mathematics

MDPI

Article

SlowFast Multimodality Compensation Fusion Swin Transformer Networks for RGB-D Action Recognition

Xiongjiang Xiao [1], Ziliang Ren [1,*], Huan Li [1], Wenhong Wei [1], Zhiyong Yang [2] and Huaide Yang [3]

[1] School of Computer Science and Technology, Dongguan University of Technology, Dongguan 523820, China; xxiongjiang@outlook.com (X.X.); lihuan@dgut.edu.cn (H.L.); weiwh@dgut.edu.cn (W.W.)
[2] School of Artificial Intelligence, Yantai Institute of Technology, Yantai 264003, China; yzy913@sina.com
[3] School of Electronic Information, Dongguan Polytechnic, Dongguan 523109, China; yanghd@dgpt.edu.cn
* Correspondence: renzl@dgut.edu.cn

Abstract: RGB-D-based technology combines the advantages of RGB and depth sequences which can effectively recognize human actions in different environments. However, the spatio-temporal information between different modalities is difficult to effectively learn from each other. To enhance the information exchange between different modalities, we introduce a SlowFast multimodality compensation block (SFMCB) which is designed to extract compensation features. Concretely, the SFMCB fuses features from two independent pathways with different frame rates into a single convolutional neural network to achieve performance gains for the model. Furthermore, we explore two fusion schemes to combine the feature from two independent pathways with different frame rates. To facilitate the learning of features from independent multiple pathways, multiple loss functions are utilized for joint optimization. To evaluate the effectiveness of our proposed architecture, we conducted experiments on four challenging datasets: NTU RGB+D 60, NTU RGB+D 120, THU-READ, and PKU-MMD. Experimental results demonstrate the effectiveness of our proposed model, which utilizes the SFMCB mechanism to capture complementary features of multimodal inputs.

Keywords: action recognition; multimodality compensation; SlowFast pathways; swin transformer; dual-stream

MSC: 68T07

Citation: Xiao, X.; Ren, Z.; Li, H.; Wei, W.; Yang, Z.; Yang, H. SlowFast Multimodality Compensation Fusion Swin Transformer Networks for RGB-D Action Recognition. *Mathematics* **2023**, *11*, 2115. https://doi.org/10.3390/math11092115

Academic Editor: Jakub Nalepa

Received: 25 March 2023
Revised: 23 April 2023
Accepted: 26 April 2023
Published: 29 April 2023

1. Introduction

Human action recognition is widely used in computer vision research, such as intelligent video surveillance, intelligent human–computer interaction, robot control, video retrieval, pose estimation, and many other fields [1–3]. Since the environments faced by human action recognition are diverse and complex, capturing effective features for action recognition is still a challenging problem. Recently, several works focused on exploiting the complementary information provided by RGB and depth [4–6] models have made considerable progress.

Human action recognition in RGB videos has been extensively studied in the past decades. In the early years, the method was to manually extract the behavioral features in the video that can represent the temporal and spatial changes of human action. There are mainly methods involving spatiotemporal volume [7,8], spatiotemporal interest points (STIP) [9] and trajectory [10,11]. Deep learning methods have a powerful ability to learn and analyze under complex network structures, which has made them become the mainstream of current human behavior research. Two-stream-based networks [12] can capture different types of information from different input models with little computational cost, but it is difficult to learn complete action sequences. Several works [13,14] have tried dense temporal sampling to learn more motion information, but sampling frames may generate high computational costs. Extracting larger features from the spatiotemporal dimension,

the 3D convolutional neural network (CNN)-based approach achieves better performance in human behavior recognition. Its drawback is the high number of parameters and computational complexity. Transformer-based methods can achieve remarkable results in the global connection mode, but these methods rely on the pre-training of large-scale datasets and train with many parameters. Over the years, numerous studies have proposed the integration of RGB, depth, and skeleton modalities for accurate human action recognition, with works including those by [4,15–19]. Decision-level fusion methods have shown promising results that aim to capture the unique features of each modality independently and combine them to produce the final classification score. However, this method does not use neural networks to learn features from each other and cannot capture complementary information. Based on the good performance of two-stream structure and multimodal learning strategies, we propose a SlowFast multimodality compensation network for RGB-D action recognition. The framework consists of two separate networks and SFMCB to capture spatial semantics and motion information. To learn the rich color and texture information in RGB video, the fast pathway runs at high frame rates and captures motion with fine temporal resolution. In addition, the fast pathway makes it lightweight by reducing the number of channels to reduce the complexity of the network model for video recognition. The slow pathway runs at low frame rates to capture spatial semantics and distance information on depth sequences. SFMCB aims to extract compensatory and more discriminative features from fast pathway and slow pathway source data for action recognition.

The main contributions of this paper are summarized as follows:

- An efficient SlowFast multimodality compensation framework is proposed to learn complementary features. Our framework uses a single swin transformer network [20] to learn spatio-temporal features from RGB and depth sequences separately. To enhance the generalization ability of the model, the extracted features are fed into SFMCB, which effectively learns complementary features;
- We explore two fusion schemes to combine the feature from two independent pathways with different frame rates. To facilitate the learning of features from independent multiple pathways, multiple loss functions are utilized for joint optimization in the SlowFast Compensation Networks. To ensure comprehensive analysis and evaluation, we validate our proposed architecture on multiple RGB-D datasets.

The rest of this paper is organized as follows: In Section 2, we provide a brief overview of some related work. Section 3 provides an overview of our proposed framework and presents the details of the architecture. In Section 4, we present our experimental results and analyze the performance of the model. Finally, we summarize our work and draw conclusions in Section 5.

2. Related Work

At present, human behavior recognition is one of the hotspots in the field of computer vision, and its working process is mainly divided into two parts: feature extraction and behavior recognition. Feature extraction needs to extract features which represent the key information from the video, and its features directly have a decisive impact on the recognition effect. This paper briefly introduces human action recognition methods for single-modality action recognition, multimodality action recognition, and attention mechanism of video understanding.

2.1. Single-Modality Action Recognition

In videos, traditional methods have limitations which extract spatio-temporal behavioral features through manual methods, such as Histogram of Oriented Gradient (HOG) [8] and STIP [9]. In recent years, methods of deep learning have been able to learn and distinguish human action features from raw video frames, exhibiting superior representation capabilities and powerful performance. To address the challenge of capturing motion information, Ref. [21] proposed a temporal template that computes frame-to-frame differences to capture the entire motion sequence. Ref. [12] introduced a two-stream CNN model with

a spatial network and a temporal network. To reduce the cost of computing optical flow, Ref. [22] proposed a method that accelerates deep two-stream architecture by replacing optical flow with motion vector. Furthermore, the method of training 3D convolutional to explore spatiotemporal features has attracted considerable attention. In order to simultaneously understand the spatio-temporal features in videos, Ref. [23] pioneered the use of 3D convolutional networks for simultaneous spatio-temporal feature learning in action recognition. In particular, C3D [24] developed an end-to-end framework that effectively adapts to deep 3D convolutional networks for spatio-temporal feature learning from raw videos. However, C3D ignores the long-term dependence of video on spatio-temporal and performs poorly on standard benchmarks. A Long-term Temporal Convolution (LTC) is proposed in [14] to build the long-term temporal structure by reducing the spatial resolution and increasing the temporal extent of the 3D convolutional layers. Ref. [25] proposed a nolocal operation that captures long-term dependencies, which refers to modeling the correlation between any two locations in a feature map.

2.2. Multimodality Action Recognition

In complex scenes, single-modal action recognition has limitations. In order to improve the accuracy and robustness of human behavior recognition, multi-modal information is combined to learn complementary features. The two-stream structure proposed by [12] is a solution to the problem of insufficient information caused by a single modality. This framework consists of a spatial network and a temporal network, which are combined to obtain the final result by merging the prediction scores. Another approach to addressing the limitations of single-modal action recognition was proposed by [26] that developed a two-stream architecture which incorporated low-resolution RGB frames and high-resolution center crops. Since then, researchers have continued to build on these classic two-stream frameworks, exploring new ways to improve their performance. Ref. [27] proposed a temporal segment network (TSN), which performs sparse temporal sampling of the video, and fuses the classification scores of the segments. Depth sequences and RGB are treated as a single entity in [28], and scene flow information is being extracted from them. First, dynamic images are generated from feature sequences, then different dynamic images are input into two different convolutional neural networks, and finally, their classification scores are fused for human action recognition. To explore complementary information, jointly training multiple networks has attracted considerable attention. Ref. [16] improved performance from videos by converting RGB and depth sequences into two pairs of dynamic images (one pair for RGB and one pair for depth), and jointly training a single neural network to recognize human actions using both types of dynamic images. A Modality Compensation Network (MCN) is proposed [18] to explore common features between different modalities. In order to facilitate mutual learning of features extracted from dynamic images of multiple modalities, Ref. [29] proposed a novel Segment Cooperative ConvNet (SC-ConvNet), which utilizes a rank pooling mechanism [30] to construct these features. In another work, Ref. [31] introduced a cross-modality compensation block (CMCB) that improved the interaction between different modalities by jointly learning compensation features. To improve the performance of human action recognition, 3D convolutional models with two-stream or multi-stream designs are studied [32–35]. The work of [32] designs a novel 3D convolutional two-stream network, which is an Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation. Feichtenhofer [35] introduced a two-stream 3D convolutional framework, which comprises a slow pathway for low frame rate and a fast pathway for high frame rate to capture semantic and motion information.

