Exploiting Attention-Consistency Loss For Spatial-Temporal Stream Action Recognition

HAOTIAN XU, XIAOBO JIN**, and QIUFENG WANG, Xi'an Jiaotong-Liverpool University, China
AMIR HUSSAIN, Edinburgh Napier University, UK
KAIZHU HUANG*, Duke Kunshan University, China

Currently, many action recognition methods mostly consider the information from spatial streams. We propose a new perspective inspired by the human visual system to combine both spatial and temporal streams to measure their attention consistency. Specifically, a branch-independent convolutional neural network (CNN) based algorithm is developed with a novel attention-consistency loss metric, enabling the temporal stream to concentrate on consistent discriminative regions with the spatial stream in the same period. The consistency loss is further combined with the cross-entropy loss to enhance the visual attention consistency. We evaluate the proposed method for action recognition on two benchmark datasets: Kinetics400 and UCF101. Despite its apparent simplicity, our proposed framework with the attention consistency achieves better performance than most of the two-stream networks, i.e. 75.7% top-1 accuracy on Kinetics400 and 95.7% on UCF101, while reducing 7.1% computational cost compared with our baseline. Particularly, our proposed method can attain remarkable improvements on complex action classes, showing that our proposed network can act as a potential benchmark to handle complicated scenarios in industry 4.0 applications.

CCS Concepts: • **Computing methodologies → Object recognition**.

Additional Key Words and Phrases: Action Recognition, Attention Consistency, Multi-level Attention, Two-stream Structure

ACM Reference Format:
Haotian Xu, Xiaobo Jin, Qiufeng Wang, Amir Hussain, and Kaizhu Huang. 2022. Exploiting Attention-Consistency Loss For Spatial-Temporal Stream Action Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2022), 14 pages. https://doi.org/10.1145/3538749

1 INTRODUCTION

Action recognition has been a hot topic in deep learning [19] for years. One of the main focuses of this task is how to build a computation-efficient model in real-world scenarios. In contemporary industrial production, million hours of all kinds of videos are being created in the security-monitoring system, quality inspection, and automated production workshop. In industrial 4.0 applications, videos are produced under complex settings and need be processed locally on edge devices due to privacy concerns, thus these demands require accurate and efficient video understanding solutions.

Currently, mainstream approaches towards action recognition can be categorized as: 3D convolutional kernel based CNN network [10, 34] and two-stream based network [11, 25, 30, 41, 48]. The former handles motion information

*Correspondence to Xiaobo Jin and Kaizhu Huang.

Authors' addresses: Haotian Xu, haotian.xu18@student.xjtlu.edu.cn; Xiaobo Jin, xiaobo.jin@xjtlu.edu.cn; Qiufeng Wang, Xi'an Jiaotong-Liverpool University, 111 Renai Road , Suzhou, Jiangsu, China, 215000; Amir Hussain, A.Hussain@napier.ac.uk, Edinburgh Napier University, EH11 4BN, Edinburgh, , UK; Kaizhu Huang, kaizhu.huang@dukekunshan.edu.cn, Duke Kunshan University, 8 Duke Avenue , Kunshan, Jiangsu, China, 215316.

by expanding traditional 2D kernels into 3D kernels with an extra temporal dimension. Researchers believe that those sequential 2D filters can naturally encode human movements with the assistance of stacked 2D kernels [21]. On the other hand, two-stream structures separately process spatial information and temporal information based on different types of inputs such as normally RGB and precomputed optical flow [2] using 2D CNN backbones. Also, many work [4, 35] in two-stream approaches has employed 3D kernels in the temporal stream to facilitate motion capturing.

The earlier two-stream architecture largely relies on the spatial stream [30] and does not exploit the full potential of the temporal stream. To solve this problem, a common practice of the two-stream architectures is to fuse feature information between temporal stream and spatial stream at different stages of the backbones via various fusion operations. For example, sum fusion computes the sum of the two feature maps at the exact spatial locations; concatenation fusion connects the two feature maps at the same spatial location; and convolution fusion refines the features between two branches. It has been found in the literature that fusion between temporal stream and spatial stream has a positive effect on the recognition accuracy [12].

Despite its success, we argue in this paper that the present fusion process would hamper the spatial stream to correctly recognize actions and inevitably increase the computational cost. Such limitation is more prominent especially under some sophisticated circumstances where there are many irrelevant moving objects facing complex foregrounds and backgrounds. First, repeating more convolution operations between two branches in the hidden layers of the network may lead to an optimization dilemma [17]. Second, the temporal stream which encodes the action information based on optical flow [28] or dense frames extraction [11] is prone to be more sensitive to subtle movements. Varies in nature, spatial stream to a large extent relies on spatial details of still images. Thus, we believe that the fusion operation might disturb both streams from focusing on the most discriminative action region accurately.

To this end, we propose a new perspective to combine both spatial and temporal streams by promoting the *Action Attention Consistency*. Two stream action recognition structures are in principle kin to observations of the human primate visual system, where Magnocellular (M-cells) and Parvocellular (P-cells) respond to fast movements and spatial details [6, 20, 27], respectively. Inspired by the working mechanism of the human optic nerve that both P-cells and M-cells receive the same optic nerve stimulation [36], we design a novel method to imitate this visual process. Namely, we design an attention consistency loss enabling the temporal stream to concentrate on the consistent discriminative regions with the spatial stream in the same period. The attention consistency loss measures the distance between the attention heatmaps of spatial stream and temporal stream, and is then combined with the cross-entropy loss to enhance the visual attention consistency through network training. As shown in Figure1, our network demonstrates superiority in localizing high-value action regions precisely. Compared with our baseline[11], we can achieve more localized and comprehensive activated areas under the situation of multiple coexisting fast-moving objects.

