Neural Computing and Applications (2021) 33:16439–16450
https://doi.org/10.1007/s00521-021-06239-5

ORIGINAL ARTICLE

Local-aware spatio-temporal attention network with multi-stage feature fusion for human action recognition

Yaqing Hou[1] · Hua Yu[1] · Dongsheng Zhou[2] · Pengfei Wang[1] · Hongwei Ge[1] · Jianxin Zhang[3] · Qiang Zhang[1]

Received: 13 January 2021 / Accepted: 13 June 2021 / Published online: 11 July 2021

Abstract
In the study of human action recognition, two-stream networks have made excellent progress recently. However, there remain challenges in distinguishing similar human actions in videos. This paper proposes a novel local-aware spatio-temporal attention network with multi-stage feature fusion based on compact bilinear pooling for human action recognition. To elaborate, taking two-stream networks as our essential backbones, the spatial network first employs multiple spatial transformer networks in a parallel manner to locate the discriminative regions related to human actions. Then, we perform feature fusion between the local and global features to enhance the human action representation. Furthermore, the output of the spatial network and the temporal information are fused at a particular layer to learn the pixel-wise correspondences. After that, we bring together three outputs to generate the global descriptors of human actions. To verify the efficacy of the proposed approach, comparison experiments are conducted with the traditional hand-engineered IDT algorithms, the classical machine learning methods (i.e., SVM) and the state-of-the-art deep learning methods (i.e., spatio-temporal multiplier networks). According to the results, our approach is reported to obtain the best performance among existing works, with the accuracy of 95.3% and 72.9% on UCF101 and HMDB51, respectively. The experimental results thus demonstrate the superiority and significance of the proposed architecture in solving the task of human action recognition.

Keywords Spatio-temporal attention networks · Spatial transformer network · Feature fusion · Human action recognition

1 Introduction

Research on human-robot interactions has attracted increased attention during the past few years. In the interaction system, the robot needs to recognize actions of the human from the available video data, which can be commonly categorized into 3D skeletons [25, 31] and RGB video datasets [1, 5, 21, 32]. Notably, the task on RGB video datasets is usually more difficult since the visual content is significantly more complicated compared to that in 3D skeleton datasets. In recent decades, conventional approaches, i.e., hand-engineered descriptor algorithms including HOG [20], HOF [24], SIFT [28], MBH [3], have been carried out for extracting features on RGB datasets. Nevertheless, these approaches require manual feature extraction and usually fail in recognition problems with multiple classes and large-scale training data.

Artificial neural networks, as the biologically inspired computing paradigm, provide an alternative methodology for automatically recognizing the underlying relationships from the available data. With the growth of computing power from GPUs and distributed computing, deep neural networks, particularly of the convolutional type (deep CNNs), have sparked a revolution in artificial intelligence, triggering many research studies and practical applications.

Yaqing Hou and Hua Yu are joint first authors.

✉ Yaqing Hou
houyq@dlut.edu.cn

✉ Qiang Zhang
zhangq@dlut.edu.cn

[1] School of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2] School of Software Engineering, Dalian University, Dalian, China

[3] School of Computer Science and Engineering, Dalian Minzu University, Dalian, China

More recently, CNNs have been widely applied in human–computer interaction for human action recognition, where they have significantly outperformed the conventional machine learning methods. Existing approaches [14–16, 33] have shown good performance in capturing the appearance information of objects from images. Nevertheless, a CNN often fails to model videos precisely based on the spatial information obtained from actions [19, 35].

Unlike static images, the temporal information of human actions in videos provides additional cues. Long short-term memory (LSTM) has been introduced into CNN structures due to its capacity to preserve long-term information over time. For example, the long-term recurrent convolutional network (LRCN) was proposed to model temporal sequences by connecting the outputs of CNN to the multi-layer LSTM model [5]. One of the methods of detecting the activities is modeling the activity using the Activation Spreading Network (ASN) [27]. This method was inspired in hierarchical task networks, which is a way to represent the long-term relationships of a process or activity. Shi et al. [30] replaced the fully connected layer with the convolutional layer in LSTM and proposed ConvLSTM networks to capture the temporal dynamics. Unfortunately, these approaches fail to consider the interactions between the spatial and temporal information in feature extraction. The 3D CNN (C3D) is another extension of the CNN in the temporal domain and employs 3D convolutional kernels to extract the temporal evolution information across video frames [18]. It has shown better performance in extracting spatial–temporal features than 2D CNNs. For example, Tran et al. [37] conducted empirical studies on C3D with multiple configurations of 3D convolutional kernels and obtained promising performance with a kernel size of $3\times3\times3$. Still, existing C3D approaches are often reported to suffer from poor scalability and high computational cost [13, 18, 37].

Compared to the approaches based on LSTM and C3D, two-stream CNNs can easily utilize the new CNN structures such as residual networks (ResNet) [14] and BN-Inception [16]. These two-stream CNNs can decompose a video into spatial and temporal streams for capturing the appearance and motion features, respectively [32]. Based on two-stream CNNs, Wang et al. [40] introduced temporal segment networks, which enforced consensus over different short snippets. Qiao et al. [39] proposed a trajectory-pooled deep-convolutional descriptor (TDD) with a sum-pooling method to leverage the hand-engineered and deep-learned information in two-stream CNNs, wherein the spatial and temporal networks of TDD were trained separately. Furthermore, Rohit et al. [11] devised an attention pooling approach and reported better performance than first-order pooling in traditional CNNs. Deva et al. [12] proposed ActionVLAD pooling to aggregate the information across the entire temporal span of videos. Different strategies have been investigated to integrate the signals from two streams, including concat fusion, early fusion, and late fusion.

However, there remain issues in two-stream CNNs. First, existing approaches have not yet tackled the precise relationship between local and global features of human actions since different regions of the human body have different degrees of saliency during human action. This could be exaggerated when similar actions are being recognized. For example, a "high jump" is a composite action that could be easily confused with a series of actions such as "running", "jumping", and "tumbling". Existing two-stream CNN methods usually fail since the impact of fine-grained differences between human movements is neglected. Therefore, in our study, we introduce an attention model, the spatial transformer network (STN) [17], to locate attentional local regions of the human body, and then aggregate these local regions with the global information captured from the spatial input image during the task of recognizing human action.