2.3. Attention Mechanism for Video Understanding

In recent years, attention-based neural networks to explore video understanding work have attracted considerable attention, such as person reidentification [36] and video object segmentation [37]. The work of [38] proposes using the Transformer model, originally designed for natural language processing, for image recognition tasks. To address the

limitation of handling long token sequences in videos, Ref. [39] proposed a transformer-based approach which decomposes the different components of the transformer encoder along the temporal and spatial dimensions. In order to improve attention efficiency, Ref. [40] proposes a novel directed-attention mechanism to understand human actions in exact order. A trajectory attention block (trajectory attention block) is proposed [41] to enhance the robustness of human action recognition in dynamic scenes which generates a set of specific trajectory markers along the spatial dimension and performs pooling operations along the temporal dimension. The work of [42] proposes a multi-view transformer for video recognition that laterally connects multiple encoders to efficiently fuse mutual information from different features within the video. A self-supervised Video Transformer (SVT) is designed by [43], which allows learning cross-view information and the dependencies between motion from video clips of different spatial extents and frame intervals. In addition, a lot of work has proposed effective methods for the memory and computational overhead issues in Transformer-based action recognition. To reduce the memory and computation constraints, a Shifted Chunk Transformer (SCT) was designed by [44], which involves dividing each frame into several local patches and inputting them into the image blocks of Locality-Sensitive Hashing (LSH). A Recurrent Visual Transformer (RViT) was proposed by [45] to reduce memory, which utilizes an attention gate mechanism and is operated in a recurrent manner.

3. Slowfast Multimodality Compensation Block

Our network structure consists of two separate slow and fast pathway networks of swin transformers, which are designed to capture finer spatio-temporal features in RGB-D modalities. A SlowFast multimodality compensation block fused with a swin transformer network is proposed to learn compensated information from RGB and depth modalities. With this approach, the network can enhance the robustness of the introduced slow and fast pathway networks for action recognition. The proposed framework for human recognition contains two important components, as illustrated in Figure 1: independent Swin-S and Swin-B network designs in the upper and lower parts, and the SlowFast multimodality compensation block in the middle.

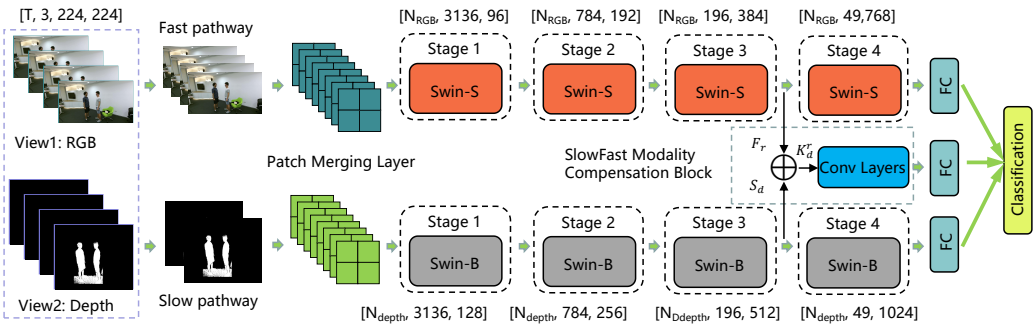


Figure 1. SlowFast multimodality. Compensation fusion swin transformer. Network has two separate pathways and a SlowFast multimodality Compensation block. The slow pathway learns 2D spatiotemporal features in deep images at a low frame rate. The fast pathway captures color and texture information in RGB images at high frame rates. The SlowFast multimodality Compensation block learns compensation information from RGB and depth images by using laterally connected fusion. We use F_r, S_d and K_r^d to describe the processing of information for RGB, depth, and SlowFast multimodality compensation, respectively. Swin-S illustrates a small version of the swin transformer. Swin-B illustrates a base version of the swin transformer.

3.1. Baseline Mode

Based on the dual-stream framework, two neural networks operating at different frame rates are designed to capture spatio-temporal features with different properties from the entire RGB and depth sequence. The proposed two-stream architecture for learning spatio-temporal information consists of two horizontal and independent pathways: one pathway

processes depth sequences of low frame rate through the Swin-S network; the other pathway processes RGB sequences of high frame rate through the Swin-B network. In addition, SlowFast multimodality compensation block is designed to fuse data from two independent paths with different frame rates into a single convolutional neural network, with the aim of achieving higher performance gains for the model. The complete structure of the proposed network can be seen in Figure 1.

The slow path processes the depth dynamic images of segment τ, while weakening its temporal modeling ability and enhancing its spatial modeling ability. The fast pathway processes a dynamic sequence from $\lambda \times \tau$ RGB fragments, where $\lambda > 1$ represents the frame rate ratio between the fast and slow pathways. In our experiments, a typical value of $\lambda = 2$ is used. The fast pathway has a higher input frame rate and utilizes a significantly lower channel capacity to achieve good performance for the fast pathway model. An example instantiation of the dual-stream SlowFast Swin Transformer network is shown in Table 1, where the temporal-spatial sizes are denoted by $T \times (H \times W)$, with T representing the length of the temporal dimension, and $H \times W$ representing the height and width of a video frame. As shown in Table 1, the slow pathway of $T = 4$ frames is sparsely sampled from the 64 frames raw clip and converted to 128 channels as the input of the network. Fast pathway with $T = 8$ frames and converted to 96 channels as the input of the network.

For the two pathways we proposed, which are independent of each other, one pathway does not learn the representations captured by the other pathway. To build the complementary information of the two pathways to be fused, we design two concatenation methods which enable matching of feature sizes before fusion. The characteristic shape of the slow pathway is denoted as $\{T, H, W, C\}$, and the characteristic shape of the fast pathway is denoted as $\{\lambda T, H, W, \mu C\}$. We constructed the following two transformation methods:

- Time-to-depth: We split and join $\{\lambda T, H, W, \mu C\}$ into $\{T, H, W, \lambda\mu C\}$, meaning that we split all time dimensions into λ parts, and then concatenate at the channel;
- Time-Step Sampling: We strictly sample at time step λ, so $\{\lambda T, H, W, \mu C\}$ becomes $\{T, H, W, \mu C\}$.

Table 1. An illustrative example that demonstrates the implementation of the SlowFast multimodality compensation fusion swin transformer networks. The output sizes are denoted by $T \times (H \times W)$ for temporal and spatial resolutions. Here, the frame rate ratio is $\lambda = 2$ and the frames of slow pathway τ is 4. Swin transformer blocks are shown by parentheses. *S*: slow pathway. *F*: fast pathway. Output sizes: $T \times (H \times W)$.

Stage	Slow Pathway	Fast Pathway	Slow Output	Fast Output
raw	-	-	$S : 64 \times (224 \times 224)$	$F : 64 \times (224 \times 224)$
data layer	-	-	$S : 4 \times (224 \times 224)$	$F : 8 \times (224 \times 224)$
conv	4×4, 128, stride 4×4	4×4, 96, stride 4×4	$S : 4 \times (56 \times 56)$	$F : 8 \times (56 \times 56)$
stage 1	(dim = 128, head = 4)	(dim = 96, head = 3)	$S : 4 \times (56 \times 56)$	$F : 8 \times (56 \times 56)$
stage 2	(dim = 256, head = 8)	(dim = 192, head = 6)	$S : 4 \times (28 \times 28)$	$F : 8 \times (56 \times 28)$
stage 3	(dim = 512, head = 16)	(dim = 384, head = 12)	$S : 4 \times (14 \times 14)$	$F : 8 \times (14 \times 14)$
stage 4	(dim = 1024, head = 32)	(dim = 768, head = 24)	$S : 4 \times (7 \times 7)$	$F : 8 \times (7 \times 7)$

3.2. SlowFast Multimodality Compensation Block

It is difficult for dual-stream swin transformer networks to realize each other's existence to capture compensating features and improve the recognition accuracy of the model. In action recognition, information of complementary has always been an effective method and a research difficulty. Motivated by the superior performance of the compensation networks, we propose a SlowFast Multimodality Compensation Block (SFMCB). On the baseline model, the features extracted in the third stage of the dual-stream swin transformer network are connected and added to the SlowFast Multimodality Compensation Block to learn complementary information.

