mathematics

MDPI

*Article*

# A Channel-Wise Spatial-Temporal Aggregation Network for Action Recognition

**Huafeng Wang [1,2,†], Tao Xia [2,†], Hanlin Li [1,*,†], Xianfeng Gu [3], Weifeng Lv [2] and Yuehai Wang [1]**

[1] School of Information Technology, North China University of Technology, Beijing 100144, China; wanghuafeng@buaa.edu.cn (H.W.); wangyuehai@ncut.edu.cn (Y.W.)
[2] School of Software, Beihang University, Beijing 100191, China; Xiatao21@buaa.edu.cn (T.X.); lwf@buaa.edu.cn (W.L.)
[3] Department of Computer Science, State University of New York at Stony Brook, New York, NY 11794, USA; gu@cs.stonybrook.edu
* Correspondence: 2019311020134@mail.ncut.edu.cn; Tel.: +86-188-1309-9160
† These authors contributed equally to this work.

**Abstract:** A very challenging task for action recognition concerns how to effectively extract and utilize the temporal and spatial information of video (especially temporal information). To date, many researchers have proposed various spatial-temporal convolution structures. Despite their success, most models are limited in further performance especially on those datasets that are highly time-dependent due to their failure to identify the fusion relationship between the spatial and temporal features inside the convolution channel. In this paper, we proposed a lightweight and efficient spatial-temporal extractor, denoted as Channel-Wise Spatial-Temporal Aggregation block (CSTA block), which could be flexibly plugged in existing 2D CNNs (denoted by CSTANet). The CSTA Block utilizes two branches to model spatial-temporal information separately. In temporal branch, It is equipped with a Motion Attention Module (MA), which is used to enhance the motion regions in a given video. Then, we introduced a Spatial-Temporal Channel Attention (STCA) module, which could aggregate spatial-temporal features of each block channel-wisely in a self-adaptive and trainable way. The final experimental results demonstrate that the proposed CSTANet achieved the state-of-the-art results on EGTEA Gaze++ and Diving48 datasets, and obtained competitive results on Something-Something V1&V2 at the less computational cost.

**Keywords:** action recognition; channel-wise; spatial-temporal; video



**Citation:** Wang, H.; Xia, T.; Li, H.; Gu, X.; Lv, W.; Wang, Y. A Channel-Wise Spatial-Temporal Aggregation Network for Action Recognition. *Mathematics* **2021**, *9*, 3226. https://doi.org/10.3390/math9243226

Academic Editors: Ezequiel López-Rubio and Ioannis G. Tsoulos

Received: 25 October 2021
Accepted: 29 November 2021
Published: 14 December 2021



## 1. Introduction

In recent years, video understanding has attracted increasing attention from the academic community [1–3]. The accurate recognition of human action in videos is a key step for most video understanding applications. In the literature, several surveys [4–8] have proposed to make the task of the human action or activity recognition more effective by extracting new semantic or visual features. For the exploration of semantic features, researchers have tried to conduct research in terms of poses [9] and poselets [10], objects [11], scenes [12], skeletons [13,14], key-frames [15], attributes [16] and inference methods [17], etc. For example, the Dense Trajectory Features (DTF) [18] and the Improved Dense Trajectories (IDT) [19] showed their high potential for effectiveness. Recently, with the rise of deep learning, methods such as using convolutional neural networks, modeling of temporal features and designing multi-branch networks have also been proposed to address the challenges existing in the human action recognition task [5]. Further, to capture spatial and temporal features for video analysis, a 3D convolution is proposed by Ji et al. [20]. However, due to the layer-by-layer stacking of 3D CNNs, 3D CNN models cause higher training complexity as well as higher memory requirements [21]. In particular, many researchers put their efforts into the motion and temporal structures modeling [22] recently. Accordingly,

the short-term motion modeling method based on optical flow technique was proposed in [23]. Temporal Segment Network(TSN) [24] introduced a sparse segment sampling strategy in the entire video to model long-term temporal features. Furthermore, the frameworks based on 3D CNN [20,25] utilized convolution operation along both the temporal and the spatial dimension for the sake of making the model directly learn the relation between the temporal and the spatial features. More recently, several approaches suggested using (2+1)D CNN instead of 3D CNN to achieve more explicit temporal modeling [26,27].

Though some actions can be inferred by relying solely on spatial information named "scene-related" actions, such as "diving", "ride a horse", while other actions are more closely related to temporal information named "temporal-related" actions, such as "moving something from left to right", "moving something approaching something", etc., and the latter is more challenging. According to this observation, we can deduce that the importance of temporal and spatial features to the recognition of different actions is not uniform, and there should be competitive relationship between these two types of features when a model trying to recognize the action. Therefore, we believe that a good spatial-temporal features extractor should have the ability to determine what feature would be the dominant for building a better model.

In the literature, the development of spatial information(or feature) extraction methods is relatively mature, and how to efficiently extract the temporal information of a video has become one of the key challenges in the current action recognition research. We may regard the video as a fourth-dimensional representation ($W \times H \times T \times C$). In particular, $W$ and $H$ are the width and height of a single video frame, respectively, which can also be collectively referred to as the spatial representation of the video frame, denoted as $S$. $T$ and $C$, which represent temporal and channel accordingly.

It should be mentioned that most of the previous work tends to focus on the two dimensions of $S$ and $T$, but this study focuses on the channel dimension ($C$) along with deep learning structure. As a breakthrough, the proposal of the Temporal Shift Module (TSM) network [28] made researchers realize that using 2D convolution and 3D convolution to extract time information from the video has a certain equivalence, but the 2D network requires much fewer parameters and calculations. In consideration of these essential facts, we propose an efficient convolutional network for action recognition with the following main contributions.

Firstly, in view that the existing 2D structures perform not very well on the extracting of temporal context features (denoted as foreground changes) compared with the spatial feature extraction, we propose to use a motion attention module (MA) to obtain enhanced temporal features.

Secondly, for better utilization of the obtained features, a spatial-temporal channel attention (STCA) mechanism is used to perform a trainable feature fusion.

Finally, this research reveals that for temporal-related action recognition video datasets, spatial-temporal features have a complex relationship of cooperation and competition inside the convolution channel. The rest of paper is organized as follows: first, some related works are introduced in Section 2; second, the detailed training process for the proposed approach is depicted in Section 3; third, some experimental results are reported in Section 4; finally, conclusions and discussions are given in Section 5.

## 2. Related Works

Compared with single image recognition [7], the main difference of video understanding is that it possesses temporal information. In recent years, researches on deep mining of spatial information by using convolutional neural networks have achieved fruitful results. However, the research on how to effectively model the temporal information is still one of the most important challenges in the current action recognition research. For this reason, Feichtenhofer et al. [29] suggested that current *ConvNet* architectures are not able to take full advantage of temporal information and their performance is consequently often dominated by spatial (appearance) recognition. Overall, according to most research progress in

the current literature, the video-based action recognition task mainly faces the following three challenges.