Class Activation Mapping (CAM) [50] is very popular to compute the attention heatmaps for each given frame. However, it can only highlight the most distinctive areas of an object. In a gesture to better understand a complete action, more expansive attention areas are needed to cover the whole course of movements. Therefore, we design a post-processing module to generate the spatial CAM to better describe the attention of spatial streams.

We evaluate the proposed method for action recognition tasks on two benchmark datasets: Kinetics400 [23] and UCF101 [31]. Despite its apparent simplicity, our proposed attention consistency method on par with our baseline attains very encouraging results, i.e. 75.7% top-1 accuracy on Kinetics400 and 95.7% on UCF101, while reducing 7.1% computational cost. More importantly, our network exhibits substantial improvements on recognizing actions in sophisticated scenarios, large elevation in top-1 accuracy on complex classes. This is especially critical for many cases in industrial 4.0 applications because modern manufacturing involves frequent motion variation in background settings.

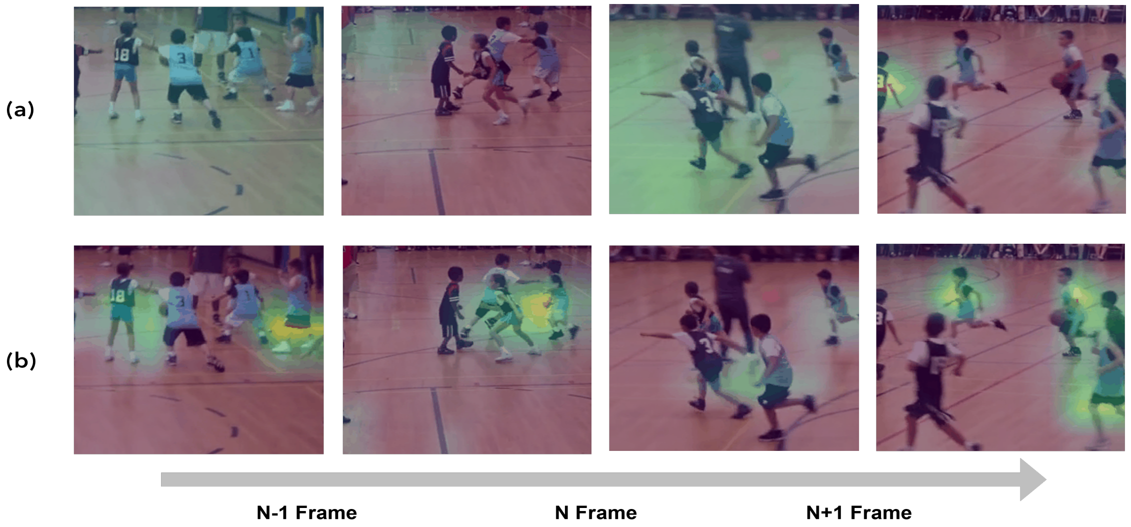

Fig. 1. **An illustration of activated regions** of class "playing basketball" in Kinetics400. The upper row (a): the heatmap of our baseline slowfast with T-conv fusion. (b): the heatmap of our method without fusion operations in the middle layers. For fair comparison, the backbone for both methods is slowfast R50 4×16. The time dimension increases from left to right.

We also conduct further studies to verify the improvement gained by incorporating our proposed attention consistency over the state-of-the-art two-stream methods experimentally. The main contributions of this paper can be summarized as follows:

1. We have designed a branch-independent two stream network with proposed action attention consistency method; we also verify by experiments that information change during middle stages is not indispensable for two stream structure.
2. Our network achieves compelling accuracy on public datasets Kinetics400 and UCF101. Especially, we largely elevate the action recognition ability on complex scenarios which is a critical ability in industrial 4.0 applications.

2 RELATED WORK

2.1 Action Recognition in Videos

Action recognition in deep learning has been formulated as a task of modeling joint spatial-temporal information. Recent approaches can be broadly categorized as two-stream structures, 3D CNNs and transformer-based networks. 3D CNNs[5, 7, 32, 42] expanding 2D kernels with an extra temporal dimension spontaneously endow the network with the ability to model hierarchical motion patterns. The presence of large action datasets [15, 16, 22] helps such parameter-heavy models acquire ideal performance. Several work [26, 35, 49] attempts to decouple the spatial and temporal convolution into 3D CNNs to achieve better temporal modeling on tracking feature points, body joints, and human movements. Transformer-based networks [1, 3, 9, 14, 29] in computer vision tasks have aroused much interest among researchers. In action recognition, the working scheme of the self-attention stems mainly from decomposing the video into a sequence of frame-level patches and then feeds linear embeddings of these patches as the input to a transformer.

For two-stream action recognition models, there are many exquisitely designed structures trying to utilize temporal information in different approaches. For early designs, the temporal stream is formed with the optical flow [30], the dense trajectory [38, 39] and the RGB difference [33, 41]. TSN [41] models long-range temporal structures with a sparse segment sampling in the whole video during training, which applies stacked optical flow fields to model short-term motion patterns. More recently, Feichtenhofer et al. [11] construct the two-stream structure based on a different frame sampling strategy, where sparse frame sampling and dense frame sampling are used for slow and fast pathway respectively. Most of these work focus on dealing with the motion and static information jointly by fusing two branches and ignores the unique characters of two separate branches largely. In this work, we provide a new insight specific to the nature of the action recognition problem itself and show how it requires an increasing sensitivity to attention regions. Different from the methods above, our work is explicitly designed for the action classification with complex backgrounds. In particular, the proposed approach is inspired by recent advances in the attention mechanism of the image recognition. We believe that this work could shed light on the understanding of generic action classification.

