mathematics

MDPI

Article

Interactive Learning of a Dual Convolution Neural Network for Multi-Modal Action Recognition

Qingxia Li [1], Dali Gao [2,3], Qieshi Zhang [4], Wenhong Wei [5] and Ziliang Ren [5,*]

[1] School of Computer and Information, Dongguan City College, Dongguan 523419, China
[2] School of Mathematics and Computer Science, Quanzhou Normal University, Quanzhou 362000, China
[3] Fujian Provincial Key Laboratory of Data-Intensive Computing, Quanzhou 362000, China
[4] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[5] School of Computer Science and Technology, Dongguan University of Technology, Dongguan 523808, China
* Correspondence: renzl@dgut.edu.cn

Abstract: RGB and depth modalities contain more abundant and interactive information, and convolutional neural networks (ConvNets) based on multi-modal data have achieved successful progress in action recognition. Due to the limitation of a single stream, it is difficult to improve recognition performance by learning multi-modal interactive features. Inspired by the multi-stream learning mechanism and spatial-temporal information representation methods, we construct dynamic images by using the rank pooling method and design an interactive learning dual-ConvNet (ILD-ConvNet) with a multiplexer module to improve action recognition performance. Built on the rank pooling method, the constructed visual dynamic images can capture the spatial-temporal information from entire RGB videos. We extend this method to depth sequences to obtain more abundant multi-modal spatial-temporal information as the inputs of the ConvNets. In addition, we design a dual ILD-ConvNet with multiplexer modules to jointly learn the interactive features of two-stream from RGB and depth modalities. The proposed recognition framework has been tested on two benchmark multi-modal datasets—NTU RGB + D 120 and PKU-MMD. The proposed ILD-ConvNet with a temporal segmentation mechanism achieves an accuracy of 86.9% and 89.4% for Cross-Subject (C-Sub) and Cross-Setup (C-Set) on NTU RGB + D 120, 92.0% and 93.1% for Cross-Subject (C-Sub) and Cross-View (C-View) on PKU-MMD, which are comparable with the state of the art. The experimental results shown that our proposed ILD-ConvNet with a multiplexer module can extract interactive features from different modalities to enhance action recognition performance.

Keywords: convolutional neural network; rank pooling; feature interactive learning; action recognition

MSC: 65Z05

Citation: Li, Q.; Gao, D.; Zhang, Q.; Wei, W.; Ren, Z. Interactive Learning of a Dual Convolution Neural Network for Multi-Modal Action Recognition. *Mathematics* **2022**, *10*, 3923. https://doi.org/10.3390/math10213923

Academic Editor: Teng Li

Received: 20 September 2022
Accepted: 19 October 2022
Published: 22 October 2022

1. Introduction

As an important means of information and intelligent human–computer cooperation, action recognition has a broad application prospect in video monitoring, video retrieval, virtual reality, human–computer interaction, etc. [1–5]. Recently, deep convolutional neural networks (ConvNets) have strong feature learning and model generalization ability by automatically learning feature information from the bottom to the top, which have been applied to many fields, such as video understanding [6–10]. However, it is difficult to further improve the performance of models because of the limited ability of information representation and discrimination feature learning.

In recent years, most approaches mainly focus on designing deep ConvNets to learn spatial-temporal features, which improved action recognition performance through extracting the discriminative features from different modalities [11–15]. Most of those approaches are designed to work on a single frame or a single stack of frames in a short clip for extracting spatial-temporal features, while some methods obtain rich information by compressing

action videos [16–20]. Although these ConvNets based methods have achieved amazing performance in action recognition, the performance improvement of the models is limited due to the inherent limitations, e.g., some video frame sparse sampling methods are easy to cause the loss of temporal information and most single-stream ConvNets cannot jointly learn interactive features of different modalities.

Inspired by the video compression representation method proposed in [17,21,22], in this paper, we use the compression method to generate visual dynamic images (VDIs) and depth dynamic images (DDIs) to obtain rich information from those two modalities, and a multi-modal feature interactive learning model (ILD-ConvNet) is provided to extract the interactive features of RGB and depth modalities in order to improve recognition performance.

The main contributions of this paper are as follows:

1. An effective 2D-ConvNet framework is proposed to extract interactive features from different modalities to recognize human action, which jointly learn features from different modalities for human action recognition.
2. We experimentally demonstrate that the constructed VDIs and DDIs can effectively represent spatial-temporal information from the whole RGB-D sequences.
3. The proposed ILD-ConvNet recognition framework has been evaluated on large-scale human action datasets. The results show that our proposed ILD-ConvNet achieved better recognition performance and generalization performance, which demonstrated the effectiveness of our ILD-ConvNet.

2. Related Work

The extraction of underlying features is mainly divided into spatial and temporal features as for RGB and depth sequences, which can well describe the details of human actions. Early low-level feature information extraction mainly includes feature extraction algorithms with prior knowledge, geometric features, motion information, and Histogram of Oriented Gradient (HOG). [23,24]. In addition, some recognition methods used to represent time information mainly focus on spatial configuration modeling, dynamic time axis (DTW), Hidden Markov model (HMM), and the recognition methods based on hand-draft are summarized in [25,26].

In the past two decades, with the development of deep learning networks, action recognition methods based on RGB and depth modalities have made great progress. Simonyan et al. proposed a two-stream ConvNet recognition framework to fuse RGB and optical flow features, which achieved the effect of combining spatial-temporal information [27]. Wang et al. [28] present a series of good training practices for deep action recognition, Temporal Segment Networks (TSN), which included a sparse temporal sampling strategy and video-level supervision. However, it is difficult to ensure that the interactive information between different streams can be differentiated and explored. Zhang et al. [29] designed a cross-stream ConvNet to investigate syndicated information from multiple modalities. Wang et al. [30] provided a Temporal Difference Network (TDN) to learn multi-scale temporal information. In addition, 3D ConvNets can obtain more comprehensive spatiotemporal feature information, such as C3D, I3D, and X3D [31–33]. However, the amount of computation is too large and requires a lot of computing resources. In addition, transformer-based methods achieved remarkable results due to globally connected patterns across spatial and temporal dimensions [34–36].

Recently, some researchers have designed a series of ConvNets to realize human action recognition based on depth sequences, because it can reflect the comprehensive three-dimensional geometric information. Wang et al. [37,38] expressed the depth sequences to dynamic images through rank pooling and applied 3D point cloud mapping transformation to improve recognition performance, which achieved relatively good recognition results on multi-modal datasets. In addition, Recurrent Neural Networks (RNNs) can capture the feature information of temporal well, which have developed rapidly in action recognition [39–41].