More effective strategies for fusing spatial attentional region features and temporal features need to be explored. Many researchers have investigated how to capture the interactions between the two streams. In our present study, we reconsider this problem and introduce a fusion method, namely compact bilinear pooling (CBP) [9], into two-stream CNNs to instruct the interaction between spatial and temporal information. The CBP algorithm was initially proposed to fuse spatial features in fine-grained recognition tasks. Here, we investigated the efficacy of fusing the spatial and temporal features through compact bilinear pooling. This is possible due to the fusion method's ability to markedly reduce the training parameters over competitive methods such as element-wise sum and concatenation (both of which are investigated in [8]).

In summary, the specific interest of our work lies in aggregating the beneficial information across entire videos by proposing a novel spatio-temporal attention network with multi-stage feature fusion. Our proposed architecture is shown in Fig. 1. The key contributions of the present work can be summarized as follows.

1. We propose a local-aware spatio-temporal attention network with multi-stage feature fusion for human action recognition. In our approach, we employ a spatial transformer network (STN) as our attention model for capturing the meaningful regions from video frames. Then, the network outputs the discriminative descriptors for human action recognition through feature fusion.
2. We improve the feature fusion method by introducing the enhanced compact bilinear pooling. The feature fusion is conducted three times in our architecture. The

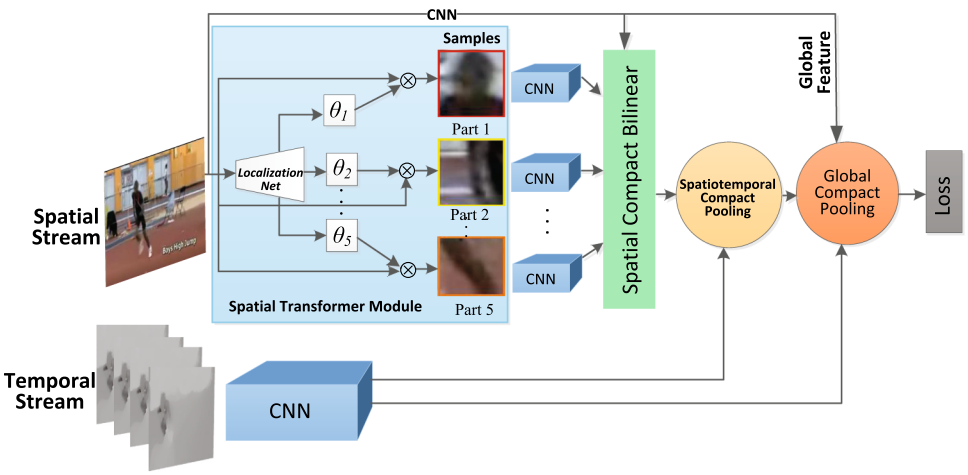

Fig. 1 Overview of our model architecture

first feature fusion is proposed to strengthen the human action representation through combining the local attentional regions obtained by the STN with global features from the original spatial input. We perform the second feature fusion by fusing the output of the first feature fusion and the temporal features. In this way, the pixel-level feature correspondences between the spatial and temporal streams are learned successfully, while parameters of our architecture can be reduced. The third feature fusion is employed to generate the global descriptors of human action, bringing together three outputs: the spatial features, temporal features, and spatio-temporal interaction information.

3. We compare the proposed feature fusion method with existing alternatives, such as sum, conv, and concatenation. The overall architecture is evaluated on two standard human action recognition datasets: HMDB51 [22] and UCF101 [34]. Experiments demonstrate that the proposed architecture leads to superior performance over both the traditional machine learning methods as well as the state-of-the-art deep architectures.

2 Related work

2.1 Attention models

The attention mechanism has been widely applied to existing CNNs over the past decades [31, 41]. Due to its significance, recent studies have explored the performance of attention models in various forms for the human action recognition task. For example, Girdhar et al. [11] devised an attention pooling model to replace the last pooling layer of their baseline architecture, but its performance on the RGB dataset of HMDB51 did not surpass that of the I3D model. Wang et al. [41] presented non-local operations to capture long-range dependencies. Specifically, they computed the interactions at any two positions, regardless of their positional distance. Another way of exploring attention models was proposed in [6], where the authors combined an attention structure of spatio-temporal networks with an RNN to recognize human actions from videos. However, their attention models failed to capture the fine-grained difference between video frames, which have a deep impact on recognizing human actions that share high similarities. Ge et al. [10] proposed an attention mechanism based on the convolutional LSTM network to improve the performance of human action recognition by embedding an LSTM module into a spatial transformer network (STN). Kuen et al. [23] devised a recurrent attentional convolutional–deconvolution network and combined the STN [17] with recurrent network units to detect the fine-grained differences from static images. Despite this progress, the correlations between the local and global regions of objects have yet to be considered in the literature.

Therefore, in this work, we investigate the performance of the STN in more complex tasks, such as video-level human action recognition, with the aim of capturing the regions of the human body that are expected to be useful for fine-grained action recognition. The STN is capable of extracting the attentional regions of the input with different spatial transformation methods, including rotation, cropping, transition, and scaling. In addition, we enhance the human action representation through combining the local and global regions of objects that are obtained by the STN and the original spatial input, respectively. The proposed attention model is expected to yield better performance in distinguishing human actions that share a high similarity and to capture the correlations between different frames.

2.2 Fusion methods

After extracting spatial and temporal features from videos, we perform feature fusion for human action recognition. In the past, researchers have studied the effectiveness of different fusion methods across streams in two-stream networks. Simonyan et al. [32] proposed fusing the two streams through averaging the Softmax approximation of the spatial and temporal streams by late fusion (i.e., fusing fully connected layers of the two streams), and then applying a multi-class linearity SVM on L2 regularization. Although higher recognition accuracies were reported compared to hand-crafted models, this research neglected the feature correspondence between the two streams when extracting the spatio-temporal features. Following this, Feichtenhofer et al. [8] improved two-stream CNNs through Conv fusion to stack the spatio-temporal feature maps by extracting the spatio-temporal features with 3D convolution and 3D pooling. However, this work came with the risk of increasing training parameters in the later stage of the network. Meanwhile, Pinz et al. [7] investigated the fusion of the two streams by multiplicative motion gating functions, aiming to gain leverage on multiplicative interactions by utilizing the cross-stream residual connections. Unfortunately, this method did not yield much improvement in recognition accuracy.

