Multimedia Systems (2023) 29:487–498
https://doi.org/10.1007/s00530-022-00961-3

SPECIAL ISSUE PAPER

Multi-head attention-based two-stream EfficientNet for action recognition

Aihua Zhou[1,2] · Yujun Ma[3] · Wanting Ji[4] · Ming Zong[5] · Pei Yang[1,2] · Min Wu[6] · Mingzhe Liu[7]

Received: 7 January 2022 / Accepted: 29 May 2022 / Published online: 24 June 2022

Abstract

Recent years have witnessed the popularity of using two-stream convolutional neural networks for action recognition. However, existing two-stream convolutional neural network-based action recognition approaches are incapable of distinguishing some roughly similar actions in videos such as sneezing and yawning. To solve this problem, we propose a Multi-head Attention-based Two-stream EfficientNet (MAT-EffNet) for action recognition, which can take advantage of the efficient feature extraction of EfficientNet. The proposed network consists of two streams (i.e., a spatial stream and a temporal stream), which first extract the spatial and temporal features from consecutive frames by using EfficientNet. Then, a multi-head attention mechanism is utilized on the two streams to capture the key action information from the extracted features. The final prediction is obtained via a late average fusion, which averages the softmax score of spatial and temporal streams. The proposed MAT-EffNet can focus on the key action information at different frames and compute the attention multiple times, in parallel, to distinguish similar actions. We test the proposed network on the UCF101, HMDB51 and Kinetics-400 datasets. Experimental results show that the MAT-EffNet outperforms other state-of-the-art approaches for action recognition.

Keywords Action recognition · Multi-head attention · Two-stream network

1 Introduction

Action recognition aiming to recognize human actions has been highlighted in vision computing [3, 27, 50, 53]. Action recognition has been widely applied in elderly behaviour monitoring, surveillance systems, human–computer interaction, video retrieval, public opinion monitoring and many other applications [59, 60, 63, 64] related to action recognition [1, 5, 28].

Convolutional Neural Networks (CNNs) have achieved great performance in many research fields such as speech processing [46] and natural language processing [47, 62]. Early efforts on action recognition utilized some well-known CNNs such as AlexNet [6], VGGNet [7] and ResNet [8] to recognize actions in videos. Google proposed an

✉ Yujun Ma
yma1@massey.ac.nz

Aihua Zhou
zhouaihua@geiri.sgcc.com.cn

Wanting Ji
wanting.ji@lnu.edu.cn

Ming Zong
zongming@pku.edu.cn

Pei Yang
404648438@qq.com

Min Wu
wm_lucy@sina.com

Mingzhe Liu
liumz@cdut.edu.cn

[1] State Grid Smart Grid Research Institute CO., LTD, Beijing, China

[2] State Grid Key Laboratory of Information and Network Security, Nanjing, China

[3] School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

[4] School of Information, Liaoning University, Shenyang, China

[5] National Engineering Research Center for Software Engineering, Peking University, Beijing, China

[6] Bejing Institute of Computer Technology and Applications, Beijing, China

[7] State Key Laboratory of Geohazard Prevention and Geo-Environment Protection, Chengdu University of Technology, Chengdu, China

EfficientNet [10] in 2019, which used all dimensions of the recombination coefficient unified scaling CNN models to obtain the highest accuracy. EfficientNet [10] had a great performance in all aspects compared with previous CNNs in classification related tasks [10]. However, videos contain complex spatial–temporal structures [4]. These CNN-based approaches [7, 10, 33] only extracted the spatial features in videos, while ignoring the temporal features.

To extract both spatial features and temporal features from videos [39], two important types of action recognition approaches were proposed: (i) 3D CNN-based approaches [9], and (ii) two-stream network-based approaches [1]. Differing from previous CNN-based approaches, 3D CNN-based approaches perform 3D convolutions over stacked video frames for feature extraction. For example, Carreira et al. [11] proposed an Inflated 3D CNN (I3D) to initialize 3D CNNs by inflating deep CNNs to recognize actions in videos. However, 3D CNN-based action recognition approaches usually include abundant parameters and need to be pre-trained on a large-scale video dataset.

In contrast, the training process of two-stream network-based approaches is similar to the training process of CNN-based approaches. In general, two-stream network-based approaches consisted of a spatial stream and a temporal stream. The two streams extracted features from videos, in which the spatial stream adopted RGB video frames [26] as the input and the temporal stream adopted the multi-frame optical flow of a video as the input. Each stream employed a CNN as the backbone network, and the softmax layer scores of the two streams were fused by late average fusion to calculate the final recognition results.

Recently, attention mechanisms have shown remarkable performance in capturing effective features [49] from videos [22, 23]. Various action recognition approaches utilized attention mechanisms [16, 29, 48] to capture action information from videos. Sharma et al. [14] proposed a soft attention-based network for action recognition, which concentrated on the key information of video frames for action recognition. Wang et al. [34] proposed a Cascade multi-head Attention Network (CATNet) for action recognition, which constructed the process of CNN feature extraction with a multi-head attention mechanism in an end-to-end fashion. However, CATNet only utilized a multi-head attention mechanism to extract 3D CNN-based motion information from video frames rather than to extract motion features from optical flow frames, which could capture motion information directly.

In this paper, we propose a Multi-head Attention-based Two-stream EfficientNet (named MAT-EffNet for short) for action recognition. The proposed network contains two streams, i.e., a spatial stream and a temporal stream, which extract the spatial and temporal features [2] from videos using EfficientNet. We utilize EfficientNet-B0 [10] as the baseline network since EfficientNet [10] has shown remarkable performance on the image classification task. The main contributions of our MAT-EffNet approach are summarized as follows:

- Existing approaches [1–3] only use general CNN to extract the spatial and temporal features in videos, which ignore the key action information (e.g., objects and motion) in videos. To address this issue, we propose a multi-head attention mechanism-based two-stream network to capture the key action information from the extracted features in videos. Thus, MAT-EffNet can focus on the key action information at different frames to distinguish similar actions. The EfficientNet is applied as a feature extractor because of the high parameter efficiency and speed.
- We conduct experiments on three widely used action recognition datasets (i.e., UCF101 [35], HMDB51 [36] and Kinetics [51]) to verify the performance of our approach. The experimental results show that the MAT-EffNet approach achieves the best classification results compared with several state-of-the-art methods. The rest of this paper is organized as follows. We review the existing two-stream network-based approaches and the attention mechanism-based approaches in Sect. 2. Section 3 presents the details of the proposed MAT-EffNet approach. Experimental results are presented in Sect. 4. Finally, Sect. 5 is the conclusion of this paper.

2 Related works

In this section, we first review the existing two-stream network-based action recognition approaches in Sect. 2.1, and then, the attention mechanism-based approaches are reviewed in Sect. 2.2.

