedure named RESOUND to quantify and minimize representation biases of action datasets. In our approach, we introduce dynamic pose features in complementary with appearance features to obtain a discriminative representation of action.

B. Pose-based Action Recognition

Different from visual appearance, human poses concentrate on the on-going actions and are very suitable to represent temporal action dynamics. Many action recognition studies [27], [64] take 3D pose skeletons as inputs and achieve excellent accuracy in indoor datasets [20]. Since the 3D poses they used are mostly obtained by depth sensors such as Microsoft Kinect [20], [27] or Motion Capture (MoCap) systems [28], [29] under constrained environments, which greatly hinders the realistic application of 3D pose-based methods. Even with the implementation of 3D pose estimators such as LCR-Net [29], 3D poses are still validated to be vulnerable to the noise in unconstrained videos [23]. With the development of 2D pose estimation methods [21], [22], 2D pose can be obtained easier than 3D pose and presents higher stability and accuracy, making it a potential choice for action recognition. Luvizon et al. [13] aggregate pose estimation and recognition into a unified framework. They leverage two-stream CNNs to process the extracted pose skeletons and the visual featuremaps from the pose estimation part, and combine their classification results to make final prediction. ST-GCN [27] represents each joint with 2D coordinates and confidence scores derived from OpenPose [21], and models the pose sequence with graph convolutional network. However, due to the varying video capturing environments, challenging problems like joints occlusion, wrong detection, and body truncation inevitably appear in the pose estimation stage, which may cause great degradation to the action recognition performance. Instead of using specific coordinates, recent studies utilize mid-level features/heat maps of pose estimators for action recognition. RPAN [14] decomposes the joints heatmaps into several body parts and obtains a discriminative pose feature by attentional joint selection and part-pooling operation. Potion [16], PA3D [15] and STAR-Net [17] encode pose representations of different modalities through temporal aggregation, and re-input them into 3D CNNs for action recognition. These methods are mostly built upon specific 2D pose estimators, and collaborate with some other independent strong 3D CNNs (e.g., two-stream I3D [6]) to achieve performance gains with auxiliary appearance representations.

In this work, 2D poses derived from videos is used as the pose input. Thus, various pose estimators [21], [29], [22] can be adopted as pose-detection tools. Through the relational modeling strategy, pose-stream and appearance-stream complement each other effectively. Therefore, PARNet is not limited to visual contexts or dynamic poses, but has



(a)

Parts	Keypoints
Right arm	(j₁, j₂, j₃)
Left arm	(j₄, j₅, j₆)
Head	(j₁, j₄, j₁₃, j₁₄)
Right leg	(j₇, j₈, j₉)
Left leg	(j₁₀, j₁₁, j₁₂)

(b)

Fig. 2: (a) 14 human keypoints (joints) detected by 2D pose estimator. (b) Human parts and corresponding keypoints.

a comprehensive understanding of on-going action, which reduces the recognition bias.

III. Methods

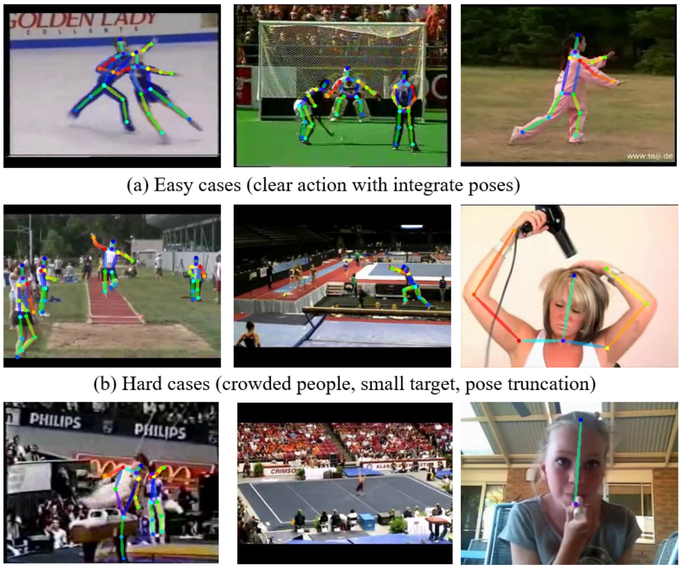
Firstly, we give a brief introduction of the Multi-Person 2D Pose Estimator adopted in our approach. Then we present the architecture of the Temporal Attention LSTM, which is the basic component of RNN layers in PARNet. After that, three main modules are introduced separately, i.e., the Temporal Multi-Pose RNN Module, the Spatial Appearance CNN Module, and the Pose-Aware RNN Module.

A. Muti-Person 2D Pose Estimation

The 2D pose estimator proposed in [21] is employed for its real-time property and robustness in multi-person scenarios. It works in a bottom-up way with Part Affinity Fields to associate the detected body parts. There are 14 keypoints for each person (shown in Figure 2(a)) instead of 18 keypoints utilized in the original method. Figure 3 shows estimated pose examples in diverse action videos. The first row lists the easy cases with clear actions and complete poses. The second row shows the hard cases, including crowed environments, small targets, and pose truncation. The third row presents several failure cases which include wrong detection and miss detection due to background clutters, small targets and truncated human body. The unstable estimated poses become a great challenge to the pose-based action recognition methods.

B. Temporal Attention LSTM (TA-LSTM)

Long-Short Term Memory (LSTM) [30] is a proper choice for action recognition due to its sequential modelling ability. In order to capture long-term contextual information in action videos, a temporal attention mechanism is adopted to enhance the memory of LSTM cell. As presented in [31], [32], the previous tokens within a time window are adaptively selected and merged by the attention mechanism to build a memory injected representation, which is in complementary with the current step input. Thus, the impact of negative inputs, e.g., incomplete poses or visual features cluttered with noise is effectively reduced. Therefore, the memory-enhanced Temporal Attention LSTM improves the robustness of the network while maintaining the advantage in long-term action recognition.



(a) Easy cases (clear action with integrate poses)

(b) Hard cases (crowded people, small target, pose truncation)

(c) Failure cases (wrong detection, miss detection, joints missing)

Fig. 3: Pose estimation examples of diverse action videos.

The TA-LSTM architecture is shown in Figure 4. At time step t, the attention vector h'_{t-1} from the last iteration step is concatenated with the current input x_t to form the input of the LSTM cell. Specifically, the hidden state h_t and the cell state c_t are calculated by:

$$x'_t = W_i[x_t; h'_{t-1}] \tag{1}$$

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W \cdot [x'_t; h_{t-1}] \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{3}$$

$$h_t = o_t \odot tanh(c_t) \tag{4}$$

Where W_i, W are parameters of the Fully-Connected layers (abbreviated as FC in Figure 4), σ and $\odot$ refer to the sigmoid activation and the element-wise product function. Given the concatenation of h_t and c_t as a query, the attention distribution $u^t = \{u^t_{t-n}, \cdots, u^t_{t-1}\}$ can be calculated by attention mechanism [33] upon the previous n steps outputs $\tilde{H}_{t-1} = (\tilde{h}_{t-n}, \cdots, \tilde{h}_{t-1})$. Softmax is used to normalize u^t to get the attention map a^t. Through the attentional selection and aggregation over $\tilde{H}_{t-1}$, the current attention vector h'_t is obtained. The whole process can be given as:

$$u^t_i = v^T tanh(W'_1 \tilde{h}_i + W'_2[h_t; c_t]) \tag{5}$$

$$a^t_i = softmax(u^t_i) \tag{6}$$

$$h'_t = \sum_{i=t-n}^{t-1} a^t_i \tilde{h}_i \tag{7}$$

The output at time step t is obtained through a linear transformation of the concatenation of h_t and h'_t:

$$\tilde{h}_t = W_o[h_t; h'_t] \tag{8}$$

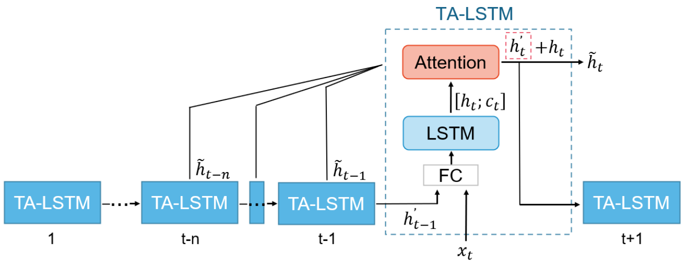


Fig. 4: Architecture of TA-LSTM layer. We use TA-LSTM cell at time step t as an illustration. The attention vector from the last time step $h_{t-1}^{'}$ and the current step input x_t are concatenated to form the input of LSTM cell. Then the calculated hidden state h_t and cell state c_t are used to perform attentional selection on the previous n steps outputs and generate current attention vector $h_t^{'}$, which is combined with h_t to form the output $\tilde{h}_t$ of TA-LSTM cell.

and the adjacent outputs within the sliding time window is updated as: $\tilde{H}_t = (\tilde{h}_{t-n+1}, \cdots, \tilde{h}_t)$. The vector v and matrices $W_1^{'}, W_2^{'}, W_o$ are trainable parameters. For convenience, the calculation process from equation 5-7 can be denoted as:

$$h_t^{'} = Attention([h_t; c_t], \tilde{H}_{t-1}) \tag{9}$$

It should be noted that the attention vector $h_t^{'}$ plays multiple roles in the Temporal Attention LSTM. Besides composing the current step output, $h_t^{'}$ is exported to the input of the next recurrent step, which strengthens information transmission over the whole action period.