- Effective extraction of spatial and temporal features. Especially the extraction of temporal features is the focus of current research and is one of the great challenges.
- The systematic integration of spatial and temporal features. From the early score [23] fusion to the recent pixel-level fusion [29], researchers have proposed many novel spatial-temporal feature fusion methods, but at present, there is still much room for improvement in the effectiveness of each method.
- Improving the efficiency of action recognition network. This is one of the important constraints on whether the proposed algorithm can be applied in practice. It is also the focus of current research and will be discussed in the following sections.

In the literature, researchers have proposed a few advanced spatial-temporal convolution structures. For intuitive comparison, the spatial-temporal convolution structure proposed by other researchers and the new structure we proposed in this paper are put together here and they can be formally divided into six categories: C2D [30,31], C3D [25,32], cascade 3D [27,33], a reversed cascade 3D (derived from [27]), parallel [34] and DTP.

Figure 1a illustrates a C2D Residual Block (TSN [30] and Temporal Relational Reasoning Network (TRN) [31] belong to this category); Figure 1b shows a C3D network; Figure 1c indicates a cascade 3D network, which decomposes a standard 3D kernel with a 1D temporal convolution followed by a 2D spatial convolution; Figure 1d demonstrates a reversed cascade 3D architecture, which exchanges the order of temporal and spatial convolution in cascade 3D; Figure 1e displays a parallel architecture, which models the spatial and temporal information in two independent branches; Figure 1f shows a new channel-separate temporal convolution proposed to replace the standard temporal convolution in Figure 1e, named as DTP.

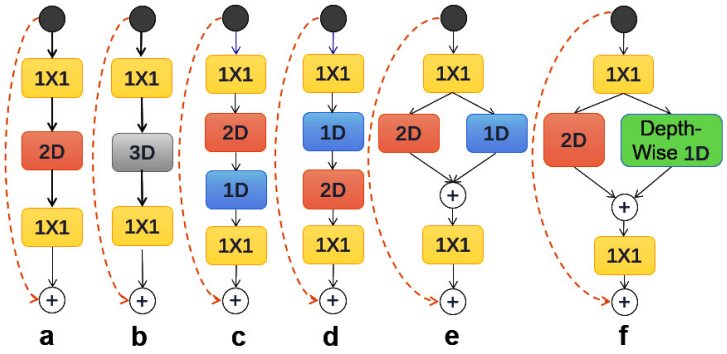


**Figure 1.** Typical spatial-temporal convolutional structures. (**a**) C2D; (**b**) C3D; (**c**) 3D; (**d**) reversed 3D; (**e**) parallel architecture; (**f**) DTP.

As for the parallel structure, such as SlowFast [35], ArtNet [34], which independently extracts the temporal and spatial information in a video. In this paper, based on the parallel structure, we decompose the 3D convolution into a spatial and a temporal convolution. The core contribution is that we no longer use the cascading method to perform the two kinds of convolutions but extract spatial-temporal information, respectively, (as shown in Figure 1e). By doing so, the network will not be disturbed by spatial information when extracting temporal information. Moreover, we separate the original temporal convolution in the parallel structure, and technically use a depth-wise temporal convolution (the characteristic of depth-wise temporal convolution is to decouple the temporal and channel dimensions, and perform temporal convolution on a depth-by-depth basis) instead of the standard temporal convolution. We call this structure as DTP (depth-wise temporal parallel), which is shown in Figure 1f.

Currently, the existing structures can be roughly summarized into 5 categories: CNN+Pooling [29,30], CNN+RNN [31,36,37], C3D [25,32], Efficient 3D [33,38], and new C2D(NC2D) [3,28]. We will introduce these different architectures separately and the representative results of them on the Something-Something V1 dataset, which is the most widely used dataset in action recognition tasks.

**CNN+Pooling** [29,30,39,40]: Karpathy et al. [39] achieved the initial success in applying CNN to the video field for the first time. Then a two-stream structure proposed by Simonyan et al. [40] surpassed the manual feature method for the first time in terms of accuracy. Later, TSN network [30] uses the redundancy between video frames on the basis of the "two-stream" network and proposes a sparse sampling strategy that can greatly reduce the computational cost of the network compared to the dense sampling. Although the computational complexity of 2D networks is lower than that of 3D networks, the "frame-by-frame" processing has caused many difficulties in utilizing the full temporal information contained in the video. Based on the reports of their experimental results, the CNN+Pooling method (represented by TSN [30]) can obtain 19.7% Top-1 accuracy at the computational cost of 33G floating point operations(Flops) on the Something-Something V1 dataset.

**CNN+RNN** [31,36,37]: Researchers have proposed several CNN+RNN methods for the temporal modeling. However, this category of method has an inherent flaw: when CNN is used for encoding, the temporal information of the underlying features is often directly lost. Besides, it only considers the temporal information in the top-level features, so its recognition accuracy is not significantly improved compared to C2D networks such as the two-stream networks. Based on the reports of their experimental results, the CNN+RNN method (represented by TRN [31]) can obtain 34.4% Top-1 accuracy at the computational cost of 33G Flops on the Something-Something V1 dataset.

**C3D** [25,32]: The 3D convolution network can simultaneously extract the spatial-temporal features of a video. Tran et al. [25] first elaborated on how to apply C3D network to the field of action recognition. However, the lack of large-scale video datasets for pre-training makes the 3D network training very difficult. Although the recognition accuracy of this method is not much improved compared with the previous methods, it opens up a new horizon for the application of 3D convolution in the field of action recognition. Carreira et al. [32] proposed the I3D network, which introduced a deep 3D network for the first time, and creatively used the weights of the 2D-InceptionV1 network pre-trained on ImageNet to initialize the network through the "inflated" operation. The proposal of the 3D network has greatly improved the recognition accuracy, but its large amount of parameters and calculations make it difficult to deploy on mobile devices with low computational cost. Based on the reports of their experimental results, the C3D method (represented by I3D [32]) can obtain 41.6% Top-1 accuracy at the computational cost of 306G Flops on the Something-Something V1 dataset.

**Efficient 3D** [33,41–43]: Since the high computational complexity of 3D convolution is mainly due to the simultaneous convolution of the three dimensions of the video (S, T, C), researchers proposed to decouple the two dimensions of time and space, and decompose the original 3D convolution into cascaded spatial convolution and temporal convolution [33,42,43]; Another way to reduce the computational overload is to use a hybrid spatial-temporal convolution structure similar to "2D-3D" or "3D-2D" to replace the part of 3D convolution in the 3D network kernel [44]. However, the structure, like "CNN+RNN", may lose part of the temporal information of the feature. In addition, many experiments have proved that the spatial-temporal features (low-level and high-level) in different layers of the fusion network are meaningful for action recognition [27]. More recently, X3D [38] reveals that networks with thin channel dimension and high spatial-temporal resolution can be effective for video recognition. This is also highly consistent with the point held by this research, i.e., making full use of the spatial-temporal channel can improve the performance of the task of action recognition. Based on the reports of their experimental results, the Efficient 3D

method (represented by S3D [33]) can obtain 47.3% Top-1 accuracy at the cost of 66G Flops computational power on the Something-Something V1 dataset.