2.2 Attention Mechanism for Classification

It is widely accepted that attention plays a significant role in human natural perception. There are many attempts [24, 47] incorporating attention to enhance the performance of CNNs. Wang et al. [37] propose the residual attention network, which uses an encoder-decoder attention module to refine features that is robust to noise. Hu et al. [18] introduce the compact model to characterize inter-channel relationship, where a squeeze-and-excitation module with the global average is used to exploit channel-wise attention. Woo et al. [44] propose the CBAM module which utilizes both average pooling and max pooling on features maps to achieve optimal performance. For video understanding, one straightforward task related to the attention regions is the action segmentation [13, 43]. However, the pixel-level annotation makes these methods less feasible. Benefiting from the above work, we manage to inspect the attention mechanism from a different view. Though heatmaps of the temporal stream are activated with most changing action regions and the spatial stream is sensitive to semantic spatial details, both streams should focus on the same distinctive timestamp and discriminative regions when an action happens in a video clip. Therefore, we propose the concept *Action Attention Consistency*. With the proposed consistency loss, we try to guide the network to localize the true action in an indirect but workable way.

2.3 Fusion between Two Streams

Feature fusion between the temporal and spatial stream has long been a standard operation for two stream action recognition models. In fact, such a fusion operation can be applied at any stage of the backbone. Researchers have explored numerous fusion techniques [12, 41] to ensure that the response at the same pixel position from the channels of two streams are handled correspondingly. Classic fusion methods includes the sum fusion that makes an element-wise summation of the two feature maps at the same spatial locations i, j and feature channel d. Max fusion simply takes the maximum value of two feature maps. Considering that the channel is arbitrarily numbered, this operation defines a primitive correspondence between two streams. The concatenation fusion stacks the feature maps of two streams at the same spatial locations i, j across the feature channels d in order to avoid matching the channels randomly. It does not define a correspondence between two steams but leaves it to the subsequent convolutional layer for learning suitable kernel weights. Another approach is the convolution fusion, which stacks feature maps at the same spatial locations i, j across the feature channels d and then convolves the stacked features with a bank of filters f and biases b. Such a fusion utilizes a set of trainable filters to combine feature maps. Feichtenhofer et al. [11] achieve the state-of-the-art accuracy for action recognition, which constructs a two-stream structure network by adopting two different frequencies in the frame sampling and fusing the information between two pathways at the end of every “stage". In spite of the effectiveness, we argue that the convolution fusion inevitably leads to excessive parameters and computational cost. We tackle this problem by proposing a new consistency loss and achieve convincing accuracy with the smaller model size.

3 PROPOSED METHOD

In this section, we describe the intuition and details of the design for the proposed network and elaborate on the proposed branch-independent algorithm. We focus on the discriminative action regions of spatial and temporal streams under which the visual attention should be consistent. We design the network based on this observation.

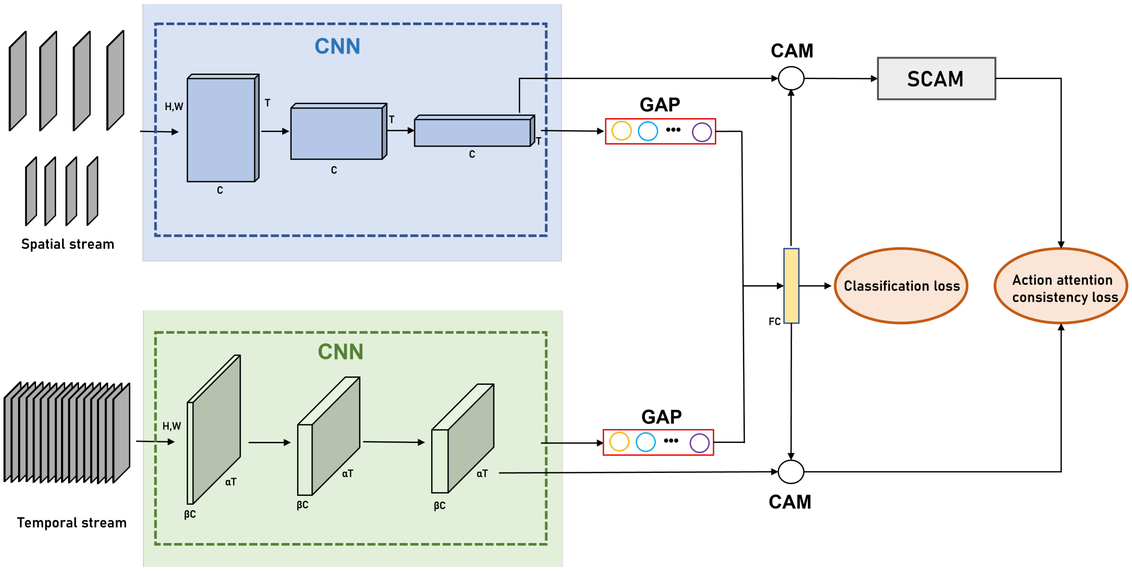

Fig. 2. **Overall architecture of our proposed method**. We illustrate the design with a 3D ResNet in which the spatial stream takes 4 frames as inputs with two scales. The temporal stream takes 16 frames as inputs with the same input size. SCAM are adopted at the spatial stream only.