Based on the analysis above, Temporal Attention LSTM cell can be summarized as:

$$\tilde{h}_t, S_t = TA\text{-}LSTM(x_t, S_{t-1}) \tag{10}$$

$$S_t = ((h_t, c_t), h_t^{'}, \tilde{H}_t)$$

S_t denotes the state set at time step t. It is worth mentioning that the temporal memory capacity of $\tilde{H}_t$ is controlled by the sliding window size n, which is an important hyperparameter affecting model's performance and computational complexity.

C. Temporal Multi-Pose (TMP) RNN Module

For the pose stream, a Multi-Pose Encoding Layer is used to encode the 2D pose skeletons. Then the multiple encoded poses are adaptively selected during the temporal evolution of the Multi-Pose Attention RNN Layer. Both pose representation and motion representation are calculated to capture the dynamic information.

1) Multi-Pose Encoding Layer: Given 2D coordinates of keypoints as input, the Multi-Pose Encoding Layer is deployed to generate high-level pose features based on the physical body structure. The maximum person number is set to N in each frame. Thus data clipping and zero padding are utilized to adapt the multiple human poses to a fixed size of $N \times K \times 2$, where K refers to the keypoint number, which is 14 in our approach, and 2 corresponds to the dimension of (x,y) coordinates. For each person, the skeleton keypoints are grouped into five body parts according to the semantic relationships (Figure 2(b)). Then, as shown in Figure 5, the body parts $l = \{l_i\}_{i=1}^5$ are passed through the parameter-sharing Parts Encoding Layer with Multi-Layer Perceptron (MLP). Finally, the encoded parts feature $l_e = \{l_{e_i}\}_{i=1}^5$ are concatenated and linearly transformed by the Pose Encoding Layer to get a pose vector $\mathbf{p_m}$.

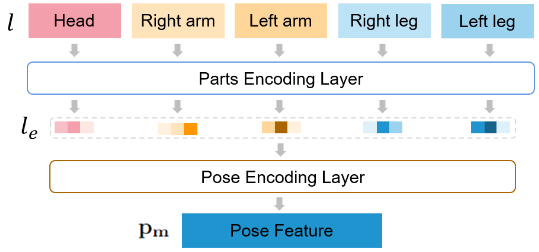


Fig. 5: Pose encoding architecture.

$$l_{e_i} = tanh(W_2^l(relu(W_1^l l_i)) + b^l) \tag{11}$$

$$\mathbf{p_m} = W_3^l Concat(l_{e_1}, l_{e_2}, ..., l_{e_5}) \tag{12}$$

Here the body parts are transformed into 32 and 100 dimensions by W_1^l and W_2^l, respectively. b^l refers to the bias parameter. Activation functions of $tanh$ and $relu$ are leveraged to imply non-linear transformation on the parts encoding process. W_3^l in the Pose Encoding Layer is used to transform pose feature into a 512-dim vector. Thus, the multi-pose set P_m can be represented as $P_m = (\mathbf{p_{m_1}}, \mathbf{p_{m_2}}, ..., \mathbf{p_{m_N}})$

2) Multi-Pose Attention RNN Layer: The Pose RNN layer is comprised of the TA-LSTM as basic cell. At each iteration step, the previous output $\tilde{h}_{t-1}^p$ is used to perform the attentional selection on the current multi-pose set $P_{m,t}$. Thus, the pose-fusion vector P_t is generated by:

$$P_t = Attention_p(\tilde{h}_{t-1}^p, P_{m,t}) \tag{13}$$

Here $Attention_p$ is the attentional pose selection function which works in the same way as Equation 9.

The Pose RNN Layer is fed with P_t at each iteration, which can be given by:

$$\tilde{h}_t^p, S_t^p = TA\text{-}LSTM(P_t, S_{t-1}^p) \tag{14}$$

where S_t^p denotes the pose state set.

In our model, the Pose RNN employs a bidirectional structure, in which the forward/backward outputs at the last time step T are concatenated to form $\tilde{h}_T^p = [\tilde{h}_T^{p_f}; \tilde{h}_T^{p_b}]$ as the pose representation.

3) Motion RNN Layer: To further model pose dynamics in human actions, a motion representation is implemented by performing temporal difference in the pose-fusion vector sequence, and another motion RNN layer is built on the temporal differences:

$$P_t^{'} = P_t - P_{t-1}, \quad t = 2, ..., T \tag{15}$$

$$\tilde{h}_t^v, S_t^v = TA\text{-}LSTM(P_t^{'}, S_{t-1}^v) \tag{16}$$

Similar to the pose RNN layer, the motion RNN layer also employs a bidirectional structure and $\tilde{h}_T^v = [\tilde{h}_T^{v_f}; \tilde{h}_T^{v_b}]$ is taken as the motion representation.

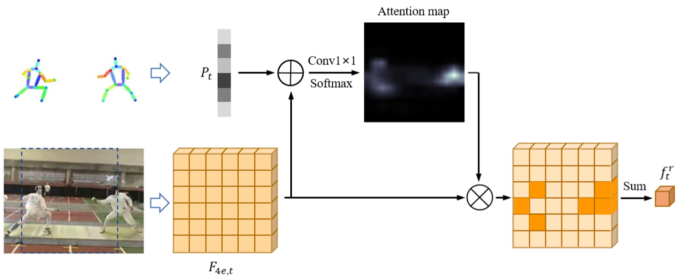


Fig. 6: The attention process to generate the pose-aware local appearance feature f_t^r. The pose-fusion vector P_t is utilized to perform attentional selection on the relevant mid-level featuremap $F_{4e,t}$. This process is performed iteratively by feeding f_t^r into the Relation RNN Layer at each time step.

D. Spatial Appearance (SA) CNN Module

For the appearance stream, 2D convolutional network is utilized to extract spatial features from the frame sequence. The *BN-Inception* architecture is employed [34] considering both efficiency and accuracy. The outputs of the 2D CNN model include two-stage featuremaps with different resolutions. Since earlier convolutional layer keeps richer spatial information, the mid-level feature sequence $F_{4e} \in \mathbb{R}^{T\times 14\times 14\times 256}$ from *inception-4e* layer is exported to the Pose-Aware Module. Meanwhile, the high-level feature sequence $F_{5b} \in \mathbb{R}^{T\times 7\times 7\times 1024}$ from final convolutional layer is used to generate the global appearance feature F through:

$$F = Avg\_pool(Conv(v_s^T F_{5b}, W_t)) \tag{17}$$

Here, F_{5b} is firstly transformed by v_s to reduce the channel into 512. $W_t \in \mathbb{R}^{3\times 1\times C_{in}\times C_{out}}$ is the parameter of the temporal convolution layer. The kernel size along temporal dimension is set to 3, which improves correlation between adjacent frames. Then the temporal-correlated feature sequence is squeezed in both spatial and temporal dimensions by average-pooling operation.