The above feature learning methods have achieved success in improving the human action recognition performance. However, the single-modality data cannot provide the comprehensive information required for action recognition. Based on the multi-stream recognition framework, some strategies are proposed to capture distinguishing features, such as sparse temporal sampling [14], motion image [33] and dynamic image representation [17,21,22]. In [42], a cooperative ConvNets is designed to jointly learn the middle-level feature information of RGB and depth sequences, which obtained good recognition performance on the existing RGB-D datasets such as NTU RGB + D 60 [43]. In addition, the joint learning mechanisms are designed to learn single-modal and cross-modal information for action recognition tasks, such as c-ConvNet [42], SC-ConvNets [44] and J-ResNet-CMCB [45]. In addition, some approaches improved recognition performance through fusing RGB, depth and skeleton in multi-stream feature learning, such as Modality Compensation Network [46] and multi-modal action recognition model [47].

To overcome the lack of the above multi-stream ConvNets, with the advent of an increasing amount of multi-modal data, we are motivated to design more efficient recognition frameworks to represent spatial-temporal information from entire action sequences and extract complimentary features from different modalities.

3. The Proposed Method

The proposed ILD-ConvNet architecture as shown in Figure 1, which has three parts: the ILD-ConvNet, network input representation and joint optimization.

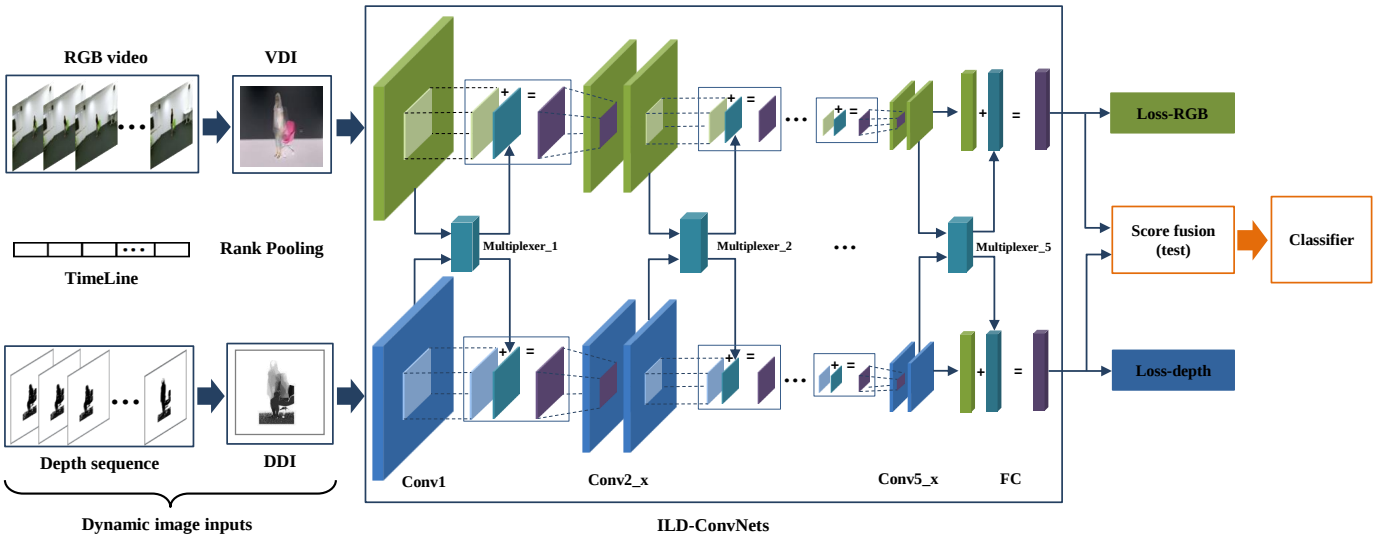

Figure 1. Overview of the proposed recognition framework.

3.1. ILD-ConvNet Framework

By considering 2D ConvNets, 3D ConvNets and RNNs, a dual ConvNets framework is designed based on the multi-modal feature interactive learning and fusion strategy. The designed multi-modal interactive learning dual flow network ILD-ConvNet consists of two separate paths, multiplexer modules and two losses. Given an RGB and depth sequences, the rank pooling method converted them into a pair of dynamic images $\langle VDI, DDI \rangle$. Then, a designed single ResNet with multiplexer modules is used to extract the discriminative features from $\langle VDI, DDI \rangle$. Then, the two losses and effective fusion strategy are used to jointly optimize and fuse multi-modal features.

Due to the high modularity, ResNet is selected as the basic module for RGB and depth information flow learning [48]. The ILD-ConvNet recognition framework includes two independent information flow learning networks. Each network includes 7×7 convolution layer, four bottleneck blocks, a full connection layer and softmax mapping layer. The network adopts the same strategies as ResNet. The network framework based on the

ResNet$_{50}$ model is summarized in Table 1. In the ResNet$_{50}$ model, the input maps of the 4 bottleneck blocks are 64, 256, 512 and 1024. To reduce the number of parameters of the multiplexer module, the network is placed before each bottleneck block to learn the interaction characteristics of different modalities.

Table 1. The network architecture of ILD-ConvNet-based ResNet$_{50}$. Down sampling is performed by conv3_x, conv4_x and conv5_x with a stride of 2.

Layer Name	Output Size	ILD-ConvNet$_{50}$	
		RGB Stream	Depth Stream
Conv1	112×112	14×14, 64, stride 2	7×7, 64, stride 2
Multiplexer_1	112×112	$\begin{bmatrix} 1 \times 1, & 64 \times 2 \\ 1 \times 1, & 64 \end{bmatrix} \times 1$	
Conv2_x	56×56	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$
Multiplexer_2	56×56	$\begin{bmatrix} 1 \times 1, & 256 \times 2 \\ 1 \times 1, & 256 \end{bmatrix} \times 1$	
Conv3_x	28×28	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$
Multiplexer_3	28×28	$\begin{bmatrix} 1 \times 1, & 512 \times 2 \\ 1 \times 1, & 512 \end{bmatrix} \times 1$	
Conv4_x	14×14	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 23$
Multiplexer_4	14×14	$\begin{bmatrix} 1 \times 1, & 1024 \times 2 \\ 1 \times 1, & 1024 \end{bmatrix} \times 1$	
Conv5_x	7×7	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$
Multiplexer_5	7×7	$\begin{bmatrix} 1 \times 1, & 2048 \times 2 \\ 1 \times 1, & 2048 \end{bmatrix} \times 1$	
FC	1×1	FC-RGB	FC-depth

3.2. Network Inputs

Represent RGB or depth sequences as $\langle I_1, I_2, I_3, \ldots, I_T \rangle$. $\varnothing(I_t) \in R^d$ is the feature vector extracted therefrom of I_t, and the average feature vector at the time t is defined as

$$V_t = \frac{1}{t} \sum_{\tau=1}^{t} \varnothing(I_\tau), \tag{1}$$

where the sorting score function of V_t at the time t is as follows.