2.1 Two-stream network-based action recognition approaches

Two-stream network-based approaches are the mainstream for action recognition since they can extract both spatial and temporal CNN features from videos [32, 61]. Simonyan et al. [1] first proposed a two-stream CNN, which consisted of a spatial stream and a temporal stream for action recognition. Specifically, given a video, RGB video frames were fed into the spatial stream and the dense optical flow frames were used as the input to the temporal stream. Then, the outputs of the two streams were fused to recognize the actions in the videos.

Based on the two-stream network proposed [1], Wang et al. [13] proposed a temporal segment network that could extract snippets from videos using sparse sampling rather than dense sampling. The short snippets were fed into the temporal stream and the spatial stream, respectively, and then the classification scores of the two streams were fused to obtain a video-level prediction. Feichtenhofer et al. [2] proposed a convolutional two-stream fusion approach for action recognition which utilized a convolutional fusion layer and a temporal fusion layer to capture short-term information in videos but did not increase the number of parameters remarkably. Zhu et al. [14] proposed an extra pre-trained layer (MotionNet) for motion information generation. The output of MotionNet was fed into a temporal stream, which projected the motion information onto the target action recognition labels. However, these two-stream network-based approaches cannot distinguish the key information for action recognition from the videos. Thus, attention mechanisms were introduced for action recognition.

2.2 Attention mechanisms

Attention mechanisms focus on specific parts of the input, which were first developed for machine translation [19]. Later, since attention mechanisms achieved good performance in machine translation, they have been widely introduced to machine reading [20], image captioning [21], meta-learning [44] and many other tasks [57].

Bahdanau et al. [19] proposed a soft attention mechanism-based approach for machine translation, which could capture the alignments between raw text and target words using the proposed soft attention mechanism. Cheng et al. [20] proposed a self-attention mechanism-based approach for machine-reading, which learned the correlation between previous parts of a sentence and new generated words using the proposed self-attention mechanism. Xu et al. [21] proposed an attention-based network for image captioning, which utilized CNNs to extract features and utilized visual attention-based recurrent neural networks to generate words to describe image contents.

Meanwhile, some approaches adopted attention mechanisms for action recognition. Wang et al. [40] proposed a hierarchical attention mechanism-based network for action recognition, which used a multi-step spatial–temporal attention mechanism to capture important spatiotemporal information from videos. Tran et al. [41] proposed a two-stream flow-guided convolutional attention network for action recognition, which added a cross-linked layer between two streams. This approach focused on the foreground of the object rather than the background. Girdhar et al. [15] proposed an attentional pooling layer to extract attended features for action recognition. The proposed attentional pooling layer focused on the specific part of the input frames. This approach added the human pose as intermediate supervision to train the attention mechanism. Peng et al. [42] proposed a spatial–temporal attention-based two-stream collaborative approach for video classification, which could exploit the complementarity between spatial and temporal information. Girdhar et al. [18] proposed a video action transformer network for action recognition, which focused on faces and hands that were discriminative cues for action recognition.

However, existing attention mechanism-based two-stream networks did not perform well in distinguishing roughly similar actions. To solve this problem, in this paper, we propose a multi-head attention-based two-stream EfficientNet model that can focus on the key action information in videos via a multi-head attention mechanism.

3 Multi-head attention-based two-stream EfficientNet

As shown in Fig. 1, the proposed MAT-EffNet is based on a two-stream network, which consists of two streams: a spatial stream and a temporal stream. Input videos are first decomposed into RGB video frames and stacked optical flow frames for extracting spatial and temporal features. In the spatial stream, RGB frames of the input video are fed into an EfficientNet. Stacked optical flow frames are the input of the temporal stream. Then, a multi-head attention mechanism is used on both streams to capture the key action information from videos. The outputs of each stream are combined via a late average fusion to compute the final predictions (i.e., the action labels of the input video).

In this section, we first introduce EfficientNet (EffNet) [10] in Sect. 3.1. The multi-head attention mechanism is presented in Sect. 3.2. Lastly, the detailed architecture of the proposed MAT-EffNet is presented in Sect. 3.3.

3.1 EfficientNet

EfficientNet [10] was a novel CNN network with high parameter efficiency and speed. EfficientNet [10] used a simple and compound scaling method to scale up the CNN models in a more structured way by uniformly scaling the network dimensions such as depth, width, and resolution. EfficientNet [10] was used as the spatial feature extraction network in classification tasks. The EfficientNet [10] family contained seven CNN models which were named EfficientNet-B0 to EfficientNet-B7. With the same input size, EfficientNet-B0 [10] could surpass Resnet-50 [8] with less parameter number and FLOPs (floating-point operations per second) accuracy, which indicated that EfficientNet-B0 [10] has an efficient feature extraction capability. The detailed structure of EfficientNet-B0 is shown in Fig. 2, which can be divided into seven blocks based on several channels, striding and convolutional filter size.

The main building block of EfficientNet-B0 is the mobile inverted bottleneck (MBConv), which is based on the concept of MobileNet [54, 55]. As shown in Fig. 3, MBConv consists of two convolutional layers(k1 × 1), a depthwise convolutional layer, a Squeeze and Excitation (SE) [54, 55] block, and a dropout layer. The first convolutional layer is used to expand the channels. The depthwise convolution is used to reduce the number of parameters. The SE block can focus on the relationship between channels and give a different weightage to each channel instead of computing them all equally. The second convolutional layer is used to compress the channels.