Our present study explores an alternative fusion method to capture the correlations between the spatial and temporal streams while reducing the training parameters. Specifically, we introduce compact bilinear pooling [9], which was initially proposed to fuse spatial features in fine-grained recognition tasks. We further study its effectiveness for fusing the spatial attentional features and the temporal features. In particular, the feature fusion is performed in a multiplicative manner to achieve the interactions across the spatial and temporal streams.

3 Approach

This section provides the detailed description of our proposed approach by firstly introducing the overall neural architecture. In what follows, we discuss the details of STN for attentional region localization and the feature fusion method for spatio-temporal information interaction.

3.1 Proposed architecture

In the proposed architecture, we first introduce the STN to locate the meaningful regions of human actions in the spatial stream. Then, we combine the spatial attentional features with the original spatial input by compact bilinear pooling to strengthen the human action representation between local and global regions. We consider this process as the first feature fusion. The fusion layer is placed between the last convolutional layer and the fully connected layer of the backbone network. At the same time, the optical flow of the temporal stream is pre-computed to capture the action motions, thereby extracting the trajectory of human action. Furthermore, the spatial and temporal features are fused together to achieve the spatio-temporal interaction. After the second feature fusion, we design two convolution layers to produce the weights for each grid location and then employ a softmax layer to generate the output feature map for global compact bilinear pooling. Finally, we aggregate information to obtain the global descriptors, which encompass all three outputs: the spatial features, temporal features, and spatio-temporal interaction information. The three inputs are fused into a single fixed-length vector through compact bilinear pooling.

3.2 STN for attentional region localization

In our present study, the STN [17] is introduced to search for the discriminative regions of the human body in the recognition task. Notably, the STN is a differentiable attention network and can be integrated into the CNN directly without additional supervision. It is capable of learning the scales of discriminative attentional regions and cropping them out from video frames automatically. The STN was originally proposed for handwritten character recognition in static images [17]. Here, we apply it to video-level human action recognition by reconstructing its structure. For a clearer understanding, a detailed description of the STN is provided before diving deep into our network.

The STN consists of three parts: Localization Net, Grid Generator, and Sample. The schematic diagram is shown in Fig. 3. The design of Localization Net in the STN is shown in Fig. 4. It consists of a network structure with two convolutional layers, two pooling layers, and two fully connected layers.

Localization Net generates the matrix of parameters A_θ for spatial affine transformation:

$$A_\theta = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \tag{1}$$

where s_x, s_y, t_x, and t_y denote the varying parameters that are used for attention cropping, translation, and isotropic scaling.

Grid Generator generates a matrix $T_\theta(G)$, which represents the mapping matrix from the input feature map (U) to the output feature map (V). We assume that the coordinate

of each pixel of U and V is (x_i^s, y_i^s) and (x_i^t, y_i^t), respectively. We now have the point-wise coordinate transformation from U to V. The calculation process of matrix $T_\theta(G)$ is thereby defined as in Equation (2):

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \tag{2}$$

Lastly, according to the input feature map U and the mapping matrix $T_\theta(G)$, Sample generates the output feature map V for the subsequent feature extraction.

Human actions may vary greatly according to the intra-class differences in different situations. For example, when different people perform the same action (e.g., the high jump), their behavior varies due to their different body sizes. Human actions are detected by dividing different parts of the human body, including the head, arms, and legs, which all possess different degrees of saliency in human action recognition. To capture the fine-gained behavior difference across frames in the process of human action, multiple STNs are employed in a parallel manner. The performances of these STNs are also investigated in our study.

Moreover, to enhance human action representation, we perform the feature fusion through connecting the local attentional regions obtained by the STN with global features from the original spatial input to complement the information about human action, e.g., the relationship of local regions of the human body. Details of this building block are outlined in Fig. 2.

3.3 Spatio-temporal information interaction

In this section, we determine the interaction between the spatial attentional features and temporal features. Note that an effective interaction method should preserve spatial and temporal information maximally and enable the feature correspondences between the spatial and temporal streams at the same pixel. Taking this cue, we introduce bilinear pooling [26], which enables the spatial and temporal features in different dimensions to interact with each other in a multiplicative way. Bilinear pooling is suitable for our work because the information for feature fusion consists of spatial attentional features, spatio-temporal features, and temporal features, which could be in different dimensions.

Firstly, we describe the process of bilinear pooling. Given the two initialized vectors $\boldsymbol{x} \in \boldsymbol{R}^n$ and $\boldsymbol{y} \in \boldsymbol{R}^d$, which represent the spatial and temporal features, respectively. The formula for bilinear pooling is denoted by $\mathbf{z} = vec(x \otimes y)$, where $\otimes$ represents the outer product of $\boldsymbol{x}$ and $\boldsymbol{y}$, and *vec* indicates its vectorization. Notably, when the dimension of the features for fusion becomes rather large, the effectiveness of this algorithm will inevitably suffer. Higher dimensionality of parameter representation (e.g., when the length of $\boldsymbol{x}$ and $\boldsymbol{y}$ are particularly large) will eventually limit the algorithm's effectiveness.