$$S(t|d) = \langle \mathrm{d}, V_t \rangle, \mathrm{d} \in R^d, \tag{2}$$

where d is the learned parameter vector. The score $S(t|d)$ reflects the ordering of each image frame in the sequence, which satisfies $q > t \rightarrow S(q|d) > S(t|d)$.

$$E(d) = \frac{\lambda}{2}||d||^2 + \frac{2}{T(T-1)} \sum_{q>t} \max\{0, 1 - S(q|d) + S(t|d)\}, \tag{3}$$

where the first and second terms are the quadratic term of SVM optimization and the error accumulation of score ranking, respectively. The interval between q and t is set as 1 to

distinguish the score ranking. The mapping function from video to parameter feature d^* vector is further defined as [49]:

$$d^* = \rho(I_1, \ldots, I_T; \varnothing) = argmin_d E(d), \tag{4}$$

where d^* is the optimal estimate of d. d^* contains all the ordering information of the video sequence, which is equal to the number of pixels of a single frame of the sequence. Therefore, it can be seen as a descriptor of the videos and interpreted as a standard image frame. For depth sequence, we first normalize them into single-channel images, and then use the introduced rank pooling method (d^*) to compress them into one dynamic image (DDI).

3.3. Multiplexer Modules

In terms of network architecture design, ResNet is selected as the basic network based on the highly modular residual unit. The proposed ILD-ConvNet is divided into two paths to process the input RGB and depth information flow characteristics, respectively. High-level interactive learning is conducted through the multiplexer module, as shown in Figure 2.

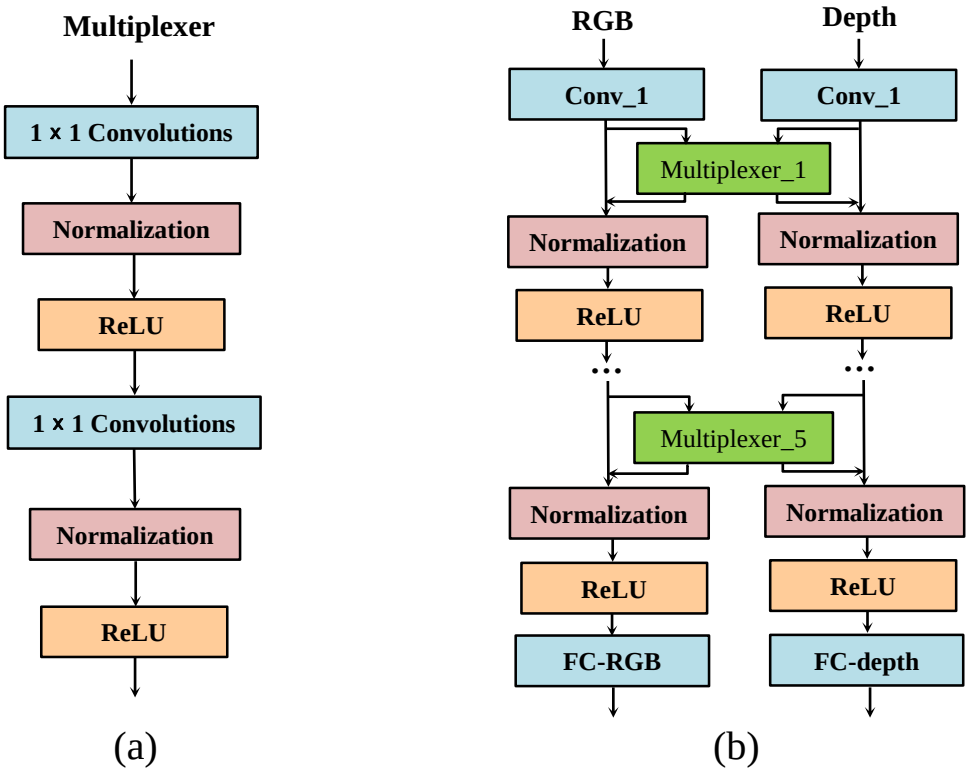

Figure 2. Multiplexer module and information stream. (**a**) Multiplexer module. (**b**) Two-stream information.

The multiplexer module combines the characteristics of two-path information flow. After weighted redirection interactive learning, it is input back to RGB and depth information flow, respectively. The characteristic output function of the multiplexer module can be defined as

$$F_{multiplexer}(x) = ReLU(BN(f_{1\times1}(ReLU(BN(f_{1\times1}(w * x)))))), \tag{5}$$

where $BN(\cdot)$ and $ReLU(\cdot)$ denote the batch normalization and rectified linear units activation function, respectively, and $f_{1\times1}$ is a learnable 1×1 convolution kernel. The information flow input is defined as:

$$X_{RGB} = \lambda X_{RGB} + (1-\lambda) F_{multiplexer}\Big(X_{RGB} + X_{depth}\Big), \tag{6}$$

$$X_{depth} = \lambda X_{depth} + (1-\lambda) F_{multiplexer}\Big(X_{RGB} + X_{depth}\Big), \tag{7}$$

where X_{RGB} and X_{depth} represent the output of the information flow, respectively. $\lambda \in (0,1)$ is the adjustment coefficient.

3.4. Optimization Learning

Different from the ConvNets, which is trained on a single frame, the designed ILD-ConvNet has paired sample objects $\langle VDI, DDI\rangle$ for training. In order to eliminate the deviation between two different modalities, a dual loss function is used to jointly optimize RGB and depth information flows. In the meantime, standard cross-entropy loss functions are used in the training process.

The information flow of the training sample $\langle VDI, DDI\rangle$ is expressed as $X_{\langle VDI,DDI\rangle}$. The scores of VDI and DDI are obtained by using the softmax function. The classification probability score can be estimated as follows:

$$prob_c = \frac{\exp(W_c X + b_c)}{\sum_{c_i} \exp(w_{c_i} X + b_{c_i})}, \tag{8}$$

where W_c and b_c are the weights and offsets of the softmax layer. The loss function of the two streams can be described as:

$$L_{RGB}(y, C) = -\sum_{c=1}^{C} y_c(\log(prob_{c-RGB})), \tag{9}$$

$$L_{depth}(y, C) = -\sum_{c=1}^{C} y_c(\log(prob_{c-depth})), \tag{10}$$

where C is the classification number of actions, y_c is the ground truth label, $prob_{c-RGB}$ and $prob_{c-depth}$ are the classification probabilities, respectively.

