applied sciences

MDPI

Article

Human Action Recognition Based on Improved Two-Stream Convolution Network

Zhongwen Wang [1], Haozhu Lu [2], Junlan Jin [1] and Kai Hu [1,3,*]

[1] School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China; 201983240003@nuist.edu.cn (Z.W.); jjl0610@nuist.edu.cn (J.J.)
[2] Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool L69 7ZX, UK; sghlu8@liverpool.ac.uk
[3] Jiangsu Provincial Collaborative Innovation Center for Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing 210044, China
* Correspondence: 001600@nuist.edu.cn; Tel.: +86-137-7056-9871

Abstract: Two-stream convolution network (2SCN) is a classical method of action recognition. It is capable of extracting action information from two dimensions: spatial and temporal streams. However, the method of extracting motion features from a spatial stream is single-frame recognition, and there is still room for improvement in the perception ability of appearance coherence features. The classical two-stream convolution network structure is modified in this paper by utilizing the strong mining capabilities of the bidirectional gated recurrent unit (BiGRU) to allow the neural network to extract the appearance coherence features of actions. In addition, this paper introduces an attention mechanism (SimAM) based on neuroscience theory, which improves the accuracy and stability of neural networks. Experiments show that the method proposed in this paper (BS-2SCN, BiGRU-SimAM Two-stream convolution network) has high accuracy. The accuracy is improved by 2.6% on the UCF101 data set and 11.7% on the HMDB51 data set.

Keywords: action recognition; two-stream convolution network; BiGRU network; SimAM attention mechanism

Citation: Wang, Z.; Lu, H.; Jun, J.; Hu, K. Human Action Recognition Based on Improved Two-Stream Convolution Network. *Appl. Sci.* **2022**, *12*, 5784. https://doi.org/10.3390/app12125784

Academic Editor: Andrea Prati

Received: 25 April 2022
Accepted: 2 June 2022
Published: 7 June 2022

1. Introduction

In the era of big data, more and more videos are shared. How to quickly extract information from massive video resources has high research and application value. Human action recognition in videos has gradually become a major research hotspot in the field of computer vision, and has been widely used in public video surveillance, scientific cognition, medical rehabilitation and other fields.It has broad application prospects in the fields of human-computer interaction [1], robot teleoperation [2], etc.

The method of human action recognition has experienced the development from manual feature extraction to deep learning feature extraction. Compared with manual feature extraction, deep learning feature extraction puts forward higher requirements for the computing power of the computer. Therefore, for a long time, researchers carry out action recognition by manually extracting features.

In the early stages, people recognized actions through the geometry of the human body, that is, the action recognition method based on a template. Bobick, A. et al. [3] first proposed using motion energy image (MEI) for action recognition. Then Weinland, D. et al. [4] proposed using motion history image (MHI) for action recognition. Afterwards, people developed the feature representation method from global feature representation to local feature representation. Laptev, I. et al. [5] used optical flow histogram for local feature representation. Wang et al. [6,7] successively proposed DT and IDT algorithms, which further developed the local feature representation into an action recognition method based on joint and skeleton trajectory. Hu, K. et al. [8] proposed a multi-scale skeleton action recognition method, so that the skeleton motion model has deep spatio-temporal features.

Yang, X. et al. [9] proposed a skeleton action recognition method based on 3D depth data, which achieved good results in online action recognition scenes.

However, the manual feature extraction method is sensitive to the direction, position, and background environment of human action, and needs an accurate action template as support. It has high accuracy only for simple actions in limited scenes. Deep learning can automatically extract features when the model is uncertain, and has been widely used in object detection [10], automatic control [11], remote sensing image processing [12], system parameter identification [13] and other fields. In recent years, with the continuous improvement of computer computing power, deep learning has been deeply developed, and human action recognition algorithms based on deep learning have gradually emerged.

At present, the more commonly used deep learning network is the classical convolutional neural network (CNN). CNNs are widely used in many image semantic segmentation tasks [14]. The derivative networks based on CNN include AlexNet [15], VGG network [16], GoogleNet [17], ResNet [18] etc. These networks can extract features from a single image and classify them and have achieved good results on ImageNet. The action recognition algorithms based on these neural networks are mainly divided into three categories: (1) Single-stream network model. In this network model, the three-dimensional convolutional neural network proposed by Ji, S. et al. [19], namely 3D CNN is widely used. Its convolution network structure includes a hard connection layer, three 3D convolution layers, two pooling layers, and a full connection layer. (2) Two-Steam network model. Inspired by the two-stream hypothesis of neuroscience, simonyan et al. [20] creatively proposed the two-stream network model for the first time in 2014. (3) Multi-Stream network model. This network model is an extension of the Two-Steam network model. Based on the Two-Steam network model, other deep learning networks are added to extract different features.

Among the three kinds of action recognition algorithms, the Two-Steam network model method has good generalization and expansibility. Therefore, people have proposed many improved models based on this model, such as adding the attention mechanism module to the Two-Steam network model to form the structure of the "Two-Steam network model + attention mechanism". The existing attention mechanisms include circular attention mechanism [21], global attention mechanism [22], SE attention mechanism [23], CBAM attention mechanism [24], DA attention mechanism [25], SAM attention mechanism [26], multi-head attention mechanism [27], etc.

In early studies, people proposed classical recurrent neural network (RNN) [28] and long short-term memory network (LSTM) [29]. Afterwards, Cho, K. et al. [30] proposed a gated recurrent unit (GRU). Abhisek et al. [31] proposed Bidirectional Gate Recurrent Unit (BiGRU) based on GRU for supervised and language-independent context-sensitive lemmatization. This composite deep neural network structure exhibits an outstanding ability to capture contextual information for a given word. These neural networks with the ability of memory can well complete the tasks of emotion analysis and named entity disambiguation in the field of NLP combined with context. Similar to lemmatization being context-sensitive, an action is also sensitive to its state before and after it, which is a higher-dimensional "context". We intend to introduce BiGRU into the field of action recognition to improve the neural network's ability to perceive pre- and post-action states (the coherence feature of action appearance). In terms of attention mechanism, Yang, L. et al. [32] proposed the attention mechanism SimAM based on neuroscience theory. Compared with the two-dimensional attention weights provided by the existing channel attention and spatial attention modules, SimAm provides 3D attention weights for the feature maps in the network, which is consistent with the feature dimension of action recognition tasks. At the same time, SimAm can effectively improve the accuracy of the neural network. Another advantage of SimAm attention is that most of the operators are chosen according to the solution of the defined energy function, thus avoiding excessive structural tuning work [32].

However, many action recognition algorithms still face some problems:

(1) The main task of a spatial stream network in the classical Two-Stream network model is to extract the appearance features of actions, and its feature extraction method is only single frame recognition (still frame). However, the appearance features of actions may have great differences in different stages. The way of single frame recognition will make the neural network unable to learn these coherence features of appearance.
(2) At present, there is still room for improvement in the accuracy and stability of the classical two-stream network for action recognition. In addition, the classical Two-Stream network treats each pixel equally, which will lead the network to extract features weakly related to the action recognition task, such as video background. Therefore, a neural network needs a method to filter irrelevant information.

In this paper, we construct a new network structure (BS-2SCN) based on a two-stream convolution network to solve the two problems mentioned above. To solve these two problems, we add a bidirectional gated unit to the spatial stream network model of a two-stream convolution network to form a BiGRU network. For a video clip, we no longer only recognize a single frame, but use the method of uniform sampling to input multiple frames of images into the BiGRU network in order, and use its memory ability to improve the perception ability of a neural network to the coherence characteristics of action appearance. Due to the BiGRU network, the neural network can make use of the past and future information at the same time. In addition, we also inserted the SimAM attention mechanism into the ResNet unit, which further improves the accuracy of action recognition.

To sum up, this paper has made the following contributions:

(1) Propose a new network structure based on the two-stream convolution network model and combined with a bidirectional gated recurrent unit. This network structure can well solve the shortcomings of the original neural network model in the perception of motion appearance coherence features.
(2) Combine the SimAM attention mechanism, which is based on the spatial inhibition effect of neurons, with the ResNet network organically to improve the accuracy and stability of action recognition.

The overall framework of this paper is as follows: Section 1 introduces the research status in the field of action recognition; Section 2 briefly introduces the traditional two-stream convolution network, the attention mechanism, and the neural network with memory capability; In Section 3, the improved two-stream convolution network structure is introduced in detail; In the Section 4, ablation experiments and comparative experiments on two representative public data sets, UCF101 and HMDB51, verify the effectiveness and stability of the network; Section 5 is the conclusion of this paper.