To project the outer product of the two vectors to a low dimensional space, we propose an enhanced version of bilinear pooling, called spatial–temporal compact bilinear pooling. We introduce a projection function, namely the count sketch function ϕ [2], to project vector $\boldsymbol{m} \in \boldsymbol{R}^n$ to $\boldsymbol{m}' \in \boldsymbol{R}^d$, where n $\gg$ d. First, we initialize two vectors $\boldsymbol{s} \in \{-1, 1\}^n$ and $\boldsymbol{h} \in \{1, ...d\}^n$, and two indices j and k, where $\boldsymbol{s}$ is an index of either 1 or $-$ 1, and $\boldsymbol{h}$ maps each index j in the input $\boldsymbol{m}$ to an index k in the output $\boldsymbol{m}'$. All $\boldsymbol{s}$ and $\boldsymbol{h}$ obey the uniform distribution and remain constant

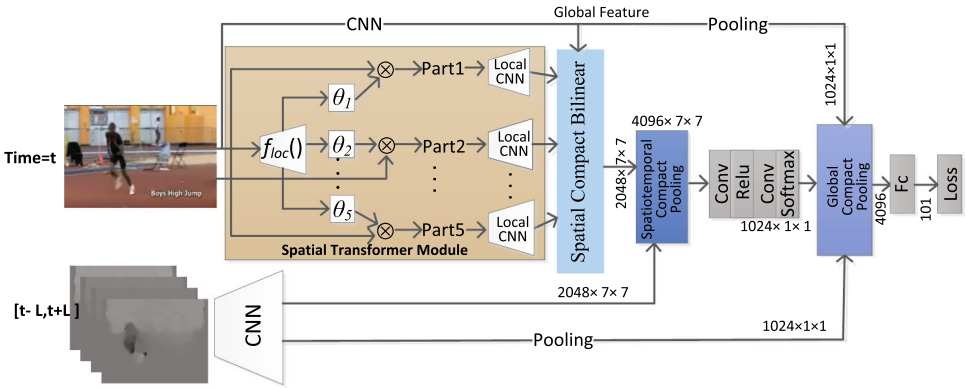

Fig. 2 We use UCF101 as an example. For spatial stream, we employ multiple STNs to locate the attentional regions of the human body; the local features and the spatial input image are fused by compact bilinear pooling. For temporal stream, we extract the optical flow to track the human action trajectory. Then, the spatio-temporal features are fused through compact bilinear pooling for further feature extraction. The final human action representations come from three sources: spatial information, temporal information, and spatio-temporal attention information

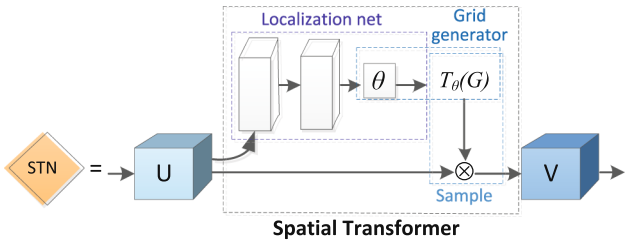

Fig. 3 Schematic diagram of the STN

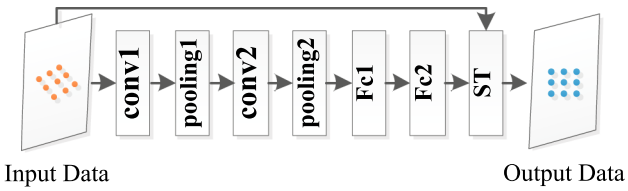

Fig. 4 Localization Net structure

for future invocations of the count sketch, and $\boldsymbol{m}'$ is initialized as a zero vector. Then, for each element $\boldsymbol{m}[j]$, its destination is $\boldsymbol{m}'[k]$, and the index k $=\boldsymbol{h}[j]$ is looked up using $\boldsymbol{h}$. We assume the number of feature pathways is v for feature fusion. Details are shown in lines 1−12 of Algorithm 1. Compared to the bilinear pooling, our proposed method

Algorithm 1: Spatial-Temporal Compact Bilinear

Input: Spatial and/or temporal features $\{m_i \in R^{mi}\}_{i=1}^{v}$
Output: $\Phi(\{m_i\}_{i=1}^{v}) \in R^d$
For i←1 to v do
  If h_i, s_i not initialized then
    For j←1 to n_i do
      Sample h_i[j] from {1,…,d}
      Sample s_i[j] from {-1,1}
    End
    m'_i =[0,…0]
    For j←1 to n_i do
      $m'_i[h_i[j]] = m'_i[h_i[j]] + s_i[j] \cdot m_i[j]$
    Return m'_i
  End
  $\Phi(\{m_i\}_{i=1}^{v}) = FFT^{-1}(FFT(m'_i) \odot_{i=1}^{v} FFT(m'_n))$
Return Φ

The above process projects the outer product of two vectors to a lower dimensional space, which can reduce the number of training parameters significantly. In addition, and to avoid computing the outer product of two vectors explicitly, we introduce the convolution computation

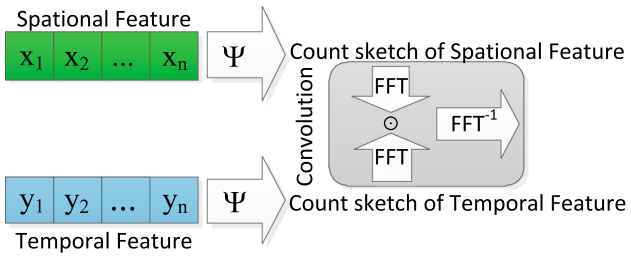

Fig. 5 Details of fusing spatio-temporal features through compact bilinear pooling

approach of count sketches [29] to represent the count sketch of the outer product of the two vectors:

$$\psi(\boldsymbol{x} \otimes \boldsymbol{y}, \boldsymbol{h}, \boldsymbol{s}) = \psi(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{s}) * \psi(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{s}) \quad (3)$$

where * denotes the convolution operator. According to the convolution theorem in the temporal domain, x*y can be rewritten as FFT^{-1} (FFT (x) $\odot$ FFT (y)), where the operator $\odot$ represents the element-wise product. The entire process denotes the compact bilinear representations of the spatial and temporal information. Details of this process are outlined in Fig. 5 and described in Algorithm 1.