Different from training on a single RGB or depth frame, ILD-ConvNet operates on $\langle VDI, DDI\rangle$ and obtains two-stream features. To make the proposed ConvNets more discriminative and eliminate the multi-modal variance between the RGB-D information streams, a dual loss function is adopted to jointly optimized ILD-ConvNet. In the training process, the standard cross-entropy loss is used to jointly optimize ILD-ConvNet. The overall loss function can be expressed as:

$$L(y, C) = L_{RGB} + L_{depth}, \tag{11}$$

The feature vectors v_{RGB} and v_{depth} of the dual stream network can be obtained in the testing process. The final score feature vectors of RGB and depth sequences can be obtained through feature fusion:

$$v_{fusion} = v_{RGB} \circ v_{\text{depth}}, \tag{12}$$

where "$\circ$" represents the maximum, minimum or inner product operation. Then, the maximum value of v_{fusion} is used as the classification result of RGB and depth sequences.

3.5. Datasets Implementation Details

3.5.1. Datasets

NTU RGB + D 120 [50] is the largest multi-modal human action dataset currently released. NTU RGB + D 120 includes RGB, depth, infrared IR and 3D skeleton modalities, and two evaluation protocols are defined. For the Cross-Subject (C-Sub) evaluation protocol, samples collected by subjects with camera numbers of 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38, 45, 46, 47, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 70, 74, 78, 80, 81, 82, 83, 84, 85, 86, 89, 91, 92, 93, 94, 95, 97, 98, 100, and 103 were used for training (63,360 samples). The rest were selected for testing (51,120 samples). For the Cross-Setup (C-Set), the samples collected by the even ID cameras are divided into training (54,720 samples) and the rest for testing (59,760 samples).

The PKU-MMD dataset [51] includes multi-modal data such as RGB, depth, infrared radiation and skeleton modalities, and each modality includes 1074 long video sequences, and two action recognition evaluation protocols are designed for the dataset. For the Cross-Subject (C-Sub) evaluation protocol, it is divided into 18,134 training and 2600 testing samples. For the Cross-View (C-View) evaluation protocol, the training set have 13,815 samples and the testing have 6919 samples.

3.5.2. Implementation Details

Network input. The RGB and depth sequences are compressed to obtain the input of the ILD-ConvNet by adopting the rank pooling method proposed above. In order to adapt to the dual flow network training model, $\langle VDI, DDI \rangle$ is scaled to 224×224. The DDI sorted by depth sequence is scaled to $[0, 255]$. In addition, the training samples are incrementally processed by random clipping and dithering.

Model training and testing. The parameters of ILD-ConvNet can be initialized through the pre-trained weights of ResNet. In the training and testing process, we follow the good practice proposed in [16], the methods of mini batch, momentum, random gradient descent and data augmentation methods are used for feature learning, where the batch and momentum are set as 64 and 0.8, respectively. The initial learning rate is set as 0.01, which decreases to 0.1 times every 30 generations. In order to overcome the overfitting phenomenon, dropout technology was further adopted. The dropout rate was set to 0.5. Inspired by the good practices proposed in [16], we further verified the impact of segmentation strategy on recognition performance in our experiments.

4. Experimental Results and Analysis

4.1. Effectiveness of the Proposed ILD-ConvNet

4.1.1. Effectiveness

In order to evaluate the effectiveness of the proposed recognition framework, comparative experiments of different modalities are conducted on the NTU RGB + D 120 dataset, including RGB, depth and RGB + depth dual modalities. The parameter λ was set to 0.1 and the average fusion method is used in the experiments, Table 2 shows the experimental results. The recognition rate of VDI in ResNet_{50} model is 16.1% and 13.0% higher than that of RGB, and the DDI is 28.6% and 30.1% higher than that of depth on the basis of the evaluation protocols of the NTU RGB + D 120 dataset C-Sub and C-Set. The improvement of the recognition rate shows that the proposed video frame ranking method can improve the behavior recognition performance of the training mode. The results proves that this method can effectively represent the spatial-temporal information of samples.

Table 2 shows that the recognition rates of ResNet-50 model achieve 51.4% and 51.3%, respectively, by fusing RGB and depth modalities. The recognition rates of ILD-ConvNet_{50} achieve 55.4% and 55.0%, respectively. The recognition rates of ILD-ConvNet50 are 4.0% and 3.7% higher than those of ResNet_{50}. The recognition rates of ResNet_{50} model achieve 73.8% and 74.0% by fusing VDI and DDI modalities. ILD-ConvNet_{50} can obtain 75.3% and 75.2% recognition rates. The proposed ILD-ConvNet_{50} are 1.5% and 1.2% higher than those of ResNet_{50}. Furthermore, we can see that the recognition rate is greatly improved by

fusing RGB and depth modalities, which indicates the interactive characteristics between RGB and depth sequences. Furthermore, the experimental results of our ILD-ConvNet show that learning the interactive characteristics between two different modalities can better improve recognition performance. The reason of the improvement of recognition rate is that the designed multiplexer module has learned the interaction characteristics of RGB and depth.

Table 2. The results of ResNet_{50} and ILD-ConvNet.

Methods	Modality	C-Sub	C-Set
ResNet_{50}	*RGB*	40.6%	39.5%
ResNet_{50}	*depth*	43.6%	42.2%
ResNet_{50}	*RGB + depth*	51.4%	51.3%
ResNet_{50}	VDI	56.7%	52.5%
ResNet_{50}	DDI	72.2%	72.3%
ResNet_{50}	$\langle VDI, DDI \rangle$	73.8%	74.0%
ILD-ConvNet_{50}	$\langle RGB, depth \rangle$	55.4%	55.0%
ILD-ConvNet_{50}	$\langle VDI, DDI \rangle$	75.3%	75.2%

4.1.2. Comparison of Different Fusion Methods

The average fusion method was adopted to achieve the classification results in Table 2. To obtain better recognition performance, the weighted average, maximum and product score fusion methods are evaluated on the NTU RGB + D dataset, and the results are shown in Table 3. It can be seen from the experimental results that the recognition rates of the three feature fusion methods are similar, and the average method is relatively good. Therefore, the average fusion method is selected as the feature fusion method for the following experiments.

Table 3. Comparative accuracies of the proposed ILD-ConvNet_{50} with $\langle VDI, DDI \rangle$ on the NTU RGB + D 120 dataset.

Methods	C-Sub	C-Set
Average	75.3%	75.2%
Product	75.2%	75.0%
Max	74.9%	74.9%

4.1.3. Effectiveness of Parameter λ

The parameter λ is varied from 0 to 0.5 to evaluate recognition performance. The comparison results are summarized in Figure 3, and we can see that increasing λ will lead to worse performance. The results implies that increasing λ does not help obtain richer information from the other modality. Thus, $\lambda = 0.1$ is set in the following experiments.