2. Related Works

2.1. Two-Stream Convolution Network

The two-stream hypothesis assumes that the visual cortex has ventral and dorsal pathways. The ventral pathway is sensitive to the shape and colour of the target, and the dorsal pathway is sensitive to the spatial transformation caused by the movement of the target. The two-stream network architecture imitates the visual cortex to establish the temporal information path and spatial information path. The independent parallel CNN network is used to extract the temporal and spatial features of the video. Finally, fusing the features. The structure of the two-stream convolution network is shown in Figure 1. After sampling, scaling, and clipping, the input video will be input into the spatial stream convolution network and temporal stream convolution network respectively. The input of the spatial stream network is a video single frame image (after RGB three-channel decomposition), and the input of the temporal stream network is a stacked optical stream image. The dense optical stream can be regarded as a set of displacement vector fields

between continuous frames t and $t+1$. Note the point (u, v) in frame t, and the optical stream in frame t is I_t, The calculation formula is as follows:

$$I_t(u, v, 2k-1) = d^x_{t+k-1}(u, v) \quad (1)$$

$$I_t(u, v, 2k) = d^y_{t+k-1}(u, v) \quad (2)$$

In the above formula, $u = [1; w], v = [1; h], k = [1; L]$.

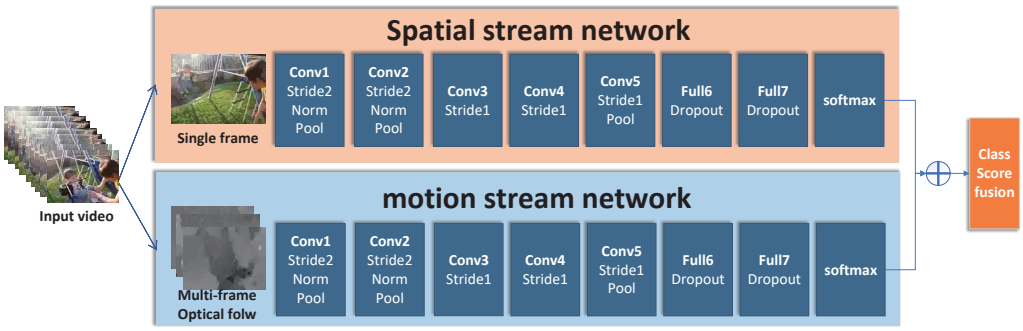


Figure 1. Classical two-stream convolution network structure.

The two-stream convolution network adopts the VGG16 network. After 5-layer convolution, 3-layer pooling, and 2-layer full connection, the input image is input into the SoftMax classifier to complete the classification. Finally, the two-stream network is fused by the weighted addition or support vector machine (SVM) method to obtain the final classification result. In this paper, we use a two-stream convolution network as the framework and the effective ResNet network to replace the VGG16 network at the same time.

2.2. Attention Mechanism

Attention mechanism is a useful tool applied in the field of deep learning in recent years, which originated from the human visual attention mechanism. Allport [33] proposed the concept of visual-spatial attention. Humans scan the global image to obtain target areas that need to be focused on, and then invest more attention resources in these areas while ignoring other unimportant information. Through this visual attention mechanism, the limited attention can be used to quickly filter out high-value information from a large amount of information. The attention mechanism in neural networks is a resource allocation scheme that allocates computing resources to more important tasks and solves the problem of information overload in the case of limited computing power. The problem of information overload refers to the use of too many parameters to improve the expression ability of the model, resulting in an excessive amount of information stored in the model. By introducing an attention mechanism, the neural network can pay attention to more important information among many input information, while reducing the attention to other information, and even filtering irrelevant information. The neural network can solve the problem of information overload and improve the processing efficiency and accuracy of the neural network.

In 2014, Bahdanau, D. et al. [21] first introduced the attention mechanism into NLP, built a neural machine translation model, and achieved excellent results. Subsequently, Luong, M.T. et al. [22] improved the circular attention model, proposed the concept of the global attention mechanism, and expanded the calculation method of the attention mechanism. In 2017, Vaswani, A. et al. [27] and others proposed a transformer architecture based entirely on attention mechanisms and achieved standout results in NMT and other tasks.

In 2018, Hu, J. et al. [23] proposed the SE attention mechanism, which has achieved excellent performance on the existing CNN model. Cheng, X. et al. [34] proposed a new network structure based on the SE attention mechanism, which improved the performance of deep architecture. Jin, X. et al. [35] improved the SE attention mechanism and proposed a

new spatial pooling method. In terms of application, SE attention is widely used in phonetic recognition [36], aquaculture [37], medical lesion identification [38], remote sensing imagery recognition [39], communication [40] and other fields.

Based on SE attention mechanism, Woo, S. et al. [24] proposed CBAM attention mechanism, which can capture more important information between indistinguishable objects [41]. Hou, Q. et al. [42] combined SE attention with CBAM attention to improve the performance. Many scholars combined the CBAM attention mechanism with neural network and gave birth to convolutional attention residual network (CARNET) [43], CBAM confrontation network (CBAM-GAN) [44], etc. In terms of application, the CBAM attention of Wang, S.H. et al. [45] is combined with the classical VGG network to identify COVID-19. In addition, CBAM attention has been well applied in face detect [46], facial expression recognition [47], ocean target detection [48], agriculture [49], vessel detection [50].

Based on the research on the human brain's attention [51], Yang, L. et al. [33] proposed a parameterless attention mechanism, SimAM. This is an attention mechanism with full three-dimensional weights, and the weights are calculated by an energy function. SimAM attention has been widely used in complex tasks due to its simplicity, non-participation, and plug and play characteristics, such as pathologic portrait recognition [52,53], extreme-exposure image fusion [54]. In this paper, we intend to use SimAM attention as a powerful method to improve the accuracy of action recognition.

2.3. Neural Network With Memory

Neural networks with memory are widely used in NLP. Tasks such as text sentiment analysis, machine translation, and named entity disambiguation often need to be analyzed in combination with the language environment (the context of the object to be analyzed). In 1991, Elman, J.L. et al. [28] proposed the classical recurrent neural network (RNN), which is characterized by having a memory module to store the output of the previous time and use it as the network input together with the input of the next time. However, the problems of gradient disappearance and gradient explosion of RNN limit its application. The long short-term memory network (LSTM) proposed by Hochreiter, S. et al. [29] in 1997 solves this problem well. LSTM is still widely used by researchers. Recently, Hu, K. et al. [55] proposed an enhanced input differential feature based on LSTM—Spatio-Temporal Differential Long Short-Term Memory (ST-D LSTM), and added a differential link to the LSTM network to effectively extract the dynamic characteristics of the target. In 2014, Cho, K. et al. [30] proposed a gated recurrent unit (GRU), which simplifies the structure and improves the speed of calculation with the same effect as LSTM. In 2017, Abhisek, et al. [31] proposed a BiGRU network with a strong ability to mine context features. Its structure is shown in Figure 2.

In Figure 2, given an input word w, the goal is to obtain a high-dimensional vector representing the syntactic structure of w. w is represented as a sequence of characters $c_1, c_2 \ldots c_m$, where m is the word length. Each character c_i is defined as a one-hot encoding vector 1_{c_i}, and each one-hot encoding vector obtains its corresponding projection vector e_{c_i} through the embedding layer. For a series of projection vectors $e_{c_1}, e_{c_2} \ldots e_{c_m}$, the state sequences h_m^f and h_1^b are generated by the forward GRU unit and the reverse GRU unit. f represents the forward sequence, and b represents the reverse sequence. Finally, the two state sequences are connected to obtain a high-dimensional vector e_w^{syn} representing the syntactic structure of w. At this time, e_w^{syn} has the context feature of w.

In this paper, we intend to use a bidirectional gated recurrent unit (BiGRU) in a spatial stream network. Compared with a unidirectional gated recurrent unit, BiGRU can remember not only the past action characteristics but also the future action characteristics, which will give our neural network higher accuracy.

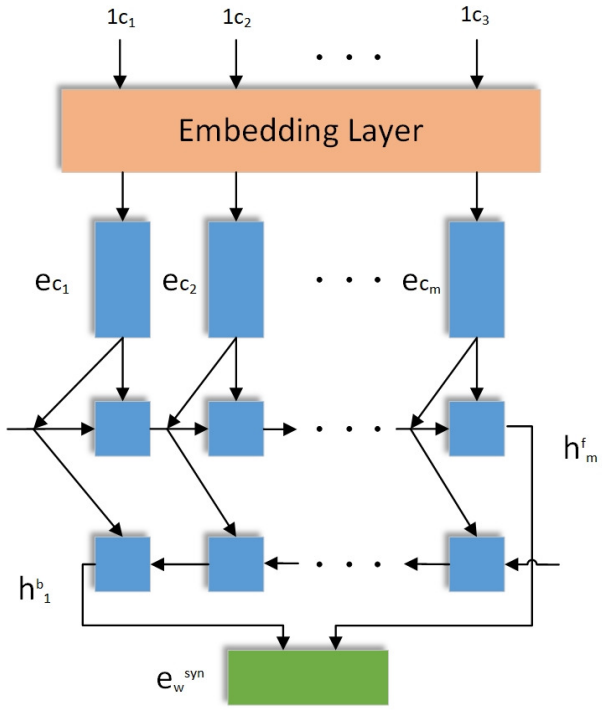


Figure 2. BiGRU network structure for mining contextual features.