4 Experiment

We carried out comprehensive experimental studies to investigate the efficacy of the proposed architecture for human action recognition. First, the details of experimental datasets are introduced. This is followed by an analysis of the effectiveness of different backbone networks, including VGG-19, ResNet-152, and BN-Inception, for feature extraction. Then, we discuss the impact of employing multiple spatial transformers in a parallel way to locate the fined-grained regions of the human body. Next, the performances of multiple spatio-temporal feature fusion methods in human action recognition are studied. Multi-stage feature fusion can enhance the performance of our approach while maximizing the interaction of the spatio-temporal information. Therefore, we conducted ablation studies to test the effectiveness of the proposed module. Finally, we compared our approach with existing state-of-the-art approaches.

Table 1 Information of the datasets used in our experiments

Dataset	Video clips	Training clips	Testing clips
UCF101	13320	9590	3730
HMDB51	5100	3762	1338

4.1 Datasets and implementation details

We tested the proposed architecture on two standard human action datasets: HMDB51 and UCF101. UCF101 contains 101 kinds of human actions, and each action is composed of 130 videos with different backgrounds and characters; the dataset contains 13,320 videos in total. HMDB51 contains 51 kinds of human actions. Each action is composed of about 100 videos, and the dataset contains a total of 5100 videos. The information of the datasets is listed in Table 1. We clipped a single frame with a size of 300×224 as the input for the spatial network, while the temporal network took 20 consecutive optical flow frames as input. We pre-trained the two networks of the proposed architecture on ImageNet [4] and fine-tuned them on standard datasets. The mini-batch stochastic gradient descent algorithm was used for optimization, and the batch size was set to 128. The learning rate was initially set as 0.01 and then decreased by 10 times every 20 epochs. We stopped training at 10K iterations. We performed data augmentation by random flipping and cropping on video frames to avoid overfitting. All experiments were implemented on TensorFlow. In our present study, we classified the video in a single forward pass, which took ≈200 ms on two Nvidia Titan V cards (the computational time for one video sequence is about 200 ms).

4.2 Comparison of different backbone networks

We investigated the effectiveness of different backbone networks in terms of the accuracy of human action recognition. VGG-19, ResNet-152, and BN-Inception were chosen as the backbone networks for video-level human action recognition as they have reported substantial gains in image classification [14, 16, 28]. We evaluated their performance on our proposed architecture and present the results in Table 2.

For the spatial network, recognition accuracy benefitted the most with an increased number of convolutional layers; meanwhile, the performance decreased slightly for the temporal network. This is because the CNN has a higher capability to learn spatial features, and deeper CNNs typically lead to higher recognition accuracy of the network, due to its discriminative representation of the original input in image recognition tasks [26, 36]. However, the CNN fails to capture the long-term information and cannot learn the optical flow field since temporal data have a different data distribution. BN-Inception obtained the best performance among all the backbone networks, reporting a recognition accuracy of 85.9% for the spatial network and 83.3% for the temporal network. In light of these results, we selected BN-Inception as our target backbone network for the human action recognition task.

Table 2 Classification accuracy of spatial and temporal networks with different backbone networks (i.e., VGG-19, ResNet-152 and BN-Inception) on UCF101

Backbone network	Spatial (%)	Temporal (%)
VGG-19	81.2	82.5
ResNet-152	85.1	81.9
BN-Inception	85.9	83.3

4.3 Comparison of accuracy based on different numbers of STNs

We verified the effectiveness of the attention models in capturing the fine-grained regions of the human body. Specifically, we employed different numbers of STNs in parallel to capture the information of human action. As we can see from Table 3, with the increase in the number of STNs, the accuracy of recognition improved gradually. However, when the number of STN reaches a certain level (seven in our present study), the accuracy decreased slightly. We conjecture that this result is due to the additional layers in the localization network of the STN, which yield more trainable parameters. When the number of attention regions reaches seven, the lack of a gradient signal leads to reduced accuracy and, hence, limits the ability of the STN to capture the meaningful local information. Therefore, we set the number of STNs to five and placed them in the same layer.

The feature fusion between the local regions and global features obtained from the original input improved the overall performance of our approach. As shown in Table 3, the accuracy of $5\times$STNs with global features was better than that without global features, and better performance was reported (0.2 points higher in accuracy). The results further highlight the efficacy of the global features obtained from the original input being used to provide more global spatial information of each local region.

Lastly, we discuss the visualization of our attention model predicted by $5\times$STNs on the human body after performing 10K iterations. As shown in Fig. 6, one STN (shown in blue) learned to detect the head of the human body, while the others fixated on the rest of the human body. We observed some overlaps among several regions that contained useful information of human actions. Our approach is capable of locating meaningful regions of the human body to capture the fine-grained differences between human actions.

Table 3 Effectiveness of placing multiple STNs on UCF101 dataset for the spatial input and recognition accuracy of feature fusion

Number of STNs	Accuracy−without global features(%)	Accuracy−with global features(%)
1	86.2	86.3
3	86.8	86.3
5	87.3	87.5
7	87.0	87.3