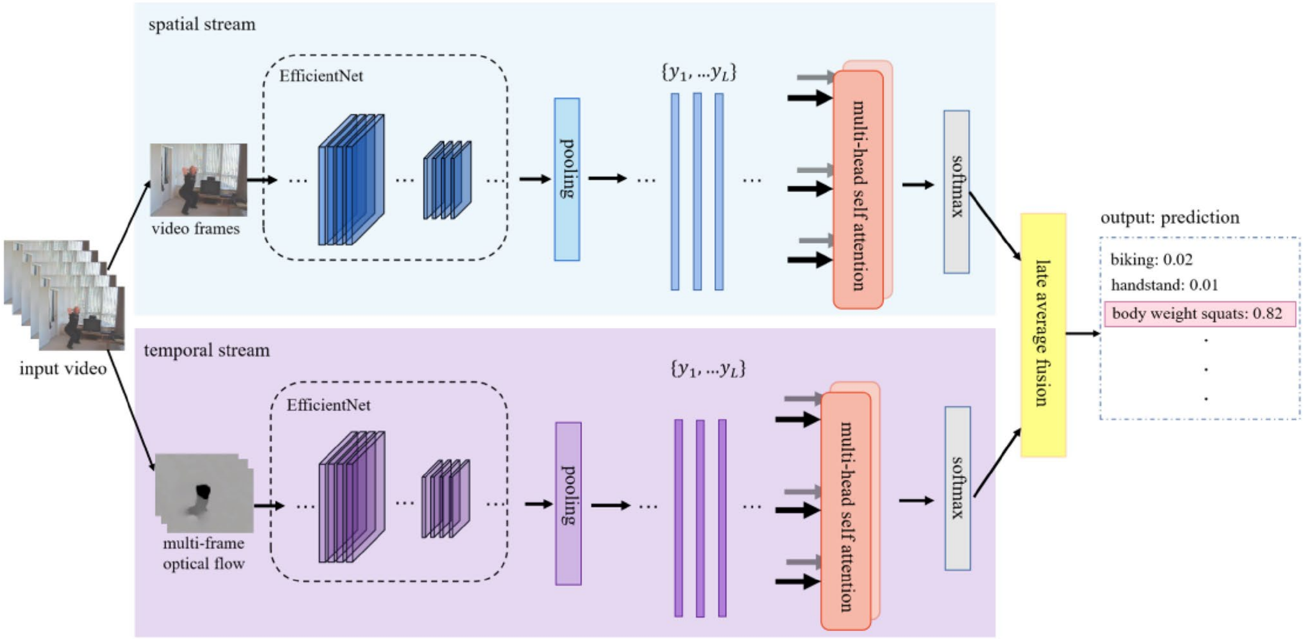

Fig. 1 The framework of the proposed MAT-EffNet, which consists of two streams: a spatial stream and a temporal stream. Each stream contains an EfficientNet [10] and a multi-head attention layer. The final prediction is obtained via a late average fusion

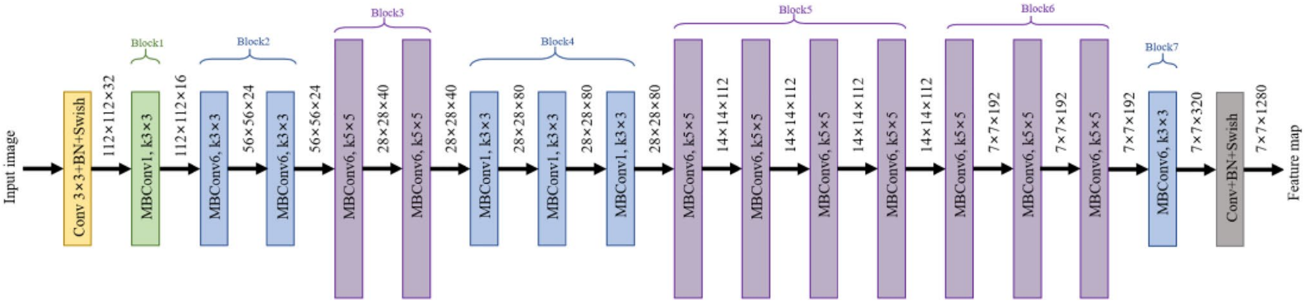

Fig. 2 The detailed architecture of EfficientNet-B0. EfficientNet-B0 consists of seven blocks which are shown in different colours. The basic building block of EfficientNet-B0 is a mobile inverted bottleneck convolution (MBConv), while each MBConv block is shown with the corresponding kernel filter size

Additionally, EfficientNet-B0 used a new activation function Swish [10], which is defined as:

$$f_{Swish} = \frac{1}{1 + e^{-\beta x}}, \tag{1}$$

where β is a parameter that can be learned during the training of the CNN.

In this paper, we utilize EfficientNet-B0 [10] for feature extraction since it provides a good balance between computational resources and accuracy. The multi-head attention layer is added between the pooling layer and the softmax layer of EfficientNet-B0.

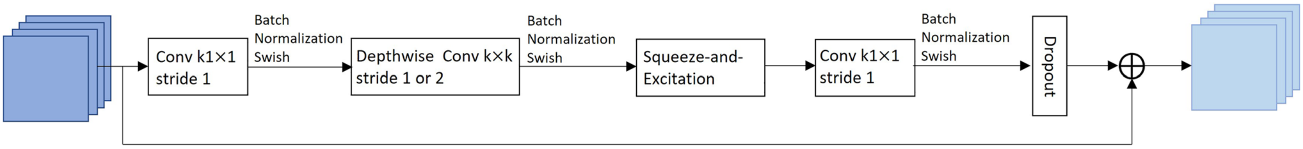

Fig. 3 The structure of the MBConv block

3.2 Multi-head self-attention mechanism

In this paper, we utilize a multi-head self-attention mechanism [30] to capture key information from videos. Figure 4 illustrates the structure of the multi-head self-attention mechanism, which processes the scaled dot-product attention mechanism [30] multiple times in parallel. The outputs of each scaled dot-product attention mechanism are concatenated. The dimension of the concatenated results is linearly transformed into the expected dimension, where *h* denotes the number of the scaled dot-product attention mechanism.

This multi-head self-attention mechanism strengthens a network to concentrate on the key information in different frames, which offers the network numerous “representation subspaces”. The self-attention mechanism can analyze the different influences of the different positions of the pixels and set different weights for the classification.

In this paper, the multi-head self-attention layer is added between the pooling layer and the softmax layer of EfficientNet-B0. We use an $L \times N$ matrix Y to represent a set containing L N-dimensional features. Y is the output of the pooling layer, and each row of Y is an independent feature vector y_i:

$$Y = \left(y_1, y_2, \ldots, y_L\right), \tag{2}$$

where Y is the input of the multi-head self-attention layer, which can be used to create three vectors: queries Q, keys K, and values V. These vectors can be regarded as abstractions for attention calculation. The output vector of the attention is a weighted sum of V, where the weight specified for each value is identified by the dot products of the query with all the keys, which can be defined as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{n}}\right)V, \tag{3}$$

where n denotes the dimension of K and V.

The multi-head self-attention linearly processes Q, K, and V multiple times via different weight matrices. Then the multi-head self-attention can be defined as:

$$\mathrm{MultiHead}(Q, K, V) = \left[\mathrm{head}_1, \ldots, \mathrm{head}_h\right]W^O, \tag{4}$$

$$\mathrm{head}_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{5}$$

where h denotes the total number of heads, and W^* denotes weight matrix. In our proposed network, we set $h = 2$. The dimension of the output of the attention layer is 512.