3. Improved Network Structure

3.1. General Network Structure

Due to the shortcomings of the classical two-stream convolution network in obtaining the coherence features of action appearance, an improved two-stream convolution network is proposed in this paper. The original spatial stream input changes from a single video frame to a multi-video frame sampled at equal intervals. In order to improve the performance of the network, the network used in the feature extraction process is replaced by a VGG16 network with deeper ResNet. The network used in the motion stream has the same change. Moreover, this paper adds an advanced SimAM attention mechanism to the ResNet network, which greatly improves its feature extraction ability. The overall structure of the improved two-stream convolution network proposed in this paper is shown in Figure 3.

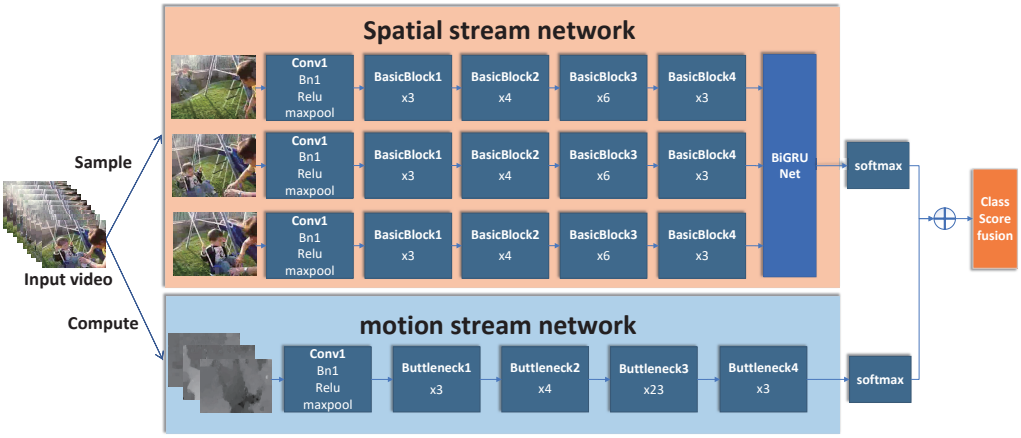


Figure 3. Improved two-stream convolution network structure.

To make the network work better, for an input video, first carry out equal spacing sampling, optical flow calculation, and other preprocessing work. The calculation formula of optical flow has been given by Formulas (1) and (2).

The number of spatial stream input frames can be adjusted according to the actual situation, but the input images need to be arranged in strict order and equidistant order. Figure 3 shows the network structure and the process of forwarding propagation by taking the input three frames as an example. These three frames are sampled at equal intervals from the input video. Therefore, for a video, the three frames represent the appearance of the action at different times in a continuous action.A frame of images first passes through a convolution layer, a normalization layer, an activation layer, and a max-pooling layer. Then enter a continuous number of ResNet network basic block layers (BasicBlock). Each of the three frames goes through the same network.The feature map generated by three consecutive frames of images passing through the above network is input into the BiGRU network in sequence, and finally, the recognition result is obtained through the SoftMax layer.

The input of the motion stream is the stacked optical flow image. After passing through the same front-end network as the spatial stream, the optical flow image enters multiple continuous basic unit layers of ResNet (Buttleneck), and finally enters the SoftMax layer to obtain the recognition result. The motion stream does not use the bidirectional Gru network.

The recognition results of spatial stream and motion stream are fused by the classical weighted fusion method to obtain the final action recognition result.

3.2. Network Structure of BasicBlock & Buttleneck

With the increase in the difficulty of classification tasks and the complexity of classification scenes, people must increase the depth of neural networks to adapt to the change. The traditional AlexNet, VGG, GoogleNet, and other networks have different degrees of degradation after increasing the network depth. With the increase in depth, the recognition accuracy does not increase but decreases. Kaiming He et al. proposed a powerful ResNet in 2015, which effectively solved the problem of network degradation and can extract deep image information [56]. ResNet is composed of many residual blocks, and the basic residual block structure is shown in Figure 4.

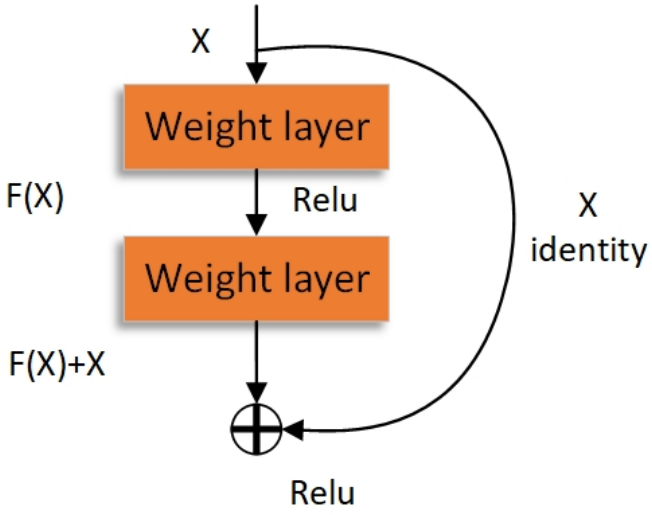


Figure 4. Structure of basic residual block.

The input x is superimposed with the original input after being activated by two weight layers. The output retains a certain original gradient and characteristic information to make the network easier to optimize and provide a guarantee for the increase of the number of network layers. The network model proposed in this paper uses ResNet instead of VGG to achieve better performance. Figure 5 shows the structure of BasicBlock and Buttleneck network.

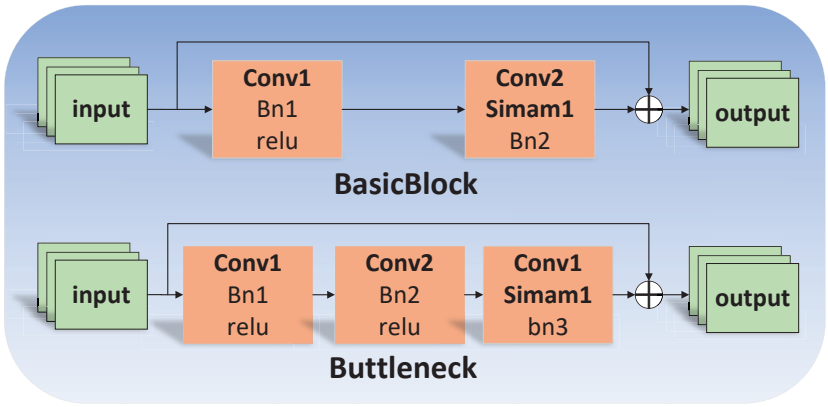


Figure 5. Structure of BasicBlock and Buttleneck.

The input feature maps will pass through the convolution layer, normalization layer, and activation layer in BasicBlock in turn. Buttleneck thickens the convolution layer, normalization layer, and activation layer compared with BasicBlock. In this paper, the SimAM layer is added between the last convolution layer and the normalization layer. SimAM is a nonparametric attention mechanism of convolutional neural network proposed by Yang, L. et al. [32] based on the spatial inhibition effect of neurons. It studies the importance of each neuron by calculating the energy function of each neuron, forming the attention of the corresponding neuron, and using the energy function en_r indicates. The calculation formula is as follows:

$$en_r(we_r, b_r, y, q_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(we_r q_i + b_r))^2 + (1-(we_r r + b_r))^2 + \lambda we_r^2 \quad (3)$$

where r represents the target neuron in a single input channel; q_i represents other neurons in the input channel, and i is the serial number; we_r and b_r is the linear conversion of weight and offset; μ_r is the average value; σ_r^2 indicates variance; M is the number of other neurons on the channel; y is the variable; λ Is the coefficient. The calculation formula of $we_r, b_r, \mu_r, \sigma_r^2$ is:

$$we_r = -\frac{2(r-\mu_r)}{(r-\mu_r)^2 + 2\sigma_r^2 + 2\lambda} \quad (4)$$

$$b_r = -\frac{1}{2}(r+\mu_r)we_r \quad (5)$$

$$\mu_r = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \quad (6)$$

$$\sigma_r^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i-\mu_r)^2 \quad (7)$$

In neuroinformatics, Webb et al. [57] found that the most informative neurons are usually those that display different firing patterns from other surrounding neurons. On this basis, Yang, L. et al. [32] found that the easiest way to find these neurons is to measure the linear separability between one target neuron and other neurons, find the target neuron and all other neurons in the same channel The linear separability of neurons is equivalent to minimizing Formula (3), so the smaller the energy of each neuron, the more important that neuron is compared to other neurons in the same channel. Yang, L. et al. [32] marked the minimum neuron energy as en_r^*, and used the reciprocal $1/en_r^*$ of the minimum neuron energy to represent the weight of the neuron. Calculate the minimum neuron energy en_r^* using the following formula:

$$en_r^* = \frac{4(\sigma_r^2 + \lambda)}{(r-\mu_r)^2 + 2\sigma_r^2 + 2\lambda} \quad (8)$$

The energy of all neurons of a single channel constitutes the energy matrix e of the channel. The attention weight matrix E' of the channel is obtained after the reciprocal of each element in the energy matrix E is normalized by the sigmod function. The calculation formula is:

$$E' = sigmod(\frac{1}{E}) \tag{9}$$

Finally, the feature map is fused with the attention weight of the channel to calculate the fused feature map S_i' (S_i is the original feature map), and its calculation formula is:

$$S_i' = S_i \cdot E' \tag{10}$$

The feature map processed by a BasicBlock or Buttleneck will be superimposed with the original feature map as the final output of the network.