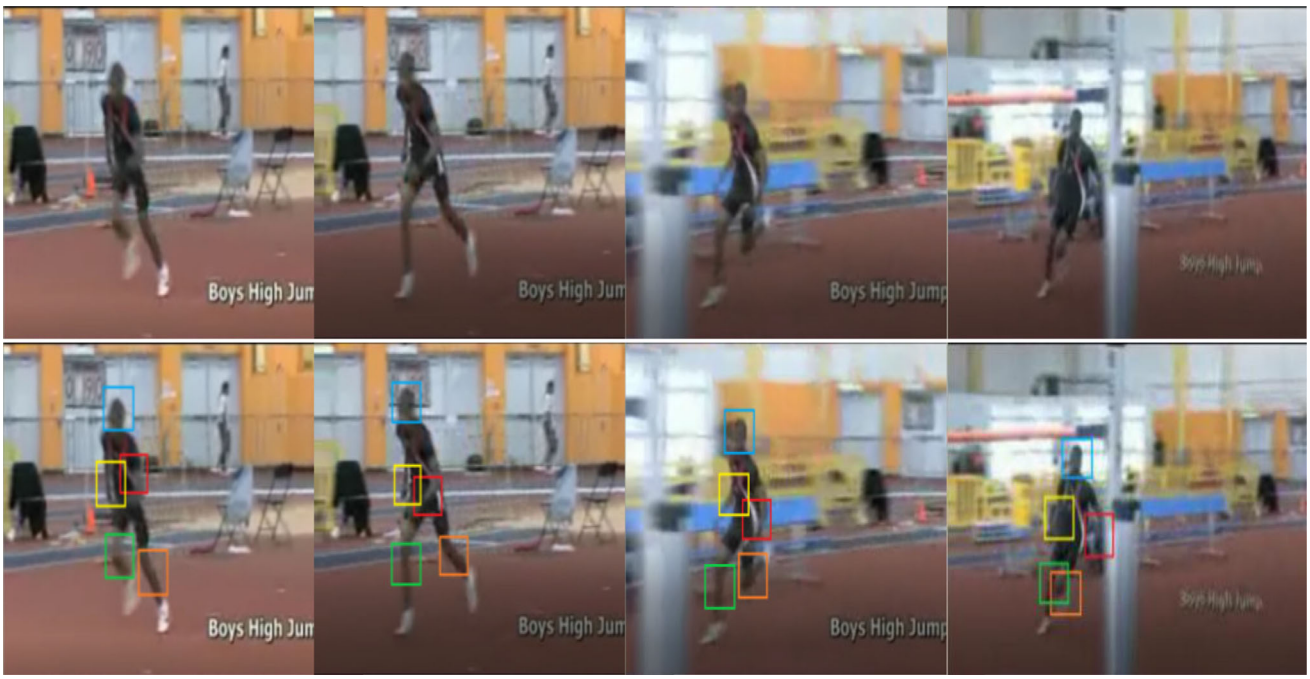

Fig. 6 Visualization of the attentional regions using our method

Table 4 Accuracy (%) of different fusion methods on UCF101 dataset

Fusion method	Accuracy (%)	Training parameters (M)
Sum	90.7	201.06
conv	91.4	220.48
Concatenation	91.2	340.48
CBP (d=1024)	93.1	150.16
CBP (d=2048)	93.6	169.06
CBP (d=4096)	94.2	174.48
CBP (d=8192)	94.2	198.96
CBP (d=16000)	94.1	213.06

4.4 Different spatio-temporal fusion methods

We studied the performance of different feature fusion methods in obtaining information interaction between spatial and temporal features. Note that the fusion layer in our present study was placed between the last convolution layer and softmax layer of the backbone network. This is because the network with a late fusion performed better than the network with the fusion layer in an earlier stage since there was much more information of deeper feature extraction in the CNN [42]. We investigated the efficacy of fusing the spatial and temporal features through compact bilinear pooling. As shown in Table 4, compact bilinear pooling with 1024-d improved the average accuracy by 2.0% compared with the other fusion methods, such as sum, conv, and concatenation. Therefore, we can conclude that our method achieves significant improvements compared with other fusion methods.

We also validated that different dimensions of compact bilinear pooling greatly affect recognition accuracy. As shown in Table 4, we observed that with the increase in the output's dimension, recognition accuracy improved accordingly, as more information of the spatial and temporal features was captured. However, a larger dimensionality will produce a much larger number of training parameters. As we can see, when the number of dimensions reached 16,000, the training parameters increased dramatically compared with 4096-d (174.48M vs 213.06M), thereby incurring additional difficulties in the training

Table 5 Results of ablation studies on UCF101 (Split1)

Fusion method	Original two-stream CNNs[6]	Temporal stream	Spatial stream	Two-stream fusion	Final result
Spatial compact bilinear(♯1)	–	–	✓	✓	✓
Spatial compact bilinear(♯2)	–	–	–	✓	✓
Global compact bilinear(♯3)	–	–	–	–	✓
Accuracy	88%	82.6%	90.8%	93.1%	95.3%

process. To sum up, the 4096-d vector turns out to be an appropriate choice for compact bilinear pooling in our present study. The obtained results highlight the efficiency of the feature fusion method.

4.5 Ablation results

Ablation studies were conducted to testify the effectiveness of our proposed architecture. The complete experimental results are shown in Table 5. First, we measured the performance of spatial and temporal streams to perform the task of human action recognition on UCF101. The results show that the spatial stream and temporal stream achieved accuracy of 90.8% and 82.6%, respectively. By taking into consideration the fusion of the two streams, our approach boosted the classification performance significantly and reported a higher accuracy of 95.3%. Moreover, the effectiveness of multi-stage feature fusion was estimated in detail. Specifically, the first stage feature fusion, denoted by "spatial compact bilinear", reported a higher accuracy of 90.8% and was 2.8 points higher than the 88% of the original two-stream CNNs [32]. After involving the second stage feature fusion, the "spatio-temporal compact bilinear", which integrates both temporal and spatial features, obtained an increased accuracy of 2.3 points compared with that of the first stage feature fusion. Finally, by aggregating the spatial features, temporal features, and the output of the spatio-temporal compact bilinear pooling, the performance of the "global compact bilinear" was improved by 2.2 points, and the overall highest accuracy rate was obtained (95.3%).

Thus, the experimental results verify the effectiveness of our proposed architecture for video-level human action classification.

Table 6 Average classification accuracy of state-of-the-art methods on HMDB51 and UCF101

Method	UCF101 (%)	HMDB51
IDT [38]	86.4	61.7%
C3D [1]	85.2	–
C3D +IDT [1]	90.4	–
Two-stream (VGG) [32]	88.0	59.4%
TDD+IDT [39]	91.5	65.9%
Two-stream +LSTM[19]	88.6	–
Convolutional two-stream [8]	93.5	69.2%
TSN [40]	94.2	69.0%
ActionVLAD (VGG-16) [12]	93.6	69.8%
Spatio-temporal ResNet[34]	94.6	70.3%
Spatio-temporal multiplier [7]	94.9%	72.2%
Ours + SVM	92.1	70.6%
Ours + softmax	95.3	72.9%

4.6 Comparison to state-of-the-art methods

Table 6 summarizes the complete results in terms of prediction accuracy of our proposed architecture against existing works on UCF101 and HMDB51 datasets. Specifically, when comparing the trajectory-based handcrafted IDT [38] features, our approach improved the accuracy by 8.9% and 11.2% on UCF101 and HMDB51, respectively. Furthermore, considering the classical classification methods in machine learning, such as SVM and ELM, we can see that our approach with softmax provided a consistent performance gain of around 3% on both datasets. When compared with deep learning methods, such as C3D [38], Two-stream + LSTM [40], TSN [40], and spatio-temporal multiplier [7], our approach obtained the highest prediction accuracies of 95.3% and 72.9% for UCF101 and HMDB51, respectively, which were 0.4% and 0.7% higher than the state-of-the-art method, spatio-temporal multiplier. This result thus clearly underlines the significance of the proposed approach.