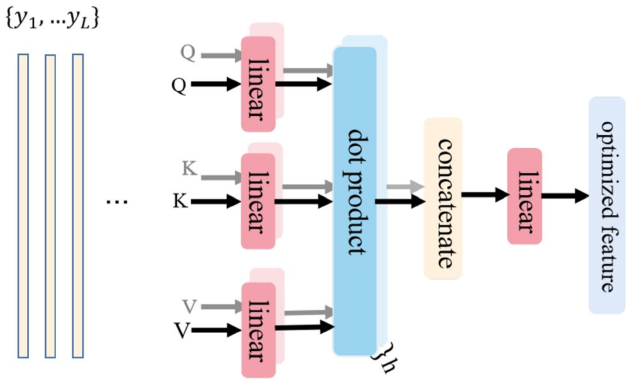

Fig. 4 The structure of the multi-head attention mechanism, which contains the *h* scaled dot-product attention mechanism

3.3 Our proposed MAT-EffNet

We propose a multi-head self-attention-based two-stream EfficientNet (MAT-EffNet) model for action recognition. The framework of MAT-EffNet is illustrated in Fig. 1. Similar to most two-stream-based networks, MAT-EffNet processes RGB video frames of the spatial/appearance stream. The temporal/motion stream aims to extract motion features from stacked optical flow frames.

Figure 5 illustrates the detailed structure of each stream in MAT-EffNet. In this work, considering the tradeoff between accuracies and complexities, EfficientNet-B0 is adopted as the backbone network to accomplish feature representations. In our proposed MAT-EffNet, parameters of the spatial stream and temporal stream are initialized by EfficientNet-B0 [10] which was pre-trained on a large-scale ImageNet dataset [43]. To present the framework of EffNet-B0, we use shorthand notations expressed as follows: Conv, MBConv1, MBConv6, where Conv is the first convolutional layer. MBConv1 and MBConv6 are convolutional layers with different sizes of the kernel and different numbers of blocks.

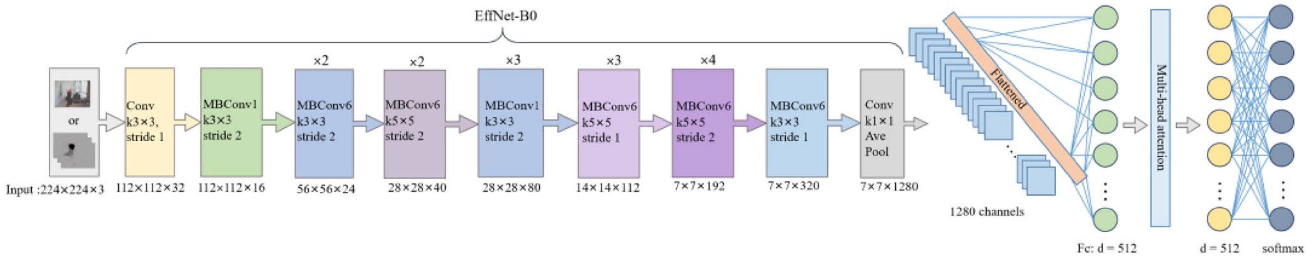

Fig. 5 The detailed structure of each stream in MAT-EffNet. The input is RGB video frames or stacked optical flow. A fully connected (Fc) layer (combined by green circles) is designed between the average pooling layer of EfficientNet-B0 and the multi-head attention layer. The output of each stream is a 512-dimensional vector

AvePool is the average pooling layer and Fc is the linear fully connected layer.

To fully explore the important spatial features and temporal features in videos, in our study, we adopt the multi-head self-attention mechanism to focus on the key information. The self-attention mechanism can determine where the important pixel's area is with large weights by computing how much a feature map corresponds to another. Thus, the network will focus on the area where the action happens and ignore the background or the irrelevant objects. This is also especially useful for fine-grained action recognition because of the subtle difference in the actions, and similar backgrounds such as sneezing and yawning.

As shown in Fig. 1, after the softmax layer of each stream, we adopt a late average fusion layer to obtain the final prediction by averaging the output scores of the softmax layer. Late average fusion fuses the spatial and temporal streams in the softmax prediction scores level, which is different from the data level in the early fusion method. The spatial stream and temporal stream are significantly varied in terms of dimensionality and sampling rate, adopting late fusion is a simpler and more flexible way than early fusion.

4 Experiments

In this section, we first introduce experimental datasets in Sect. 4.1. Then, the implementation details of the proposed MAT-EffNet are presented in Sect. 4.2. To verify the effectiveness of the proposed MAT-EffNet, we conduct ablation experiments in Sect. 4.3. The exploration of the proposed MAT-EffNet on the Kinetics dataset is presented in Sect. 4.4. Finally, we compare the proposed MAT-EffNet with several reference approaches in Sect. 4.5.

4.1 Datasets

We conduct experiments on three widely used action recognition datasets: UCF101 [35], HMDB51 [36] and Kinetics [51]. Three examples of different action classes selected from the UCF101, HMDB51 and Kinetics-400 datasets are illustrated in Fig. 6.

UCF101 dataset [35]: The UCF101 dataset is an expansion of the UCF50 dataset [45], which includes more than 13 K videos collected from YouTube with 101 action classes. The UCF101 dataset provides a multiplicity of actions collected from multi-angles such as object appearance, complex viewpoint, camera motion, cluttered background,

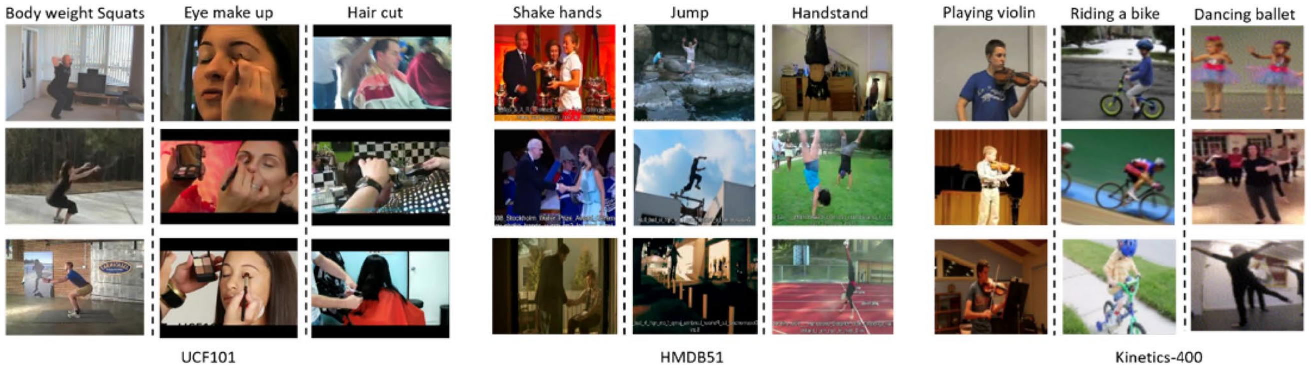

Fig. 6 Three examples of different action classes selected from the UCF101, HMDB51, and Kinetics-400 datasets

illumination circumstances, etc. The videos in each action class are sorted into 25 groups, and each group includes 4–7 videos. These action classes can be grouped into five categories: human–object interaction, human–human interaction, body-motion, sports and playing instruments. The UCF101 dataset is split into a training set containing about 9.5 K videos and a test set containing about 3.7 K videos.