To enable the spatial stream network to learn the appearance coherence characteristics of the action, the input image is input into the BiGRU network for processing after passing through the ResNet. The network structure of BiGRU is shown in Figure 6.

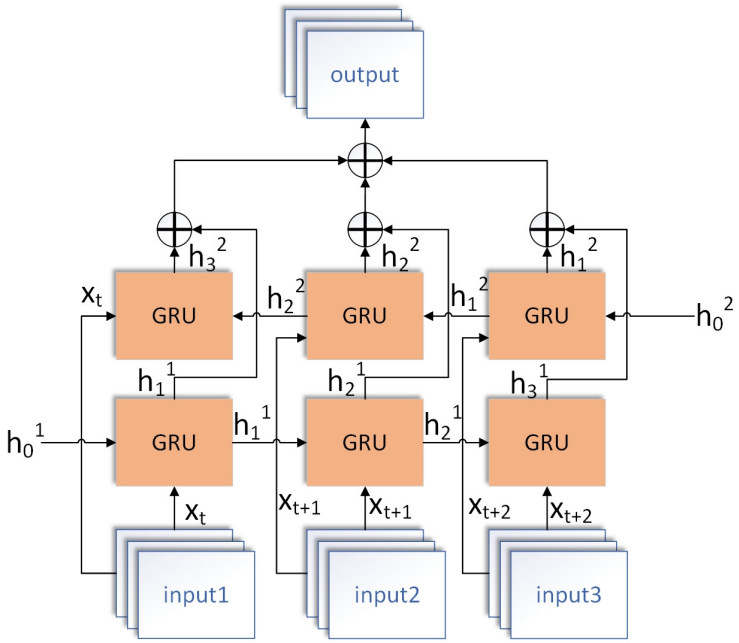


Figure 6. Network structure of BiGRU.

The input feature maps are input into the BiGRU network in strict order. The same group of feature maps is not only the input of sequential GRU but also the input of reverse GRU. GRU evolved from LSTM, which simplifies the input gate and forgetting gate in LSTM into an update gate. Its basic structure is shown in Figure 7.

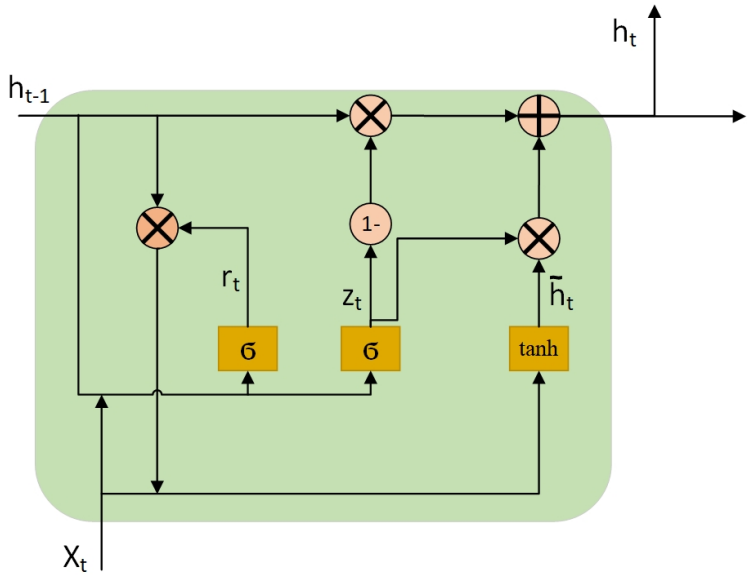


Figure 7. Structure of GRU.

3.3. Network Structure of BiGRU

The updated formula of GRU is as follows:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{11}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{12}$$

$$\tilde{h_t} = tanh(W \cdot [r_t \times h_{t-1}, x_t]) \tag{13}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h_t} \tag{14}$$

The above formulas can be expressed by $h_t = GRU(x_t, h_{t-1})$, where r_t represents the reset gate at time t, z_t represents the update gate at time t, $\tilde{h_t}$ represents the state of candidate activation at time t. The tilde represents the output of the tanh function, h_t indicates the state of activation at time t, h_{t-1} represents the state of hidden layer at time $(t-1)$, W_r, W_z, W are the weight matrixes. GRU network can reduce the amount of computation and training difficulty under the effect equivalent to that of an LSTM network, but its memory ability is unidirectional. Inspired by GRU, our network uses the BiGRU network to realize the bidirectional memory and enhance its memory ability.

BiGRU network is composed of two unidirectional GRU network in opposite directions. The hidden layer state of BiGRU network at time t can be obtained by weighted summation of forward hidden layer state $\overrightarrow{h_{t-1}}$ and reverse hidden layer state $\overleftarrow{h_{t-1}}$ with the calculation formula as follows:

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \tag{15}$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t-1}}) \tag{16}$$

$$h_t = w_t \overrightarrow{h_t} + v_t \overleftarrow{h_t} + b_t \tag{17}$$

where w_t, v_t are weight matrixes, b_t is offset. The right arrow indicates the output of the sequential GRU unit, which is specifically expressed as the output of the lower-layer GRU unit h_1^1, h_2^1, h_3^1 in Figure 5; the left arrow indicates the output of the reverse-order GRU unit, which is specifically expressed as the upper-layer GRU in Figure 5. The output of the unit h_3^2, h_2^2, h_1^2 sums the three groups of inputs of the BiGRU network to obtain the output of the whole BiGRU network.

4. Experiments

To demonstrate the performance of the neural network proposed in this paper, four groups of experimental subjects were designed for ablation experiments: basic control group (2SCN), SimAM attention mechanism group (S-2SCN), BiGRU network group (B-2SCN) and neural network group (BS-2SCN) proposed in this paper. Ten independent repeated experiments were carried out respectively. The independently repeated experiments are divided into two parts, which are run on the classic UCF101 data set and the HMDB51 data set respectively. The four groups of experimental subjects on the same data set have the same experimental conditions except for the network structure, and each group of experimental subjects carried out 10 experiments respectively. Between these two different data sets, we optimized the parameters of the neural network to fit the different datasets. Subsequently, we compare the highest accuracy of the BS-2SCN network with other action recognition algorithms on both datasets.

This paper will analyze the performance of the proposed neural network from the aspects of recognition accuracy, learning rate and loss value, and the stability of recognition results. This experiment runs on the classic UCF101 data set and HMDB51 data set.

4.1. Data Set

In the field of action recognition, there are many data sets that are used. These include UCF101 and UCF50, HMDB, KTH, Hollywood, Hollywood 2, Kinetics and more.

KTH data set [58], released in 2004, is one of the earliest published action recognition data sets. The data set includes 6 types of actions in 4 scenes, which are completed by 25 people, and the total number of videos is 2391. This data set is a milestone in the field of computer vision and has been widely used in the field of traditional action recognition by manually extracting features. However, due to its single background, few scenes, fixed perspective, and other restrictive conditions, it has not been widely used in recent years.

The Hollywood [59] and Hollywood2 [60] data sets are successively released by the IRISA Research Institute of France in 2008 and 2009. As its name suggests, the Hollywood dataset comes from the films and television works produced by Hollywood. The Hollywood2 dataset contains 12 action categories in 10 scenes, and the number of videos is 3669 in total. The feature of the data set of film and television works is that people's actions, expressions, and gestures are relatively rich, and the background factors, such as the movement of viewing angle and lighting conditions are full of changes. Therefore, it is very challenging for the action recognition algorithm. However, there are still a few action categories.