As illustrated in Figure 2, the proposed SFMCB first collects features from two independent information streams, reshapes and transposes them into a unified pathway, and then applies convolutional layers for further learning. Specifically, SFMCB consists of two 1×1 convolutional layers and one 3×3 convolutional layer, where the first convolutional layer converts the number of channels to 1024. In addition, each convolutional layer is followed by a batch normalization $(BN(\cdot))$ layer and a $(ReLU(\cdot))$ activation function. Formally, the features F_r and S_d are captured from the Fast and Slow pathways after stage 3 which are concatenated as

$$K_d^r = F_r \bigoplus S_d, \tag{1}$$

where $\bigoplus$ represents concatenation operator we designed. Furthermore, we define the input of the network in the a 1×1 convolution layer and a 3×3 convolution layer as

$$F_{1\times1}(K) = ReLU(BN(f(W_{1\times1} \times K))), \tag{2}$$

$$F_{3\times3}(K) = ReLU(BN(f(W_{3\times3} \times K))), \tag{3}$$

where K is the feature obtained by connecting the RGB and depth modalities, $W_{1\times1}$ is a standard 1×1 convolution kernel and $W_{3\times3}$ is a standard 3×3 convolution kernel. Therefore, total parameters of the entire network framework are calculated as

$$P_{total} = P_{RGB} + P_{depth} + P_{SFMCB}, \tag{4}$$

where P_{RGB} is the Swin-S network parameter that is trained on RGB Modality, P_{depth} is the Swin-B network parameter that is trained on depth Modality, and P_{SFMCB} is the parameters of the SFMCB block.

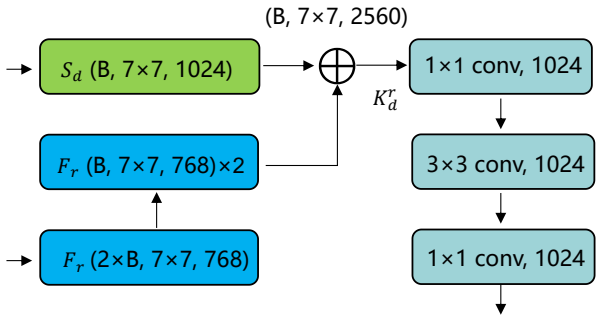


Figure 2. The structures of SFMCB. F_r and S_d represent features extracted after stage 3. $\oplus$ indicates the transformation methods which connect F_r, S_d with different channel numbers (Time-to-depth or Time-Step Sampling). Complementary features are captured with convolutional networks.

3.3. Joint Optimization and Fusion

The proposed compensation framework differs from models that solely rely on a single RGB or depth frame, as it operates on a pair of images (RGB and depth) and captures their respective features. To facilitate the learning of features from two independent pathways and capture complementary information that enhances discrimination ability, multiple loss functions are utilized for joint optimization in the SlowFast Compensation Networks. To improve the recognition effect, we apply the segmentation training strategy proposed in [27] to obtain dynamic images, and then input dynamic images of RGB and depth into the network at different frame rates. During the process of the proposed network to train, we employ cross-entropy loss function to optimize learning and class probability score is formulated as

$$pro_c = \log \frac{e^{(W_c X + b_c)}}{\sum_{c_i=1}^{C} e^{\left(w_{c_i} X + b_{c_i}\right)}}, \tag{5}$$

where weight W_c, bias b_c for the softmax layer and C represents the number of action categories. The weights and biases are optimized during the training process, with the

goal of minimizing the cross entropy loss and improving the accuracy of the classification model. The class probability scores for the SlowFast multimodality compensation fusion block (SFMCB), RGB, and depth modalities are represented by $pro_{c-SFMCB}$, pro_{c-RGB}, and $pro_{c-depth}$, respectively. Additionally, the loss functions for different modules are optimized using the following formulas:

$$L_{SFMCB}(y, C) = -\sum_{c=1}^{C} y_c(pro_{c-SFMCB}), \tag{6}$$

$$L_{RGB}(y, C) = -\sum_{c=1}^{C} y_c(pro_{c-RGB}), and \tag{7}$$

$$L_{depth}(y, C) = -\sum_{c=1}^{C} y_c\left(pro_{c-depth}\right), \tag{8}$$

where y_c is the ground-truth label for the c action.

During testing, a pair of RGB and depth sequences with different frame rates are simultaneously input into the proposed trained network. Then, the scores for each category can be obtained by compensating learning from RGB and depth modalities. Based on the analysis of various aggregation functions by [27], we choose max, average, and product for feature aggregation in our proposed architecture, resulting in the formation of the fusion vector v_{fusion}, which is denoted as follows:

$$v_{fusion} = v_{RGB} \odot v_{depth} \odot v_{SFMCB}, \tag{9}$$

where v_{RGB} and v_{depth} represent scores obtained from the independent Fast and Slow pathways using the RGB and depth modalities, while v_{SFMCB} represents the scores obtained from the complementary features learned by SFMCB, and $\odot$ is aggregation functions which are the element max, element sum, or element multiplication (Mul). To obtain the corresponding class labels, we use the score of v_{SFMCB} as the probability distribution of the test and choose the largest score as the result.

4. Experimental Results and Analysis

For evaluation, we perform experiments on the following RGB-D datasets: the NTU RGB+D 60 dataset [46], NTU RGB+D 120 dataset [47], THU-READ [48], and PKU-MMD [49]. Furthermore, we analyzed the experimental results to further probe the effectiveness of each component in our work.

4.1. Datasets

NTU RGB+D 60 is a multimodal dataset for action recognition and behavior analysis. It contains RGB and depth images captured from 3D sensors and RGB cameras, which record human behaviors for 60 action categories. Two different evaluation protocols are used to assess to generalization abilities of action recognition algorithms. For cross-subject (C-Sub), the training and testing sets have 40,320, and 16,560 samples, respectively. For cross-view (C-View), there are 37,920 and 18,960 samples in the training and testing sets, respectively.

NTU RGB+D 120 dataset is a large-scale dataset for RGB+D human action recognition, compared with the RGB+D 60 action recognition dataset version, which provides more video samples, action categories, human subjects and camera views. The dataset is jointly captured by three cameras in different orientations with 32 settings and contains 106 different subjects with a total of over 114,000 video samples and eight million frames. A total of 120 action categories are captured in this dataset, which includes daily activities, mutual activities, and health-related activities. This data provides four distinct data types: RGB, depth, skeletal joints, and infrared radiation. In the cross-subject (C-Sub) protocol, the 106 subjects were divided into separate training and testing groups, where the training group utilized samples described in [47], and the remaining samples were used for testing.

Under the cross-setup (C-Set) protocol, training and testing groups were determined based on the set ids, where samples with even set ids were utilized for training samples with odd set ids were used for testing.

PKU-MMD is a comprehensive, multimodal 3D dataset designed to facilitate a deep understanding of human behavior. The dataset encompasses 66 distinct objects and was captured from three different camera views. This dataset contains 1076 long video sequences, each video sequence contains 20 action instances of 51 action classes, and nearly 20,000 motion instances. The dataset also provides data sources for different modalities, which include depth maps, RGB images, skeletal joints, infrared sequences, and RGB videos. Action recognition was performed following a cross-subject (C-Sub) and cross-view (C-View) evaluation protocol. For C-Sub, the training and testing sets contain 18,134 samples and 2600 samples, respectively. For C-View, the training set contains videos from camera 1 and camera 3, while the test set contains videos from other cameras. The training set and test set contain 13,813 samples and 6919 samples, respectively.

THU-READ dataset record videos for egocentric action recognition, which contains 40 different actions performed by eight subjects. THU-READ defines two evaluation protocols which are cross-group (CG) and cross-subject (CS). The CG protocol categorizes video samples into three distinct groups based on the number of times each action is performed, with one group allocated for training and the remaining groups used for testing. Conversely, the CS protocol splits video samples into four groups based on the subjects featured in the footage, with three groups designated for training and the fourth used for testing. We use the two protocols described in [48] and calculated the recognition accuracy on all groups and splits.

4.2. Implementation Details

Network Inputs: To reduce redundancy, we resize both RGB and Depth sequences to 224×224 during model training. In the RGB modality, certain frames are selected from the original video to construct a dynamic image. Similarly, in the depth modality, the dynamic image is obtained by applying the time span λ based on the range of the RGB image, and is discretized into intervals within the range of $[0, 255]$ via linear transformation. To further augment the training samples, random cropping and horizontal flipping are applied to RGB and Depth sequences. Unless otherwise specified, the parameter τ is 4 and λ is fixed to 2.

Model Training: During training, stochastic gradient descent (SGD) was chosen as the optimizer with weight decay and momentum set to 0.0001 and 0.9, respectively. For the learning rate, an initial value of 0.001 was set and a Step Scheduler was used to decrease it. For THU-READ and PKU-MMD, the preset step number was set to 4 with a learning rate reduction ratio of 0.6. For NTU RGB+D 60 and NTU RGB+D 120, the preset step number was set to 20 with a learning rate reduction ratio of 0.1. To enhance the representation of robust features and reduce the risk of overfitting, dropout was employed with a regularization ratio of 0.5.