**NC2D** [3,28,45]: The previous C2D networks have the advantage of fewer parameters and a low computational cost but struggle to use temporal information effectively compared to C3D networks. To address this issue, TSM [28] uses a shift in the temporal channel so that channel information of two adjacent frames is shared. Recently, Gated Shift Module(GSM) [3] has attracted the attention of researchers, and it performs very well, while being extremely light in terms of network width and parameters. Much luckily, this category also provides our study with many aspirations. Based on the reports of their experimental results, the NC2D method (represented by GSM [3]) can obtain 49.6% Top-1 accuracy only at the cost of 33G Flops computational power on the Something-Something V1 dataset currently.

**Motivations**: In summary, it should be said that both 2D and 3D methods have their own advantages, but this research mainly focuses on the scope of the NC2D method. On the one hand, 2D and 3D methods have great equivalence in extracting video features and the 2D network requires much fewer parameters and calculations [28]; on the other hand, compared to the 3D method, NC2D methods generally extract spatial-temporal features separately, which ultimately leads to the necessity of experiencing a reasonable fusion of the two types of features. However, because the existing methods do not conduct an in-depth evaluation of the qualitative or quantitative contribution of the individual features obtained by most of NC2D methods to the classification accuracy, thus largely limits the NC2D method to achieve better classification results on time-related video datasets. However, before the contribution of spatial-temporal features can be evaluated reasonably, it is necessary to explore the relationship between them inside the convolution channel, which is the prerequisite for designing the fusion method. To this end, we start with the latest and typical NC2D structure to explore the internal relationship between spatial-temporal features inside the convolution channel. Our study further explores that compared with their spatial modeling capabilities, the current NC2D structures used in the convolution channel have relatively weakened their own temporal modeling capabilities, so we proposed a MA to improve the temporal modeling capabilities of NC2D. As a result, MA enhanced the temporal modeling capability of NC2D to some extent, and the experimental results also revealed the cooperative and competitive relationship of spatial-temporal features inside the convolution channel. Finally, in response to such cooperative and competitive relationship, we propose to use a parameter-trainable STCA module to integrate temporal and spatial features instead of using a concatenated method.

## 3. The Proposed Approach

As mentioned in the above analysis, in order to achieve the effective extraction of temporal features and better fusion of temporal and spatial features, we propose a channel-wise spatial-temporal aggregation network (CSTANet, as shown in Figure 2). Intuitively, our CSTANet looks like a stack of CSTA Blocks. As shown in the Figure 2, after sampling the frames from a video, we adopt the same strategy as described in TSN [30]. Technically, the input video is first split into $N$ segments equally and then one frame from each segment is sampled. In detail, we adopt the ResNet-50 [46] as a backbone, and replace the Residual block with our CSTA block.

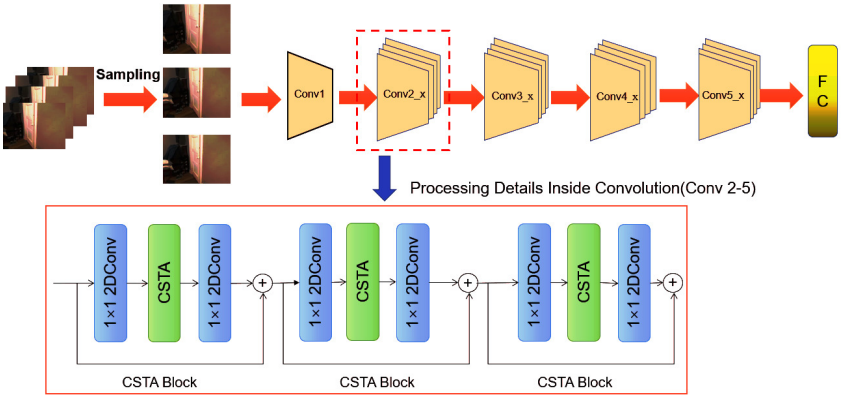


**Figure 2.** Framework of CSTANet.

As shown in Figure 3, our CSTA block contains two stages: Spatial-Temporal Feature Extraction (STE) and Spatial-Temporal Feature fusion (STF). We use DTP with MA as the spatial-temporal extractor, in which the MA is assembled to the temporal branch for enhancing the temporal context features. Next, an STF Module is used to fuse the spatial-temporal features. As for the spatial-temporal feature extraction, we use two independent branches to model the spatial-temporal information, respectively. In the spatial feature extractor module, we adopt a standard 2D convolution while applying a 1D depth-wise convolution to model temporal information. After extracting the features, in order to fuse the features extracted by the two branches, we construct a Spatial-Temporal Channel Attention (STCA) module to aggregate the spatial and temporal features channel-wisely. In contrast to Group Spatial-Temporal (GST) [47], which uses a hard-wired channel concatenation, we use a self-adaptive and trainable approach to aggregate spatial-temporal features for each block channel. Hence, the new model should be more discerning for different types of video actions. In the following paragraphs, we first introduce the details of the novel Motion Attention Module (MA) and then depict the novel Spatial-Temporal Channel Attention (STCA) module.

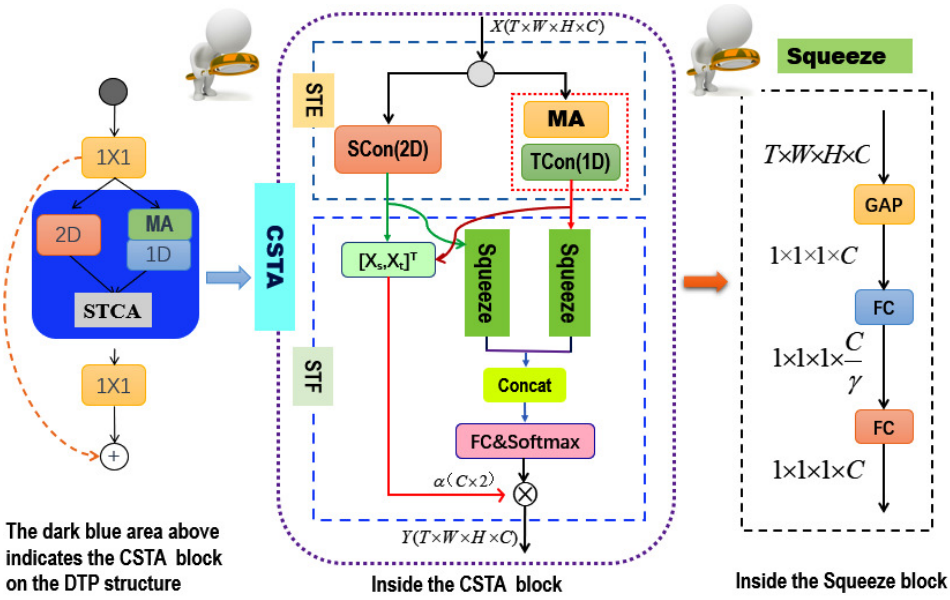


**Figure 3.** The illustration of the architecture after replacing standard convolution in Resnet block by the new proposed CSTA.

### *3.1. Motion Attention Module (MA)*

From experience, a clip often contains a moving foreground and a relatively stable background. It can be observed that in a temporal-related dataset, the backgrounds of

many action categories may be closely similar. Therefore, in order to accurately distinguish these actions, we must pay more attention to the moving part of the video, which is denoted as the foreground. However, most current action recognition networks often ignore the distinction between foreground and background, and simply extract foreground (motion) and background (relatively static) features together. In fact, we believe that temporal features should be more closely related to the motion parts of a video. In order to further strengthen the model's ability of extracting temporal information, we propose a "Parameter-Free" foreground attention mechanism, which highlights the foreground features of a motion in the video without adding additional model parameters. At the same time, the model will improve the model's ability of perceiving foreground changes.