HMDB51 dataset [36]: The HMDB51 dataset contains more than 6 K videos, most of which are collected from internet movies. The videos in the HMDB51 dataset are sorted into 51 action classes, most of which are daily actions. Each action class includes more than 101 videos. The action classes can be roughly divided into two categories: facial actions and body movements. All videos in the HMDB51 dataset are annotated with the action classes, video conditions, and meta information. The annotation contains the position of the body, the visible body, and the number of objects involved in the action. The HMDB51 dataset is split into a training set containing about 3.5 K videos and a test set containing about 1.5 K videos.

Kinetics-400 dataset [51]: The Kinetics-400 dataset is a large and well-labelled dataset, which has 400 action classes. The Kinetics-400 dataset contains 240 K training data, 40 K test data and 20 K validation data. Each class consists of more than 600 videos. The Kinetics-400 dataset includes human–object interaction actions such as riding a bike and typing as well as human–human interaction actions such as shaking hands and salsa dancing.

4.2 Implementation details

Feature extraction To capture efficient information, we utilize EfficientNet-B0 [10] to extract features from the RGB video frames and stacked optical flow frames. We pre-train the CNN models on the ImageNet dataset [43]. After initializing with the pre-trained model in the ImageNet dataset, we then use the mini-batch stochastic gradient descent algorithm to fine-tune the parameters in the proposed MAT-EffNet. Table 1 demonstrates the detailed architecture of our proposed MAT-EffNet.

RGB video frames are fed into the spatial stream and stacked optical flow frames are fed into the temporal stream. During training, all the input RGB video frames and optical flow frames are randomly cropped to 224×224 pixels with data augmentation. Our baseline networks consist of ResNet-18 [8], ResNet-34 [8], ResNet-50 [8] and EfficientNet-B0 [10], corresponding to the input resolutions 112×112, 168×168, 224×224 and 224×224, respectively. In this paper, the mini-batch size is set to 16, which is the maximum value allowed by hardware resources in all models. For both streams, the learning rate is set to 10^{-2} according to the literature [1]. The Swish activation function is adopted to our proposed MAT-EffNet. An average pooling is adopted to the pooling layer. The softmax function is used in the last layer and categorical cross-entropy is selected to be the loss function. The dropout ratio is set to 0.2 to ease the overfitting issue according to the literature [10].

Data augmentation We use data augmentation to solve the class imbalance problem [26]. We randomly reflect or flip the input frames horizontally with a 50% probability to increase the multiplicity of data.

Hardware and software The experiments are implemented in the Ubuntu16.04 Operation System. The training process of the MAT-EffNet is implemented on four NVIDIA GTX 1080Ti GPUs. Our proposed MAT-EffNet is implemented by Python.

4.3 Ablation experiments

Effectiveness of the multi-head attention mechanism: In two-stream network-based action recognition approaches, we test the performance of the two-stream network-based

Table 1 The detailed architecture of the proposed MAT-EffNet

Stage	Operator	Size	Number of layers
0	Input	224×224×3	–
1	Conv 3×3, stride 2	112×112×32	1
2	MBConv1, k3×3, stride 1	112×112×16	1
3	MBConv6, k3×3, stride 2	56×56×24	2
4	MBConv6, k5×5, stride 2	28×28×40	2
5	MBConv6, k3×3, stride 2	28×28×80	3
6	MBConv6, k5×5, stride 1	14×14×112	3
7	MBConv6, k5×5, stride 2	7×7×192	4
8	MBConv6, k3×3, stride 1	7×7×320	1
9	Conv 1×1, stride1, average pooling	1×1×1280	1
10	Multi-head attention layer	1×1×512	1
11	Fc & Softmax	512×{101 or 51}	–

Table 2 The recognition accuracy of two-stream network-based approaches for action recognition with (or without) multi-head attention mechanism on the UCF101 dataset

Training setting	Spatial stream (%)	Temporal stream (%)	Two-stream (%)
ResNet-18	76.2	79.1	81.9
ResNet-18 + Multi-head attention	78.1	81.2	83.9
ResNet-34	79.7	80.3	82.5
ResNet-34 + Multi-head attention	81.1	81.6	84.7
ResNet-50	82.5	85.6	88.5
ResNet-50 + Multi-head attention	85.1	88.7	91.7
EfficieNet-B0	87.6	89.1	91.8
EfficientNet-B0 + Multi-head attention (MAT-EffNet)	90.2	92.4	94.5

Table 3 The recognition accuracy of two-stream network-based action recognition approaches with (or without) multi-head attention mechanism on the HMDB51 dataset

Training setting	Spatial stream (%)	Temporal stream (%)	Two-stream (%)
ResNet-18	36.7	38.1	40.1
ResNet-18 + Multi-head attention	38.9	40.2	41.9
ResNet-34	37.6	39.1	43.1
ResNet-34 + Multi-head attention	40.9	44.2	49.9
ResNet-50	43.2	51.4	57.8
ResNet-50 + Multi-head attention	46.1	54.9	63.4
EfficieNet-B0	53.3	59.1	65.2
MAT-EffNet	59.3	65.3	70.9

action recognition approaches with (or without) the multi-head attention mechanism. Several CNNs are used in the two-stream network-based approaches, including ResNet-18 [7], ResNet-34 [8], ResNet-50 [8] and EfficientNet-B0 [10]. All models are pre-trained on the ImageNet [43]. Table 2 shows the recognition accuracies of these two-stream network-based approaches on the UCF101 dataset. Table 3 illustrates the recognition accuracies of these two-stream network-based approaches on the HMDB51 dataset. We compare the recognition accuracies of the spatial stream, the temporal stream, and the fused two-stream.

Tables 2 and 3 show the recognition accuracies of different two-stream network-based action recognition approaches. According to Table 2, approaches with the multi-head attention mechanism outperform the approaches without the multi-head attention mechanism. Specifically, for the UCF101 dataset, ResNet-18 with a multi-head attention mechanism approach performs better than the ResNet-18 without the multi-head attention mechanism approach (i.e., the spatial stream improves 2.2%, the temporal stream improves 1.3% and the two-stream improves 2.0%, respectively). ResNet-34 with the multi-head attention mechanism approach performs better than the ResNet-34 without the multi-head attention mechanism approach (i.e., the spatial stream improves 1.4%, the temporal stream improves 2.1% and the two-stream improves 2.2%, respectively). ResNet-50 with the multi-head attention mechanism approach performs better than the ResNet-50 without the multi-head attention mechanism approach (i.e., the spatial stream improves 2.6%, the temporal stream improves 3.1% and the two-stream improves 3.2%, respectively). The proposed MAT-EffNet performs better than EfficientNet-B0 without the multi-head attention mechanism (i.e., the spatial stream improves 2.6%, the temporal stream improves 3.3% and the two-stream improves 2.7%, respectively).