4.2. Comparison to the State of the Art

4.2.1. Experimental on the NTU RGB + D Dataset

To further demonstrate the effectiveness of RGB and depth modalities, BNInception, ResNet_{50}, ResNet_{101}, ILD-ConvNet_{50} and ILD-ConvNet_{101} were selected for further experiments, and the comparison results are shown in Table 4. RGB and depth modalities are used as inputs in this group of experiments. In addition, BNInception, ResNet_{50} and ResNet_{101} are used to train RGB, depth, VDI and DDI, respectively. Meanwhile, image pairs $\langle RGB, depth \rangle$ and $\langle VDI, DDI \rangle$ are used as the inputs of ILD-ConvNet_{50} and ILD-ConvNet_{101}.

The experimental results indicate that ILD-ConvNet_{101} achieved 82.8% and 83.6% recognition rates, respectively, on the C-Sub and C-Set protocols of the NTU RGB + D 120 dataset. ILD-ConvNet_{101} is 0.7% and 2.1% higher than ResNet_{101}, respectively, for the same input and base network, and our ILD-ConvNet_{101} is 5.6% and 6.2% higher than

$ResNet_{101}$, respectively. Furthermore, inspired by the segmentation training mechanism proposed in [16], we conduct experiments in 3 segments on the NTU RGB + D 120 dataset, which obtain 86.9% and 89.4% on two protocols. In specially, the proposed ILD-ConvNet degenerates to the plain SC-ConvNet as in [44]. Although the proposed recognition framework is only 1.7% higher than SC-ConVnet in C-Set protocol, it also shows the effectiveness of our designed multiplexer module. Moreover, some subtle actions are easier to recognize, such as "jump table tennis ball". The experimental results demonstrated that the interactive features of RGB and depth modalities learned by the designed multiplexer module can improve recognition performance.

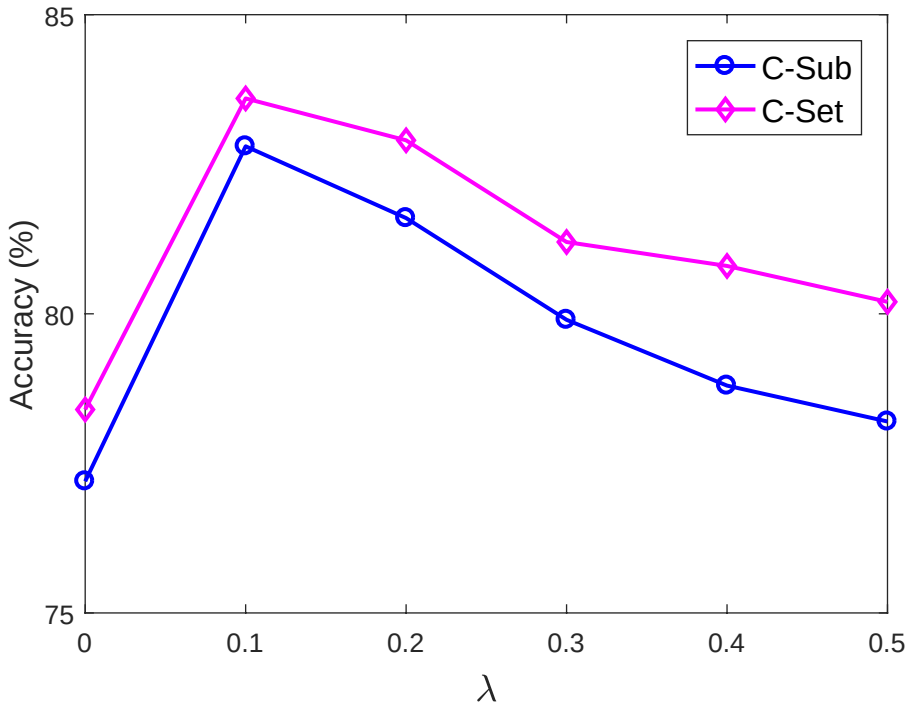

Figure 3. Comparison of weight λ for two protocols using the ILD-ConvNet101 with the inputs $\langle VDI, DDI \rangle$.

Table 4. Comparative accuracies of different networks, and the segment is set 3 in TSN [16].

Methods	Modality	C-Sub	C-Set
BNInception	$\langle RGB, depth \rangle$	52.5%	53.1%
$ResNet_{50}$	$\langle RGB, depth \rangle$	51.4%	51.3%
$ResNet_{101}$	$\langle RGB, depth \rangle$	56.5%	54.1%
BNInception	$\langle VDI, DDI \rangle$	75.4%	76.2%
$ResNet_{50}$	$\langle VDI, DDI \rangle$	73.8%	74.0%
$ResNet_{101}$	$\langle VDI, DDI \rangle$	77.2%	78.4%
ILD-$ConvNet_{50}$	$\langle RGB, depth \rangle$	55.4%	55.0%
ILD-$ConvNet_{101}$	$\langle RGB, depth \rangle$	57.2%	56.2%
ILD-$ConvNet_{50}$	$\langle VDI, DDI \rangle$	75.3%	75.2%
ILD-$ConvNet_{101}$	$\langle VDI, DDI \rangle$	82.4%	83.1%
J-ResNet-CMCB [45]	$\langle VDI, DDI \rangle$	82.8%	83.6%
TSN [16] + SC-ConvNet [44]	$\langle VDI, DDI \rangle$	86.9%	87.7%
TSN [16] + ILD-$ConvNet_{101}$	$\langle VDI, DDI \rangle$	86.9%	89.4%

4.2.2. Experimental on the PKU-MMD Dataset

To evaluate the effectiveness and generalization ability of the proposed ILD-ConvNet, we further conducted experiments on the multi-modal human action dataset PKU-MMD with different scales and application backgrounds. The basic network is $ResNet_{101}$, and λ is set to 0.1. Table 5 shows the experimental comparison results.

Table 5. Comparative accuracies of different networks on the PKU-MMD dataset; the segment is set 3 in TSN [16].

Methods	Modality	C-Sub	C-View
SA-LSTM [52]	*skeleton*	86.3%	91.4%
TA-LSTM [52]	*skeleton*	86.6%	92.3%
STA-LSTM [52]	*skeleton*	86.9%	92.6%
ResNet_{101}	VDI	77.3%	74.8%
ResNet_{101}	DDI	81.7%	79.0%
ResNet_{101}	$\langle VDI, DDI \rangle$	83.6%	82.7%
ILD-ConvNet_{101}	*RGB + depth*	80.7%	79.8%
J-ResNet-CMCB [45]	$\langle VDI, DDI \rangle$	90.4%	91.4%
TSN [16] + SC-ConvNet [44]	$\langle VDI, DDI \rangle$	92.1%	93.2%
TSN [16] + ILD-ConvNet_{101}	$\langle VDI, DDI \rangle$	92.0%	93.1%

Table 5 shows that the proposed ILD-ConvNet based on ResNet_{101} obtained 92.0% and 93.1% recognition rates, respectively, on the two protocols. The results show that the proposed ILD-ConvNet obtained generalization ability on different types and scales of datasets.