Given an input $X\{x_1, x_2, x_3, ....., x_T\}$, $x_t \in R^{w,h,c}$, the proposed MA module is mainly divided into three processing steps. As an illustration, we select the $t$-th frame and the $(t+1)$-th frame as inputs. As shown in Figure 4, a mapping function is first used to map a 3-D tensor to a 2-D tensor, which is based on the statistic of activation tensors across the channel dimension. Next, we make an element-wise subtraction between adjacent frames with an activation function followed to get an attention map for frame $t$. Finally, the attention map is then multiplied in pixel-wise manner to the frame $t$'s feature, where $\ominus$ denotes element-wise subtraction, and $\odot$ denotes element-wise multiplication. In detail, first we define the mapping function:

$$\mathrm{M}_{w,h} = \varphi(x) = \frac{1}{C}\sum_{c=1}^{C} x_{c,w,h} \tag{1}$$

where $x_{c,w,h}$ represents the pixel value at the position $(c, w, h)$. Correspondingly, $\mathrm{M}_{w,h}$ represents the value on the output 2D tensor on the position $(x, y)$. After the global features are extracted, we next calculate the changes in two adjacent frames for highlighting the moving part. So, the M-tensor of the two frames obtained in the first step is subtracted and then activated by using an activation function.

$$\text{Attention} = \text{Activate}(|\mathrm{M}_{t+1} - M_t|) \tag{2}$$

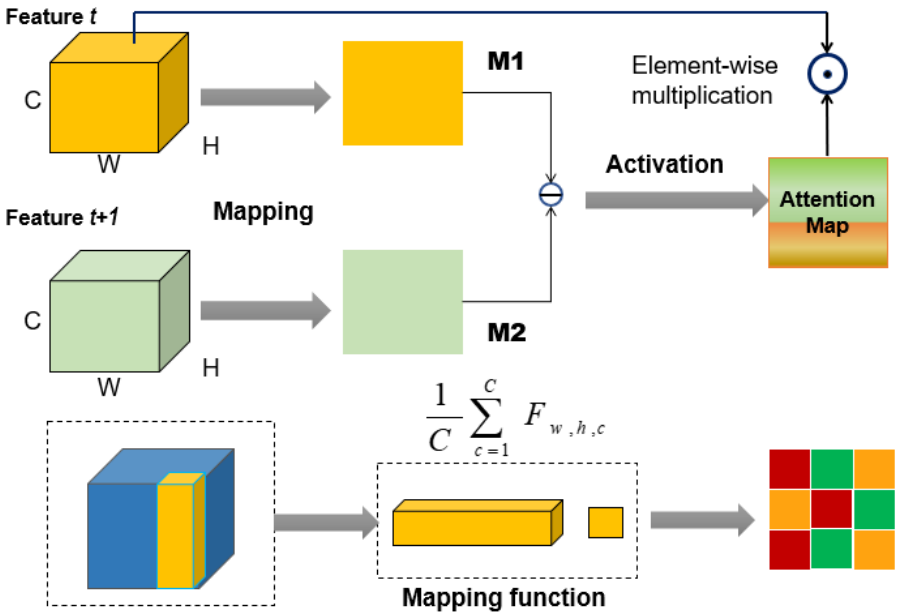


**Figure 4.** The process of MA.

We have tried a variety of activation functions such as Sigmoid, Softmax, tanh and so on, and the experiments indicate that Sigmoid has achieved the best activation effect. In the end, the attention map is still a 2D tensor, which has the same shape as $M$. In other words,

the foreground attention we proposed is "spatial-specific", i.e., the attention weights are applied to each position on the original frame. Finally, we multiply the obtained attention map with the frame $t$ pixel by pixel and obtain the following.

$$x_t^a = Attention_t \odot x_t \tag{3}$$

where $x_t^a$ represents frame $t$ after foreground attention is applied, and A$ttention_t$ represents the foreground attention of frame $t$, and $t \in [1, T-1]$. In order to ensure that the output of MA ($X^a\{x_1^a, x_2^a, x_3^a, ....., x_T^a\}$) and input $X\{x_1, x_2, x_3, ....., x_T\}$ are consistent in the temporal dimension, the last frame of a video is directly used as the last frame of $X^a$, i.e.,

$$x_t^a = x_t \tag{4}$$

### *3.2. Spatial-Temporal Channel Attention Module (STCA)*

In the literature, both 3D convolution and a cascaded spatial-temporal convolution can only perform simple fusion of spatial-temporal features. Based on the results of literature comparison and feature extraction practices, we believe that the spatial-temporal information contained in different layers and channels of the network should be treated in a different manner. Among them, some channels may be more related to temporal information, and some channels may be more related to spatial information. Therefore, we propose a spatial-temporal feature fusion module (STCA), which is tightly coupled in temporal ($T$) and spatial($S$). The structure is shown in Figure 3. For the output feature maps of spatial convolution and temporal convolution $X_s$ and $X_t$, where $X_s \in R^{T\times W\times H\times C}$ and $X_t \in R^{T\times W\times H\times C}$, both are fed into the Squeeze module. The Squeeze module consists of a global average pooling layer and two consecutive fully connection (FC) layers. First, we process $X_s \in R^{T\times W\times H\times C}$ and $X_t \in R^{T\times W\times H\times C}$ to obtain $P_s \in R^{1\times 1\times 1\times C}$ and $P_t \in R^{1\times 1\times 1\times C}$ by the global average pooling, which is followed by two consecutive fully connected layers (FC). Among them, in order to reduce the number of parameters, the first FC layer is configured with the number of channels to the original $\frac{1}{\gamma}$. In the experiments, we take $\gamma = 16$. The output of the Squeeze module is $S_s \in R^{1\times 1\times 1\times C}$ and $S_t \in R^{1\times 1\times 1\times C}$. Finally, we concatenate $[S_s, S_t]$ to form $S$, followed by a FC layer and Softmax for activation in each row. That is, the temporal and spatial features in each channel obey a probability distribution as a whole, and within a channel, the spatial-temporal features form a competitive relationship. Formally,

$$\alpha[\alpha_t, \alpha_s] = Soft\max(W([S_s, S_t]) + b) \tag{5}$$

where $\alpha \in R^{C\times 2}$ is a matrix of $C \times 2$. In each channel, two attention weights are calculated, representing temporal attention and spatial attention, respectively. In other words, we aggregate the spatial-temporal information channel by channel. Finally, we apply a matrix multiplication to $\alpha[\alpha_t, \alpha_s]$ and $[X_s, X_t]^T$ to obtain the output of CSTA block (denoted as $Y$), whose dimension is the same as input.

In summary, based on the DTP network structure, the MA module can make the action foreground in a video get special attention. Simultaneously, the temporal and spatial features can be fused in channel-wise by STCA module. In comparison to the recent work GSM [3], we still keep the backbone structure of GST while developing a new fusion technique.