The HMDB data set [61] is a data set released by Brown University in 2011. The data set contains a rich number of scenes, including some from many public data sets and some from online video databases, such as YouTube. The data set contains 51 action categories, with a total of 6849 videos. Compared with the KTH and Hollywood data sets, the HMDB data set has greatly improved the number of video types and includes network video data. Its complexity and diversity lay a foundation for improving the generalization performance of action recognition algorithms.

The UCF50 data set [62] and UCF101 data set [63] were published by Florida Central University in 2012 and 2013 respectively. The UCF101 data set contains 101 action categories, and the number of videos has reached 13,320. The data set widely contains video data from radio and television channels and the Internet. It is one of the data sets with the largest number of action categories at present. Among them, the amount of equal scale data and high video quality provide help for the application of action recognition algorithms with deep learning as a means of feature extraction.

The Kinetics data set [64] was released in 2017. The dataset has multiple series sets, such as Kinetics400, Kinetics600, and Kinetics700, including 400, 600, and 700 action categories respectively. The number of videos has reached 500,000. This almost increases the number of action categories in multiple existing datasets by one order of magnitude. The data set is born for deep learning, and its massive video data is conducive to the

training of the neural network. Unfortunately, not all devices can make full use of this data set, and the high hardware conditions limit the wide use of this data set.

Considering the hardware conditions of experimental equipment and neural network structure, the UCF101 data set and HMDB51 data set are used in our experiments.

4.2. Details of Experiments

The main network structure is shown in Figure 3. Before the image is input to the neural network, a series of preprocessing needs to be taken to enhance the stability of the network. The preprocessing method used in our experiments is the transform method under the PyTorch framework. Firstly, the input images are randomly cropped and unified in the format, which is randomly cropped to the size of 224 × 224. Then, the images are randomly flipped horizontally with a 50% probability. Then it is converted to tensor data type; Finally, standardized image data. The process of standardization is to subtract the mean value and then divide it by the standard deviation. After the preprocessing, the images will be decomposed into three channels: red, green, and blue, which will be input into the neural network in the form of a high-dimensional array.

In terms of the loss function, This paper uses the cross-entropy loss function that can effectively avoid gradient dispersion [65], whose calculation formula is:

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic} log(p_{ic}) \tag{18}$$

where M represents the number of action categories, here is 101; y_{ic} is a symbolic function, whose value is 0 or 1. If the value of sample i is equal to c, take 1, otherwise, take 0; p_{ic} represents the prediction probability that sample i belongs to category c.

This paper employs the cross-entropy loss function, which is characterized by amplifying the loss when the model effect is poor, so as to speed up the learning speed. When the model effect is good, reduce the loss, slow down the learning speed, make the model converge quickly and improve the accuracy accurately.

In terms of the optimization method, this paper adopts the random gradient descent algorithm (SGD), and the learning rate adopts the stagnant descent method to decline in segments. When the accuracy of the model verification set stops improving, the learning rate will be reduced to one-tenth of the original, and the training will continue. The initial learning rate of the spatial stream network is set to 0.0005. The initial learning rate of the motion stream network is set to 0.01.

The hardware configuration of our experiments is: an Intel Xeon E5-2678 V3 CPU; Four GeForce RTX 2080ti GPUs; 4 × 16 GB, 64 GB memory in total. The experimental software is configured as Ubuntu 16.04.6 LTS operating system (Canonical Ltd.; Isle of Man, UK); python version 3.8 (Guido van Rossum; Holland); CUDA version 11.0. (NVIDIA; Santa Clara, CA, USA)

4.3. Results of Experiments and Analysis

4.3.1. Ablation Experiments of Network Structure

To analyze the influence of different components of the neural network on the accuracy of action recognition, we conducted ablation experiments. Figures 8 and 9, respectively show the top1 accuracy and top5 accuracy of four groups of experimental objects on the spatial stream network and motion stream network in the basic control group (2SCN), SimAM attention mechanism group (S-2SCN), BiGRU network group (B-2SCN), and neural network group (BS-2SCN) proposed in this paper.

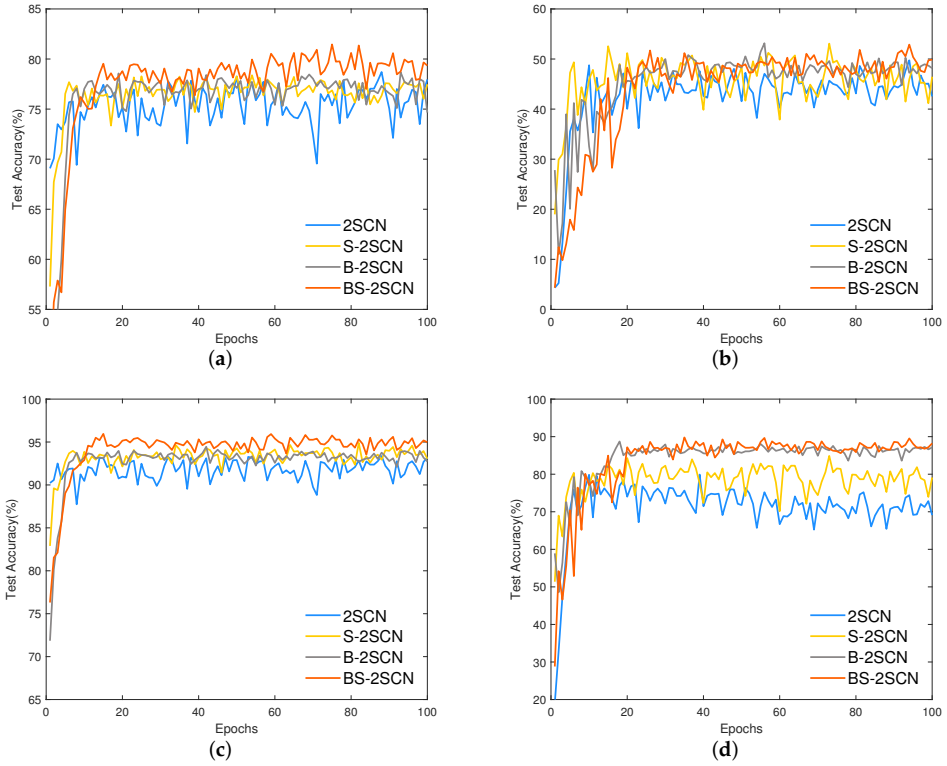


Figure 8. Accuracy of spatial stream network on two data sets. (**a**) UCF101 (top1). (**b**) HMDB51 (top1). (**c**) UCF101 (top5). (**d**) HMDB51 (top5).

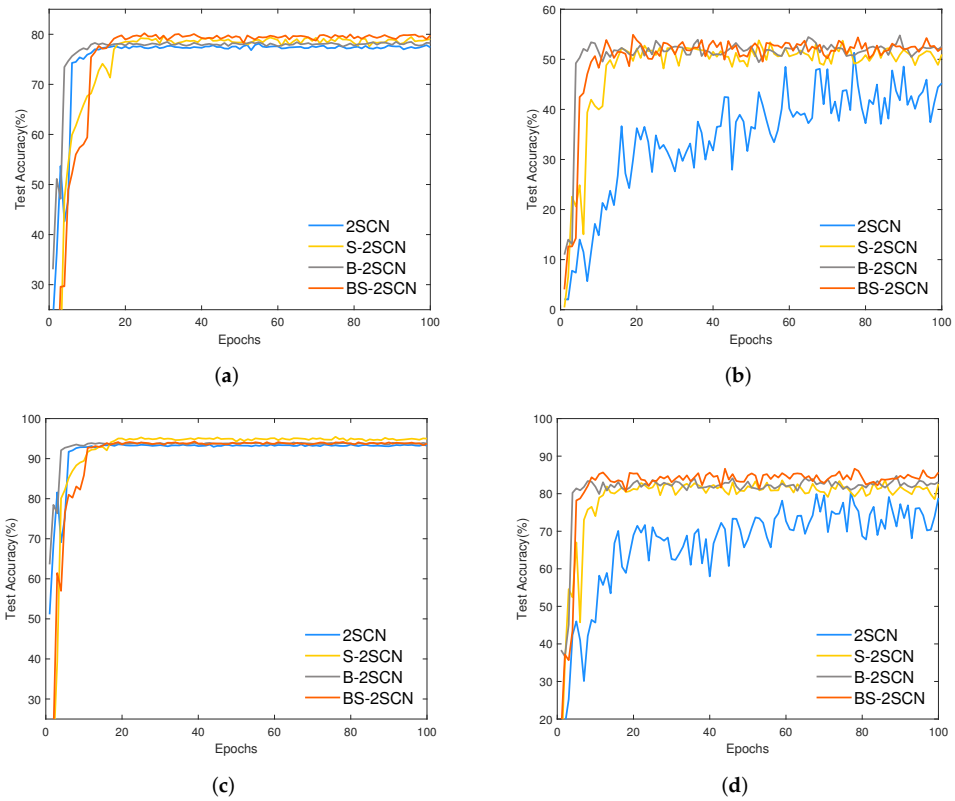


Figure 9. Accuracy of motion stream network on two data sets. (**a**) UCF101 (top1). (**b**) HMDB51 (top1). (**c**) UCF101 (top5). (**d**) HMDB51 (top5).