Model Testing: Our testing is conducted in accordance with the testing protocol proposed in [27], in which different dynamic images are constructed to evaluate the proposed network. The final recognition result averages the classification accuracies across all action categories. To ensure consistency with the training phase, the input for testing is resized to 224×224 and augmented with random cropping and horizontal flipping techniques. Unless otherwise specified, the fusion method used is sum.

4.3. Efficacy of the Proposed Method

We compare the proposed method to evaluate the effectiveness of the SlowFast Multimodality Compensation network, using several RGB-D datasets for experimentation. To balance the parameters of the proposed framework, we selected the swin transformer in the following experiments. Specific configurations vary depending on the network model, and further details are provided below:

- Swin-S: The small model in the swin transformer serves as the backbone of a TSN, which is designed to process RGB sequences for action recognition in a single-modality approach;
- Swin-B: The base model in the swin transformer serves as the backbone of a TSN, which is designed to process depth sequences for action recognition in a single-modality approach;
- F-Swin: The dual-stream architecture is used to learn features from RGB and depth modalities separately, and then combine the scores for action recognition. The backbone of the dual-stream structure is Swin-S and Swin-B, respectively;
- J-Swin: Joint optimization based on Swin-S and Swin-B;
- J-Swin-SFMCB-I: J-Swin transformer with SFMCB and concatenation methods of Time-to-depth;
- J-Swin-SFMCB-II: J-Swin transformer with SFMCB and concatenation methods of Time-Step Sampling.

(1) Joint Optimization: We evaluate the impact of joint optimization by conducting experiments to compare the performance of various network architectures on the PKU-MMD. Specifically, we use Swin-S with ImageNet pre-training and Swin-B with ImageNet21 pre-training in all experiments. It is worth noting that τ is set to 4 in the slow pathway and the frame rate ratio between slow and fast is set to 2. Specifically, the sampling lengths in RGB and depth modalities are 8 and 4, respectively. Table 2 shows the results.

Table 2. Conducted experiments on the PKU-MMD with RGB and depth as inputs to compare the accuracy of F-Swin and J-Swin. The notations for the header were specified as [RGB + depth], which corresponds to input with SlowFast ratio.

Model	Modality	C-Sub	C-View
Swin-S	RGB	76.56%	77.73%
Swin-B	depth	73.57%	74.26%
F-Swin	RGB + depth	80.34%	82.23%
J-Swin	[RGB + depth]	89.65%	90.05%

A weighted average of the two streams based on F-Swin is first taken, resulting in accuracies of 80.34% (C-Sub) and 82.23% (C-View), which outperform either single modality. Additionally, J-Swin method achieves approximately 9.31% and 7.82% improvements in accuracy on the two protocols, respectively. The observed enhancements serve as a proof of concept that the Joint Optimization approach can substantially boost the accuracy of human action recognition.

(2) Impact of parameter τ: The parameter τ controls the number of sampling frames of the depth sequence, and increasing τ can usually improve the spatiotemporal representation performance of depth sequences. In the conducted experiments, we varied the value of τ from 2 to 8, while keeping the same value of λ to evaluate the recognition performance. The experimental results are expressed and summarized in Figure 3. As τ increases, the required spatio-temporal features are more abundant, and the corresponding efficiency is also improved. Increase τ to obtain corresponding recognition performance is not absolute, which refers to τ being large enough and the spatiotemporal information contained in the dynamic image can be used for most action video sequences.

(3) Comparing various fusion techniques: Based on J-Swin-SFMCB-I, different commonly fusion methods are evaluated to obtain the final classification, which includes maximum, product, and mean score fusion. As shown in Table 3, we summarize the experimental results of three commonly used fusion methods on the PKU-MMD dataset and compare their performance. Compared with Max, average and product fusion can achieve more ideal results, where average fusion is used for remainder of experiments in this paper.

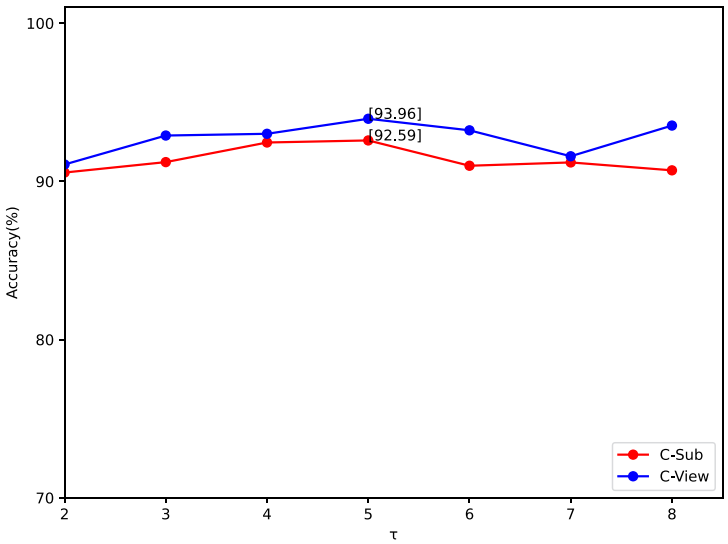
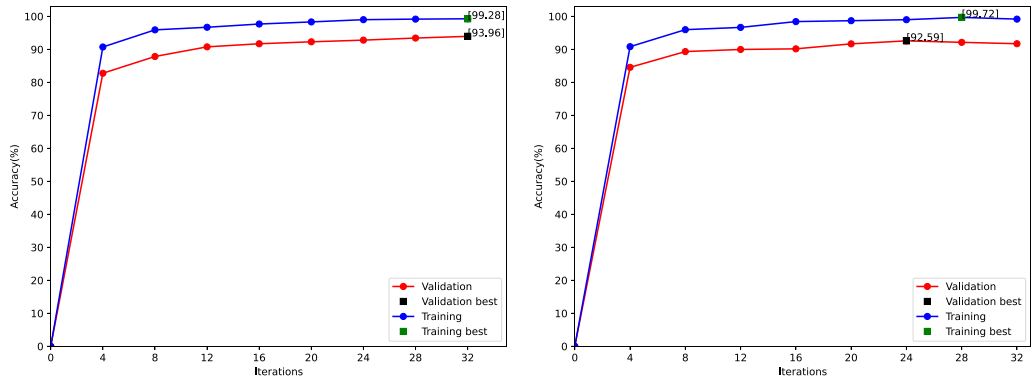


Figure 3. Compare the performance of different values of the parameter τ on the PKU-MMD dataset using J-Swin-SFMCB-II with [RGB + depth]. Specifically, the τ from 2 to 8 and keep λ constant to evaluate the recognition accuracy.

Table 3. Comparative accuracy of the proposed J-Swin-SFMCB-I using [RGB + depth] on the PKU-MMD dataset.

Fusion Methods	Modality	C-Sub	C-View
Max	[RGB + depth]	88.82%	89.41%
Product	[RGB + depth]	90.72%	91.61%
Sum	[RGB + depth]	90.52%	91.36%

(4) Convergence of J-Swin-SFMCB-II: The τ is set to 5 in the experiments. Figure 4 displays the accuracy curves of training and validation for the two evaluation protocols on the PKU-MMD using J-Swin-SFMCB-II. The model converges rapidly, with the entire training process taking approximately 32 iterations. On the C-Sub and C-View protocols, the validation accuracy is able to converge to nearly 100% and best accuracy of training obtained 92.59% and 93.96%, respectively.

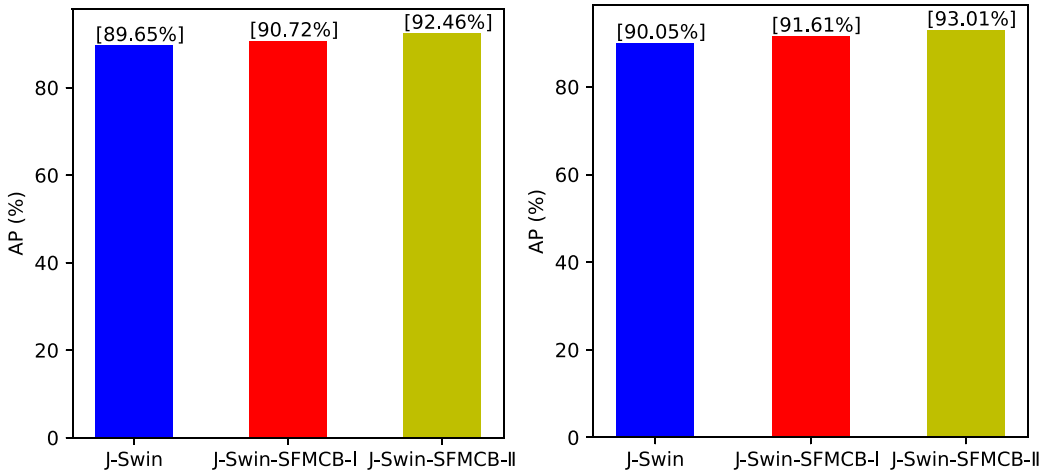


(**a**) C-Sub (**b**) C-View

Figure 4. The experiments conducted on the PKU-MMD dataset using J-Swin-SFMCB-II with [RGB + depth] as input yielded high levels of training and validation accuracy.