E. Pose-Aware (PA) RNN Module

Figure 6 shows the process of pose-aware attentional appearance selection at time step t. The pose-fusion vector P_t is used as the guidance to modulate the relevant mid-level featuremap $F_{4e,t}$. Conv1×1 is applied on the modulated features to generate a single-channel attention map, which is then normalized by the softmax operation. Thus, the local appearance feature f_t^r can be obtained by attentional weighted summation over the elements of $F_{4e,t}$. Finally, the Relation RNN Layer is built to perform temporal evolution of the local feature sequence $F_r = (f_1^r, \cdots, f_T^r)$. The whole process can be summarized as:

$$f_t^r = Attention_r(P_t, F_{4e,t}) \tag{18}$$

$$\tilde{h}_t^r, S_t^r = TA\text{-}LSTM(f_t^r, S_{t-1}^r) \tag{19}$$

The last iteration output $\tilde{h}_T^r$ is taken as the pose-aware representation. Different from the highly-compressed global appearance feature F, $\tilde{h}_T^r$ is a high-level representation generated from the dynamic evolution of action-sensitive appearance features, which provides context supplement to the outputs of TMP Module.

F. Loss function

The pose representation $\tilde{h}_T^p$, the motion representation $\tilde{h}_T^v$, the global appearance representation F, and the pose-aware representation $\tilde{h}_T^r$ are concatenated to form the comprehensive representation H,

$$H = [\tilde{h}_T^p; \tilde{h}_T^v; \tilde{h}_T^r; F] \tag{20}$$

H and the other four representations $(\tilde{h}_T^p, \tilde{h}_T^v, \tilde{h}_T^r, F)$ are transformed by five independent fully-connected layers to get classification scores and the corresponding losses. These losses are added with coefficient 1 to get the objective function, which can be given by:

$$\mathcal{L}_{total} = \mathcal{L}_H + \mathcal{L}_p + \mathcal{L}_v + \mathcal{L}_r + \mathcal{L}_F + \lambda \left\|\Theta\right\|_2 \tag{21}$$

Where $\left\|\Theta\right\|_2$ is the L^2-norm weight decay normalization applied on model parameters with the coefficient λ. $\mathcal{L}$ is the softmax cross-entropy loss which is widely used for action classification.

IV. Implementations

In the training stage, the mini-batch stochastic gradient descent (SGD) algorithm with a momentum of 0.9 is employed as the optimizer. The coefficient λ of the L^2-norm weight decay loss is set as 4e-5. The initialized learning rate is linearly increased to 1e-3 in the first 200 steps through the warm-up strategy. All the RNN layers are with the hidden size of 512. The BN-Inception architecture in the SA Module is initialized with the weights pre-trained on the ImageNet dataset [35]. Sparse sampling strategy [4] is adopted in the selection of frame/skeleton sequences. We divide the original video into T=16 segments and randomly select one frame and its relative skeleton coordinates from each segment, forming the frame groups and the skeleton groups. Random cropping and scaling are used as the augmentation methods for the frame groups, and the output size of each frame is fixed as 224 × 224. For the skeleton groups, the maximum number of persons is set as N=4, thus the skeleton groups are with the size of $T\times N\times K\times 2 = 16\times 4\times 14\times 2$. Position transfer and scaling are applied to the skeleton groups along with the change of the corresponding frame groups. Random horizontal flipping is also performed with the probability of 0.3 on frame/skeleton groups.

In the inference stage, frames and skeletons in the middle of each segment are chosen to form the testing frame/skeleton groups. The smaller dimension of frames is re-scaled to 256 and the other dimension is resized with the same ratio. Center cropping is applied on the resized frames to keep the output size of 224×224. Similar to the training stage, skeleton groups are also transformed along with the frame groups.

V. Experiments

A. Datasets

The following 7 action recognition benchmarks are leveraged to testify the proposed model, i.e., the pose-complete

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3228156

datasets of KTH, Penn-Action, UCF11, in which the extracted 2D poses are mostly intact, with small ratios of videos with joints-occlusion and pose-truncation; the pose-incomplete datasets of UCF101, HMDB51 and JHMDB which have a fraction of videos that only contain part of the body, such as "chew", "brush hair", and "cutting in kitchen" where only face (head) or hands are visible, making it even harder for pose detection and recognition; and the NTU-RGBD dataset with depth information, on which PARNet still works with 2D skeletons.

KTH [36] includes 6 kinds of actions acted by 25 subjects in four different scenarios. There are total 2391 single-person videos in KTH, most of which have static backgrounds.

Penn-Action [37] consists of 2326 videos with 15 action classes. All the videos are collected from the internet, making it a challenging dataset with various of perspectives and backgrounds.

UCF11 [38] is known as the YouTube action dataset composed of 1600 videos among 11 actions. Inference factors such as camera motion, viewpoint change, clutter background and illumination are involved in it.

UCF101 [39] is also a challenging real-world dataset with videos collected from the YouTube. It has 13320 videos from 101 categories, with large action diversities including pose/object variance and all the difficulties listed in the UCF11. UCF11 and UCF101 both have three splits.

HMDB51 [40] is a large complex dataset comprised of web videos and movie clips. It has 6766 videos with 51 action categories. It has three splits with 70 videos for training and 30 videos for testing in each category.

JHMDB [41] is a subset of HMDB51 which consists of 928 videos in 21 actions. It also has three splits with about 660 training videos and 268 testing videos.

NTU-RGBD [20] is a large action recognition dataset with 56880 clips in 60 action classes. It is captured under indoor experimental conditions with Kinect depth sensor to provide 3D skeleton annotations.

B. Ablation Study

Series of property evaluations are conducted in this part to analyse the effectiveness of the important components of PARNet. Both qualitative and quantitative analyses are used to explore how and to what extend these components affect the model's performance. The pose-complete datasets of KTH, Penn-Action and UCF11(split1) are used as the benchmarks in this section. When a specific property is analysed, the other settings of the model are kept unchanged. Models in all the experimental settings are trained from scratch.

Analysis of the TMP, SA and PA Modules We conduct experiments separately with the TMP Module and the SA Module to assess the properties of each part. As shown in row 1 & 2 in Table I, the TMP Module achieves better performance than the SA Module in KTH and Penn-Action, but the accuracy in UCF11-1 is much lower. The reason is that UCF11 is a complex YouTube dataset accompanied with more defective pose data. Therefore, the importance of the pose-stream and the appearance-stream varies in datasets with different characteristics. It is essential to fuse them in a proper way to improve the stability of performance.

TABLE I: Performanc of the TMP, SA, TMP+SA Modules and the whole PARNet on KTH, Penn-Action, and UCF11-1.

Modules	KTH	Penn-Action	UCF11-1
TMP	93.5	95.8	80.6
SA	93.2	94.7	93.8
TMP+SA	94.1	97.2	95.1
PARNet (TMP+SA+PA)	97.2	99.2	97.9

TABLE II: Performances of PARNet with different multi-pose fusion methods on Penn-Action and UCF11-1

Pose-Fusion Methods	Penn-Action	UCF11-1
Sum	98.3	96.1
Attention	99.2	97.9

But is it enough to simply combine the TMP Module and the SA Module? To validate the effectiveness of the pose-appearance relational modeling strategy, a TMP+SA architecture is constructed without the PA Module. For the TMP+SA, the comprehensive representation H in equation 20 is modified as $H = [\tilde{h}_T^p; \tilde{h}_T^v; F]$, and the classification losses of $(H, \tilde{h}_T^p, \tilde{h}_T^v, F)$ are equally added to form the target loss. The results of TMP+SA are listed in row 3 of Table I, which shows a degradation of 2% to 3.1% compared with PARNet among the three pose-complete datasets. It demonstrates that the pose and appearance streams bring positive effects to each other through the relational modelling deployed by PA Module.

Meanwhile, the confusion matrix is utilized to give a more detailed analysis of the pose and appearance modules. As shown in Figure 7, the classification results of PARNet, TMP Module and SA Module on the 15 actions of Penn-Action dataset are presented. The misclassification mainly happens between the "tennis forehand" and "tennis serve" categories for the SA Module (Figure 7(c)). These two tennis-related actions have similar backgrounds, e.g., the tennis courts and the stadiums shown in Figure 8(b), which occupy large portion of the frames. Hence, the SA Module cannot tell the subtle movements and difference from the global appearance representation. Meanwhile, there are 7 "squat" videos misclassified as "clean and jerk" for the same reason. This problem is solved in the TMP Module (Figure 7(b)) because the temporal modelling of pose skeletons can better capture the dynamic movements. However, there are also drawbacks for the pose-based method. The first one is the confusion caused by the similar motion dynamics between some action categories. For example, both of the "tennis forehand" and "bowl" in Figure 8(a) have the dynamic process of arm swing from bottom to top, leading to the misclassificaiton of 4 videos. The other one is the performance degradation caused by the defective skeleton data. Therefore, the pose and appearance streams need to complement each other effectively, thereby realizing true understanding of the on-going action. It can be seen from Figure 7(a) that PARNet shows a superior performance among all categories, indicating that the robustness and accuracy of the model are both improved.