3.1 Baseline Network

Our baseline method follows the recent approach [11] in action recognition and it consists of two pathways separately dealing with high frequency frames and low frequency frames followed by a standard global average pooling and a linear classifier. The slow pathway exploits a large temporal stride on input time dimension and processes semantic information with a standard ResNet. In parallel to the slow pathway, the fast pathway is a same convolutional model with significantly lower channel capacity, *i.e.*, light-weight and computationally efficient. Features between two pathways are exchanged during the middle stages in order to enhance the classification ability of the classifier. This pooling method is highly effective in capturing large contextual cues across the video sequence, as it aggregates information from the entire spatial-temporal volume of the video via the average pooling. It is designed to capture the gist of the full video sequence. As such, it is quite effective at separating actions that take place in different scenes, like skating vs. playing baseball. However, this pooling methodology limits the networks' ability to focus on a particular local, spatial-temporal region as the global average forces the classifier to consider all parts of the video at a time.

3.2 Spatial Class Activation Mapping (SCAM)

Class Activation Mapping(CAM) has already become a popular tool for researchers to generate attention heatmaps. The original CAM only highlights those confident action regions that are usually tiny but with high response scores on the class activation maps. However, for action classification we need comparatively broader image regions to describe a complete action due to the nature of human movements. Also because of the different sampling strategy, the amount of the frames fed into the spatial stream are 4 × fewer than the temporal stream, the temporal stream generates more feature maps and leads to broader activated regions after we sum up all heatmaps. In our attention consistency step, to avoid large differences when computing distances between two streams and to emphasize finer spatial details, we enhance the activated regions of spatial stream by adopting post-processing such as multi-scale class activation mapping. We choose S scales to enhance the post-processing and rescale the feature map of the last convolution layer of two input images to obtain S feature maps with the identical dimensions.

Given that L is the class, C is the number of channels and $\mathcal{T}(.)$ denotes the operation to resize the feature maps of multi-scale input frames to the same shape, the fully connected layer after pooling layer takes the feature map with weights $\boldsymbol{W} \in \mathbb{R}^{L\times C}$ and generate predictions for classification. Multi-scale CAM uses these weights to compute the attention heatmap at location (x, y) for the class l as the element of $\boldsymbol{S}^l$ shown in Eq.(1):

$$\boldsymbol{S}^l = \frac{1}{M}\mathcal{T}\left(\sum_{s=1}^{M}\sum_{c=1}^{C}\boldsymbol{W}(l,c)\boldsymbol{F}_{\boldsymbol{c},\boldsymbol{s}}\right) , \tag{1}$$

where $\boldsymbol{W}(l,c)$ and $\boldsymbol{F}_{c,s} \in R^{H\times W}$ represent the weight of class l at channel c and the feature map of channel c from the last convolution layer at the scale s, respectively. The final SCAM is an average of CAMs on all scales. In our experimental settings, we set $M = 2$ mainly due to the concern of the time complexity.

Using the multi-scale clones of the original images is beneficial for generating a stable CAM. We compute the initial CAMs of the training dataset following previous work. In our network, global average pooling (GAP) is applied on the last convolution layer. The GAP output of each branch is then concatenated and classified with a fully-connected layer. Finally, the fully-connected layer weights are used on the last convolution layer to obtain the heatmap for each class.

3.3 Temporal Attention Consistency

For spatial stream, we generate multiple static image heatmaps using the method described in the previous section. For temporal streams, due to the nature of two-stream structure, it is difficult to make a comparison between the dynamic process and the static image. We have experimented three methods on selecting a suitable timestamp in the temporal stream to mark a point in order to compute the attention consistency loss between the spatial stream and the temporal stream. In the first method, we compress the activated regions from the temporal streams, where the weights from the same spatial position of all the heatmaps are added up and divided by the number of the total input frames. This method achieves 75% accuracy and can get a heatmap with a broader activated area. But the complexity inherited from the time dimension and the simplicity of summing operations makes the model hard to focus on certain discriminative regions.

The second method achieves an inferior performance of 73.8% , where a frame is randomly chosen from a selected clip to generate a heatmap of the temporal stream, the lack of abundance in dynamic information may harm the representation ability of motion branch capturing actions and leads to the least difference values with the spatial class activation mapping.

The third method is to compute the weighted class activation mapping as shown in Eq.(2) :

$$\boldsymbol{T}^l = \sum_{t=1}^{T}\sum_{c=1}^{C}\boldsymbol{W}(l,c)\cdot\boldsymbol{F}_{t,c} , \tag{2}$$

where $\boldsymbol{T}^l$ is the heatmap of a consistent timestamp of class l in the temporal stream, $\boldsymbol{W}(l,c)$ denotes the weight of the fully-connected layer for the class l and the channel c, and $\boldsymbol{F}_{t,c}$ is the feature map from the last convolution layer of our backbone from the c-th channel in the t-th frame. We sparsely sample half of the frames to avoid the accumulation of invalid activated regions. In this method, the temporal heatmaps T^l give less weights on useless areas but with more weights on salient regions. We adopt this method in our final experiment setting.