## 4. Experimental Results and Analysis

This section presents an extensive set of experiments to evaluate CSTANet. First, we will conduct experiments on various spatial-temporal convolution architectures. Then, ablation analysis on MA and STCA is also provided. Please note that all experiments are conducted on single modality (RGB frames).

### *4.1. Datasets*

We evaluate the CSTANet on three action recognition benchmarks, including Something-Something V1&V2 [48,49] (abbreviated as SthV1, SthV2), Diving48 [50], and EGTEA Gaze++ [51]. SthV1 version contains 108,502 clips, and the V2 version contains 220,847 clips, which contains a total of 174 types of fine-grained actions. Performance is reported on the validation set. Diving48 dataset contains around 18K videos with 48 fine-grained dive classes, and we report the accuracy on the official train/val split. EGTEA Gaze++ contains 15,484 samples with 106 activity classes. We use the first split as described in [51], which contains 8299 training and 2022 testing instances. In addition, in order to understand the impact of specific parameters in the ablation experiments, due to hardware constraints, we randomly selected about 1/4 of the data in the Something-SomethingV1 dataset (about 2W+ clips) to make a Mini-Something-Something (MSth) dataset, which will keep the ratio of the number of samples among classes roughly unchanged compared with original dataset.

### *4.2. Implementation Detail*

We adopt ResNet-50 [46] pretrained on ImageNet [52] as the backbone. For the temporal dimension, we use the sparse sampling method described in TSN [30]. Furthermore, for spatial dimension, the short side of the input frames are resized to 256 and then cropped to 224 $\times$ 224. We do random cropping and flipping as data augmentation during training. We train our model with RTX 2080TI GPUs, and each GPU processes a mini-batch of 8 video clips (when T = 8) or 4 video clips (when T = 16). Furthermore, we optimize the model by using SGD with an initial learning rate of 0.01. During the training process, the learning rate decay at the 31-th, 41-th, and 46-th epoch is 1/10 of the previous epoch, respectively. The total training epochs are about 50, and the dropout ratio is set to 0.3.

For inference, in addition to the single-clip method used during training, we also refer to the strategy of TSM [28] using randomly sample 2 different clips from the video and get the final prediction by averaging the scores of all the clips. If not specified, we use just the center crop during inference.

### *4.3. Comparisons on Various Network Structures*

Different spatial-temporal convolution structures are conducted on our MSth dataset, and the experimental results are shown in the Table 1.

**Table 1.** Performance of different structures on MSth.

| Method | Params | GFlops | Accuracy (%) |
|---|---|---|---|
| C2D | 23.9 M | 33 G | 3.0 |
| C3D | 46.5 M | 62 G | 25.9 |
| Cascade 3D | 27.6 M | 37 G | 26.9 |
| Reversed Cascade 3D | 27.6 M | 40.6 G | 27.8 |
| Parallel | 27.6 M | 40.6 G | 31.7 |
| Our DTP | **23.9 M** | **33 G** | **32.5** |

The accuracy of the C2D structure is only 3.0%, which proves that the 2D network's ability to utilize temporal information is far lower than the 3D network. On the one hand, this conclusion has been also drawn by many previous works; on the other hand, it has been proved that the action recognition on the Something-Something benchmark needs to fully capture the temporal information of the video. Compared with C3D, cascade 3D and reversed cascade 3D, we can find that the decomposed 3D convolution kernel is not only beneficial to reduce the number of network parameters and calculations, but also has the ability to achieve better results, which is consistent with the conclusions proposed in [27]. After comparing and analyzing cascade 3D and reversed cascade 3D, we found that compared to the commonly used cascade structure (they generally perform spatial convolution first and then temporal convolution), the method of performing temporal

convolution followed by a spatial convolution seems to obtain better performance. It is worth noting that although the parallel structure has the same parameters as cascade 3D and reversed cascade 3D, the accuracy rate exceeds the reversed cascade 3D structure by 3.9%, and surpasses the cascade 3D structure by nearly 4.7%. This strongly proves the rationality of modeling spatial-temporal information separately.

Finally, comparing parallel and DTP, we find that temporal convolution and spatial convolution perform differently in the channel dimension. Compared to spatial convolution (standard 2D convolution), channel fusion is not necessary for 1D temporal convolution. Furthermore, we found that the channel-separated temporal convolution not only uses fewer parameters but also achieves a higher accuracy rate than the original temporal convolution. Observing the loss curve on the left side of the Figure 5, the train loss of the DTP structure is significantly lower than that of the parallel structure, while the val loss is only slightly lower than the parallel structure. Next, we further use a deeper network structure for comparison experiments. Specifically, we use ResNet101 as the backbone and conduct experiment on the SthV1 dataset. The training curve is shown on the right side of Figure 5 and we found that in a deeper network, the gap between the train loss of the DTP structure and the parallel structure is reduced, while the gap of the val loss becomes larger than that of the shallow network. This proves that the DTP structure can effectively combat overfitting, and this manifestation is more obvious in much deeper networks. From Figure 6, we can observe that our DTP structure can achieve a good trade-off between accuracy and computational cost compared with other structures.

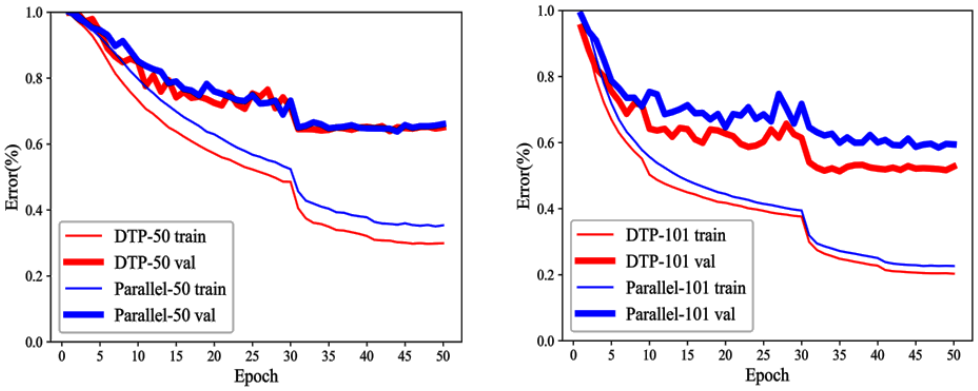


**Figure 5.** Training and testing errors for DTP and Parallel on Something-Something V1.

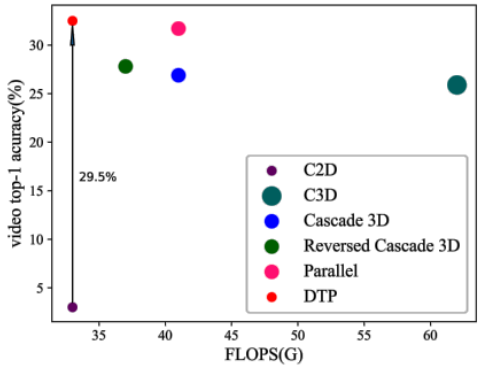


**Figure 6.** Accuracy vs. computational cost for various structures, and the size of dots denotes the amount of parameters.