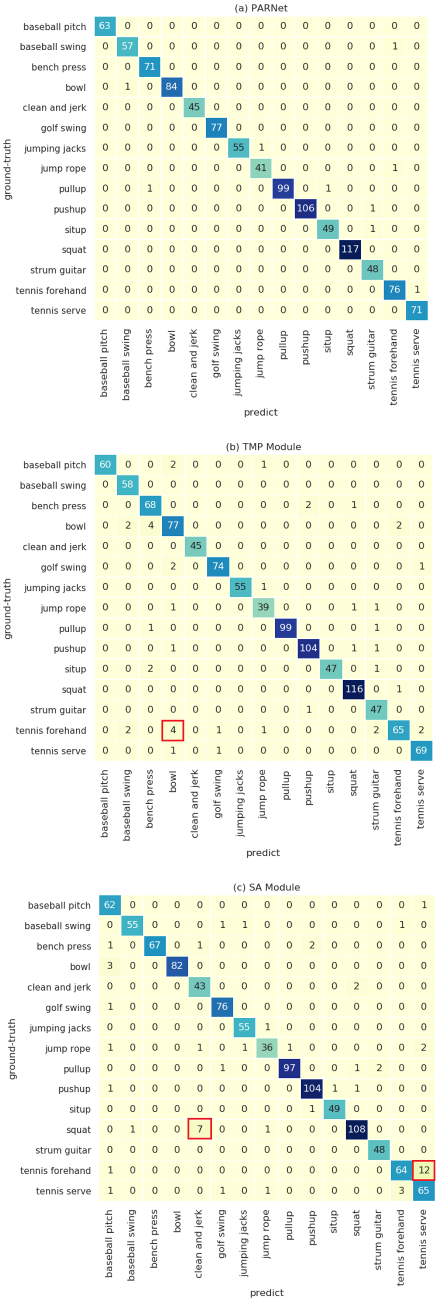
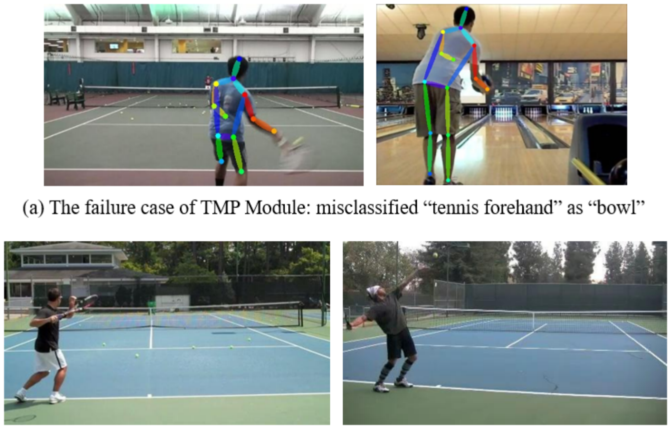


Fig. 7: Confusion matrices of (a) the full PARNet, (b) the TMP Module, (c) the SA Module on the Penn-Aciton dataset.

Attentional Selection in Pose and Appearance Attention mechanism is deployed in both the multi-pose selection of TMP Module and the pose-aware appearance selection of PA Module. In the Attentional Pose Selection process, the multiple poses are assigned with different attention weights according to their relevance with the pose representations generated at each iteration step. Since the temporal evolution of the pose-stream is directly corresponding to the dynamic change of actions, the model tends to pay more attention to the moving subjects. The first rows of two actions in Figure 9 illustrate the attention distribution of multiple poses in each frame. Take the action "baseball swing" in Figure 9(a) as an example. As the action changes from baseball-throwing to baseball-hitting, the attention transfers from the pitcher (left) to the batter (right). For the action "high jump" in Figure 9(b), there are irrelevant persons appearing in the first two frames. The attention mechanism correctly distinguishes the target person from the surrounding people, which improves the discrimination of the generated pose-fusion vector. To make a quantitative illustration of the effect of multi-pose attentional selection operation, we conduct comparative experiments which directly sum the multiple encoded poses to form the pose-fusion vector P_t. The comparison results are shown in Table II. Since the KTH dataset is captured under single-person scenarios, it is not introduced to the comparison experiments. The models equipped with Attentional Pose Selection show higher accuracy compared with the ones with direct-sum operations, especially in the UCF11 dataset which has more multi-person videos.


(a) The failure case of TMP Module: misclassified "tennis forehand" as "bowl"

(b) The failure case of SA Module: misclassified "tennis forehand" as "tennis serving"

Fig. 8: The misclassification cases of TMP Module and SA Module. (a) Since the actions of "tennis forehand" and "bowl" both have a bottom-up arm swing process, it brings confusion to the TMP Module concentrating on dynamic poses. (b) The videos of "tennis forehand" and "tennis serving" mostly have similar backgrounds occupying large portion of frames, leading to misclassification of the appearance-based SA Module that is biased towards scenes.

Different from features generated from specific pose skeletons, the poses-aware local appearance feature contains visual information of the target person, the surrounding scenes or the related objects. Such action-sensitive information provides context supplement to the pose features. The second rows of the two actions in Figure 9 show the dynamic changing of active regions in the Attentional Appearance Selection process. In the "baseball swing" example, the lighting areas are mainly around two players, especially the one who is performing the dynamic movements. For the "high jump" action with fast-moving subject and dynamically changing backgrounds, the model attends to the person and the surrounding scenes such as the athletic track and the rail, and filters out unrelated information like the background audiences. Figure 10 shows the visualized attention maps of the sampled frames, which contain objects that are important parts of the actions and interact with people. For the "walking with dog" example, the attention areas concentrate on two moving targets, i.e., the person and the dog. As for the "playing daf" action with a static background, the model mainly focuses on the musical instrument, and the attention weights become much larger once human hand slaps the daf. Taking into account the quantitative study about the effectiveness of PA Module in the previous section, we can draw a conclusion that the pose-aware attentional selection not only assists in the performance enhancement, but also improves the interpretability of the model.

Sliding time window in TA-LSTM The TA-LSTMs are used in three parts of the PARNet: Pose RNN layer and Motion RNN Layer in the TMP module, Relation RNN layer in the PA module. Since the size of time window plays a key role in the memory-enhanced TA-LSTM cell, we present the model results varying with different window sizes in Figure 11. When the window size equals to 0, the TA-LSTM functions as a normal LSTM cell. We can see that PARNet achieves the best performance when the window size is set to 5 in all three benchmarks, while the accuracy declines when the window size becomes larger. Overall, the models equipped with TA-LSTM (window size>0) perform better than the normal LSTM (window size=0) ones. The accuracy trend in Figure 11 proves that the aggregated information from the previous states benefits the current prediction, but the information needs to be constrained within a proper range. In our experiments, the sliding window size is set as 5.

C. Comparison with the State-of-the-arts

We compare PARNet with other state-of-the-art methods under the pose-complete datasets (KTH, Pann-Action, UCF11), the pose-incomplete datasets (UCF101, HMDB51, JHMDB), and the NTU-RGBD dataset. Since approaches with different solutions have been evaluated on these types of benchmarks, we can have a more comprehensive understanding of the pros and cons of PARNet. It is worth noting that the optical flow is not used in our experiments. There are two reasons for this. One is that the pose and motion representations have been deployed to capture the dynamic information, which is a more efficient and concise method compared with optical flow; the other is to avoid the high computational cost of extracting optical flow. Two-stream methods [4], [6], [53] usually adopt TV-L^1 algorithm [59] to extract optic flow. However, the TV-L^1 algorithm is unable to meet the real-time requirements in video-based action recognition tasks ($\geq$ 25fps). Recently, many deep learning based methods have been proposed and achieved a significant increase in the computational speed of optical flow (up to 10-150fps), such as FlowNet [62], FlowNet 2.0 [60]. However, the speeds of these methods are still slower than the 2D pose estimation methods (28-180fps), such as OpenPose [21], Pose Proposal Networks [61]. Moreover, human keypoint coordinates generated by pose estimators have much smaller size compared with optical flow, which largely reduces memory consumption and improves I/O efficiency for the downstream action recognition models.