3.4 Our Network

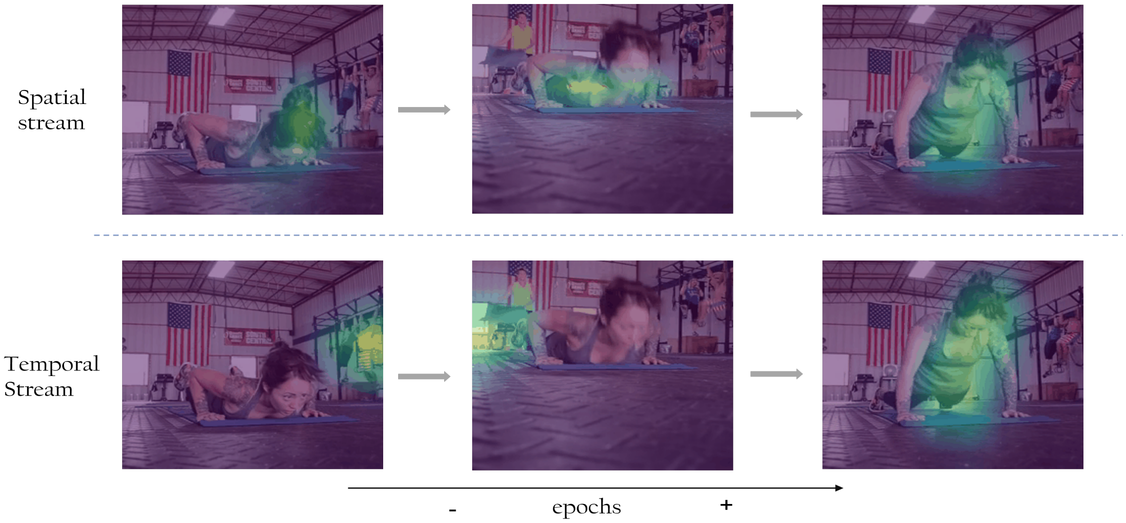

Fig. 3. **An illustration of action attention consistency**. The activation visualization of class "push up" shows the process of action consistency. The activated regions of the temporal stream in the first column and second column are misguided by the backgrounds. The third column demonstrates the consistent activation of both streams after a complete training process.

In general, the plausibility of attention heatmaps can reflect the performance of the CNN classifier. If the attention heatmaps highlight the regions that are semantically relevant to the considered action, we can expect better CNN classification performance. As demonstrated in Figure 2, our model adopts the classic spatial-temporal stream structure as the baseline. Unlike most of the other work that takes optical flow as the temporal stream, we follow the design of the slowfast to construct a temporal stream with the dense frame sampling in a gesture to prevent the additional computation. We then separately handle the spatial and temporal information from static inputs and motion inputs without parameter-sharing backbones.

In our design, We build an end-to-end two stream model that cuts off complex feature map exchange between two streams without need to store extra feature information, this could accelerate the training process and reduce GPU memory usage. The only feature-level information exchange happens at the last stage of the network. Therefore, we separately generate class activation maps for both the streams. Two-stream networks take motion and static information as their inputs. Significantly different in nature, it is critical to specify a certain timestamp between dynamic and static information with the ordered temporal information to facilitate the action-consistent comparison. Not until the last stage of the network, we compute the consistency loss from the feature maps of spatial-temporal streams extracted from the hidden layers, where the action attention consistency is described as (W and H are the width and height of CAM):

$$\ell_a = \frac{1}{NHW}\sum_{n=1}^{N}\sum_{l=1}^{L}\left\|T_n^l - S_n^l\right\|_2, \tag{3}$$

where T_n^l and S_n^l are the temporal CAM and the spatial CAM from the n-th example, separately. The final loss is defined as:

$$\ell = \ell_c + \lambda\ell_a, \tag{4}$$

where ℓ_c is the cross entropy loss and λ is set to 1.5 by default.

While our discriminative spatial branch provides fine-grain local cues, it still needs temporal information to classify actions. Here, two issues are discussed. First, the feature volume has been down-sampled to such a high degree that the motion filters cannot learn the finer details. Second, this feature volume is shared between the classifier and the discriminative filters, meaning that it cannot be specialized for two modalities. In order to overcome these issues and improve the feature diversity, we utilize the last stage of our backbone, where one motion branch is used for our average pooling classifier and the other for our discriminative filters. Such processing allows the motion branch to specialize on context while the spatial branch to specialize on finer details. Next, as we seek the sensitivity to distinguish finer details, we add a bilinear upsampling operation in the spatial branch in charge of computing the discriminative classifiers and a skip connection from the features. These modules provide a specialized and detailed feature volume for the discriminative filters and further enrich the information that the classifiers can learn.

As demonstrate in Figure 3, with the process of training, our method has coerced the two branches to gradually focus on the same region of the network. There are usually many moving objects coexisting in a video clip, only the most noteworthy action can lead the classifier to make the correct prediction. Through the process of the attention consistency, it can assist the network to ignore the misleading backgrounds and wrong actions in the video and concentrate on the valid action region.

4 EXPERIMENTS

We compare our approach with other state-of-art on two video recognition datasets including Kinetics-400 [23] and UCF101 [31] with standard evaluation protocols.

4.1 Datasets

Kinetics-400 dataset is a action video collection extracted from Youtube, where the action videos are categorized as 400 classes and most of them are 10 seconds long. The well-balanced dataset is pre-splitted into 240,000 training videos and 19,800 validation videos. It is a standard practice to report performance on the validation set since the labels of the testset are withheld for the challenges. Kinetics-400 is a popular dataset for evaluation owing to its very large scale, large intra-class variability, and the extensive coverage. UCF-101 dataset contains 13K videos (180 frames/video on average) annotated into 101 action classes, where each UCF-101 split contains 9.5K training videos. We follow the evaluation metrics and report the accuracy over the first test split of UCF-101.

4.2 Implementation Details

Our models on Kinetics-400 are trained from the random initialization (“from scratch") without using any pre-training model. We take synchronized SGD training following the recipe of the slowfast [11] with the backbone R50 4×16. For the temporal domain, fast pathway exploits a 4× smaller temporal stride than slow pathway. For the spatial domain, we randomly generate a 224×224 pixels crop from a video, or its horizontal flip, with the shorter side randomly sampled in $[256, 320]$ pixels.