4.3. Analysis

Compared with other ConvNet-based methods, the experimental results on the NTU RGB + D and PKU-MMD datasets have shown the effectiveness of our ILD-ConvNet. It can be seen that the proposed dual-stream framework can achieve better recognition performance. The experimental results also implied that the introduced rank pooling mechanism can represent spatial-temporal information from entire RGB and depth sequences, and the designed multiplexer module can enhance recognition performance by extracting the interactive features from different modalities.

Figures 4 and 5 shows the confusion matrix under the two evaluation protocols. It can be seen that the designed ILD-ConvNet can effectively identify most of the actions, but some refined actions, such as "make an OK sign" and "snap fingers". On the C-Sub test protocol, highly similar actions of human–object interaction are easy to be confused, such as "put on glasses" and "take off glasses", "reading" and "writing". In addition, it is difficult to effectively distinguish actions with similar behavior trajectories, such as "rub two hands" and "clasping".

Figures 6 and 7 are the confusion matrix under the two test protocols in the PKU-MMD dataset. The PKU-MMD is a long video sequence multi-modal dataset. Each video sequence contains a series of actions, which needs to be preceded by segmentation. We can see from Figures 6 and 7 that the proposed ILD-ConvNet can distinguish most human actions after segmentation. However, some actions cannot be effectively identified, such as "eat meal/snack", "drink water", "tear up paper" and "put on glasses".

The hyper parameter λ controls the weights of the previous convolution layers in ILD-ConvNet. Adjusting λ should improve the recognition performance of the proposed ILD-ConvNet. We first roughly get the range of λ through analyzing the contribution of different item in Equations (6) and (7). The higher λ forces the RGB and depth streams to obtain more interactive features from different modalities. However, more noise will be introduced at the same time, which will affect the recognition performance of the proposed network model. Therefore, we should select the optimal λ through verifying the results in practice.

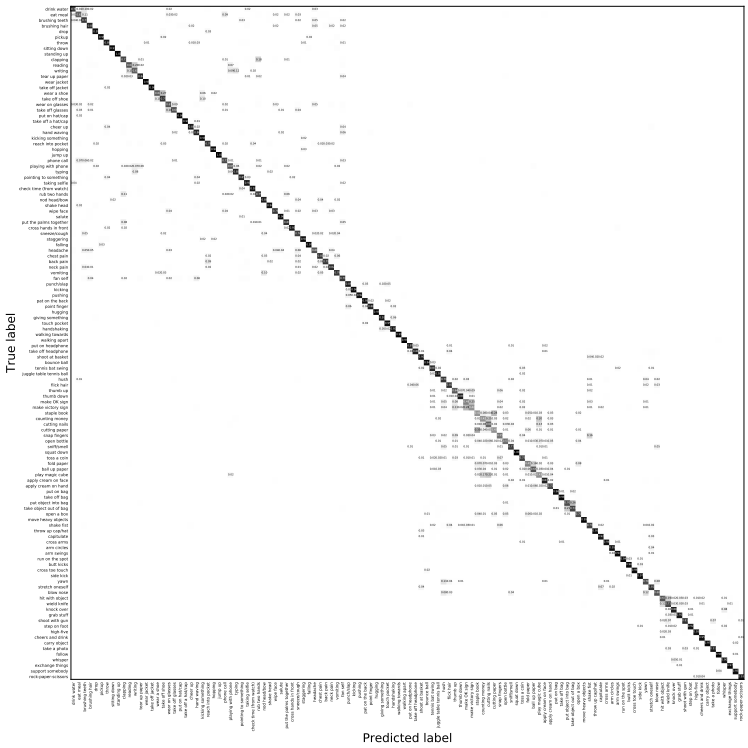

Figure 4. Confusion matrix of the ILD-ConvNet$_{101}$ using C-Sub protocol on NTU RGB + D 120.

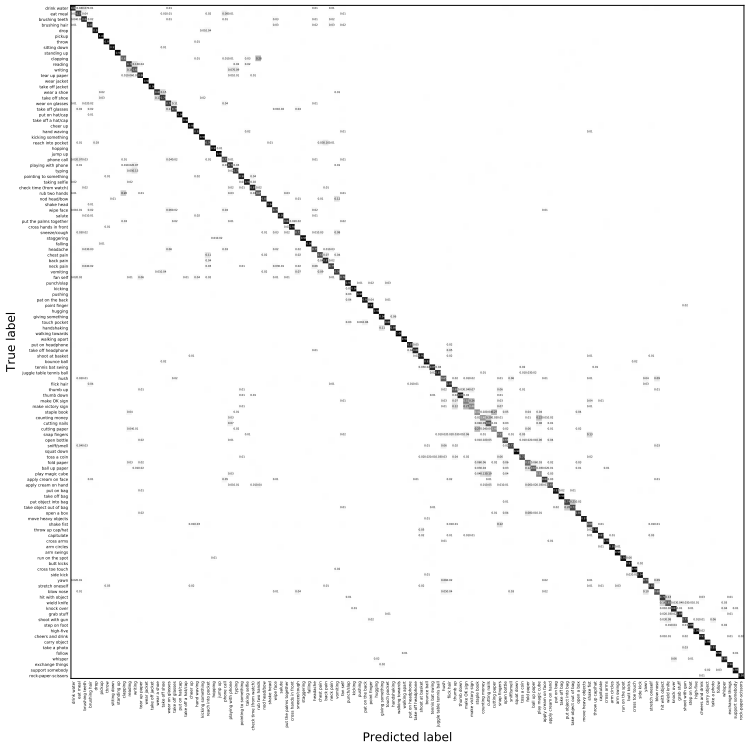

Figure 5. Confusion matrix of the ILD-ConvNet$_{101}$ using C-Set protocol on NTU RGB + D 120.