TABLE III: Comparison results of PARNet with other state-of-the-art (SOTA) methods on pose-complete datasets of KTH, Penn-Action and UCF11 (over 3 splits)

Methods	KTH	Penn-Action	UCF11
DT(2011) [2]	94.2	-	84.2
Visual Attention(2015) [24]	-	-	85.0
Two Stream LSTM(2017) [10]	-	-	94.6
RPAN(2017) [14]	-	84.8	-
3-stream CNN(RGB+Flow+Trajectory)[58]	96.8	-	-
Multitask(2018) [13]	-	97.4	-
Attention Again(2018) [9]	-	-	90.1
ST-GCN(OpenPose)*(2018) [27]	-	71.6[1]	-
DA-Potion(2020) [19]	-	97.2	-
Two-Branch(2020) [54]	**98.3**	-	-
SIP-Net(2021) [23]	-	93.5	-
SIP-Net+3DResNeXt101(2021) [23]	-	98.9	-
Three Schemes(2021) [42]	97.0	-	95.6
PARNet (Ours)	97.2	**99.2**	**97.0**

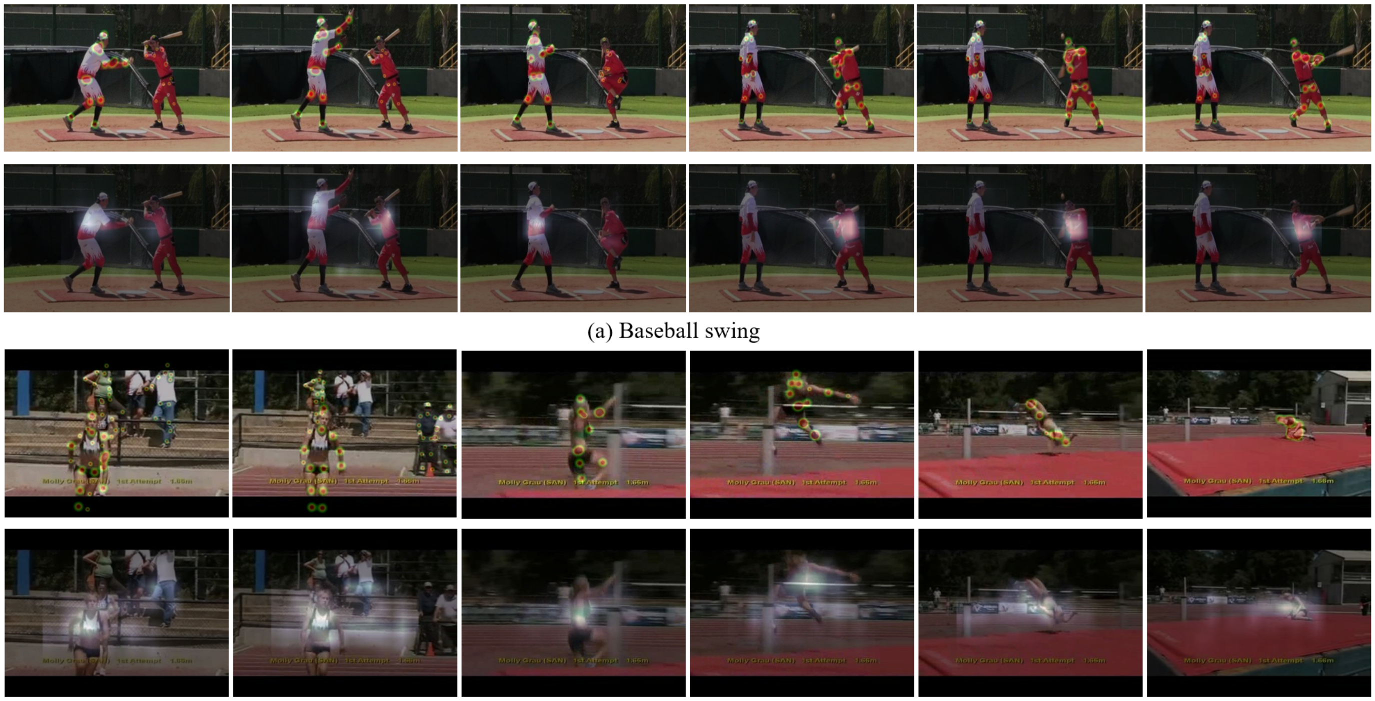
* is quoted from the reproduced experiment result in [23]

Results in pose-complete datasets We first conduct experiments on the pose-complete datasets. The comparison results are summarized in Table III. A variety of approaches have been performed on these three datasets, including the methods based on handcrafted features [2], [42] and encoded pose features [14], [19], [23], the visual feature attentional selection methods [24], [9]. Compared with Multitask [13], which combines the 2D/3D pose estimator and action recognizer into an unified framework and makes prediction based on both the generated pose skeletons and visual featuremaps modulated by joint heatmaps, our PARNet shows a higher accuracy of 1.8% on the Penn-Action dataset. PARNet also exceeds the SIP-Net+3DResNeXt101 [23] which leverages 3D ResNeXt101 to learn appearance features together with pose features. Compared with Three Schemes [42] using three different methods and complex handcrafted image features for action recognition, PARNet exceeds it by 1.4% on UCF11 and 0.2% on KTH. PARNet obtains a comparable performance to Two-Branch [54] on KTH dataset.

Results in pose-incomplete datasets In this part, experiments are conducted on the pose-incomplete datasets and all the results are averaged over 3 splits. Considering that the model pre-trained on the Kinetics dataset will show a significant performance improvement, we make extra comparison experiments using Kinetics pre-trained model to further demonstrate the potential and advantage of our framework.

We first compare PARNet with the methods employing the ImageNet pre-trained CNNs and the MSCOCO pre-trained pose estimators in their feature extraction parts in Table IV. Compared with the appearance-based methods in the first part of the table, such as TSN [4] and ECO [5], both of which use the same sparse sampling strategy and 2D CNN structure with our method, PARNet shows a significant improvement over the three datasets. The gap is even larger (up to 14.8% in HMDB51) compared with ImageNet pre-trained I3D [6]. Compared with STAN [43] and R-STAN [12] which conduct attentional selection on both spatial and temporal dimensions, PARNet outperforms them by up to 8.3% on UCF101 and 9.5% on HMDB51.

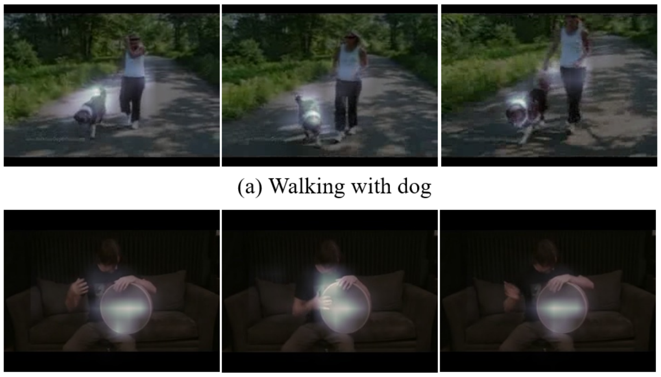
The second part in Table IV lists the pose-based methods,



(a) Baseball swing

(b) High jump

Fig. 9: Video examples of (a) baseball swing and (b) high jump. We select 6 representative frames from the sampled videos to illustrate our motivation. The upper frames of each action illustrate the dynamic multi-pose attentional selection process, in which the person with larger keypoint dots refers to the target subject with larger attention weights. The lower frames show the attention maps of the pose-aware appearance selection process, in which the brightness reveals the attention strength.



(a) Walking with dog

(b) Playing daf

Fig. 10: Visualized attention maps of pose-aware appearance selection in (a) walking with dog and (b) playing daf.