At the inference time, we uniformly sample 10 clips from a video along its temporal axis following common practice. For each clip, we scale its size into 256×256 and take its three crops to cover the spatial dimensions following the code of slowfast, as an approximation of fully-convolutional testing. We average the softmax scores for the prediction and report the final result.

4.3 Performance on Complex Classes

Our method has achieved convincing performance on complex classes. As observed in Figure 4, the first row (a) is a set of video clips with crowded backgrounds. Under this situation, background distraction occupies a large proportion of the video both in temporal length and spatial distribution. Our stream-independent structure can alienate the action regions from the background noise. The second row (b) is a set of video clips with fast-moving actions existed in both distant view and up-close view. The targeted action regions varies in spatial scales and has a low temporal resolution. Our structure-independent design can urge spatial stream to better focus on semantic information without invalid temporal feature fusion. The third row (c) is a set of video clips with slow actions. As demonstrated in Figure 4, the speed and frequency differences of an action do not weaken the ability of our network to accurately localize actions.

Complex Classes	Improved Accuracy(%)
applauding	0.6
bartending	1.1
beatboxing	0.5
climbing ladder	0.7
playing basketball	1.1
celebrating	2.2
dropkicking	1.2
overall improvement	1.1

Table 1. **Complex classes with crowds**.

Complex Classes	Improved Accuracy(%)
breakdancing	2.0
contact junggling	1.2
playing baseball	0.9
capoeria	1.7
cheerleading	2.2
skateboarding	1.3
archery	0.5
overall improvement	1.4

Table 2. **Complex classes with fast moving actions**.

From the experiments as conducted on public datasets, we find that the stream structure with two independent branches has its own advantage on the action recognition problems under complex scenarios. most present designs on the two-stream network usually facilitate the spatial stream with the feature information extracted from motion stream. However, saliency map shows that the motion stream is prone to broader active areas including all action regions, which might mislead the spatial stream to focus on backgrounds incorrectly. Our network design naturally avoids this confusion by merging two streams only at the end stage of the network, where SCAM and action attention consistency serve as complementary tools to enhance the recognition performance as shown in Figure 2. In our branch-independent design, two pathways are allowed to manage two kinds of input separately, where the slow pathway tends to focus on smaller action regions with more spatial details and the fast pathway pays more attention to fast moving regions.

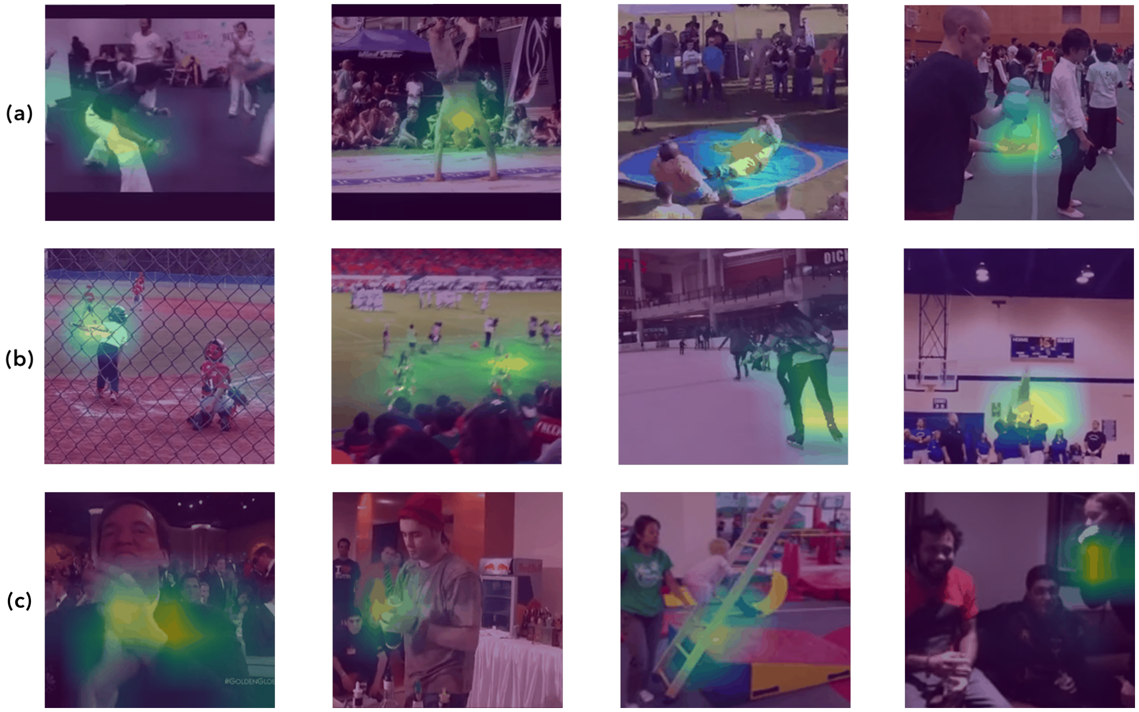

Fig. 4. **A demonstration of action heatmap of complex classes**: All videos are selected from Kintecis400. (a) is a collection of fast actions with crowded background, whose classes are “capoeira, breakdancing, dropkicking and contact juggling" from the left to the right; (b) is a collection of fast actions with foreground distraction, whose classes are “playing baseball, group dance, ice skating, cheerleading" ; (c) is a collection of subtle movements with multiple objects in the scene, where the classes are “applauding, bartending, climbing, beatboxing".