### *4.4. Ablation Analysis*

In this part, we verify the validity of the MA and STCA modules. The four structures shown in the Figure 7 are used for module effectiveness analysis. First, we verify the effectiveness of MA on the CSTA block that uses STCA as spatial-temporal feature fusion. We conduct experiments both on the MSth dataset and the SthV1 dataset. Intuitively,

based on the results in Table 2, we can observe that the structure of Figure 7c performs better than Figure 7a. Both structures use the same spatial-temporal features as input, and the direct superposition method will result in performance loss. Then, it can be inferred that the contribution weight of the instant empty feature on the convolution channel is non-uniform. In other words, the effectiveness of STCA proves to some extent that the extracted spatial-temporal features have a cooperative and competitive relationship in the convolution channel.

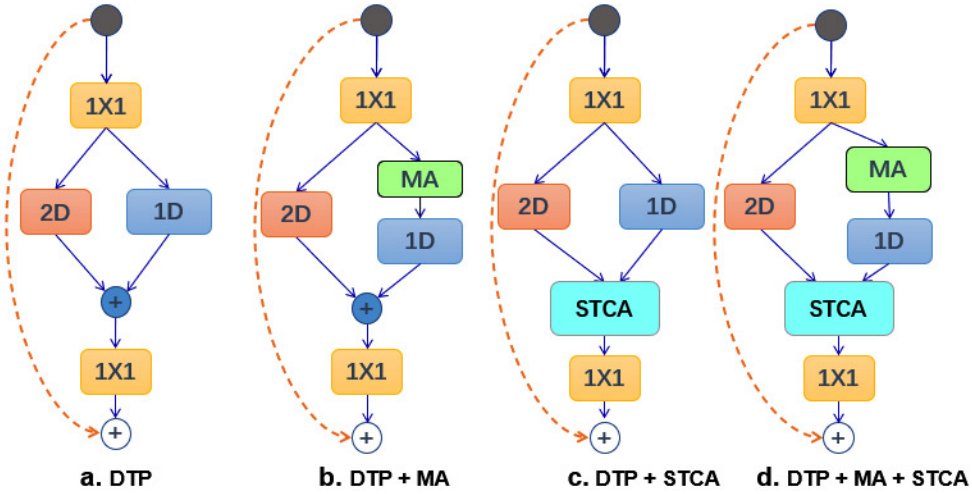


**Figure 7.** The four kinds of structure for ablation analysis. (**a**) DTP; (**b**) DTP+MA; (**c**) DTP+STCA; (**d**) DTO+MA+STCA.

**Table 2.** Performance of different structures on MSth and SthV1.

| Structure | MSth (%) | SthV1 (%) |
|---|---|---|
| DTP | 32.5 | 45.1 |
| DTP+MA | 33.0 | 46.2 |
| DTP+STCA | 33.1 | 46.5 |
| DTP+MA+STCA | **34.2** | **47.4** |

### *4.5. Results on Something-Something Dataset*

Table 3 illustrates the accuracy, parameter and complexity trade-off computed on the validation set of SthV1 dataset. Among them, our new model only uses RGB as the input and exceeds the TRN using dual-stream [47] as much as 5.4 percent. Furthermore, compared with ECO-RGB [44], it has achieved a higher accuracy rate with fewer parameters and less calculations. This proves that the temporal information of the low-level features also contributes to the classification. Intuitively, CSTANet performs slightly better than GSM [3] when the frame number is 8f, but accuracy is nearly the same as that of GSM when the frame number is 16f. We should stress here that both GSM and CSTANet are modifications of GST [47], in which GSM replaced the backbone of GST with a lightweight backbone BN-inception focusing on splitting while CSTANet keeps the same backbone as GST focusing on the fusion of spatial-temporal features. In addition, our model still achieves a higher accuracy rate compared to heavyweight networks such as I3D with non-local [32,53].

Table 4 shows the performance of our model on the SthV2 dataset. One can see that the new model using 8 frames can outwin TSM using the same number of frames with nearly 0.9 points, even surpassing the TSM model using 16 frames with only one-tenth of the calculation cost.

**Table 3.** The state-of-the-art results on the Something-Something V1 validation set.

| Models | Backbone | Frame Number | Params | GFLOPS | Top-1 Acc (%) |
|---|---|---|---|---|---|
| TSN [30] (ECCV'16) | ResNet-50 | 8 | 23.9 M | 33 | 19.7 |
| I3D-RGB [32] (CVPR'17) | ResNet-50 | 32 × 2 clips | 28 M | 153 × 2 | 41.6 |
| TRN-2stream [31] (ECCV'18) | BN-Inception | 8 | - | - | 42.0 |
| ECO-RGB [44] (ECCV'18) | BN-Inception | 8 | 47.5 M | 32 | 39.6 |
| | | 16 | 47.5 M | 64 | 41.4 |
| S3D [33] (ECCV'18) | BN-Inception | 64 | - | 66 | 47.3 |
| NL I3D-RGB [53] (CVPR'18) | 3D-ResNet-50 | 32 × 2 clips | 28 M | 117 × 2 | 44.4 |
| TSM [28] (ICCV'19) | ResNet-50 | 8 | 23.9 M | 33 | 43.4 |
| | | 16 | 23.9 M | 65 | 44.8 |
| GST [47] (ICCV'19) | ResNet-50 | 8 | 21.0 M | 29.5 | 46.6 |
| | | 16 | 21.0 M | 59 | 48.6 |
| STM [45] (ICCV'19) | ResNet-50 | 8 × 30 | - | 33.2 × 30 | 49.2 |
| | | 16 × 30 | - | 66.5 × 30 | 50.7 |
| GSM [3] (CVPR'20) | BN-Inception | 8 | - | 16.5 | 47.24 |
| | | 16 | - | 33 | 49.56 |
| CSTANet | ResNet-50 | 8 | 24.1 M | 33 | 47.4 |
| | | 8 × 2 clips | 24.1M | 33 × 2 | 48.6 |
| | | 16 | 24.1 M | 66 | 48.8 |
| | | 16 × 2 clips | 24.1 M | 66 × 2 | **49.5** |

**Table 4.** State-of-the-art results on the Something-Something V2: * denotes that 5 crops are used.

| Model | Backbone | Frame | Top-1 Acc (%) |
|---|---|---|---|
| TRN [31] | BN-Inception | 8f | 48.8 |
| TSM [28] (*) | ResNet-50 | 8f | 59.1 |
| | | 16f | 59.4 |
| GST [47] | ResNet-50 | 8f | 58.8 |
| TRN-2Stream [31] | BN-Inception | 8f | 55.5 |
| TSM-2Stream [28] | ResNet-50 | 16f | 63.5 |
| CSTANet | ResNet-50 | 8f | 60.0 |
| | | 16f | 61.6 |

### *4.6. Results on Diving48 and EGTEA Gaze++*

First, we sample 16 frames from each video clip on Diving48 and only report results using 1 clip per video. As shown in Table 5, by only employing a lightweight backbone ResNet-18, our model can outperform all the previous approaches. Second, we conduct similar experiment on EGTEA Gaze++ with results shown in Table 6. Because MA is sensitive to the foreground motion, but there are a lot of background motions in EGTEA Gaze++ dataset, MA is no longer active, but STCA module is still effective, so only by relying on STCA module can the new model achieve the state-of-the-art result. Furthermore, this experiment enlightens us that it is very important to distinguish the movement of foreground and background.