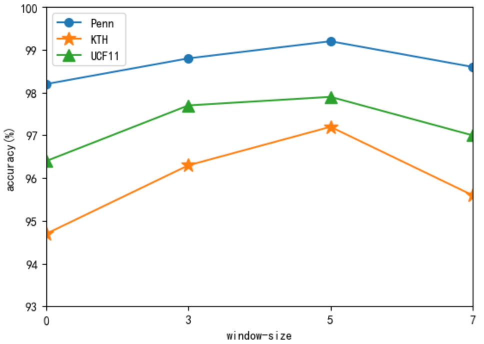


Fig. 11: The performance of PARNet with different sliding window sizes on KTH, Penn-Action, UCF11 (split1). PARNet achieves the highest accuracy when the sliding time window is set to 5 on all three datasets. Models equipped with TA-LSTM $(window\text{-}size > 0)$ outperform the ones with normal LSTM $(window\text{-}size = 0)$.

which are built upon the mid-level features of pose estimators pre-trained on the MSCOCO or other datasets. Except for STAR-Net, most of the other methods fix the parameters of pose estimators during training. Various specially-designed methods are developed for the extraction and processing of the pose features. For example, PA3D [15] leverages the joint heatmaps, part affinity fields and CNN features for action recognition. Dynamic Motion [18] applies a dynamic encoder on the joint heatmaps to capture body movements. For a better comparison, we also present the results of the pose-related part of PARNet, which consists of the TMP Module and the PA Module. The pose-related part (presented as TMP+PA in the table) outperforms the other pose-based methods by a large margin, especially on the two larger action datasets UCF101 and HMDB51. On JHMDB, it still gets a comparable performance with PA3D [15] and Joint-Aware [44] which use either complex input modalities or pose encoding strategy. Overall, from the comparisons with the competing appearance-based methods and pose-based methods on all three complex benchmarks, the superior performances of PARNet illustrate the advantage of our pose-appearance relational modeling strategy.

In order to further explore the impact of optical flow, the two-stream architectures which take RGB frames and optical

TABLE IV: Comparison results of PARNet with other SOTA methods on pose-incomplete datasets of UCF101, HMDB51 and JHMDB (over 3 splits). Methods listed in the first and second parts respectively use the ImageNet pre-trained backbones and the MSCOCO pre-trained pose estimators in their feature extraction parts, which can also be classified as the appearance-based methods and the pose-based methods.

Methods	pre-training	UCF101	HMDB51	JHMDB
TSN (RGB)(2016) [4]	ImageNet	86.4	53.7	-
I3D (RGB)(2017) [6]	ImageNet	84.5	49.8	-
Two Stream LSTM(2017) [10]	ImageNet	-	-	69.0
Attention Again(2018) [9]	ImageNet	85.8	52.6	-
ECO-16F*(2018) [5]	ImageNet	86.4	52.9	58.2
STAN (RGB)(2018) [43]	ImageNet	82.8	-	-
R-STAN-101 (RGB)(2019)[12]	ImageNet	86.2	55.1	-
TSN+TSM (RGB)(2019) [56]	ImageNet	-	55.2	-
SMART+ResNet-152(2020) [63]	ImageNet	75.5	-	-
Early fusion(2020) [52]	ImageNet	84.7	56.2	-
Potion(2018) [16]	MSCOCO	65.2	43.7	57.0
PA3D(2019) [15]	MSCOCO	-	55.3	69.5
STAR-Net(2019) [17]	MSCOCO	-	-	64.3
Dynamic Motion(2020) [18]	MSCOCO	63.5	49.1	60.2
Joint-Aware(2020) [44]	MSCOCO	-	52.1	68.3
SIP-Net(2021) [23]	Synth. data	66	51.2	62.4
TMP+PA (ours)	ImageNet	87.7	61.2	67.9
PARNet (ours)	ImageNet	**91.1**	**64.6**	**71.6**

* is our reproduced result of ECO with 16 frames input.

TABLE V: Comparison results of PARNet with other SOTA methods consisting of RGB and optical flow streams. All the listed models are equipped with ImageNet pre-trained backbones. Since most of the compared methods present the results on UCF101 and HMDB51 (over 3 splits), for the convenience of comparison, we keep it consistent.

Methods	pre-training	UCF101	HMDB51
TSN(RGB+Flow)(2016) [4]	ImageNet	94.0	68.5
I3D(RGB+Flow)(2017) [6]	ImageNet	93.4	66.4
3-stream CNN(RGB+Flow+Trajectory)[58]	ImageNet	92.2	65.2
STAN(RGB+Flow)(2018) [43]	ImageNet	92.8	-
TSN+TSM(RGB+Flow)(2019) [56]	ImageNet	94.3	**72.7**
TS-LSTM(RGB+Flow)(2019) [55]	ImageNet	94.1	69.0
R-STAN-101(RGB+Flow)(2019) [12]	ImageNet	**94.5**	68.7
PARNet(ours)	ImageNet	91.1	64.6

TABLE VI: Comparison results of the appearance-enhanced PARNet with other SOTA methods. All the listed methods are initialized with the models pre-trained on Kinetics.

Methods	pre-training	UCF101	HMDB51
T3D(2017) [7]	Kinetics	91.7	61.1
ResNeXt-101(2018) [45]	Kinetics	94.5	70.2
ARTNet(2018) [47]	Kinetics	93.5	67.6
ECO-EN(2018) [5]	Kinetics	94.8	72.4
TSM(2019) [53]	Kinetics	95.9.	73.5
SAST-EN(2019) [26]	Kinetics	96.4	75.1
3D-ResNet-18+debiased(2019) [68]	Mini-Kinetics 200	84.5	56.7
TSN (RGB)(2016) [4]	ImageNet+Kinetics	91.1	-
I3D (RGB)(2017) [6]	ImageNet+Kinetics	95.6	74.8
Disentangling(2018) [48]	ImageNet+Kinetics	95.9	-
StNet(2019) [50]	ImageNet+Kinetics	94.3	-
STM(2019) [11]	ImageNet+Kinetics	96.2	72.2
TMP+PA+I3D(RGB) (ours)	ImageNet+Kinetics	**97.2**	**76.7**

TABLE VII: Comparison results of the appearance-enhanced PARNet with other SOTA methods fed with RGB frames and optical flow as inputs. All the listed methods are initialized with the models pre-trained on Kinetics. It should be noted that our method still does not use optical flow.

Methods	pre-training	UCF101	HMDB51
TSN(RGB+Flow) (2016) [4]	ImageNet+Kinetics	97.0	-
I3D(RGB+Flow) (2017) [6]	Kinetics	97.9	80.2
HAF+BoW/FV(RGB+Flow) (2019) [57]	Kinetics	-	82.4
Early fusion+I3D(RGB+Flow) (2020) [52]	ImageNet+Kinetics	98.2	81.1
SMART+TSN(RGB+Flow) (2020) [63]	ImageNet+Kinetics	**98.6**	**84.3**
Potion+I3D(RGB+Flow) (2018) [16]	MSCOCO+Kinetics	98.2	80.9
PA3D+I3D(RGB+Flow) (2019) [15]	MSCOCO+Kinetics	-	82.1
Dynamic Motion+I3D(RGB+Flow) (2020) [18]	MSCOCO+Kinetics	98.4	84.2
Joint-Aware+I3D(RGB+Flow) (2020) [44]	MSCOCO+Kinetics	-	80.8
TMP+PA+I3D(RGB) (ours)	ImageNet+Kinetics	97.2	76.7

flow as inputs are listed in Table V. These methods show higher accuracy of 1.1%∼3.4% on UCF101 and 0.6%∼8.1% on HMDB51 than PARNet which does not use the optical flow modality. However, optical flow stream which has the same architecture as RGB stream, requires separate training, which greatly increases the computational cost and storage space.

PARNet has the model complexity of 64.9 GFLOPs. The computation cost is comparable with ECO-16F [5](64 GFLOPs) and TSM-16F [53](65 GFLOPs), and much lower than I3D-RGB-64F [6](222 GFLOPs) and R-STAN-101-240F [12](1819 GFLOPs). Figure 12 shows the runtime analysis on UCF101 dataset. Experiments are conducted on 1 NVIIDA Titan X GPU with 1 video per batch. Videos per second(vps) is utilized for the video-based action recognition task. Compared with two-stream I3D [6] and TSN [4], PARNet is 2.3%∼2.9% lower in accuracy but 2.9x∼16.4x faster in speed. PARNet also outperforms their RGB streams in both accuracy and speed. Compared with RGB-based ECO-16F [5] which is designed for online video understanding, PARNet is lower in speed but 4.7% higher in accuracy. These performance comparisons prove that optical-flow stream introduces heavy computational burden and reduces the inference efficiency. PARNet achieves the trade-off between speed and accuracy.

Furthermore, to explore the flexibility of our model framework, we fine-tune the Kinetics pre-trained I3D on UCF101 and HMDB51 datasets, and combine it with the pose-related part of PARNet by averaging the classification scores. In other words, the SA Module based on 2D CNN is replaced with a more powerful Kinetics pre-trained 3D CNN to extract the appearance representation. Comparison results are listed in Table VI. All the presented methods are under similar experimental conditions, i.e., pre-trained on Kinetics and without extra optical flow modality. The appearance-enhanced PARNet, which is listed as the TMP+PA+I3D fusion model achieves the best performance among all the comparison methods, demonstrating that the proposed pose-related part can complement well with 3D CNN, and make feasible cooperation with CNNs in various architectures.