(5) Effectiveness from SFMCB: To interpret the benefits of SFMCB in capturing complementary features, we conduct rigorous experiments on the PKU-MMD dataset. Figure 5

expresses the experimental results. According to the analysis of the results, the proposed framework of J-Swin achieves recognition performances of 89.65%(C-Sub) and 90.05%(C-View). Meanwhile, under the work of SFMCB, the recognition performance of the framework on J-Swin-SFMCB-I has also been improved by 1.07% and 1.56%, respectively. Compared with J-Swin, the recognition accuracy of the J-Swin-SFMCB-II framework are also increased by 2.81% and 2.96%, respectively. The improvement factor can be attributed to the compensation function captured by SFMCB from the slow and fast pathways.



(**a**) C-Sub (**b**) C-View

Figure 5. Comparisons with different benefits for C-Sub and C-View protocols on the PKU-MMD dataset using the designed model with different structures.

4.4. Comparison to the State-of-the-Art

In this section, we conduct multiple experiments on the THU-READ, PKU-MMD, NTU RGB+D 60 dataset and NTU RGB+D 120 datasets to evaluate the effectiveness of the proposed framework and compare our model in optimal settings with state-of-the-art approaches with different input modalities.

The experimental comparison results on the THU-READ dataset are shown in Table 4, where the experimental parameter $\tau = 2$. We can see from Table 4 that J-Swin-SFMCB-II achieves 87.0% and 94.2% for the CS and CG protocols, respectively. The confusion matrix of the CG protocol on the THU-READ dataset is shown in Figure 6 with group 3 used for training and the remaining groups used for testing. The proposed recognition framework performs perfectly on "bounce_ball", "cut_fruit", "fetch_water", "fold", " wear_watch", etc. However, some actions did not achieve accurate prediction results, such as "open_door", "push_button", and "thumb". This is largely due to the similarity in appearance and action of the recognized objects.

Table 4. Comparing our method using [RGB + depth] with previous action recognition methods on the THU-READ dataset.

Method	Modality	CS	CG
Two-stream [12]	RGB + flow	55.1%	89.0%
J-ResNet-CMCB [31]	⟨VDI, DDI⟩	77.2%	92.3%
DSCMT [6]	⟨VDI, DDI⟩	76.6%	92.0%
J-Swin-SFMCB-I	[RGB + depth]	85.6%	93.1%
J-Swin-SFMCB-II	[RGB + depth]	87.0%	94.2%

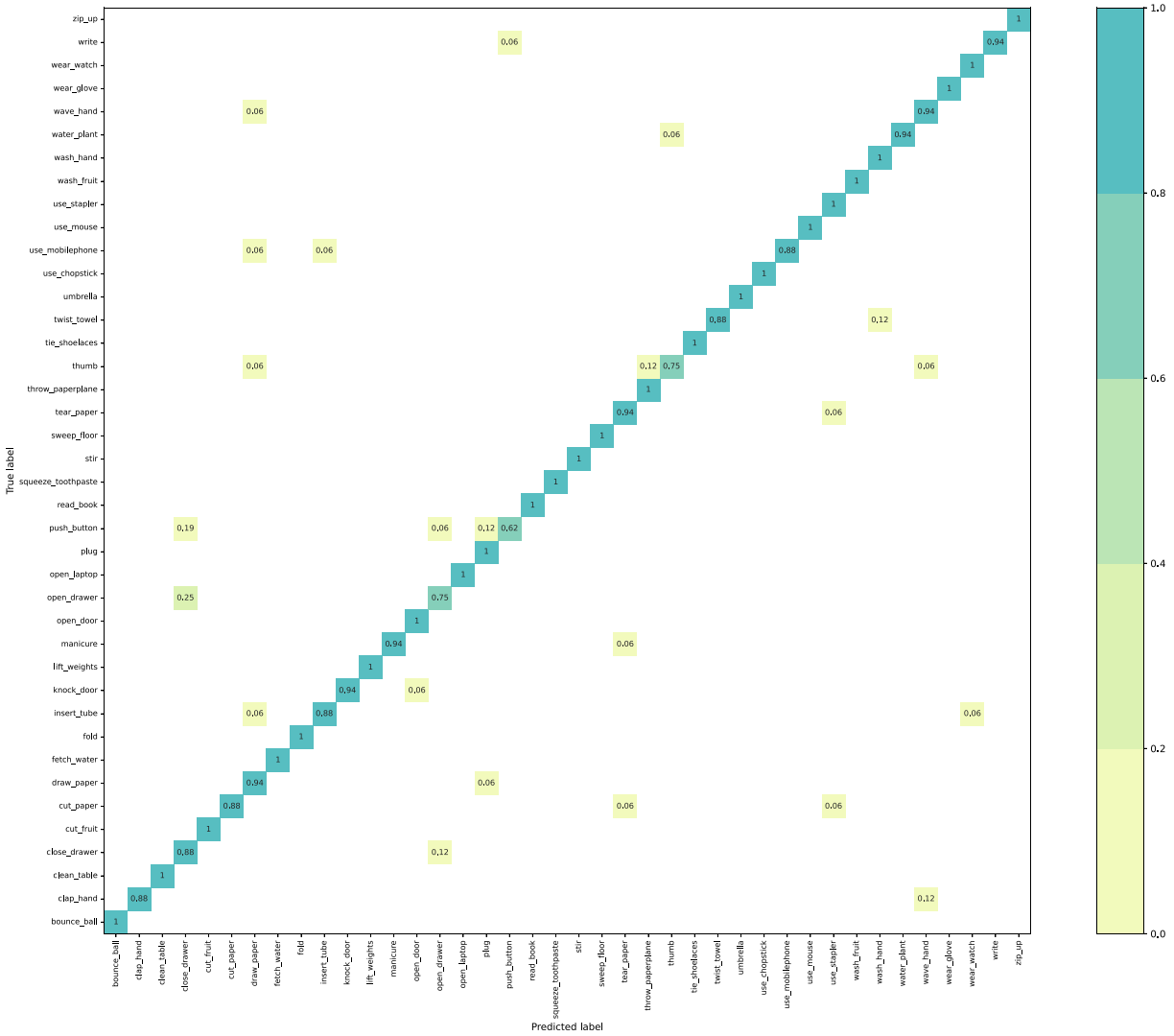


Figure 6. The confusion matrix for J-Swin-SFMCB-II was obtained through the CG protocol on the THU-READ dataset with input [RGB + depth], where group 3 was used for training and the remaining two groups were used for testing.

By considering the effect of the parameter τ on the results, we compare our model in $\tau = 5$ with several approaches, and the comparison results for the C-Sub and C-Set protocols on the PKU-MMD are shown in Table 5. Specifically, the proposed J-Swin-SFMCB-I with joint training can be observed to outperform the previous work J-ResNet-CMCB [32] by 1.2% and 0.9%, respectively. Furthermore, the proposed J-Swin-SFMCB-II model achieves recognition accuracies of 92.6% and 94.0% which outperforms all compared methods on the C-Sub and C-Set protocols, respectively.

Figure 7 displays the confusion matrices for the C-Sub protocols on the PKU-MMD dataset. The J-Swin-SFMCB-II model effectively differentiates actions with high similarity on the protocols. In particular, the C-Sub protocol achieves accurate recognition of many typical actions with 100% accuracy, such as "hand shaking", "jump up", and "brushing hair". In addition, actions such as "putting something inside a pocket", "taking something out from a pocket", "saluting", and "throwing" are often confused. This misclassification can be mainly attributed to the similarity in motion between the two interactions or high similarities in object appearance. The results demonstrate that the proposed J-Swin-SFMCB-II can distinguish actions and improve recognition performance by learning complementary information from jointly optimized networks.

Table 5. Comparing our method using [RGB + depth] with previous action recognition methods on the PKU-MMD dataset.