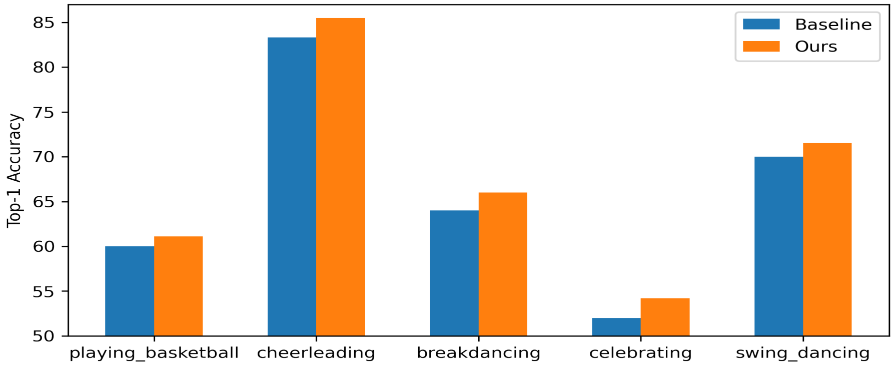

Fig. 5. **Performance comparisons** with the baseline on complex action classes.

4.4 Main Results

We compare our results with previous ones on Kinetics400. An interesting observation is on the potential benefit of adopting two-stream structure, an averaging improvement of 3% over pure 3D network. Existing two-stream models

model	Flow	Pre-train	top-1	top-5	GFLOPs×views
I3D [4]		ImageNet	72.1	90.3	108×N/A
Two-Stream I3D [4]	√	ImageNet	75.7	92.0	216× N/A
R(2+1)D FLOW [35]	√	-	67.5	87.2	152×115
STC [7]		-	68.7	88.5	N/A
S3D [46]		-	69.4	89.1	66.4×N/A
ECO [51]		-	70.0	89.4	N/A
R(2+1)D [35]	√	-	73.9	90.9	304×115
SlowFast 4×16, R50* [11]		-	73.5	90.3	34.2×30
SlowFast 4×16, R50 [11]		-	75.6	92.1	36.1×30
Ours 4×16, R50		-	75.7	92.3	34.2×30

Table 3. **Comparisons with the state-of-the-art methods on Kinetics-400**. In the last column, we report the inference cost with a single "view" (temporal clip with spatial crop) × the numbers of such views used. Our model are with the same input sampling strategy with the baseline. SlowFast 4×16, R50* indicates there are no fusion operations between slow and fast pathway. "N/A" means the numbers are not available for us.

such as ECO and two-steam I3D have achieved competing accuracy while sampling dense frames from video clips. Our method has acquired equal performance with much less frame used. Accordingly, our baseline on the fusion of multiple stages received 0.1% top-1 accuracy slightly lower than our method. It is noticed that the two-stream methods with optical flow + multi-stage fusion can double the computational cost whilst our simplified stream-independent structure is lightweight. Specifically, our model achieves 75.7% without pre-training on the ImageNet dataset, which is 3.6% higher than I3D but 6.3% higher than S3D which uses separable 3D CNN backbone. For further acknowledgement, we observe an improvement from 2% to 4 % by other algorithms with the pre-trained model on the Kinetics-600 dataset and transformer module.

Methods	Backbone	Pre-train	Top-1
TSN [41]	Inception V2	ImageNet	86.4
TSN [41]	Inception V2	ImageNet+Kinetics	91.1
T3D [8]	DenseNet3D	Kinetics	91.7
ECO [51]	BNInception+3D ResNet-18	Kinetics	94.8
C3D [34]	VGGNet-11	Sports-1M	82.3
I3D RGB [4]	3D Inception V1	ImageNet+Kinetics	95.1
ResNeXt[45]	ResNeXt-101	Kinetics	94.5
ARTNet with TSN [40]	ResNet-18	Kinetics	94.3
Slowfast 4×16 [11]	ResNet-50	Kinetics	95.0
Ours	ResNet-50	Kinetics	95.7

Table 4. **Results on UCF101**. Top-1 accuracy is reported based on the three test splits of the dataset. Kinetics refers to kinetic400. We adopt the same frame sampling strategy and spatial crops with our baseline.

For the UCF101 dataset, our approach achieves 95.7% over three test splits, which outperforms the traditional two-stream structures with a large margin and improves upon the baseline by 0.7%. Benefiting from our baseline that adopts the different frame sampling strategy with multiple spatial crops, our approach leads the traditional two-stream models by more than 10%. In comparison with the models that utilize big backbones and 3D convolutions, the advantage of our approach in accuracy still leading by 0.5% to 1.0%.

	Pre-train	top-1	top-5
Slow-Only	ImageNet	73.1	90.9
Slow (SCAM only)	ImageNet	73.4	91.0
Ours (AC only)	-	75.2	91.1
Ours (AC+SCAM)	-	75.7	92.3

Table 5. **Different combination of the network.** AC refers to our attention consistency loss. Slow refers to the spatial branch of slowfast. All experiments are conducted on Kinetics400 and use three spatial crops at test time.

In overall, our approach reduces the computation cost by 7.1% without compromising the accuracy compared with our baseline network. Further experiments show that our model outperforms other approaches in complex scenes such as running in front of the crowd, clapping in a concert and moving in a rapidly changing background.

4.5 Ablative Study

In this section, we conduct extensive ablation experiments to illustrate the performance of each module of our network. The slow only structure utilize merely the slow pathway and is pre-trained from ImageNet. Our two stream structure is trained from random initialization. The sampling strategy and training recipe are the same as described in the experiment section. All comparisons are based on ResNet-50 and use three spatial crops at test time.