To elaborate, Fig. 7 shows some typical actions from UCF101 and presents the comparison results obtained using the STM network and our proposed architecture. As we can see from the comparisons, target actions with similar backgrounds can easily lead to misclassification with the STM network. For example, the STM network mistook the action "high jump" for "pole vault" since the image backgrounds are both stadiums. However, these human actions with ambiguous spatial features can be

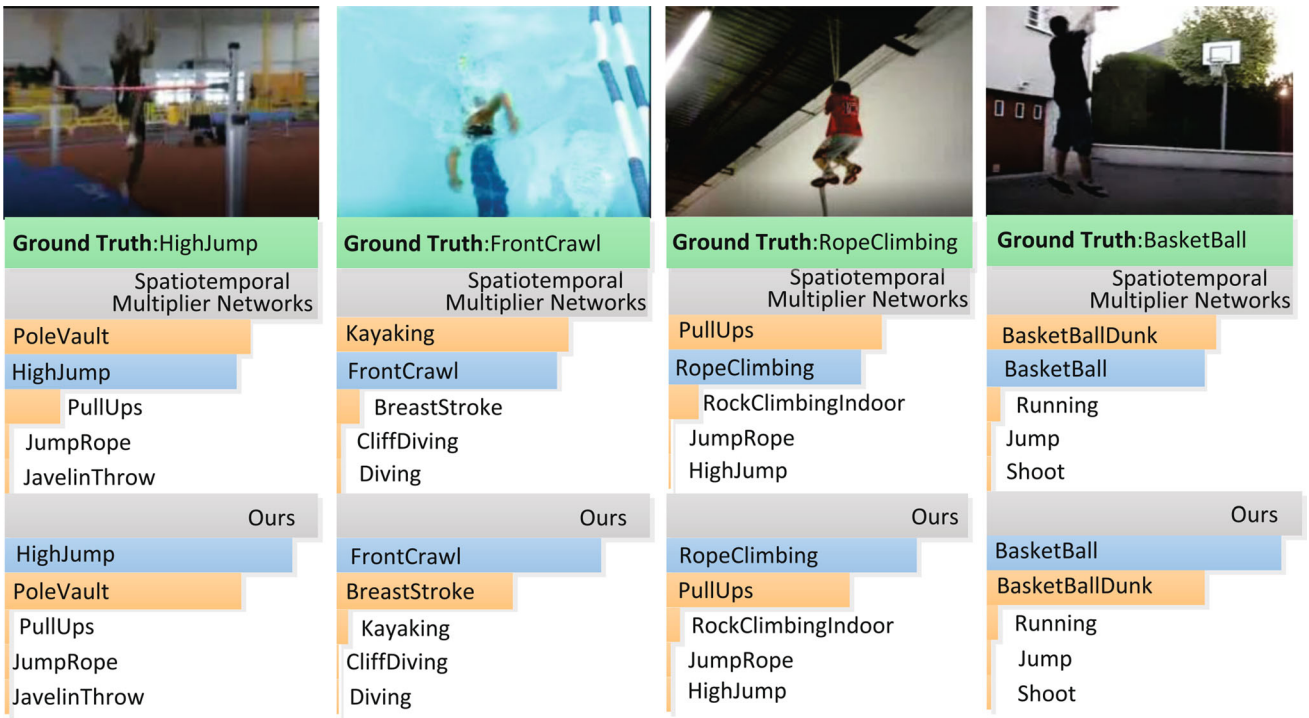

Fig. 7 We compare our approach with existing state-of-the-art approaches and test several actions that are easily misclassified on the UCF101 dataset. Finally, we select the first five classification results for presentation: the blue area represents correct classification, the orange area represents error classification, and the length represents the confidence of the classification

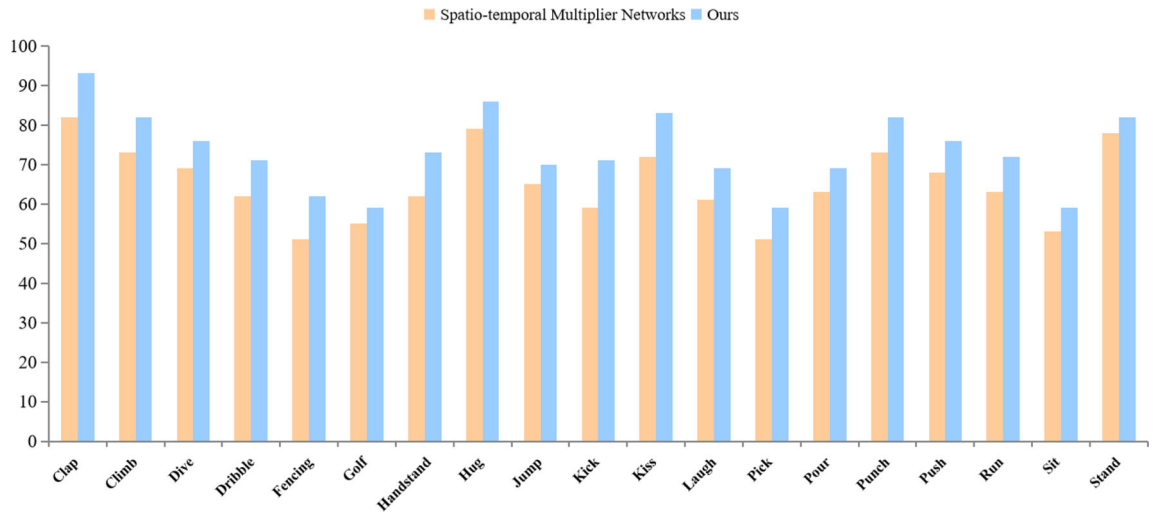

Fig. 8 Comparison between our approach and the STM network on the first 20 samples of the HMDB51 dataset. The vertical axis represents the recognition accuracy, and the horizontal axis represents the action type

easily classified through our attentional mechanism to locate the fine-grained difference. Moreover, owing to our multi-stage feature fusion, the proposed architecture can generate global descriptors of human action over long-term action sequences and can correctly distinguish the actions that have similar spatial features in the short term.