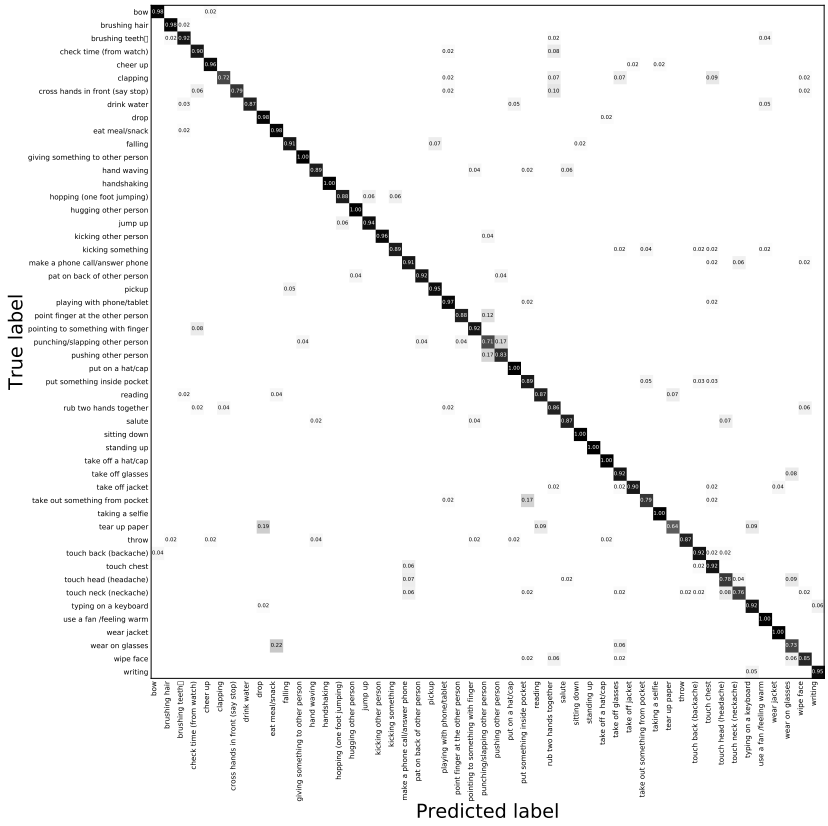

Figure 6. Confusion matrix using the C-Sub protocol on the PKU-MMD dataset.

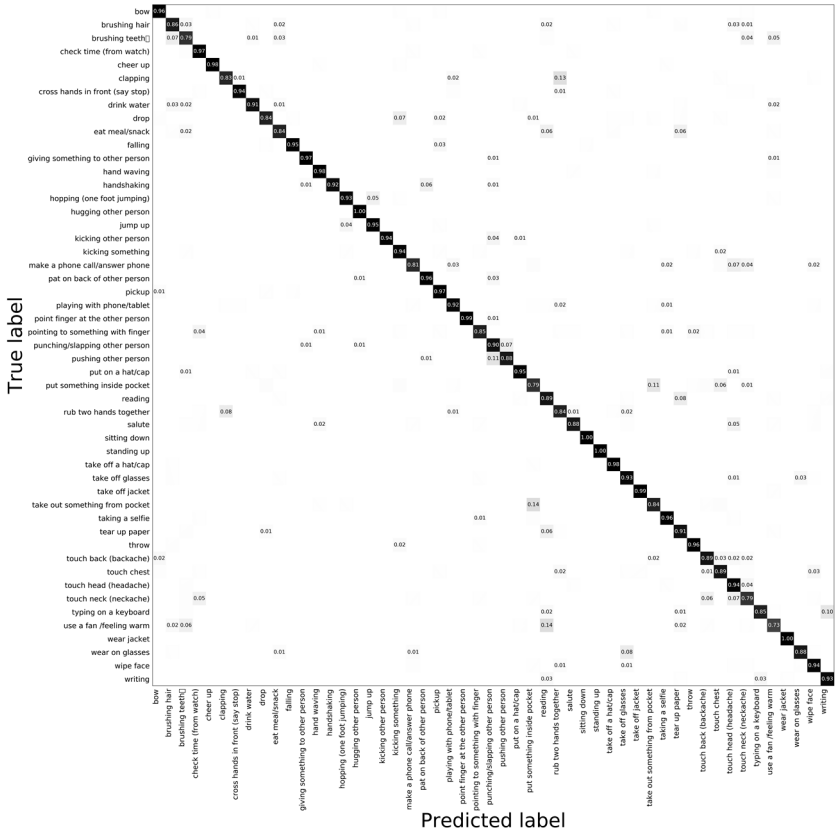

Figure 7. Confusion matrix using the C-View protocol on the PKU-MMD dataset.

5. Conclusions

A human action recognition method based on interactive learning of multi-modal features is proposed in this paper. The rank pooling method is used to construct $\langle VDI, DDI \rangle$ to obtain spatial-temporal information, and a dual-stream framework, ILD-ConvNet, with multiplexer modules is then employed to learn the interactive features from the RGB-D modalities for improving recognition performance. The experimental results illustrate the effectiveness of the rank pooling method and a multiplexer module. Compared with the state-of-the-art methods, our proposed recognition framework achieves comparable results on two multi-modal RGB-D datasets. The future research of this work is to extend the 2D ConvNet framework to 3D, and explore more effective feature fusion methods and reduce the number of model parameters.

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

Funding: This work was supported by the Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China (No. 2018AAA0101301), the National Natural Science Foundation of China (No. 61972090), the Key Projects of Artificial Intelligence of High School in Guangdong Province (No. 2019KZDZX1011), Innovation Project of High School in Guangdong Province (No. 2018KTSCX314), Dongguan Science and Technology Special Commissioner Project (No. 20221800500362), and Dongguan Social Development Science and Technology Project (No. 20211800904722).

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank everyone who contributed to this article.

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

References

1. Liu, Y.; Lu, Z.; Li, J.; Yang, T.; Yao, C. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Trans. Image Process. TIP* **2020**, *29*, 3168–3182. [CrossRef] [PubMed]
2. Zhang, J.; Shen, F.; Xu, X.; Shen, H.T. Temporal reasoning graph for activity recognition. *IEEE Trans. Image Process. TIP* **2020**, *29*, 5491–5506. [CrossRef] [PubMed]
3. Rao, H.; Wang, S.; Hu, X.; Tan, M.; Guo, Y.; Cheng, J.; Liu, X.; Hu, B. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell. TPAMI* **2021**, *44*, 6649–6666. [CrossRef] [PubMed]
4. Zhang, L.; Zhang, S.; Jiang, F.; Qi, Y.; Zhang, J.; Guo, Y.; Zhou, H. BoMW: Bag of manifold words for one-shot learning gesture recognition from Kinect. *IEEE Trans. Circuits Syst. Vid. Technol. TCSVT* **2017**, *28*, 2562–2573. [CrossRef]
5. Ji, X.; Zhao, Q.; Cheng, J.; Ma, C. Exploiting spatio-temporal representation for 3D human action recognition from depth map sequences. *Knowl.-Based Syst.* **2021**, *227*, 107040.
6. Ji, Y.; Yang, Y.; Shen, F.; Shen, H.T.; Zheng, W.-S. Arbitrary-View Human Action Recognition: A Varying-view RGB-D Action Dataset. *IEEE Trans. Circuits Syst. Video Technol. TCSVT* **2021**, *31*, 289–300. [CrossRef]
7. Lin, J.; Gan, C.; Han, S. TSM: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7083–7093.
8. Li, Y.; Ji, B.; Shi, X.; Zhang, J.; Kang, B.; Wang, L. TEA: Temporal excitation and aggregation for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 909–918.
9. Song, X.; Lan, C.; Zeng, W.; Xing, J.; Sun, X.; Yang, J. Temporal–Spatial Mapping for Action Recognition. *IEEE Trans. Circuits Syst. Video Technol. TCSVT* **2020**, *30*, 748–759. [CrossRef]
10. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slow-fast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6202–6211.
11. Song, Y.-F.; Zhang, Z.; Shan, C.; Wang, L. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Trans. Circuits Syst. Vid. Technol. TCSVT* **2021**, *31*, 1915–1925.
12. Ijjina, E.P.; Chalavadi, K.M. Human Action Recognition in RGB-D Videos using Motion Sequence Information and Deep Learning. *Pattern Recognit.* **2017**, *72*, 504–516.