According to Table 3, in detail, for the HMDB51 dataset, ResNet-18 with the multi-head attention mechanism approach performs better than the ResNet-18 without the multi-head attention mechanism approach (i.e., the spatial stream improves 2.2%, the temporal stream improves 2.1% and the two-stream improves 1.8%, respectively). ResNet-34 with the multi-head attention mechanism approach performs better than the ResNet-34 without the multi-head attention mechanism approach (i.e., the spatial stream improves 3.3%, the temporal stream improves 5.1% and the two-stream improves 6.8%, respectively). ResNet-50 with the multi-head attention mechanism approach performs better than the ResNet-50 without the multi-head attention mechanism approach (i.e., the spatial stream improves 3.1%, the temporal stream improves 3.5% and the two-stream improves 5.6%, respectively). The proposed MAT-EffNet performs better than EfficientNet-B0 without the multi-head attention mechanism

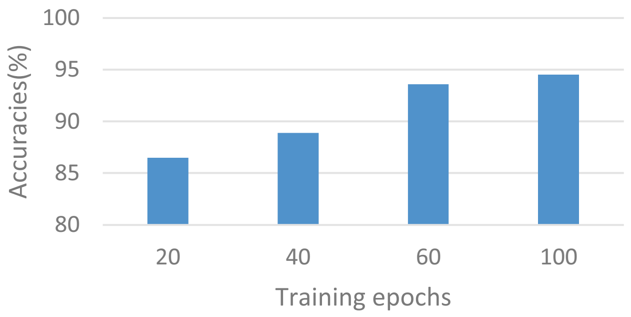

Fig. 7 Validation accuracies for different epochs on the UCF101 dataset

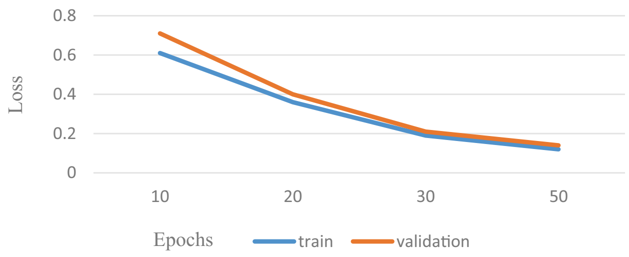

Fig. 8 The training loss of the proposed MAT-EffNet on the UCF101 dataset, the number of the epoch is 50

Table 4 Comparison of the baseline network on Kinetics-400 validation set

Methods	Top-1 accuracy (%)	Top-5 accuracy (%)
EfficientNet-B0 (baseline)	71	89.5
MAT-EffNet	72.6	90.8

Table 5 The accuracies of different action recognition approaches on the Kinetics-400 dataset

Approaches	Top-1 (%)	Top-5 (%)
Two-stream [12]	62.2	–
ConvNet + LSTM [12]	63.3	–
I3D-RGB [12]	72.1	90.3%
ARTNet [52]	70.7	89.3%
MoViNet-A5 [66]	71.7	–
VidTr-L [67]	70.2	89%
MAT-EffNet	72.6	90.8%

(i.e., the spatial stream improves 6.0%, the temporal stream improves 6.2% and the two-stream improves 5.7%, respectively).

In addition, the changes in the validation accuracy with different numbers of training epochs are shown in Fig. 7 and the training loss of the proposed MAT-EffNet on the UCF101 dataset is shown in Fig. 8.

4.4 Exploration of MAT-EffNet on the Kinetics-400 dataset

In this section, we compare our proposed MAT-EffNet to the baseline network with default settings. We use the EfficientNet-B0 [10] as the baseline network and train it on the Kinetics-400 training set from scratch, based on [35]. We use the same setup as in Sect. 4.2 when training from scratch. As shown in Table 4, our baseline network obtains 71% in top-1 accuracy and 89.5% in top-5 accuracy. To evaluate the effect of the proposed MAT-EffNet, we conduct experiments on the Kinetics-400 dataset and the experimental result outperforms the baseline network by 1.6% in top-1 accuracy and 1.3% in top-5 accuracy. The proposed MAT-EffNet obtains 72.6% in top-1 accuracy and 90.8% in top-5 accuracy.

As shown in Table 5, we compare our MAT-EffNet with several reference approaches. The proposed MAT-EffNet outperforms Two-stream [12] by 10.4%, outperforms ConvNet + LSTM [12] by 9.3%, outperforms ARTNet [52] by 1.9% and slightly outperforms I3D-RGB [12] by 0.5%, respectively. This indicates that the multi-head attention mechanism is useful for recognizing the action, and the proposed MAT-EffNet is a competitive network for action recognition.

Table 6 The accuracies of different action recognition approaches on the UCF101 dataset and HMDB51 datasets

Approaches	Input Modalities	UCF101 (%)	HMDB51
LRCN [38]	RGB + optical flow	82.9	–
C3D [26]	RGB only + 3D CNNs	85.2	–
IDTs [32]	RGB only + 3D CNNs	85.9	57.2%
Two-stream [1]	RGB + optical flow	88.0	59.4%
FSTCN [39]	RGB + optical flow	88.1	59.1%
P3D-199 [65]	RGB + 3D CNNs	89.2	62.9%
TDD [34]	RGB + optical flow	90.3	63.2%
STS-network [17]	RGB + optical flow + others	90.1	62.4%
R-M3D [11]	RGB only + 3D CNNs	93.2	65.4%
STDAN + RGB difference [58]	RGB + optical flow + others	91.0	60.4%
TSN Corrnet [55]	RGB + optical flow	94.4	70.6%
MSM-ResNets [56]	RGB + optical flow + others	93.5	66.7%
R-STAN-50 [68]	RGB + optical flow	91.5	62.8%
3D ResNeXt-101 + Confidence Distillation [69]	RGB + 3D CNNs	91.2	–
MAT-EffNet	RGB + optical flow	94.8	71.1%