In addition, Fig. 8 presents the comparison results of human action recognition accuracy between the STM network and our approach on the first 20 samples of the HMDB51 dataset. We can see that our approach produces a consistent improvement in accuracy of around 2% on these human actions. For some human actions that are difficult to differentiate, such as “clap” and “dribble”, the performance of our approach is superior to that of the STM network in terms of action classification accuracy. The improved performance obtained by our approach thus

demonstrates the effectiveness of combining an attentional model and multi-stage feature fusion. In addition, the multi-stage feature fusion provides more useful cues for human action recognition. The consistent improvements indicate the superiority of our proposed architecture.

5 Conclusion

In this paper, we propose a local-aware spatio-temporal attention network with multi-stage feature fusion for video-level human action recognition; this network is trainable in an end-to-end manner. First, to capture the meaningful regions of human actions from video frames, we utilize multiple STNs to locate the attentional regions of the human body, and then feature fusion is carried out between local and global features to enhance the human action representation. The experimental results, which showed an accuracy of 90.8% on the UCF101 dataset, demonstrate that the proposed attention model can significantly improve the recognition performance of our proposed architecture. Moreover, we employ compact bilinear pooling with a dimensionality of the 4096-d vector to learn the feature correspondences between the spatial and temporal streams. Our approach boosts the classification performance significantly and reports an accuracy of 93.1%. Finally, when we aggregated the spatial features, temporal features, and the output of the spatio-temporal compact bilinear pooling, we obtained accuracy rates of 95.3% on UCF101 and 72.9% on HMDB51. These results verify the superiority of our proposed architecture for performing the task of human action recognition. Still, the existing work could be time-consuming when preprocessing the input data of temporal network for human action recognition. Therefore, our future work might focus on exploring more efficient approaches to further reduce the time complexity while avoid loss in accuracy.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61906032, the NSFC-Liaoning Province United Foundation under Grant U1908214, the LiaoNing Revitalization Talents Program, No. XLYC2008017, the National Key Research and Development Program of China under Grant 2018YFC0910500, the National Natural Science Foundation of China under Grant 61976034, and the Liaoning Key Research and Development Program under Grant 2019JH2/10100030.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Chéron G, Laptev I, Schmid C (2015) P-cnn: Pose-based cnn features for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 3218–3226
2. Dai H, Shahzad M, Liu AX, Zhong Y (2016) Finding persistent items in data streams. Proceedings of the VLDB Endowment 10(4):289–300
3. Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. Springer, Berlin
4. Deng J, Dong W, Socher R, Li L, Li K, Feifei L (2009) Imagenet: a large-scale hierarchical image database pp. 248–255
5. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, Saenko K (2015) Long-term recurrent convolutional networks for visual recognition and description pp. 2625–2634
6. Du W, Wang Y, Yu Q (2017) Recurrent spatial-temporal attention network for action recognition in videos. IEEE Trans Image Process 27(99):1347–1360
7. Feichtenhofer C, Pinz A, Wildes RP (2017) Spatiotemporal multiplier networks for video action recognition pp. 7445–7454
8. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition pp. 1933–1941
9. Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling pp. 317–326
10. Ge H, Yan Z, Yu W, Sun L (2019) An attention mechanism based convolutional lstm network for video action recognition. Multim Tools Appl 78(14):20533–20556
11. Girdhar R, Ramanan D (2017) Attentional pooling for action recognition pp. 34–45
12. Girdhar R, Ramanan D, Gupta A, Sivic J, Russell BC (2017) Actionvlad: learning spatio-temporal aggregation for action classification pp. 3165–3174
13. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet pp. 6546–6555
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition pp. 770–778
15. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks pp. 630–645
16. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv: Learning
17. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks pp. 2017–2025
18. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
19. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Feifei L (2014) Large-scale video classification with convolutional neural networks pp. 1725–1732
20. Klaser A, arszalek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients pp. 1–10

21. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks pp. 1097–1105
22. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: a large video database for human motion recognition pp. 2556–2563
23. Kuen J, Wang Z, Wang G (2016) Recurrent attentional networks for saliency detection pp. 3668–3677
24. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies pp. 1–8
25. Li C, Zhong Q, Xie D, Pu S (2017) Skeleton-based action recognition with convolutional neural networks
26. Lin T, Roychowdhury A, Maji S (2015) Bilinear cnn models for fine-grained visual recognition pp. 1449–1457
27. Mohammad S, Mircea N, Monica N, Banafsheh R (2015) Intent understanding using an activation spreading architecture. Robotics 4(3):284–315
28. Perronnin F, Sanchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification 6314:143–156
29. Pham N, Pagh R (2013) Fast and scalable polynomial kernels via explicit feature maps pp. 239–247
30. Shi X, Chen Z, Wang H, Yeung D, Wong W, Woo W (2015) Convolutional lstm network: a machine learning approach for precipitation nowcasting pp. 802–810
31. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition
32. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos pp. 568–576
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition
34. Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. Computer ence
35. Srivastava N, Mansimov E, Salakhudinov R (2015) Unsupervised learning of video representations using lstms pp. 843–852
36. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions pp. 1–9
37. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks
38. Wang H, Schmid C (2013) Action recognition with improved trajectories pp. 3551–3558
39. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors pp. 4305–4314
40. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: towards good practices for deep action recognition pp. 20–36
41. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks pp. 7794–7803
42. Wang Y, Long M, Wang J, Yu PS (2017) Spatiotemporal pyramid network for video action recognition pp. 2097–2106