**Table 5.** State-of-the-art results on Diving48.

| Method | Pretrain | Top-1 Acc (%) |
|---|---|---|
| C3D (64 frames) [25] | - | 27.6 |
| R(2+1)D [54] | Kinetics | 28.9 |
| R(2+1)D+DIMOFS [54] | Kinetics + PoseTrack | 31.4 |
| C3D-ResNet18 [25] (from [47]) | ImageNet | 33 |
| P3D-ResNet18 [42] (from [47]) | ImageNet | 30.8 |
| CSTANet-ResNet18 (ours) | ImageNet | 35.3 |
| C3D-ResNet50 [25] (from [47]) | ImageNet | 34 |
| P3D-ResNet50 [42] (from [47]) | ImageNet | 32.4 |
| GST-ResNet50 [47] | ImageNet | 38.8 |
| CorrNet [55]) | - | 37.7 |
| Attentive STRL [2]) | ImageNet | 35.64 |
| CSTANet-ResNet50 (ours) | ImageNet | **39.5** |
| CSTANet-ResNet50 ($\times$2 clips) | ImageNet | **40.0** |

**Table 6.** State-of-the-art results on EGTEA Gaze++.

| Method | Pretrain | Top-1 Acc (%) |
|---|---|---|
| I3D-2Stream [51] | - | 53.3 |
| R34-2Stream [56] | - | 62.2 |
| P3D-R34 [42] (from [47]) | - | 58.1 |
| CSTANet-R34 (ours) | ImageNet | 59.0 |
| P3D-R50 [42] (from [47]) | - | 61.1 |
| GST [47] | ImageNet | 62.9 |
| CSTANet-R50 (ours) | ImageNet | **66.5** |
| CSTANet-R50 without MA (ours) | ImageNet | **67.6** |

### *4.7. Spatial-Temporal Feature Distribution Analysis*

In this section, we will analyze the characteristics of spatial-temporal feature distributed in each layer of the network and discuss the spatial-temporal feature contribution of different layers of the network. We use the attention weight calculated in each CSTA block to indicate the importance of the features. Figure 8 illustrates the spatial-temporal features' distribution among all CSTA blocks based on ResNet-50. We can draw the conclusion that the temporal feature plays a very important role in Something-Something dataset. In contrast, on Diving48, the contribution of spatial and temporal nearly "half to half". As for EGTEA Gaze++, the contribution of spatial surpasses that of temporal. We also can find that in all datasets, the contribution of temporal feature rises, while that of spatial declines from shallow blocks to deep ones. This suggests that our model tends to learn temporal representation in high-level semantic features, which has been verified by many previous works [25,33].

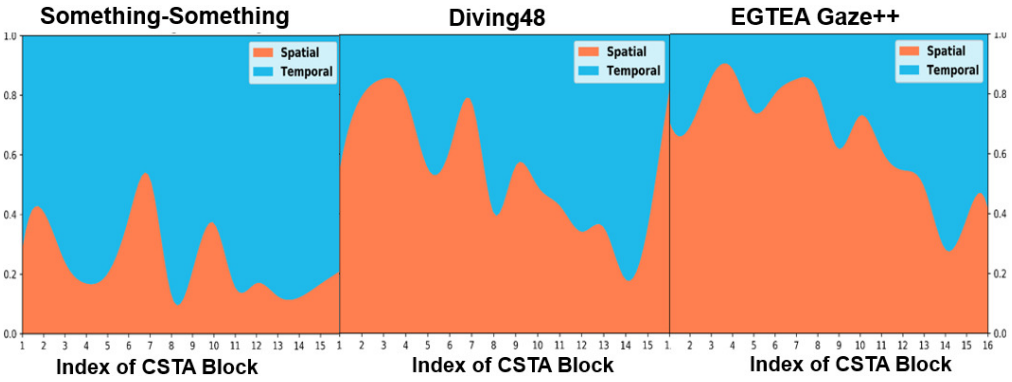


**Figure 8.** The spatial-temporal features distribution among CSTA blocks.

Furthermore, in order to analyze the spatial-temporal features in local areas, we use sigmoid as our activate function in STCA module. As shown in Figure 9 (right), we find that in shallow blocks the spatial and temporal are less distinguishable, and the performance is worse than left. It suggests that in local areas, it is beneficial to make spatial-temporal features compete with each other.

In fact, we believe that the current understanding of the characteristics of time and space is far from satisfactory. The reason is that the spatial and temporal characteristics are both complex and abstract concepts, which include at least the background, object, human body, and changes in the temporal dimension of all three.

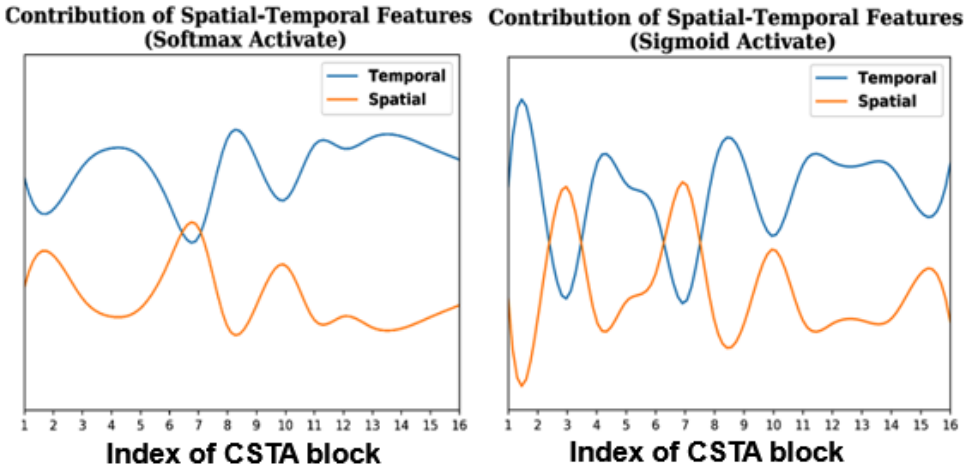


**Figure 9.** The activation function in STCA module.

## 5. Conclusions

In short, this paper mainly explores a possible best way to extract spatial-temporal information in channel-wise in a video, and presents a new spatial-temporal feature architecture composed of DTP+MA+STCA. Based on the empirical fact that the temporal and spatial characteristics are both cooperative and competitive in the channel dimension, this paper proposes to replace the hard wired fusion method often used in previous studies with an adaptive feature fusion method. The proposed CSTANet integrates the above research inspirations and provides a reference for academic researchers in action recognition in videos.

In addition, we found that in some samples, such as "pushing something from left to right" and "moving something closer to something", the two are very similar in the path of movement, and sometimes it is difficult for humans to distinguish. In fact, we believe that the current understanding of the characteristics of time and space is far from satisfactory. The reason is that the spatial and temporal characteristics are both complex and abstract concepts, which include at least the background, target, human body, and changes in the

temporal temporal dimension of all three. Therefore, it is necessary for us to further analyze in detail the contribution of each feature to action recognition, which will be the direction of our next work.