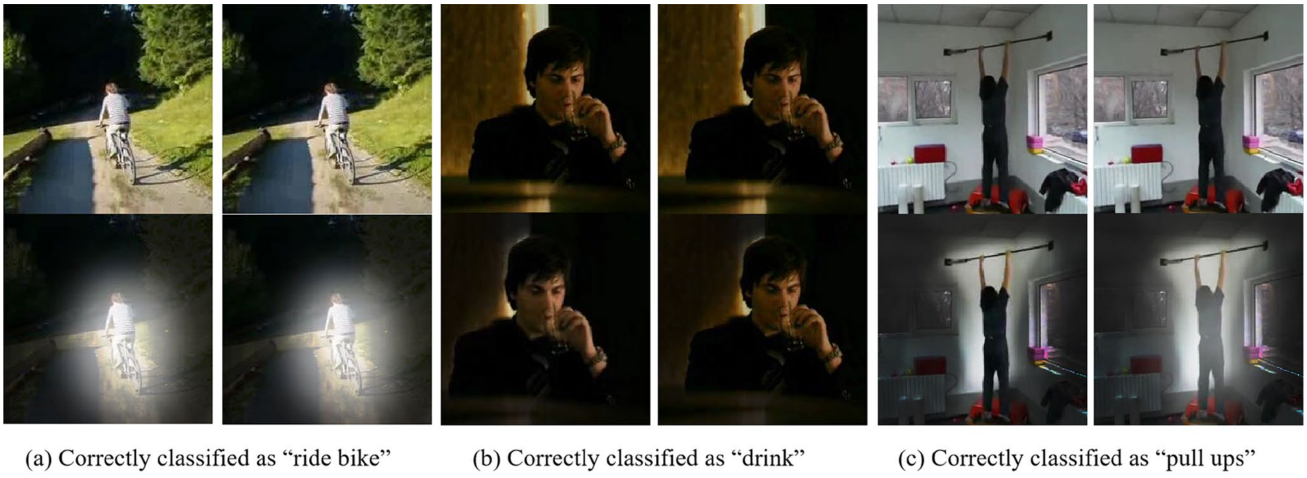
(a) Correctly classified as "ride bike" (b) Correctly classified as "drink" (c) Correctly classified as "pull ups"

Fig. 9 Visualization of multi-head self-attention attention of MAT-EffNet. The attention can focus on the informative regions

4.5 Exploration of MAT-EffNet on the UCF101 and HMDB51 datasets

The above experimental results have verified that the multi-head attention mechanism improves the recognition accuracy of two-stream network-based action recognition approaches. As shown in Table 6, we compare MAT-EffNet with several reference approaches on the UCF101 and HMDB51 datasets. We compare our approach with both conventional approaches and deep learning-based approaches such as long-term recurrent convolutional networks (LRCN) [37], 3D convolutional networks (C3D) [25], improved trajectories (iDTs) [31], two-stream neural network (Two-stream) [1], factorized spatio-temporal convolutional network (FSTCN) [38], trajectory-pooled deep-convolutional descriptors (TDD) [33], spatiotemporal saliency-based multi-stream network (STS-network) [24], multi-cue-based 3D residual network (R-M3D) [11], motion saliency-based multi-stream multiplier ResNets (MSM-ResNets) [56] and correlational convolutional LSTM network (TSN Cornet) [55].

According to Table 6, the proposed MAT-EffNet outperforms other action recognition approaches on both two datasets. Specifically, for the UCF101 dataset, MAT-EffNet improves 11.9% (vs LRCN), 9.6% (vs C3D), 8.9% (vs IDTs), 6.8% (vs Two-stream), 6.7% (vs FSTCN), 4.5% (vs TDD), 4.4% (vs STS-network), 1.6% (vs R-M3D), 3.8% (vs STDAN), 1.3% (vs MSM-ResNets) and 0.4% (vs TSN Cornet), respectively. For the HMDB51 dataset, MAT-ResNet improves 9.7% (vs IDTs), 7.5% (vs Two-stream), 7.8% (vs FSTCN), 3.7% (vs TDD), 4.5% (vs STS-network), 1.5% (vs R-M3D), 10.7% (vs STDAN), 0.5% (vs TSN Cornet), 0.4% (vs MSM-ResNet), 3.3% (vs R-STAN) and 3.6% (vs 3D ResNeXt-101), respectively. The proposed MAT-EffNet achieves 94.8% accuracy on the UCF101 dataset, and 71.1% accuracy on the HMDB51 dataset, respectively. By computing the attention multiple times, multi-head attention mechanism can improve the ability of the network to capture key information in videos. Also, the experimental results further demonstrate the effectiveness of multi-head attention used in the proposed MAT-EffNet.

To better demonstrate what multi-head self-attention mechanism has improved, we visualized some examples of self-attention weights on the validation data of HMDB51 in Fig. 9. We can observe that the self-attention attention mechanism of our MAT-EffNet can highlight representative action areas and ignore irrelevant objects and static background.

5 Conclusion

In this paper, we propose a Multi-head Attention-based Two-stream EfficientNet (MAT-EffNet) deep learning model for action recognition, which contains a spatial stream and a temporal stream. For each stream, we use EfficientNet-B0 to extract spatial features and temporal ssfeatures from videos, and then a multi-head attention mechanism is used to capture the key action information from the extracted features in videos. The final prediction is obtained via a late average fusion, which computes the softmax score of spatial and temporal streams for different classes. We test the proposed MAT-EffNet on three widely used action recognition datasets. Experimental results show that the MAT-EffNet outperforms several state-of-the-art approaches for action recognition.

For future work, we intend to develop novel attention mechanisms for a two-stream-based network to extract more discriminative spatial–temporal representations. We also intend to apply unsupervised learning into action recognition, which can make full use of the abundant unlabelled videos on the Internet.

Acknowledgements This work is supported by the Science and Technology Project of State Grid Corporation of China (Research on Intelligent Data Management and Feature Engineering Technology for Middle Platform, No. 5700-202058480A-0-0-00).