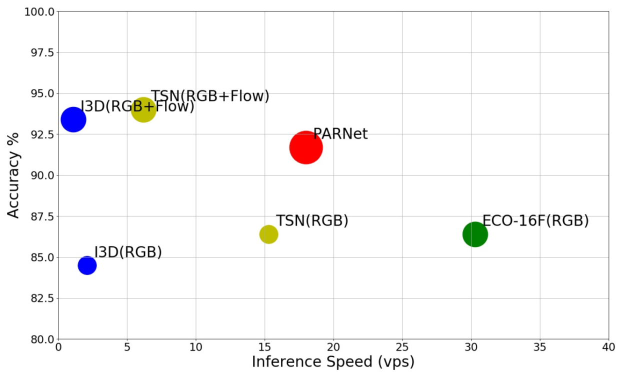


Fig. 12: Comparison of speed and performance with SOTA approaches on UCF101. The runtime is reported without considering I/O cost. The larger/smaller circles of I3D and TSN refer to their two-stream/only RGB stream variants, respectively.

Similarly, we present the results of the Kinetics pre-trained methods combined with optical flow streams in Table VII. The first and second blocks list the appearance-based methods and pose-based methods, respectively. The appearance-enhanced PARNet (without optical flow) achieves comparable performance with the SOTA approaches such as SMART [63], Early fusion [52] and Joint-Aware [44] on UCF101, but lower accuracy of 4.1%∼7.6% on HMDB51. It is worth noting that PARNet shows higher performance than the original structure of these models of up to 15.6% on UCF101 and 12.5% on HMDB51 in table IV, which proves that optical flow has a great influence on improving model accuracy. However, its disadvantages should not be ignored (as in our previous analysis).

Results in NTU-RGBD dataset Table VIII shows the comparison results with the architectures based on RNNs [67], [54], CNNs [65], [13] and GCNs [27], [64] on the cross-subject (cs) and cross-view (cv) sets of NTU-RGBD dataset. The proposed PARNet with 2D skeletons outperforms some 3D pose-based models [67], [65]. Compared with ST-GCN [27], PARNet shows comparable results with its 3D version and outperforms its 2D version (reproduced by [23]) by 8.3% on the cross-subject set. Since NTU-RGBD is collected under experimental environments with similar backgrounds, the contextual information provided by the interactions between human and objects/scenes is very limited. Depth information, however, is very informative for action classification in such scenarios. Thus, there is a gap compared with the highest scores on NTU. Considering the limitations of 3D pose acquisition analysed in Section II, the proposed 2D pose-based model still has its advantages in practical indoor/outdoor applications.

TABLE VIII: Comparison results of PARNet with other SOTA methods on NTU-RGBD dataset

Methods	NTU(cs)	NTU(cv)
ST-LSTM(2016) [67]	69.2	77.7
Joint-Distance(2017) [65]	76.2	82.3
Multitask(2018) [13]	85.5	-
ST-GCN(2018)[27]	81.5	88.3
ST-GCN(OpenPose)*	71.6	-
2s-AGCN(2019) [64]	88.5	95.1
MSG3D(2020) [66]	**91.5**	**96.2**
SIP-Net(2021) [23]	64.8	-
PARNet (Ours)	79.9	85.0

* is quoted from the reproduced experiment result in [23]

VI. Conclusion

In this paper, a Pose-Appearance Relational Network (PARNet) is proposed for robust action recognition. Our approach consists of three modules to benefit from both human pose and image appearance. The pose-stream is oriented to multi-person scenarios and can adaptively adjust the importance of multiple poses. Through the relational modeling strategy, the pose and appearance streams complement each other. Thus, PARNet has a comprehensive understanding of on-going action, which significantly reduces the recognition bias towards specific visual contexts or dynamic poses in videos. We also evaluate the performance of the pose-related part and the appearance-enhanced PARNet for better comparison with the state-of-the-arts under different experimental settings. Our approach outperforms the competitors on the RGB video-based benchmarks and shows robustness towards diverse challenges, e.g., the incomplete pose skeletons and pose/scene similarities of videos from different categories. Evaluations on NTU-RGBD validate the stable performance of PARNet on large datasets without rich contextual information.

In the future, we will continue to improve the architecture design of PARNet, e.g., using transformer network as the main backbones. Due to the fact that the self-attention mechanism of transformer ensures connections between all temporal tokens, transformer can be used to generate pose/appearance representations for the action videos with long-term front-to-back correlations. Moreover, many efficient transformers have been proposed to reduce the large memory consumption in vanilla transformer, e.g., Reformer [70] with parameter-sharing and reversible residual layers, which can preserve the efficiency of the proposed PARNet. In addition to action recognition in videos, the pose-appearance relational modeling strategy can also be applied in other video-based tasks, e.g., activity understanding [46], video captioning [49] and video parsing [51], where interactions between human and scenes/objects contribute to a higher-level understanding of video sequence.

Acknowledgment

The authors would like to thank the associate editor and reviewers for their helpful comments to improve this manuscript. This work was supported in part by the National Natural Science Foundation of China (Grant No. 61976214, 62071468, 62006225, 61806197, 6207146) and the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA27040700)

References

[1] I. Laptev, “On space-time interest points,” Int. J. Comput. Vis. (IJCV), vol. 64, no. 2, pp. 107–123, Sep. 2005.

[2] H. Wang, A. Klaser, C. Schmid, and C-L. Liu, ”Action recognition by dense trajectories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2011, pp. 3169–3176.

[3] H. Wang and C. Schmid. "Action recognition with improved trajectories." in Proc. IEEE Int. Con. Comput. Vis. (ICCV), Jan. 2013, pp. 3551–3558.

[4] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision", in Porc. Eur. Con. Comput. Vis. (ECCV), Sep. 2016, pp. 20–36.

[5] M. Zolfaghari, K. Singh, and T. Brox, "Eco:Efficient convolutional network for online video understanding." in Proc. Eur. Con. Comput. Vis. (ECCV), Sep. 2018, pp. 695–712.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.

[7] A. Diba et al., "Temporal 3d convnets: New architecture and transfer learning for video classification," 2017, arXiv:1711.08200. [Online]. Available: https://arxiv.org/abs/1711.08200

[8] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2625–2634.

[9] H. Yang, J. Zhang, S. Li, J. Lei, and S. Chen, "Attend it again: Recurrent attention convolutional neural network for action recognition," Applied Sciences, vol. 8, no. 3, pp. 383, Jan. 2018.

[10] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2017, pp. 177–186.

[11] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "Stm: Spatiotemporal and motion encoding for action recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Nov. 2019, pp. 2000–2009.

[12] Q. Liu, X. Che, and M. Bie, "R-STAN: Residual spatial-temporal attention network for action recognition," IEEE Access, vol. 7, pp. 82246–82255, Jun. 2019.

[13] D. C Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 5137–5146.

[14] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3725–3734.

[15] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "Pa3d: Pose-action 3d machine for video recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7922–7931.

[16] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7024–7033.

[17] W. McNally, A. Wong, and J. McPhee, "Star-net: Action recognition using spatio-temporal activation reprojection," in Proc. IEEE Conf. Comput. and Robot Vis. (CRV), May 2019, pp. 49–56.

[18] S. A. Esfeden, M. Sznaier, and O. Camps, "Dynamic motion representation for human action recognition," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 557–566.

[19] M. Segu, F. Pirovano, G. Fumagalli, and A. Fabris, "Depth-aware action recognition: Pose-motion encoding through temporal heatmaps," 2020, arXiv: 2011.13399. [Online]. Available: https://arxiv.org/abs/2011.13399

[20] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1010–1019.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multiperson 2d pose estimation using part affinity fields," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 7291–7299.

[22] Y. Chen, et al., "Cascaded pyramid network for multi-person pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7103–7112.

[23] P. Weinzaepfel and G. Rogez, "Mimetics: Towards understanding human actions out of context," Int. J. Comput. Vis. (IJCV), vol. 129, no. 5, pp. 1675–1690, 2021.

[24] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention". 2015, arXiv:1511.04119. [Online]. Available: https://arxiv.org/abs/1511.04119

[25] W. Kay, et al., "The kinetics human action video dataset". 2017, arXiv:1705.06950. [Online]. Available: https://arxiv.org/abs/1705.06950

[26] F. Wang, G. Wang, Y. Huang, and H. Chu, "Sast: learning semantic action-aware spatial-temporal features for efficient action recognition," IEEE Access, vol.7, pp. 164876–164886, Nov. 2019.