Figure 8 shows the top1 accuracy and top5 accuracy of four groups of experimental objects on the spatial stream network. As shown in Figure 8, the action recognition accuracy of the proposed network on the UCF101 and HMDB51 data sets is the highest when reaching the stable stage. It is worth noting that the experiment with the BiGRU network has low initial accuracy and slow convergence speed, which is due to the improvement in the complexity of the network. This phenomenon is not obvious in the experiment with SimAM, which is due to the advantage of the SimAM attention mechanism without neural network parameters.

Figure 9 shows the top1 accuracy and top5 accuracy of four groups of experimental objects on the motion stream network. In Figure 9a,b the accuracy of the basic control group is low and the fluctuation is obvious, but the fluctuation of the training process has been improved after adding the SimAM attention mechanism and the BiGRU network alone. After the superposition of SimAM and BiGRU network, its stability characteristics are retained, and the accuracy of action recognition is improved. The trend shown in Figure 8 is roughly the same as that shown in Figure 9. Combining Figures 8 and 9, the network model proposed in this paper has achieved satisfactory results in terms of accuracy and stability in training on the two datasets.

4.3.2. Comparative Experiment and Analysis

To verify the performance of the neural network proposed in this paper, it is compared with other action recognition algorithms. The results are shown in Table 1.

Table 1 shows some advanced action recognition algorithms on UCF101 data set and HMDB51 data set. We can see that the action recognition accuracy of the TLE network proposed by Diba, A. et al. [66] on the UCF101 data set has reached an amazing 95.6%, which is due to its expansion of the advanced 2D CNN network architecture into 3D CNN network, which can extract more in-depth Spatio-temporal features. Other works that achieve greater accuracy use more input [67] (75 frames) or a new network architecture [68]. The lower half of Table 1 shows the accuracy of the action recognition algorithm based on a two-stream convolution network. We can observe that our BS-2SCN network achieves the best performance on both data sets, and the accuracy on the UCF101 data set is 89.9%. The accuracy of the HMDB51 data set has reached 71.3%, which is the highest accuracy among all the comparison methods in Table 1.

Table 1. Accuracy comparison of action recognition algorithms on UCF101 dataset and HMDB51 dataset (%).

Method	Data Set	
	UCF101	HMDB51
Slow Fusion [69]	65.4	-
BiLSTM [70]	70.0	39.8
ST-D LSTM [55]	75.7	44.4
P3D ResNet [71]	88.6	-
Transformations [68]	92.4	62.0
MiCT-Net [67]	94.7	70.5
TLE [67]	**95.6**	71.1
Method based on Two Stream Convolutional Networks (2SCN)		
2SCN (VGG) [20]	86.9	58.0
2SCN (ResNet)	87.5	59.6
2SCN (Conv Pooling) [72]	88.2	-
2SCN (LSTM) [72]	88.6	-
2SCN (Fusion) [73]	89.6	57.6
2SCN (Hidden) [74]	89.8	-
BS-2SCN (proposed)	**90.1**	**71.3**

4.3.3. Experimental Overall Analysis

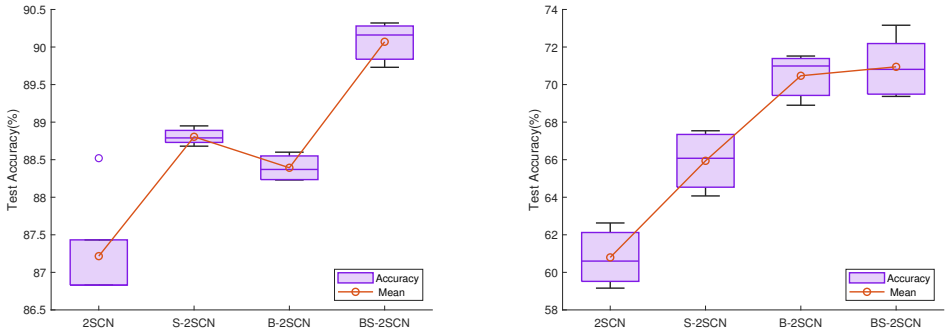
Table 2 shows the average and the best accuracy (both top1 accuracy) of the four groups of subjects on the UCF101 data set and HMDB51 data set in ten independent repeated experiments.

Table 2. Comparison of action recognition accuracy (%).

Data Set	Network	Accuracy	Model			
			2SCN	S-2SCN	B-2SCN	BS-2SCN
UCF101	Spatial	Average	78.75	78.90	78.70	**80.52**
		Best	78.80	79.34	78.70	**81.47**
	Motion	Average	78.99	78.18	78.33	**80.00**
		Best	79.83	79.67	78.51	**80.21**
	Fusion	Average	87.47	88.83	88.45	**90.07**
		Best	88.52	88.95	88.68	**90.32**
HMDB51	Spatial	Average	78.75	78.90	78.70	**80.52**
		Best	78.80	79.34	78.70	**81.47**
	Motion	Average	78.99	78.18	78.33	**80.00**
		Best	79.83	79.67	78.51	**80.21**
	Fusion	Average	87.47	88.83	88.45	**90.07**
		Best	88.52	88.95	88.68	**90.32**

It can be seen from Table 2 that both the BiGRU network and SimAM attention mechanism can effectively improve the action recognition accuracy of the two-stream convolution network. The combination of the two can greatly improve the accuracy of action recognition.

As shown in Figure 10, the fusion network accuracy of ten independent repeated experiments is summarized. Figure 10 shows the accuracy distribution of the neural network proposed in this paper in ten repeated experiments on two data sets. The box graph is the summary of the accuracy of each experimental object, and the broken line graph represents the average accuracy of each experimental object.



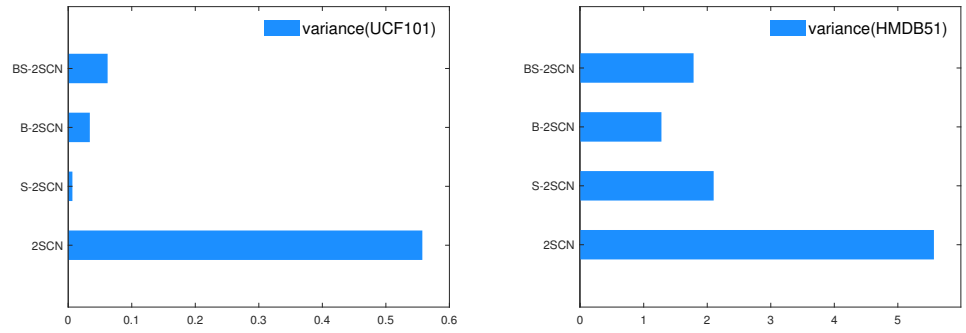
(**a**) (**b**)

Figure 10. Accuracy distributions and means for ten independent repeated experiments of four groups of subjects on two data sets (**a**) UCF101. (**b**) HMDB51.

Figure 10 shows that, from the upper edge value, the action recognition accuracy of the neural network (BS-2SCN) proposed in this paper is the highest. It is worth noting that in Figure 10a, the highest accuracy of the 2SCN network is equivalent to the average accuracy of the S-2SCN network and B-2SCN network, but this accuracy is represented by outliers and is not referential. From the perspective of interquartile distance (IQR), the interquartile distance of the 2SCN network is the largest, which shows that the accuracy distribution of the 2SCN network is relatively discrete. On UCF101 data set, the interquartile distance of the S-2SCN network is the smallest. On the HMDB51 data set, the interquartile distance of the B-2SCN network is the smallest. In addition, the red broken line diagram in Figure 10 shows the average accuracy of each neural network. By comparing the mean and median of each neural network (the purple horizontal line in the box), we can find that the two values

are relatively close, which shows that in many experiments, the accuracy distribution of each neural network is relatively uniform and the degree of dispersion is low.

To analyze the stability of the action recognition accuracy of each neural network in multiple experiments, we compared the accuracy variance of each neural network in ten independent repeated experiments, and the results are shown in Figure 11.