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

1. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199.
2. Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1933–1941).
3. Zheng, Z., An, G., Wu, D., Ruan, Q.: Spatial-temporal pyramid based convolutional neural network for action recognition. Neurocomputing **358**, 446–455 (2019)
4. Jing, C., Wei, P., Sun, H., Zheng, N.: Spatiotemporal neural networks for action recognition based on joint loss. Neural Comput. Appl. **32**(9), 4293–4302 (2020)
5. Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., Sebe, N.: Spatio-temporal attention networks for action recognition and detection. IEEE Trans. Multimed. **22**(11), 2990–3001 (2020)
6. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.
7. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).
9. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
10. Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105–6114). PMLR.
11. Zong, M., Wang, R., Chen, Z., Wang, M., Wang, X., Potgieter, J.: Multi-cue based 3D residual network for action recognition. Neural Comput. Appl. **33**(10), 5167–5181 (2021)
12. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).
13. 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. In European conference on computer vision (pp. 20–36). Springer, Cham.
14. Zhu, Y., Lan, Z., Newsam, S., & Hauptmann, A. (2018). Hidden two-stream convolutional networks for action recognition. In Asian conference on computer vision (pp. 363–378). Springer, Cham.
15. Girdhar, R., & Ramanan, D. (2017). Attentional pooling for action recognition. arXiv preprint arXiv:1711.01467.
16. Zheng, Z., An, G., Wu, D., Ruan, Q.: Global and local knowledge-aware attention network for action recognition. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 334–347 (2020)
17. Liu, Z., Li, Z., Wang, R., Zong, M., Ji, W.: Spatiotemporal saliency-based multi-stream networks with attention-aware LSTM for action recognition. Neural Comput. Appl. **32**(18), 14593–14602 (2020)
18. Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 244–253).
19. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
20. Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733.
21. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048–2057). PMLR.
22. Chen, Z., Wang, R., Zhang, Z., Wang, H., Xu, L.: Background–foreground interaction for moving object detection in dynamic scenes. Inf. Sci. **483**, 65–81 (2019)
23. Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., & Wen, S. (2018). Attention clusters: Purely attention based local feature integration for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7834–7843).
24. Ji, W., Wang, R.: A multi-instance multi-label dual learning approach for video captioning. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **17**(2s), 1–18 (2021)
25. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489–4497).
26. Song, L., Weng, L., Wang, L., Min, X., & Pan, C. (2018). Two-stream designed 2d/3d residual networks with lstms for action recognition in videos. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 808–812). IEEE.
27. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: a survey. Image Vis. Comput. **60**, 4–21 (2017)
28. Nayak, R., Pati, U.C., Das, S.K.: A comprehensive review on deep learning-based methods for video anomaly detection. Image Vis. Comput. **106**, 104078 (2021)
29. Du, W., Wang, Y., Qiao, Y.: Recurrent spatial-temporal attention network for action recognition in videos. IEEE Trans. Image Process. **27**(3), 1347–1360 (2017)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
31. Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In Proceedings of the IEEE international conference on computer vision (pp. 3551–3558).
32. Yu, Y., Gao, Y., Wang, H., Wang, R.: Joint user knowledge and matrix factorization for recommender systems. World Wide Web **21**(4), 1141–1163 (2018)
33. Wang, L., Qiao, Y., & Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4305–4314).
34. Wang, J., Peng, X., Qiao, Y.: Cascade multi-head attention networks for action recognition. Comput. Vis. Image Underst. **192**, 102898 (2020)
35. Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
36. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. In 2011 International conference on computer vision (pp. 2556–2563). IEEE.
37. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2625–2634).
38. Sun, L., Jia, K., Yeung, D. Y., & Shi, B. E. (2015). Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4597–4605).
39. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1510–1517 (2017)
40. Hu, H., Zhou, W., Li, X., Yan, N., & Li, H. (2020). MV2Flow: Learning motion representation for fast compressed video action recognition. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(3s), 1–19.
41. Tran, A., & Cheong, L. F. (2017). Two-stream flow-guided convolutional attention networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 3110–3119).
42. Peng, Y., Zhao, Y., Zhang, J.: Two-stream collaborative learning with spatial-temporal attention for video classification. IEEE Trans. Circuits Syst. Video Technol. **29**(3), 773–786 (2018)
43. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255).
44. Liu, L., Zhou, T., Long, G., Jiang, J., & Zhang, C. (2019). Learning to propagate for graph meta-learning. arXiv preprint arXiv:1909.05024.
45. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
46. Qiu, Y., Wang, R.: Adversarial latent representation learning for speech enhancement. Proc. Interspeech **2020**, 2662–2666 (2020)
47. Hou, F., Wang, R., He, J., & Zhou, Y. (2021). Improving entity linking through semantic reinforced entity embeddings. arXiv preprint arXiv:2106.08495.
48. Tian, Y., Zhang, Y., Zhou, D., Cheng, G., Chen, W.G., Wang, R.: Triple attention network for video segmentation. Neurocomputing **417**, 202–211 (2020)
49. Zheng, H., Wang, R., Ji, W., Zong, M., Wong, W.K., Lai, Z., Lv, H.: Discriminative deep multi-task learning for facial expression recognition. Inf. Sci. **533**, 60–71 (2020)
50. Shamsolmoali, P., Zareapoor, M., Wang, R., Zhou, H., Yang, J.: A novel deep structure u-net for sea-land segmentation in remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens **12**(9), 3219–3232 (2019)
51. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
52. Wang, L., Li, W., Li, W., & Van Gool, L. (2018). Appearance-and-relation networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1430–1439).
53. Ji, W., Wang, R., Tian, Y., & Wang, X. (2021). An attention based dual learning approach for video captioning. Applied Soft Computing, 108332.
54. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4510–4520).
55. Yudistira, N., Kurita, T.: Correlation net: spatiotemporal multimodal deep learning for action recognition. Signal Process. Image Commun. **82**, 115731 (2020)
56. Zong, M., Wang, R., Chen, X., Chen, Z., Gong, Y.: Motion saliency based multi-stream multiplier ResNets for action recognition. Image Vis. Comput. **107**, 104108 (2021)
57. Zhang, Zufan, et al. Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions. Neurocomputing 410 (2020): 304–316.
58. Meng, Quanling, et al. Action recognition using form and motion modalities. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16.1s (2020): 1–16.
59. Shamsolmoali, P., Zareapoor, M., Zhou, H., Wang, R., Yang, J.: Road segmentation for remote sensing images using adversarial spatial pyramid networks. IEEE Trans. Geosci. Remote Sens. **59**(6), 4673–4688 (2020)
60. Liu, M., Zhao, F., Jiang, X., Zhang, H., & Zhou, H. (2021). Parallel Binary Image Cryptosystem Via Spiking Neural Networks Variants. Int. J. Neural Syst., 2150014–2150014.
61. Wang, L., Yuan, X., Zong, M., Ma, Y., Ji, W., Liu, M., Wang, R.: Multi-cue based four-stream 3D ResNets for video-based action recognition. Inf. Sci. **575**, 654–665 (2021)
62. Liu, Y., Yuan, X., Jiang, X., Wang, P., Kou, J., Wang, H., Liu, M.: Dilated Adversarial U-Net Network for automatic gross tumor volume segmentation of nasopharyngeal carcinoma. Appl. Soft Comput. **111**, 107722 (2021)
63. Guo, J., Yi, P., Wang, R., Ye, Q., Zhao, C.: Feature selection for least squares projection twin support vector machine. Neurocomputing **144**, 174–183 (2014)
64. R. Wang, F. Hou, S. Cahan, L. Chen, X. Jia and W. Ji. (2022) Fine-Grained Entity Typing with a Type Taxonomy: a Systematic Review. IEEE Transactions on Knowledge and Data Engineering.
65. Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.
66. D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M.Tan, M. Brown and B. Gong, Movinets: Mobile video networks for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16020–16030.
67. Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli and J. Tighe, Vidtr: Video transformer without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13577–13587).
68. Liu, Q., Che, X., Bie, M.: R-STAN: Residual spatial-temporal attention network for action recognition. IEEE Access **7**, 82246–82255 (2019)
69. M. S. Shalmani, F. Chiang and R. Zheng, Efficient Action Recognition Using Confidence Distillation, 2021, arXiv preprint arXiv:2109.02137.