We first validate the impact of spatial class activation mapping (SCAM), we implement this module on spatial stream with 4-frame input. Experiment results shows no obvious improvement on accuracy which explains that spatial stream is not sensitive to naive enhanced input. Another finding as shown in the third row of Table 5, the attention consistency loss together with our two-stream structure brings up to 2.1% improvement in top-1 accuracy compared with the single spatial stream. The results confirm the significance of the two-stream structure and support the plausibility of our attention consistency loss. It demonstrates a fact that the fusion between the middle stages of the backbone is not an indispensable procedure for the two-stream recognition model. In the forth row of Table 5, SCAM that mainly enhances the representation ability of the spatial branch improves the top-1 accuracy by 0.5% and exceeds our baseline slightly by 0.1%. The SCAM module only deploys on spatial branch and facilitate the convergence of temporal consistency. It consequently has a little burden on the computation cost and the model size. The temporal stream taking a multi-frame input with extensive activation regions is not suitable for adding a SCAM module.

5 CONCLUSION AND FUTURE WORK

In overall, we investigate a new architecture to measure the attention consistency between the spatial steam and the temporal steam, where our method achieves compelling performance for video action classification. Two stream architecture that employs both motion information and static information has long been a standard protocol for the action recognition. Under the framework, we propose a stream-independent network with our action attention consistency loss. Experimental results on the open datasets including Kinetics400 and UCF101 show that we can attain better overall accuracy but outperform our baseline on complex classes with a large margin. However, Multi-scale inputs and dense sampling require higher GPU memory and the total computation cost and frame usage can be further reduced with new designs. The algorithm is not robust for all cases. As it can be seen, the recognition ability declines on easy classes like "drink" and "smoke". The success of the branch-independent structure suggests that the information exchange between the middle layers of the backbone can be replaceable, thus encouraging future work to explore more potential possibilities when designing the action recognition models.

ACKNOWLEDGEMENT

This work is partially supported by the following: National Natural Science Foundation of China under no.61876155 and 61876154; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no.BE2020006-4; Key Program Special Fund in XJTLU under no.KSF-T-06.

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691* (2021).

[2] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. 2011. A database and evaluation methodology for optical flow. *International Journal of Computer Vision* 92, 1 (2011), 1–31.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? *arXiv preprint arXiv:2102.05095* (2021).

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[5] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. 2018. Multi-fiber networks for video recognition. In *Proceedings of the European Conference on Computer Vision*. 352–367.

[6] AM Derrington and P Lennie. 1984. Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of Physiology* 357, 1 (1984), 219–240.

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. 2018. Spatio-temporal channel correlation networks for action classification. In *Proceedings of the European Conference on Computer Vision*. 284–299.

[8] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. 2017. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200* (2017).

[9] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227* (2021).

[10] Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 203–213.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 6202–6211.

[12] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1933–1941.

[13] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5958–5966.

[14] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 244–253.

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*. 5842–5850.

[16] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6047–6056.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[18] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[19] Kaizhu Huang, Amir Hussain, Qiu-Feng Wang, and Rui Zhang. 2019. *Deep learning: fundamentals, theory and applications*. Vol. 2. Springer.

[20] David H Hubel and Torsten N Wiesel. 1965. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* 28, 2 (1965), 229–289.

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2012), 221–231.

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[24] Ce Li, Chunyu Xie, Baochang Zhang, Jungong Han, Xiantong Zhen, and Jie Chen. 2021. Memory attention networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[25] Yonggang Li, Chunping Liu, Yi Ji, Shengrong Gong, and Haibao Xu. 2020. Spatio-Temporal Deep Residual Network with Hierarchical Attentions for Video Event Recognition. 16, 2s, Article 62 (June 2020), 21 pages. https://doi.org/10.1145/3378026

[26] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*. 7083–7093.

[27] Margaret Livingstone and David Hubel. 1988. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* 240, 4853 (1988), 740–749.

[28] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. 2018. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition*. Springer, 281–297.

[29] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. 2021. An Image is Worth 16x16 Words, What is a Video Worth? *arXiv preprint arXiv:2103.13915* (2021).

[30] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems*. 568–576.

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[32] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. 2020. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 625–634.

[33] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. 2015. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4597–4605.

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.

[35] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6450–6459.

[36] David C Van Essen and Jack L Gallant. 1994. Neural mechanisms of form and motion processing in the primate visual system. *Neuron* 13, 1 (1994), 1–10.

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[38] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision* 103, 1 (2013), 60–79.

[39] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*. 3551–3558.

[40] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. 2018. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1430–1439.

[41] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 20–36.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[43] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. 2020. Boundary-aware cascade networks for temporal action segmentation. In *Proceedings of the European Conference on Computer Vision*. Springer, 34–51.

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*. 3–19.

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1492–1500.

[46] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision*. 305–321.

[47] Haotian Xu, Xiaobo Jin, Qiufeng Wang, and Kaizhu Huang. 2020. Multi-scale Attention Consistency for Multi-label Image Classification. In *International Conference on Neural Information Processing*. Springer, 815–823.

[48] Junxuan Zhang, Haifeng Hu, and Xinlong Lu. 2019. Moving Foreground-Aware Visual Attention and Key Volume Mining for Human Action Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 3, Article 74 (Aug. 2019), 16 pages. https://doi.org/10.1145/3321511

[49] Yue Zhao, Yuanjun Xiong, and Dahua Lin. 2018. Trajectory convolution for action recognition. In *Proceedings of the International Conference on Neural Information Processing Systems*. 2208–2219.

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.

[51] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. 2018. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European Conference on Computer Vision*. 695–712.