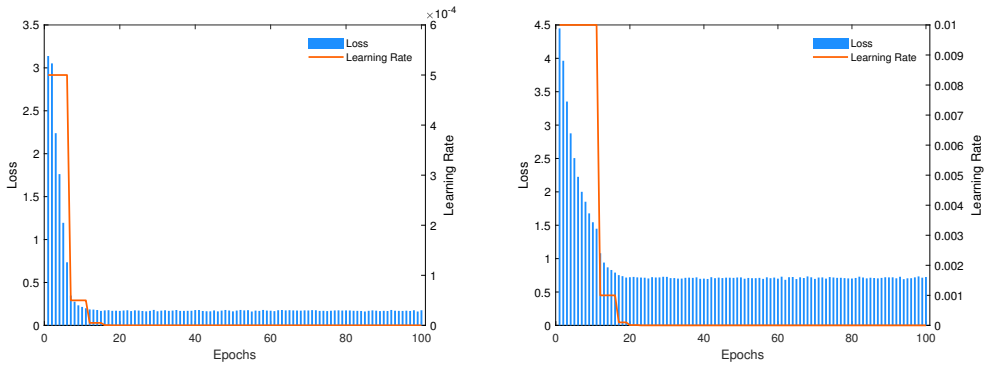
(**a**) (**b**)

Figure 11. variances of multiple experiment. (**a**) UCF101. (**b**) HMDB51.

Figure 11a shows the accuracy variance of ten independent repeated experiments of each neural network on the UCF101 data set. We can see that the variance of the 2SCN network is much larger than that of the other three networks, while the variance of the S-2SCN network is the smallest. However, as shown in Figure 11b, the variance of the B-2SCN network is the smallest, while the variance of the 2SCN network is still much larger than that of the other three networks. Overall, the neural networks with the BiGRU network reduce the variance of action recognition accuracy, which shows that the BiGRU network plays a positive role in increasing the stability of action recognition.

Learning rate is a crucial parameter in the training process of neural networks. A too small learning rate will lead to too slow convergence of the network model, or it will not converge due to the disappearance of the gradient. An excessive learning rate will cause the gradient to vibrate violently near the minimum value and it will not converge. A fixed learning rate can achieve good results in simple tasks, while in more complex tasks such as action recognition, a fixed learning rate is likely to lead to the inability of neural network convergence. Therefore, the dynamic adjustment of the learning rate is very necessary. The experiments in this paper adopt the stagnation descent algorithm to dynamically adjust the learning rate.

Figure 12 compares the learning rate and loss of spatial stream network and motion stream network. We can find that the declining trend of learning rate and loss is roughly the same, and the learning rate decreases with the stagnation of loss. The algorithm has achieved satisfactory results. The loss of motion stream is finally stable at about 0.17, and the loss value of motion steam is finally stable at about 0.7.



(**a**) (**b**)

Figure 12. Learning rate and loss of spatial stream network and motion stream network. (**a**) UCF101. (**b**) HMDB51.

As can be seen from Figure 8, the spatial stream network ablation experiments improved the accuracy by 2% on the UCF101 data set and 1.5% on the HMDB51 data set. Figure 9 shows that the motion stream network ablation experiments improved the accuracy by 2.1% on the UCF101 data set and 6.2% on the HMDB51 data set. Through the box graph and variance histogram of action recognition accuracy, we find that both BiGRU network and SimAM attention can increase the stability of action recognition. In the comparative experiment, our BS-2SCN network has the highest accuracy in the framework of a two-stream convolution network. On HMDB51 data set, our BS-2SCN network has achieved the effect of state-of-the-art. This proves that the BiGRU network with strong mining ability for contextual features helps our neural network to perceive the coherent features of appearance, and the effect of SimAm attention mechanism to improve the accuracy and stability of the neural network.

5. Conclusions

In this paper, we propose an improved two-stream convolution network. The recognition mode of a single frame of a spatial stream is changed to multi-frame image recognition by using the BiGRU network, which solves the shortcomings of a classical two-stream network in the perception of action appearance coherence features. Compared with the GRU network, the BiGRU network can record past information and predict future information at the same time, which makes the neural network more robust. Furthermore, the introduction of the SimAm attention mechanism improves the accuracy and stability of action recognition. After ablation experiments and comparative experiments, the accuracy and stability of the improved two-stream convolutional neural network (BS-2SCN) proposed in this paper have been improved on both the UCF101 dataset and the HMDB51 dataset.

Author Contributions: Conceptualization, K.H.; methodology, Z.W.; software, Z.W., H.L., and J.J.; validation, Z.W., and J.J.; formal analysis, Z.W., H.L., and J.J.; investigation, Z.W., H.L., and J.J.; resources, K.H., and Z.W.; data curation, K.H.; writing—original draft preparation, Z.W., and H.L.; writing—review and editing, Z.W., and J.J.; visualization, Z.W., and H.L.; supervision, K.H.; project administration, K.H.; funding acquisition, K.H. All authors have read and agreed to the published version of the manuscript.

Funding: Research in this article was supported by the NUIST Students' Platform for Innovation and Entrepreneurship Training Program (202110300120Y). I would like to express my heartfelt thanks to the reviewers and editors who submitted valuable revisions to this article.

Data Availability Statement: The data and code used to support the findings of this study are available from the first author upon request (201983240003@nuist.edu.cn). The data are from "https://www.crcv.ucf.edu/research/data-sets/ucf101/ (accessed on 5 December 2021)" (UCF101) and "https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/ (accessed on 15 January 2022)" (HMDB51).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GRU	gated recurrent unit
BiGRU	bidirectional gated recurrent unit
2SCN	Two-stream convolution network
S-2SCN	SimAM Two-stream convolution network
B-2SCN	BiGRU Two-stream convolution network
BS-2SCN	BiGRU-SimAM Two-stream convolution network
CNN	convlutional neural network

Parameter symbols

The following Parameter symbols are used in this manuscript:

I	optial stream	Equation (1)
t	sign of a frame	
k	sign of a frame	
d	sign of differential	
u	abscissa of a frame	
v	Ordinate of a frame	
k	Ordinate of a frame	
w	input word of BiGRU	
m	length of a word	
c	characters of a sequence	
e	projection vectors	
h	state sequence	
syn	syntactic structure	
en	energy of each neuron	Equation (3)
we	linear conversion of weight	
b	linear conversion of offset	
r	target neuron in a signal input channel	
q	other neurons in a signal input channel	
i	serial number	
M	the number of other neurons	
y	variable	
λ	coefficient	Equation (4)
μ	average value	Equation (6)
σ	indicate variance	Equation (7)
E	energy matrix	Equation (9)
E'	weight matrix of attention	
S	original feature map	Equation (10)
S'	fused feature map	
r_t	reset gate at time t	Equation (11)
z_t	update gate at time t	Equation (12)
$\tilde{h_t}$	state of candidate activation at time t	Equation (13)
W	weight matrix of GRU	
x_t	input of GRU	
$\overrightarrow{h_{t-1}}$	forward hidden layer state	Equation (15)
$\overleftarrow{h_{t-1}}$	reverse hidden layer state	Equation (16)

References

1. Xiong, P.; He, K.; Wu, E.Q.; Zhu, L.-M.; Song, A.X. Liu, P. Human-Exploratory-Procedure-Based Hybrid Measurement Fusion for Material Recognition. *IEEEASME Trans. Mechatron.* **2022**, *27*, 1093–1104. [CrossRef]
2. Xiong, P.; Zhu, X.; Song, A.; Hu, L.; Liu, X.P.; Feng, L. A Target Grabbing Strategy for Telerobot Based on Improved Stiffness Display Device. *IEEECAA J. Autom. Sin.* **2017**, *4*, 661–667. [CrossRef]
3. Bobick, A.; Davis, J. An Appearance-Based Representation of Action. In Proceedings of the 13th International Conference on Pattern Recognition, Washington, DC, USA, 25–29 August 1996; IEEE: Vienna, Austria, 1996; Volume 1, pp. 307–312.
4. Weinland, D.; Ronfard, R.; Boyer, E. Free Viewpoint Action Recognition Using Motion History Volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257. [CrossRef]
5. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning Realistic Human Actions from Movies. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
6. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.-L. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [CrossRef]
7. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
8. Hu, K.; Ding, Y.; Jin, J.; Weng, L.; Xia, M. Skeleton Motion Recognition Based on Multi-Scale Deep Spatio-Temporal Features. *Appl. Sci.* **2022**, *12*, 1028. [CrossRef]
9. Yang, X.; Tian, Y. Effective 3D Action Recognition Using EigenJoints. *J. Vis. Commun. Image Represent.* **2014**, *25*, 2–11. [CrossRef]
10. Liu, X.; Chen, H.-X.; Liu, B.-Y. Dynamic Anchor: A Feature-Guided Anchor Strategy for Object Detection. *Appl. Sci.* **2022**, *18*, 4897. [CrossRef]
11. Hu, K.; Tian, L.; Weng, C.; Weng, L.; Zang, Q.; Xia, M.; Qin, G. Data-Driven Control Algorithm for Snake Manipulator. *Appl. Sci.* **2021**, *11*, 8146. [CrossRef]
12. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. [CrossRef]
13. Xia, M.; Wang, Z.; Lu, M.; Pan, L. MFAGCN: A New Framework for Identifying Power Grid Branch Parameters. *Electr. Power Syst. Res.* **2022**, *207*, 107855. [CrossRef]
14. Gao, J.; Weng L.; Xia, M.; Lin, H. MLNet: Multichannel feature fusion lozenge network for land segmentation. *J. Appl. Remote Sens.* **2022**, *16*, 016513. [CrossRef]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
17. Szegedy, C.; Wei, L.; Yangqing, J.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
20. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *arXiv* **2014**, arXiv:1406.2199.
21. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2016**, arXiv:1409.0473.
22. Luong, M.-T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-Based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. 10. *arXiv* **2015**, arXiv:1709.01507.
24. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. ISBN 978-3-030-01233-5.
25. Xia, M.; Qu, Y.; Lin, H. PADANet: Parallel asymmetric double attention network for clouds and its shadow detection. *J. Appl. Remote Sens.* **2021**, *15*, 046512. [CrossRef]
26. Song, L.; Xia, M.; Jin, J.; Qian, M.; Zhang, Y. SUACDNet: Attentional change detection network based on siamese U-shaped structure. *Int. J. Appl. Earth. Obs.* **2021**, *105*, 102597. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
28. Elman, J.L. Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Mach. Learn.* **1991**, *7*, 195–225. [CrossRef]
29. Hochreiter S; Schmidhuber J Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
30. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.