Method	Modality	C-Sub	C-View
TSN [27]	RGB + depth	79.3%	78.2%
Bi-LSTM [50]	skeleton	86.5%	92.2%
SA-LSTM [51]	skeleton	86.3%	91.4%
J-ResNet-CMCB [31]	⟨VDI, DDI⟩	90.4%	91.4%
DSCMT [6]	⟨VDI, DDI⟩	92.4%	93.8%
J-Swin-SFMCB-I	[RGB + depth]	91.6%	92.3%
J-Swin-SFMCB-II	[RGB + depth]	92.6%	94.0%

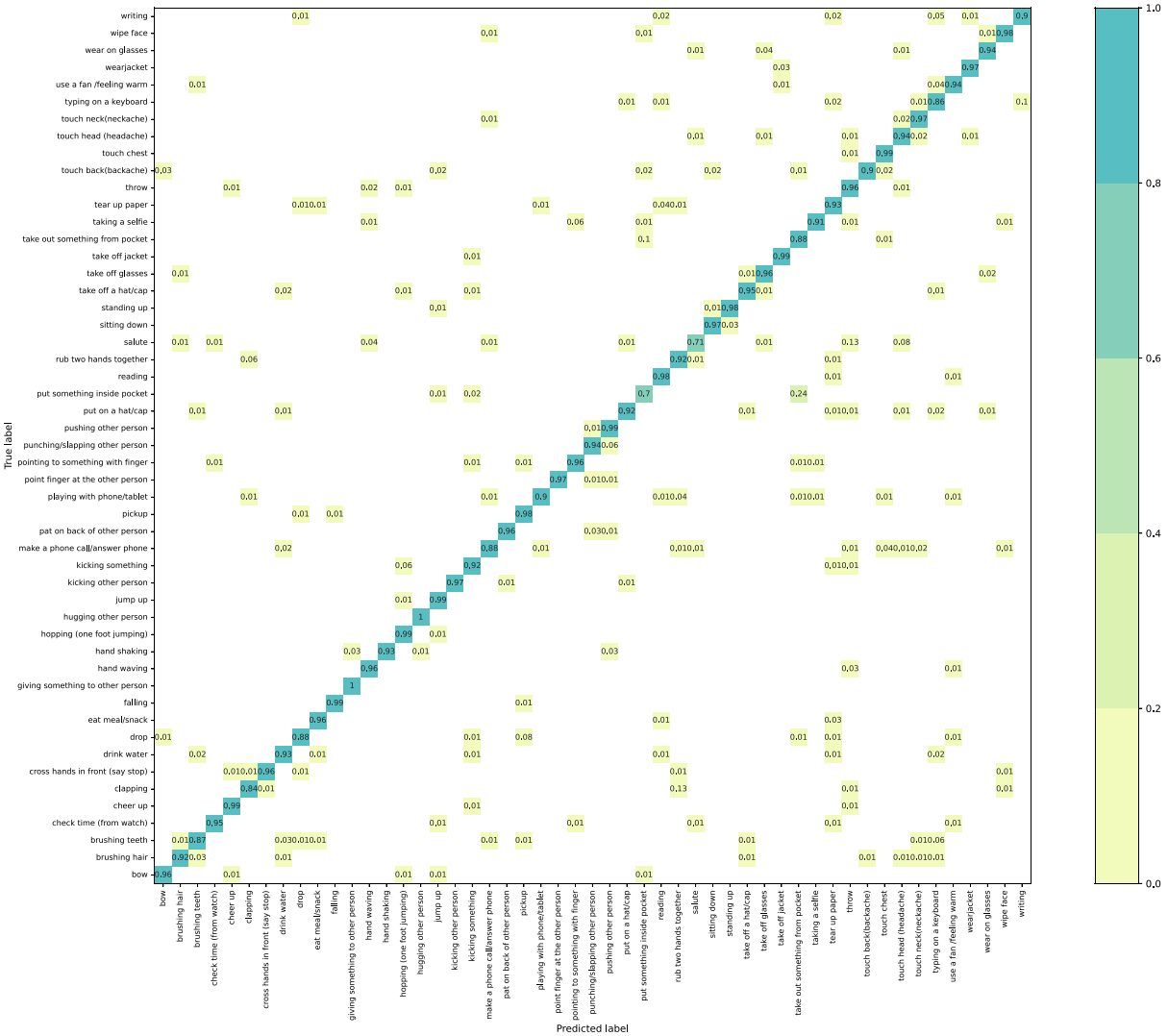


Figure 7. Confusion matrix of J-Swin-SFMCB-II obtained by C-Sub protocol on the PKU-MMD with the input [RGB + depth].

The comparison of experiments result performed on the NTU RGB+D 60 dataset is summarized in Table 6 Based on SFMCB, J-Swin-SFMCB-I achieves superior performance on the challenging NTU RGB+D 60 dataset by capturing features from the entire RGB and depth sequence. Specifically, our proposed recognition framework achieves accuracies of 90.2% and 91.5% on the C-Set and C-Sub protocols, respectively. For J-Swin-SFMCB-II, the accuracy from J-Swin-SFMCB-I is boosted by approximately 1.1% and 1.3%.

Figure 8 shows the confusion matrix matrices for C-Sub protocol on the NTU RGB+D 60 dataset. The proposed method performs perfectly in aspects such as “walking towards”,

"picking up", and "take off a hat/cap". However, it performs poorly in recognizing certain action pairs such as "rubbing two hands" and "clapping", which share very similar appearances.

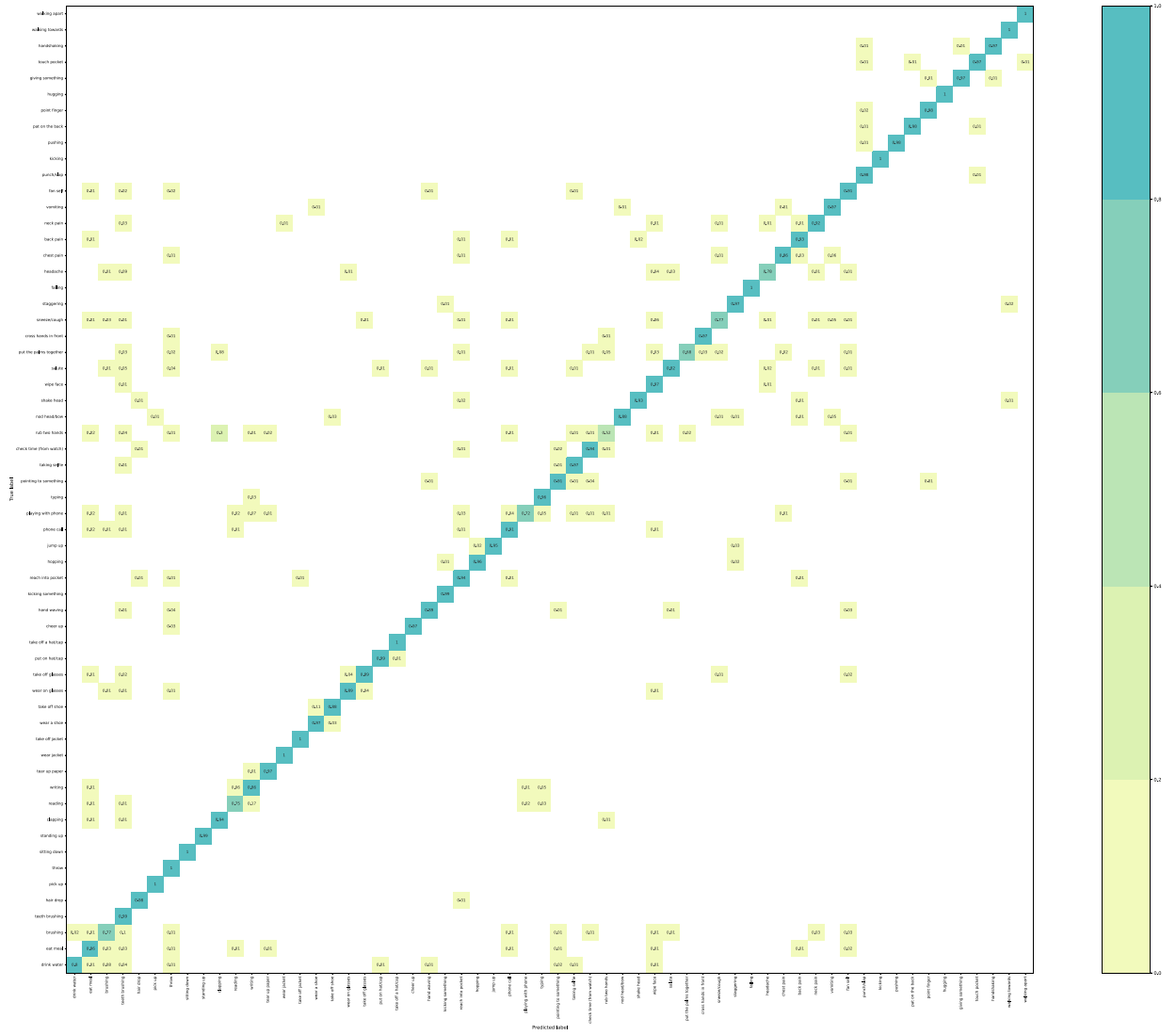


Figure 8. Confusion matrix for the C-Sub setting on NTU RGB+D 60 dataset using the J-Swin-SFMCB-II with the inputs [RGB + depth].