13. Li, C.; Hou, Y.; Wang, P.; Li, W. Multiview-Based 3D Action Recognition using Deep Networks. *IEEE Trans. Hum.-Mach. Syst. THMS* **2019**, *49*, 95–104. [CrossRef]
14. Joze, H.R.V.; Shaban, A.; Iuzzolino, M.L.; Koishida, K. MMTM: Multimodal transfer module for CNN fusion. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13286–13296.
15. Abavisani, M.; Joze, H.R.V.; Pate, V.M. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1165–1174.
16. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell. TPAMI* **2019**, *41*, 2740–2755. [CrossRef] [PubMed]
17. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank pooling for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell. TPAMI* **2016**, *39*, 773–787. [CrossRef] [PubMed]
18. Tran, D.; Wang, H.; Feiszli, M.; Torresani, L. Video classification with channel-separated convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5551–5560.
19. Wang, H.; Song, Z.; Li, W.; Wang, P. A Hybrid Network for Large-Scale Action Recognition from RGB and Depth Modalities. *Sensors* **2020**, *20*, 3305. [CrossRef] [PubMed]
20. Wu, H.; Ma, X.; Li, Y. Spatio-temporal multimodal learning with CNNs for video action recognition. *IEEE Trans. Circuits Syst. Video Technol. TCSVT* **2022**, *32*, 1250–1261. [CrossRef]
21. Fernando, B.; Anderson, P.; Hutter, M.; Gould, S. Discriminative Hierarchical Rank Pooling for Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 1924–1932.
22. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic image networks for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 3034–3042.
23. Li, N.; Cheng, X.; Zhang, S.; Wu, Z. Realistic human action recognition by fast HOG3D and self-organization feature map. *Mach. Vis. Appl.* **2014**, *25*, 1793–1812. [CrossRef]
24. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate SHIFT. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lile, France, 6–11 July 2015; Volume 1, pp. 448–456.
25. Moghaddam, Z.; Piccardi, M. Training initialization of hidden Markov models in human action recognition. *IEEE Trans. Autom. Sci. Eng. TASE* **2014**, *11*, 394–408. [CrossRef]
26. Sempena, S.; Maulidevi, N.; Aryan, P. Human action recognition using dynamic time warping. In Proceedings of the International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia, 17–19 July 2011; pp. 1–5.
27. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Proc. Adv. Neural Inf. Process. Syst. NIPS* **2014**, *1*, 568–576.
28. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
29. Zhang, J.; Shen, F.; Xu, X.; Shen, H.T. Cooperative cross-stream network for discriminative action representation. *arXiv* **2019**, arXiv:1908.10136.
30. Wang, L.; Tong, Z.; Ji, B.; Wu, G. TDN: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 1895–1904.
31. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatio-temporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
32. Jiang, S.; Qi, Y.; Zhang, H.; Bai, Z.; Lu, X.; Wang, P. D3D: Dual 3-D Convolutional Network for Real-Time Action Recognition. *IEEE Trans. Ind. Inform. TII* **2021**, *17*, 584–4593. [CrossRef]
33. Tao, L.; Wang, X.; Yamasaki, T. Rethinking Motion Representation: Residual Frames With 3D ConvNets. *IEEE Trans. Image Process. TIP* **2021**, *30*, 9231–9244. [CrossRef]
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
35. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End to end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
36. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 244–253.
37. Wang, P.; Li, W.; Gao, Z.; Tang, C.; Ogunbona, P.O. Depth pooling based large-scale 3-D action recognition with convolutional neural networks. *IEEE Trans. Multimed. TMM* **2018**, *20*, 1051–1061. [CrossRef]
38. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P.O. Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans. Hum.-Mach. Syst.* **2016**, *46*, 498–509. [CrossRef]

39. Qi, M.; Wang, Y.; Qin, J.; Li, A.; Luo, J.; Van Gool, L. StagNet: An attentive semantic RNN for group activity and individual action recognition. *IEEE Trans. Circuits Syst. Video Technol. TCSVT* **2020**, *30*, 549–565. [CrossRef]
40. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019; pp. 1227–1236.
41. Liu, J.; Wang, G.; Duan, L.-Y.; Abdiyeva, K.; Kot, A.C. Skeleton based human action recognition with global context-aware attention LSTM networks. *IEEE Trans. Image Process. TIP* **2018**, *27*, 1586–1599. [CrossRef]
42. Wang, P.; Li, W.; Wan, J.; Ogunbona, P.; Liu, X. Cooperative Training of Deep Aggregation Networks for RGB-D Action Recognition. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018; pp. 7404–7411.
43. Shahroudy, A.; Liu, J.; Ng, T.; Wang, G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
44. Ren, Z.; Zhang, Q.; Cheng, J.; Hao, F.; Gao, X. Segment spatial-temporal representation and cooperative learning of Convolution Neural Networks for multimodal-based action recognition. *Neurocomputing* **2021**, *433*, 142–153. [CrossRef]
45. Cheng, J.; Ren, Z.; Zhang, Q.; Gao, X.; Hao, F. Cross-Modality Compensation Convolutional Neural Networks for RGB-D Action Recognition. *IEEE Trans. Circuits Syst. Video Technol. TCSVT* **2022**, *32*, 1498–1509. [CrossRef]
46. Song, S.; Liu, J.; Li, Y.; Guo, Z. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Trans. Image Process. TIP* **2020**, *29*, 3957–3969. [CrossRef]
47. Xu, W.; Wu, M.; Zhao, M.; Xia, T. Fusion of skeleton and RGB features for RGB-D human action recognition. *IEEE Sens. J.* **2021**, *21*, 19157–19164.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Statist. Comput.* **2004**, *14*, 199–222. [CrossRef]
50. Liu, J.; Shahroudy, A.; Perez, M.L.; Wang, G.; Duan, L.; Chichung, A.K. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell. TPAMI* **2019**, *42*, 2684–2701. [CrossRef]
51. Liu, C.; Hu, Y.; Li, Y.; Song, S.; Liu, J. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv* **2017**, arXiv:1703.07475.
52. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Trans. Image Process. TIP* **2018**, *27*, 3459–3471. [CrossRef] [PubMed]