31. Chakrabarty, A.; Pandit, O.A.; Garain, U. Context Sensitive Lemmatization Using Two Successive Bidirectional Gated Recurrent Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 1481–1491.
32. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021.
33. Allport, A. Visual Attention. In *Foundations of Cognitive Science*; The MIT Press: Cambridge, MA, USA, 1989; pp. 631–682, ISBN 978-0-262-16112-1.
34. Cheng, X.; Li, X.; Yang, J.; Tai, Y. SESR: Single Image Super Resolution with Recursive Squeeze and Excitation Networks. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 147–152.
35. Jin, X.; Xie, Y.; Wei, X.-S.; Zhao, B.-R.; Chen, Z.-M.; Tan, X. Delving Deep into Spatial Pooling for Squeeze-and-Excitation Networks. *Pattern Recognit.* **2022**, *121*, 108159. [CrossRef]
36. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Interspeech* **2020** *2020*, 3830–3834. [CrossRef]
37. Qiu, C.; Zhang, S.; Wang, C.; Yu, Z.; Zheng, H.; Zheng, B. Improving Transfer Learning and Squeeze- and-Excitation Networks for Small-Scale Fine-Grained Fish Image Classification. *IEEE Access* **2018**, *6*, 78503–78512. [CrossRef]
38. Gong, L.; Jiang, S.; Yang, Z.; Zhang, G.; Wang, L. Automated Pulmonary Nodule Detection in CT Images Using 3D Deep Squeeze-and-Excitation Networks. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 1969–1979. [CrossRef]
39. Han, Y.; Wei, C.; Zhou, R.; Hong, Z.; Zhang, Y.; Yang, S. Combining 3D-CNN and Squeeze-and-Excitation Networks for Remote Sensing Sea Ice Image Classification. *Math. Probl. Eng.* **2020**, *2020*, 1–15. [CrossRef]
40. Wei, S.; Qu, Q.; Wu, Y.; Wang, M.; Shi, J. PRI Modulation Recognition Based on Squeeze-and-Excitation Networks. *IEEE Commun. Lett.* **2020**, *24*, 1047–1051. [CrossRef]
41. Qu, Y.; Xia, M.; Zhang, Y. Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. *Comput. Geosci.* **2021**, *157*, 104940. [CrossRef]
42. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
43. Huang, G.; Gong, Y.; Xu, Q.; Wattanachote, K.; Zeng, K.; Luo, X. A Convolutional Attention Residual Network for Stereo Matching. *IEEE Access* **2020**, *8*, 50828–50842. [CrossRef]
44. Ma, B.; Wang, X.; Zhang, H.; Li, F.; Dan, J. CBAM-GAN: Generative Adversarial Networks Based on Convolutional Block Attention Module. In *Artificial Intelligence and Security*; Lecture Notes in Computer Science; Sun, X., Pan, Z., Bertino, E., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 11632, pp. 227–236. ISBN 978-3-030-24273-2.
45. Wang, S.-H.; Fernandes, S.; Zhu, Z.; Zhang, Y.-D. AVNC: Attention-Based VGG-Style Network for COVID-19 Diagnosis by CBAM. *IEEE Sens. J.* **2021**. [CrossRef]
46. Li, Y.; Guo, K.; Lu, Y.; Liu, L. Cropping and Attention Based Approach for Masked Face Recognition. *Appl. Intell.* **2021**, *51*, 3012–3025. [CrossRef]
47. Cao, W.; Feng, Z.; Zhang, D.; Huang, Y. Facial Expression Recognition via a CBAM Embedded Network. *Procedia Comput. Sci.* **2020**, *174*, 463–477. [CrossRef]
48. Fu, H.; Song, G.; Wang, Y. Improved YOLOv4 Marine Target Detection Combined with CBAM. *Symmetry* **2021**, *13*, 623. [CrossRef]
49. Wang, Y.; Zhang, Z.; Feng, L.; Ma, Y.; Du, Q. A New Attention-Based CNN Approach for Crop Mapping Using Time Series Sentinel-2 Images. *Comput. Electron. Agric.* **2021**, *184*, 106090. [CrossRef]
50. Zhao, Y.; Zhao, L.; Xiong, B.; Kuang, G. Attention Receptive Pyramid Network for Ship Detection in SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2738–2756. [CrossRef]
51. Carrasco, M. Visual Attention: The Past 25 Years. *Vision Res.* **2011**, *51*, 1484–1525. [CrossRef]
52. IL-MCAM: An Interactive Learning and Multi-Channel Attention Mechanism-Based Weakly Supervised Colorectal Histopathology Image Classification Approach. *Comput. Biol. Med.* **2022**, *143*, 105265. [CrossRef]
53. Xie, J.; Wu, Z.; Zhu, R.; Zhu, H. Melanoma Detection Based on Swin Transformer and SimAM. In Proceedings of the 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Xi'an, China, 15 October 2021; pp. 1517–1521.
54. Zhang, J.; Zeng, S.; Wang, Y.; Wang, J.; Chen, H. An Efficient Extreme-Exposure Image Fusion Method. *J. Phys. Conf. Ser.* **2021**, *2137*, 012061. [CrossRef]
55. Hu, K.; Zheng, F.; Weng, L.; Ding, Y.; Jin, J. Action Recognition Algorithm of Spatio–Temporal Differential LSTM Based on Feature Enhancement. *Appl. Sci.* **2021**, *11*, 7876. [CrossRef]
56. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.* **2022**, 1–21. [CrossRef]
57. Webb, B.S. Early and Late Mechanisms of Surround Suppression in Striate Cortex of Macaque. *J. Neurosci.* **2005**, *25*, 11666–11675. [CrossRef] [PubMed]
58. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as Space-Time Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef]

59. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
60. Liu, J.; Luo, J.; Shah, M. Recognizing Realistic Actions from Videos "in the Wild". In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 22–24 June 2009; pp. 1996–2003.
61. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A Large Video Database for Human Motion Recognition. In Proceedings of the 2011 International Conference on Computer Vision, Washington, DC, USA, 6–13 November 2011; pp. 2556–2563.
62. Reddy, K.K.; Shah, M. Recognizing 50 Human Action Categories of Web Videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [CrossRef]
63. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv* **2012**, arxiv:1212.0402.
64. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:1705.06950.
65. Xia, M.; Zhang, X.; Weng, L.; Xu, Y. Multi-stage feature constraints learning for age estimation. *IEEE T. Inf. Foren. Sect.* **2020**, *15*, 2417–2428. [CrossRef]
66. Diba, A.; Sharma, V.; Van Gool, L. Deep Temporal Linear Encoding Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1541–1550.
67. Zhou, Y.; Sun, X.; Zha, Z.-J.; Zeng, W. MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 449–458.
68. Wang, X.; Farhadi, A.; Gupta, A. Actions Transformations. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2658–2667.
69. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 23–28 June 2014.
70. Marszalek, M.; Laptev, I.; Schmid, C. Actions in Context. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2929–2936.
71. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5534–5542.
72. Ng, J.Y.-H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond Short Snippets: Deep Networks for Video Classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
73. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
74. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A.G. Hidden Two-Stream Convolutional Networks for Action Recognition. *arXiv* **2018**, arxiv:17040.0389.