Table 6. Evaluating our model using [RGB + depth] as input on the NTU RGB+D 60 dataset in comparison with previous action recognition approaches.

Method	Modality	C-Sub	C-View
TSN [27]	RGB + depth	75.5%	78.1%
Deep Bilinear [52]	RGB + depth + skeleton	85.4%	90.7%
SC-ConvNets [29]	⟨VDI, DDI⟩	89.4%	91.2%
J-Swin-SFMCB-I	[RGB + depth]	90.2%	91.5%
J-Swin-SFMCB-II	[RGB + depth]	91.3%	92.8%

The comparison of experiments result performed on the NTU RGB+D 120 dataset is summarized in Table 7 and the parameter $\tau = 5$. Based on SFMCB, J-Swin-SFMCB-I achieves superior performance on the challenging NTU RGB+D 120 dataset by capturing complementary information. Specifically, our proposed recognition framework achieves

accuracies of 88.6% and 89.7% on the C-Set and C-Sub protocols, respectively. For J-Swin-SFMCB-II, the accuracy from J-Swin-SFMCB-I is boosted by approximately 0.7% and 0.6%.

Figure 9 shows the confusion matrix matrices for C-Sub protocol on the NTU RGB+D 120 dataset. The proposed method performs perfectly on "carry object", "bounce ball", "kicking something", "walking apart", etc. We can also note that "putting on glasses" and "taking off glasses", are frequently misidentified due to the high similarity between the objects in human–object interactions. Similarly, "counting money" and "folding paper" are also easily confused, as they have very similar appearances. These results suggest that the proposed dual-stream of SlowFast structure can effectively learn spatial-temporal feature and SFMCB strategy can capture the complementary information from multiple modalities.

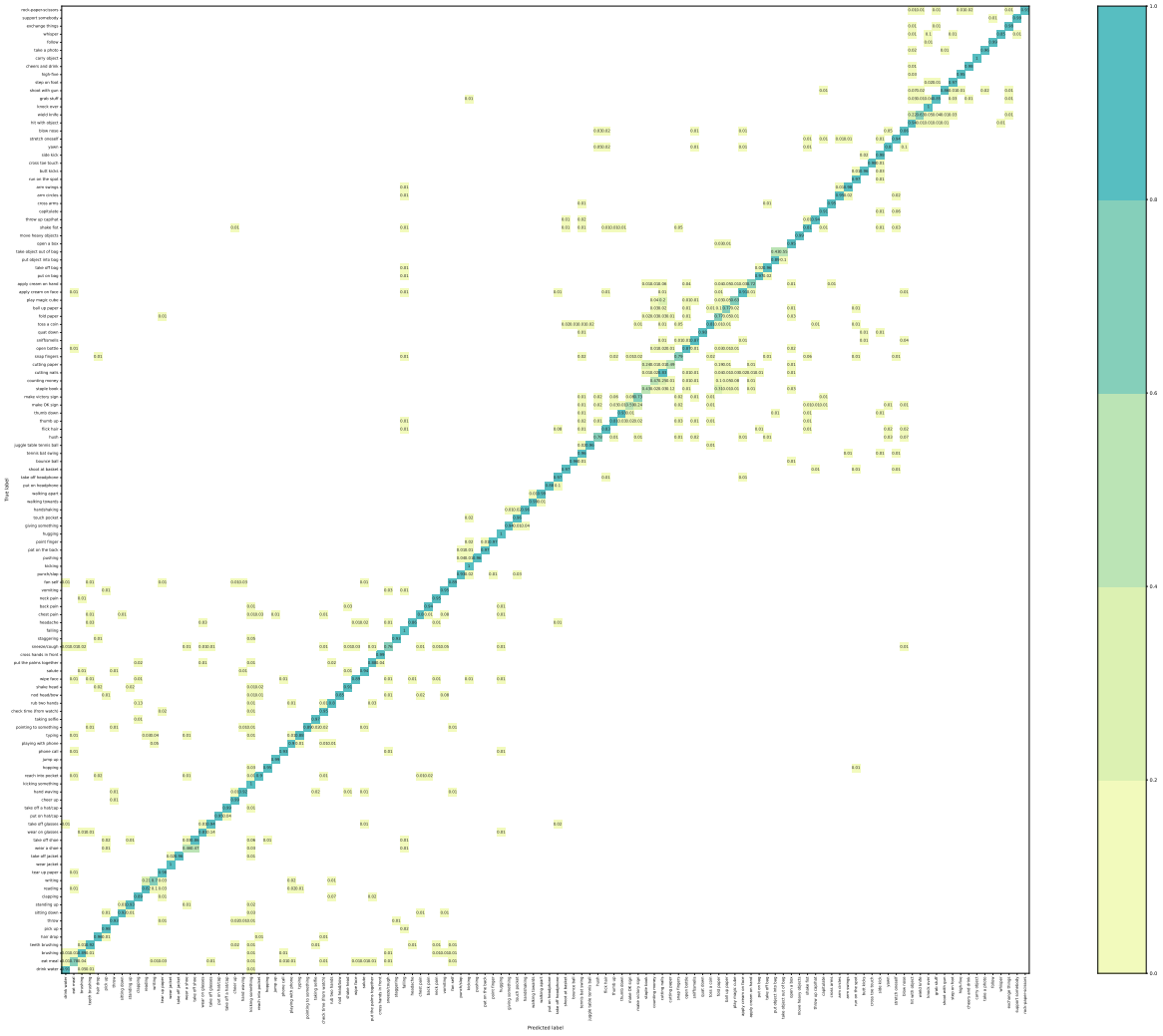


Figure 9. Confusion matrix for the C-Sub setting on NTU RGB+D 120 dataset using the J-Swin-SFMCB-II with the inputs [RGB + depth].

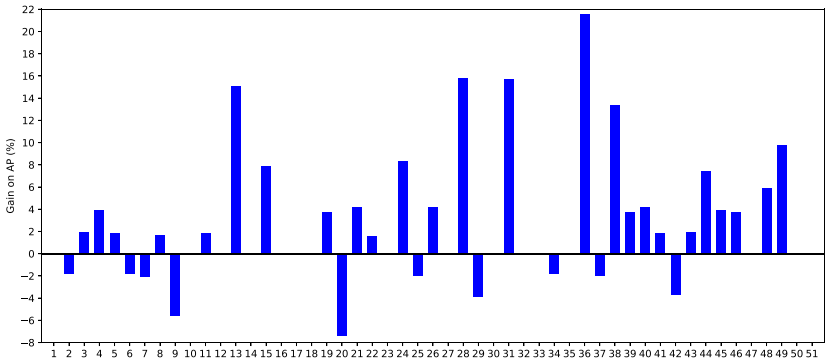
Table 7. Evaluating our model using [RGB + depth] as input on the NTU RGB+D 120 dataset in comparison with previous action recognition approaches.

Method	Modality	C-Sub	C-Set
Two-Stream [12]	RGB + depth	58.5%	54.8%
TSN + VDIs + DDIs [27]	skeleton	86.1%	86.9%
VPN [53]	skeleton	86.3%	87.8%
J-ResNet-CMCB [31]	⟨VDI, DDI⟩	82.8%	83.6%
J-Swin-SFMCB-I	[RGB + depth]	88.6%	89.7%
J-Swin-SFMCB-II	[RGB + depth]	89.3%	90.3%

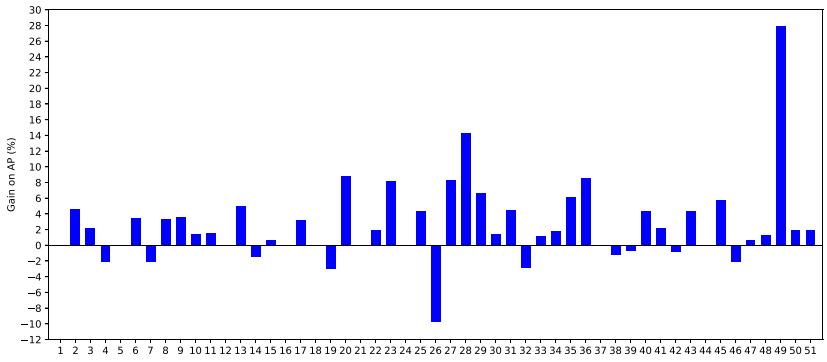
4.5. Analysis and Discussion

The proposed J-Swin-SFMCB has been extensively evaluated on multimodal datasets, outperforming current state-of-the-art RGB-D-based methods in terms of recognition performance. These results demonstrate that the dual-stream SlowFast structure can effectively capture spatio-temporal information. As seen in the analysis of the confusion matrix, our method achieves high accuracy in distinguishing most actions across different datasets of varying sizes and dimensions. While the network performs admirably in many respects, we must acknowledge that it falls short when it comes to accurately recognizing fine or subtle actions. Specifically, the proposed architecture is constrained to certain actions with very similar visual appearances and highly similar interactive behaviors. In addition, the idea behind this method is to improve performance and accuracy by fusing information from different modalities, so it can be applied to other similar problems or tasks.