**Author Contributions:** Conceptualization, H.W.; Data curation, T.X.; Formal analysis, T.X.; Methodology, W.L.; Project administration, H.W.; Software, T.X.; Supervision, X.G.; Validation, H.L.; Visualization, H.L.; Writing—original draft, H.W.; Writing—review snd editing, Y.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by 2020 Hebei Provincial Science and Technology Plan Project, grant number: 203777116D; Beijing Municipal Education Commission Scientific Research Program, grant number: KM202110009001.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank Hongyu Hao for their help on the coarse detection and thank Ao Chen and Haodu Zhang for their help on the STCA and MA models. This work was supported in part by the Beijing Municipal Education Commission Scientific Research Program under Grant KM202110009001 and the 2020 Hebei Provincial Science and Technology Plan Project under Grant 203777116D.

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

## References

1. Chen, Y.; Wang, L.; Li, C.; Hou, Y.; Li, W. ConvNets-based action recognition from skeleton motion maps. *Multimed. Tools Appl.* **2020**, *79*, 1707–1725. [CrossRef]
2. Kanojia, G.; Kumawat, S.; Raman, S. Attentive Spatio-Temporal Representation Learning for Diving Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
3. Sudhakaran, S.; Escalera, S.; Lanz, O. Gate-Shift Networks for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
4. Aggarwal, J.; Ryoo, M. Human Activity Analysis: A Review. *ACM Comput. Surv.* **2011**, *43*, 1–43. [CrossRef]
5. Kong, Y.; Fu, Y. Human Action Recognition and Prediction: A Survey. *arXiv* **2018**, arXiv:1806.11230
6. Turaga, P.; Chellapa, R.; Subrahmanian, V.; Udrea, O. Machine Recognition of Human Activities: A Survey. *Circuits Syst. Video Technol. IEEE Trans.* **2008**, *18*, 1473–1488. [CrossRef]
7. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361. [CrossRef]
8. Ziaeefard, M.; Bergevin, R. Semantic human activity recognition: A literature review. *Pattern Recognit.* **2015**, *48*, 2329–2345. [CrossRef]
9. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Action recognition via pose-based graph convolutional networks with intermediate dense supervision. *Pattern Recognit.* **2022**, *121*, 108170. [CrossRef]
10. Agahian, S.; Negin, F.; Köse, C. An efficient human action recognition framework with pose-based spatiotemporal features. *Eng. Sci. Technol. Int. J.* **2020**, *23*, 196–203. [CrossRef]
11. Ikizler-Cinbis, N.; Sclaroff, S. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010.
12. Zhang, Y.; Qu, W.; Wang, D. Action-scene Model for Human Action Recognition from Videos. *AASRI Procedia* **2014**, *6*, 111–117. [CrossRef]
13. Li, J.; Xie, X.; Pan, Q.; Cao, Y.; Zhao, Z.; Shi, G. SGM-Net: Skeleton-guided multimodal network for action recognition. *Pattern Recognit.* **2020**, *104*, 107356. [CrossRef]
14. Si, C.; Jing, Y.; Wang, W.; Wang, L.; Tan, T. Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network. *Pattern Recognit.* **2020**, *107*, 107511. [CrossRef]
15. Elahi, G.M.E.; Yang, Y.H. Online Learnable Keyframe Extraction in Videos and its Application with Semantic Word Vector in Action Recognition. *Pattern Recognit.* **2021**, *122*, 108273. [CrossRef]
16. Zhang, Z.; Wang, C.; Xiao, B.; Zhou, W.; Liu, S. Human Action Recognition with Attribute Regularization. In Proceedings of the 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, Beijing, China, 18–21 September 2012; pp. 112–117. [CrossRef]
17. Liu, L.; Wang, S.; Hu, B.; Qiong, Q.; Wen, J.; Rosenblum, D.S. Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition. *Pattern Recognit.* **2018**, *81*, 545–561. [CrossRef]

18. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558. [CrossRef]
19. Wang, H.; Oneata, D.; Verbeek, J.; Schmid, C. A Robust and Efficient Video Representation for Action Recognition. *Int. J. Comput. Vis.* **2015**, *119*, 219–238. [CrossRef]
20. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef]
21. Zhou, X.; Zhu, M.; Pavlakos, G.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K. MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 901–914. [CrossRef]
22. Martínez, B.M.; Modolo, D.; Xiong, Y.; Tighe, J. Action Recognition With Spatial-Temporal Discriminative Filter Banks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5481–5490. [CrossRef]
23. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2014; Volume 27.
24. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks for Action Recognition in Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2740–2755. [CrossRef]
25. 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, Columbus, OH, USA, 23–28 June 2014; pp. 4489–4497.
26. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 318–335.
27. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 6450–6459.
28. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7083–7093.
29. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
30. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporalsegmentnetworks: Towards good prac-tices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
31. Zhou, B.; Andonian, A.; Oliva, A.; Torralba, A. Temporal relational reasoning in videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 803–818.
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, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
33. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking Spatiotemporal Feature Learning:Speed-Accuracy Trade-offs in Video Classification. In Proceedings of the European Conference on Computer Vision, San Francisco, CA, USA, 4–9 February 2017; pp. 1–17.
34. Wang, L.; Li, W.; Li, W.; Gool, L.V. Appearance-and-Relation Networks for Video Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1430–1439.
35. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6202–6211.
36. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
37. Ng, J.Y.H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
38. Feichtenhofer, C. X3D: Expanding Architectures for Efficient Video Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 200–210. [CrossRef]
39. 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, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
40. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
41. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Mahdi Arzani, M.; Yousefzadeh, R.; Van Gool, L. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. *arXiv* **2017**, arXiv:1711.08200.
42. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.

43. Sun, L.; Jia, K.; Yeung, D.Y.; Shi, B.E. Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4597–4605.
44. Zolfaghari, M.; Singh, K.; Brox, T. Eco: Efficient convolutional network for online video understanding. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 695–712.
45. Jiang, B.; Wang, M.; Gan, W.; Wu, W.; Yan, J. STM: SpatioTemporal and Motion Encoding for Action Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019.
46. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Luo, C.; Yuille, A. Grouped Spatial-Temporal Aggregation for Efficient Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5512–5521.
48. Goyal, R.; Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fründ, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5843–5851.
49. Mahdisoltani, F.; Berger, G.; Ghar-bieh, W.; Fleet, D.; Memisevic, R. On the effective-ness of task granularity for transfer learning. *arXiv* **2018**, arXiv:1804.09235.
50. Li, Y.; Li, Y.; Vasconcelos, N. Resound: Towards action recognition without representation bias. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 513–528.
51. Li, Y.; Liu, M.; Rehg, J.M. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
52. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Jeev Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
53. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
54. Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; Torresani, L. Learning discriminative motion features through detection. In Proceedings of the IEEE International Conference on Computer Vision, Salt Lake City, UT, USA, 18–23 June 2018.
55. Wan, H.; Tran, D.; Torresani, L.; Feiszli, M. Video Modeling with Correlation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
56. Sudhakaran, S.; Lanz, O. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv* **2018**, arXiv:1807.11794.