[27] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proc. 32nd AAAI Conf. Artif. Intell., Feb. 2018, pp. 7444–7452.

[28] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1110–1118.

[29] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net: Localization-classification-regression for human pose," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 3433-3441.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[31] "Generating long-term structure in songs and stories," https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn, 2014.

[32] J. Cheng, L. Doneg, M. Lapata, "Long short-term memory-networks for machine reading." in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Nov. 2016, pp. 551-561.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate." 2014, arXiv:1409.0473. [Online]. Available: https://arxiv.org/abs/1409.0473

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn. (ICML), Jul. 2015, pp. 448–456.

[35] J. Deng et al., "Imagenet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 248–255.

[36] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in Proc. Int. Conf. Pattern Recognit. (ICPR), Aug. 2004, pp. 32–36.

[37] W. Zhang, M. Zhu, and K. G Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2013, pp. 2248–2255.

[38] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 1996–2003.

[39] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, arXiv:1212.0402. [Online]. Available: https://arxiv.org/abs/1212.0402

[40] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Nov. 2011, pp. 2556–2563.

[41] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Apr. 2013, pp. 3192–3199.

[42] C. A Aly, F. S Abas, and G. H Ann, "Robust video content analysis schemes for human action recognition," Science Progress, vol. 104, no. 2, pp. 1-21, Apr. 2021.

[43] D. Li et al., "Unified spatio-temporal attention networks for action recognition in videos," IEEE Trans. on Multimedia, vol.21, no. 2, pp. 416–428, Aug. 2018.

[44] A. Shah et al., "Pose and Joint-Aware Action Recognition," 2020, arXiv:2010.08164. [Online]. Available: https://arxiv.org/abs/2010.08164

[45] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 6546–6555.

[46] Y. Tang et al., "Learning semantics-preserving attention and contextual interaction for group activity recognition," IEEE Trans. Image Process., vol. 28, no. 10, pp. 4997-5012, May 2019.

[47] L. Wang, W. Li, W. Li, and L. V. Gool, "Appearance and relation networks for video classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1430-1439.

[48] Y. Zhao, Y. Xiong, and D. Lin, "Recognize actions by disentangling components of dynamics," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 6566-6575.

[49] J. Zhang, Y. Peng, "Video captioning with object-aware spatio-temporal correlation and aggregation," IEEE Trans. Image Process., vol. 29, pp. 6209-6222, Apr. 2020.

[50] D. He et al., "Stnet: Local and global spatial temporal modeling for action recognition," in Proc. AAAI Conf. Artif. Intell., Feb. 2019, pp. 8401–8408.

[51] L. Liu, Y. Zhou, L. Shao, "Deep action parsing in videos with large-scale synthesized data," IEEE Trans. Image Process., vol. 27, no. 26, pp. 2869-2882, Mar. 2018.

[52] Z. Zheng et al., "Global and local knowledge-aware attention network for action recognition," IEEE Trans. on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 334-347, Mar. 2020.

[53] J. Lin et al., ”Tsm: Temporal shift module for efficient video understanding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Nov. 2019, pp. 7083-7093.

[54] A. Danilo et al., ”2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs,” IEEE Trans. on Multimedia, vol. 22, no. 10, pp. 2481-2496, Oct. 2020.

[55] C. Ma et al., ”TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition,” Signal Process.: Image Communication, vol. 71, pp. 76-87, Feb. 2019.

[56] X. Song et al. ”Temporal–spatial mapping for action recognition,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 3, pp. 748-759, Jan. 2019.

[57] L. Wang et al. ”Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Nov. 2019, pp. 8698-8708.

[58] Y. Shi et al. ”Sequential deep trajectory descriptor for action recognition with three-stream CNN,” IEEE Trans. on Multimedia, vol. 19, no. 7, pp. 1510-1520, Feb. 2017.

[59] Z. Christopher et al. ”A duality based approach for realtime tv-l 1 optical flow,” In Joint pattern recognition symposium, Springer, Berlin, Heidelberg. pp. 214-223. 2007.

[60] E. Ilg et al. ”Flownet 2.0: Evolution of optical flow estimation with deep networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 2462-2470.

[61] S. Taiki. ”Pose proposal networks,” in Proc. Eur. Con. Comput. Vis. (ECCV), Sep. 2018, pp. 342-357.

[62] D. Alexey et al. ”Flownet: Learning optical flow with convolutional networks”, in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Nov. 2015, pp. 2758-2766.

[63] G. S. N, R. Marcus and S.-L. Laura, 2020, arXiv:2012.10671. [Online]. Available: https://arxiv.org/abs/2012.10671.

[64] L. Shi et al. ”Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Nov. 2019, pp. 12026-1203.

[65] C. Li et al. ”Joint distance maps based action recognition with convolutional neural networks,” IEEE Signal Processing Letters, vol. 24, no. 5, pp. 624-628, Mar. 2017.

[66] Z. Liu et al. ”Disentangling and unifying graph convolutions for skeleton-based action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 143-152.

[67] J. Liu et al. ”Spatio-temporal lstm with trust gates for 3d human action recognition,” in Porc. Eur. Con. Comput. Vis. (ECCV), Sep. 2016, pp. 816-833.

[68] J. Choi et al. ”Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition,” Advances in Neural Inf. Process. Syst. (NeurIPS), vol. 32. 2019.

[69] Y. Li et al. ”Resound: Towards action recognition without representation bias.” in Porc. Eur. Con. Comput. Vis. (ECCV), Sep. 2018, pp. 513-528.

[70] N. Kitaev et al. ”Reformer: The Efficient Transformer.” in Proc. Int. Conf. Learn. Represent. (ICLR), Sep. 2020, URL https://openreview.net/pdf?id=rkgNKkHtvB.

Mengmeng Cui received both the B.E. degree and the M.S. degree from Beijing Institute of Technology, China in 2013 and 2016. She is currently an algorithm engineer with the the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. Her main research interests include computer vision and machine learning, especially in video classification, pose estimation, and scene text recognition.

Wei Wang received the B.E. degree from the Department of Automation, Wuhan University, in 2005, and the Ph.D. degree from the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences (GUCAS), in 2011. He is currently an Associate Professor with Center for Research on Intelligent Perception and Computing and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He has published a number of papers in the leading international conferences, such as CVPR and ICCV. His research interests include computer vision, pattern recognition, and machine learning, particularly on the computational modeling of visual attention, deep learning, and multimodal data analysis.

Kunbo Zhang (Member, IEEE) received the B.E. degree in automation from the Beijing Institute of Technology in 2006 and the M.Sc. and Ph.D. degrees in mechanical engineering from the State University of New York at Stony Brook, USA, in 2008 and 2011, respectively. From 2011 to 2016, he has worked as the Machine Vision Research and Development Engineer of the Advanced Manufacturing Engineering Group, Nexteer Automotive, MI, USA. He is currently an Assistant Professor with NLPR, CRIPAC, CASIA, China. His current research interests include computational photography, biometric imaging, machine vision, and intelligent systems.

Zhenan Sun (Senior Member, IEEE) received the B.E. degree in industrial automation from the Dalian University of Technology, China, in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, China, in 2006. He is currently a Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, China. He has authored/coauthored more than 200 technical articles. His research interests include biometrics, pattern recognition, and computer vision. He is the Chair of Technical Committee on Biometrics, International Association for Pattern Recognition (IAPR) and an IAPR Fellow. He serves as an Associate Editor for the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE (T-BIOM).

Liang Wang (Fellow, IEEE) received both the B. Eng. and M. Eng. degrees from Anhui University in 1997 and 2000 respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CAS) in 2004. From 2004 to 2010, he worked as a Research Assistant at Imperial College London, United Kingdom and Monash University, Australia, a Research Fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full Professor at the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P. R. China.

His major research interests include machine learning, pattern recognition and computer vision. He has widely published at highly-ranked international journals such as IEEE TPAMI and IEEE TIP, and leading international conferences such as CVPR, ICCV and AAAI. He has obtained several honors and awards such as the Special Prize of the Presidential Scholarship of Chinese Academy of Sciences. He is a Fellow of IEEE, IAPR and CIE, as well as a member of BMVA and ACM. He is currently an associate editor of IEEE TPAMI, IEEE TIP and Pattern Recognition.