Figure 10 shows the improvement of J-Swin-SFMCB-II over J-Swin with [RGB, depth] inputs on the PKU-MMD. The proposed SFMCB significantly outperforms J-Swin on most actions, indicating that it can learn compensatory features from dual-stream networks to achieve significant performance improvements. For the C-Sub protocol, the proposed method achieves more than 10% improvement on complex actions such as "waving hand" (♯13, 15.1%), "putting on a hat/hat" (♯28, 15.7%), "rubbing two hands" (♯31, 15.6%), and "Take Off Your Glasses" (♯36, 21.5%). For the C-View protocol, many actions such as "pushing other person" (♯27, 8.3%), "putting on a hat/ca" (♯28, 14.2%), and "putting on glasses" (♯49, 27.8%) achieve significant improvements.



(**a**) C-Sub



(**b**) C-View

Figure 10. The average precision gain of J-Swin-SFMCB-II compared to J-Swin with input [RGB, depth] on the PKU-MMD dataset and the horizontal axis that represents 51 action IDs provided in [49].

5. Conclusions

In this work, an efficient SlowFast Multimodality Compensation fusion swin transformer networks was proposed for RGB-D-based human action recognition. Our network

architecture consists of two separate slow and fast pathway networks of swin transformer and employs a standard cross-entropy loss function to jointly optimize learning. To construct the complementary information of the two pathways to be fused, we designed SFMCB to learn finer spatio-temporal features in RGB-D modalities. For evaluation, we conducted experiments on four RGB-D datasets and the experimental results demonstrate the robustness and effectiveness of the proposed method compared to the state-of-the-art methods. In conclusion, the proposed approach for human action recognition has both strengths and weaknesses. The method utilizes a combination of modalities can adapt to various environmental requirements and changing human behavior, thus providing greater flexibility in interpreting human actions. However, the proposed method is limited in its ability to distinguish between highly similar actions.

Author Contributions: Conceptualization, X.X. and Z.R.; methodology, X.X. and Z.R.; software, H.L. and W.W.; validation, H.L. and W.W.; formal analysis, Z.Y. and H.Y.; investigation, X.X. and Z.R.; data curation, X.X. and Z.R.; writing—original draft preparation, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of Guangdong Province (Nos. 2022A1515140119, 2023A1515011307), Dongguan Science and Technology Special Commissioner Project (Nos. 20221800500362, 20221800500572), the National Natural Science Foundation of China (Nos. 61972090, U21A20487, U1913202), Special Projects in Key Fields of Research platforms and projects of Guangdong Universities in 2022(2022ZDZX3082), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2021(ZXF016).

Data Availability Statement: Not applicable.

Acknowledgments: We are truly grateful for the efforts of each and every person who played a part in bringing this article to fruition.

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

References

1. Wang, L.; Huynh, D.Q.; Koniusz, P. A comparative review of recent kinect-based action recognition algorithms. *IEEE Trans. Image Process.* **2019**, *29*, 15–28. [CrossRef]
2. Liu, F.; Xu, X.; Qiu, S.; Qing, C.; Tao, D. Simple to complex transfer learning for action recognition. *IEEE Trans. Image Process.* **2015**, *25*, 949–960. [CrossRef]
3. Song, X.; Lan, C.; Zeng, W.; Xing, J.; Sun, X.; Yang, J. Temporal-spatial mapping for action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 748–759. [CrossRef]
4. Shahroudy, A.; Ng, T.T.; Gong, Y.; Wang, G. Deep multimodal feature analysis for action recognition in RGB+D videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1045–1058. [CrossRef]
5. Liu, Y.; Lu, Z.; Li, J.; Yang, T. Hierarchically learned view-invariant representations for cross-view action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2416–2430. [CrossRef]
6. Liu, Z.; Cheng, J.; Liu, L.; Ren, Z.; Zhang, Q.; Song, C. Dual-stream cross-modality fusion transformer for RGB-D action recognition. *Knowl.-Based Syst.* **2022**, *255*, 109741. [CrossRef]
7. Zhang, Z.; Hu, Y.; Chan, S.; Chia, L.T. Motion context: A new representation for human action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 817–829.
8. Klaser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3D-gradients. In Proceedings of the British Machine Vision Conference (BMVC), Leeds, UK, 1–4 September 2008; pp. 1–10.
9. Das Dawn, D.; Shaikh, S.H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. *Vis. Comput.* **2016**, *32*, 289–306 [CrossRef]
10. Gaidon, A.; Harchaoui, Z.; Schmid, C. Activity representation with motion hierarchies. *Int. J. Comput. Vis.* **2014**, *107*, 219–238. [CrossRef]
11. Wang, H.; Oneata, D.; Verbeek, J.; Schmid, C. A robust and efficient video representation for action recognition. *Int. J. Comput. Vis.* **2016**, *119*, 219–238. [CrossRef]
12. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.

13. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef] [PubMed]
14. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef] [PubMed]
15. Wang, L.; Tong, Z.; Ji, B.; Wu, G. TDN: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904.
16. 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 AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018; pp. 7404–7411.
17. Khaire, P.; Kumar, P.; Imran, J. Combining CNN streams of RGB-D and skeletal data for human activity recognition. *Pattern Recognit. Lett.* **2018**, *115*, 107–116. [CrossRef]
18. Song, S.; Liu, J.; Li, Y.; Guo, Z. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Trans. Image Process.* **2020**, *29*, 3957–3969. [CrossRef]
19. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
21. 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. [CrossRef]
22. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726.
23. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef]
24. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal 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.
25. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
26. 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
27. 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), Cham, The Netherlands, 11–14 October 2016; pp. 20–36.
28. Wang, P.; Li, W.; Gao, Z.; Zhang, Y.; Tang, C.; Ogunbona, P. Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 595–604.
29. 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]
30. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A. Action recognition with dynamic image networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2799–2813. [CrossRef]
31. 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.* **2021**, *32*, 1498–1509. [CrossRef]
32. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
33. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
34. Zhou, Y.; Sun, X.; Zha, Z.J.; Zeng, W. MICT: Mixed 3D/2D convolutional tube for human action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 449–458.
35. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
36. Lu, X.; Wang, W.; Shen, J.; Crandall, D.; Luo, J. Zero-shot video object segmentation with co-attention siamese networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2228–2242. [CrossRef]
37. Wu, D.; Ye, M.; Lin, G.; Gao, X.; Shen, J. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE Trans. Inf. Forensics Secur.* **2021**, *17*, 115–126. [CrossRef]

38. 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 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
39. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 6836–6846.
40. Truong, T.D.; Bui, Q.H.; Duong, C.N.; Seo, H.S.; Phung, S.L.; Li, X.; Luu, K. Direcformer: A directed attention in transformer approach to robust action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20030–20040.
41. Patrick, M.; Campbell, D.; Asano, Y.; Misra, I.; Metze, F.; Feichtenhofer, C.; Vedaldi, A.; Henriques, J.F. Keeping your eye on the ball: Trajectory attention in video transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12493–12506.
42. Yan, S.; Xiong, X.; Arnab, A.; Lu, Z.; Zhang, M.; Sun, C.; Schmid, C. Multiview transformers for video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3333–3343.
43. Ranasinghe, K.; Naseer, M.; Khan, S.; Khan, F.S.; Ryoo, M.S. Self-supervised video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2874–2884.
44. Zha, X.; Zhu, W.; Xun, L.; Yang, S.; Liu, J. Shifted chunk transformer for spatio-temporal representational learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 11384–11396.
45. Yang, J.; Dong, X.; Liu, L.; Zhang, C.; Shen, J.; Yu, D. Recurring the transformer for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14063–14073.
46. Shahroudy, A.; Liu, J.; Ng, T.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 PatternRecognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
47. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef]
48. Tang, Y.; Wang, Z.; Lu, J.; Feng, J.; Zhou, J. Multi-stream deep neural networks for RGB-D egocentric action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3001–3015. [CrossRef]
49. 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.
50. Elias, P.; Sedmidubsky, J.; Zezula, P. Understanding the gap between 2D and 3D skeleton-based action recognition. In Proceedings of the IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 192–1923.
51. 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.* **2018**, *27*, 3459–3471. [CrossRef] [PubMed]
52. Hu, J.F.; Zheng, W.S.; Pan, J.; Lai, J.; Zhang, J. Deep bilinear learning for RGB-D action recognition. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 335–351.
53. Das, S.; Sharma, S.; Dai, R.; Bremond, F.; Thonnat, M. VPN: Learning video-pose embedding for activities of daily living. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 72–90.