Hindawi
Security and Communication Networks
Volume 2022, Article ID 5094801, 8 pages
https://doi.org/10.1155/2022/5094801

Research Article

Integrating Temporal and Spatial Attention for Video Action Recognition

Yuanding Zhou,[1] Baopu Li,[2] Zhihui Wang,[1] and Haojie Li[1]

[1]*Dalian University of Technology, Dalian 116024, Liaoning Province, China*
[2]*Baidu Research, Sunnyvale, CA 94089, USA*

Correspondence should be addressed to Zhihui Wang; zhwang@dlut.edu.cn

Received 14 February 2022; Revised 14 March 2022; Accepted 22 March 2022; Published 26 April 2022

Academic Editor: Beijing Chen

In recent years, deep convolutional neural networks (DCNN) have been widely used in the field of video action recognition. Attention mechanisms are also increasingly utilized in action recognition tasks. In this paper, we want to combine temporal and spatial attention for better video action recognition. Specifically, we learn a set of sparse attention by computing class response maps for finding the most informative region in a video frame. Each video frame is resampled with this information to form two new frames, one focusing on the most discriminative regions of the image and the other on the complementary regions of the image. After computing sparse attention all the newly generated video frames are rearranged in the order of the original video to form two new videos. These two videos are then fed into a CNN as new inputs to reinforce the learning of discriminative regions in the images (spatial attention). And the CNN we used is a network with a frame selection strategy that allows the network to focus on only some of the frames to complete the classification task (temporal attention). Finally, we combine the three video (original, discriminative, and complementary) classification results to get the final result together. Our experiments on the datasets UCF101 and HMDB51 show that our approach outperforms the best available methods.

1. Introduction

As an important communication medium, video contains a wealth of information. But this information used to be extracted and used manually, which is time-consuming and laborious. With the development of deep learning, attempts have been made to allow computers to extract information from videos. Many video-based deep learning tasks have emerged, such as video action localization [1], video captioning [2], and video question-answering [3]. The video action recognition task is to derive the behavior of a person in a video by analyzing the video content. This task is essentially a classification task where the input is a video and the output is action labels. This task has a wide range of application scenarios; most typically, it can detect violent action in surveillance videos and help police investigate and collect evidence [4].

With the development of deep learning, many excellent methods for video action recognition have emerged. Video is composed of many frames, so the understanding of video should include the relationship between frames in addition to the image frames themselves. Therefore, the classical two-stream network [5] divided the video into two parts: spatial and temporal. Spatial part is the information of video frame, for which there are many excellent 2D CNN structures available, such as ResNet [6] and Inception [7], while the temporal information comes from the association between frames, and this part was obtained by optical flow. Finally the temporal and spatial information were integrated together to classify the video. Temporality is an important feature of video; many researchers spend their efforts on how to better capture the relationship between videos in the temporal dimension [8]. In addition to the temporal dimension, the information extraction of video frames itself is also an important part. So researchers also gradually put their efforts back to the images themselves in recent years. SlowFast [9] sampled the original video at different frame rates. The slow

path learned spatial information with few frames and the fast path learned temporal information with a large number of frames and then combined it with nonlocal network to model the relationship between frames from a global perspective. Video transformer [10] used transformer instead of convolution to compute the internal relationship of the whole video. But it was too computationally intensive to compute both temporal and spatial attention for each patch of each frame. So they proposed another architecture that temporal attention and spatial attention are separately applied one after the other. They found that the latter one not only reduced the computational effort significantly, but also had a higher accuracy in the end.

We find that previous attention methods tend to favor only one of temporal or spatial attention or treat all video frames with the same attention strategy, like the different frame rates of SlowFast [9], the transformer used by [10]. Inspired by previous approaches [11], we find that temporal and spatial attention can complement each other to improve the final classification. So in this paper, we first propose a spatial attention mechanism that extracts discriminative regions in video frames and resamples them into two new videos. These two videos are like a data augmentation of the original video. We then feed these two videos together with the original video into a temporal attention network with a frame selection strategy to filter out the most useful frames for classification task. Finally our network learns the most discriminative regions in these most useful frames, resulting in a more accurate result. At the same time, since the network where we extract spatial attention and the network that finally completes the classification task are the same, our extraction of discriminative regions in image frames is also getting accurate as well as the final classification accuracy. A positive beneficial cycle is formed to continuously improve our classification results.

The main innovations and contributions of this paper are as follows: (1) We propose a novel sparse attention mechanism for extracting important regions from video frames, and the method extracts discriminative regions while preserving contextual information. We leverage this proposed method as spatial attention. (2) We combine the spatial attention with our previously proposed frame selection strategy [12] to jointly form a novel network structure containing both temporal and spatial attention. (3) Experiments on two datasets commonly used for video action recognition, UCF101 and HMDB51, show that our approach outperforms the best available methods.

In Section 2, the structure of the proposed network, loss function, and other related contents will be introduced. In Section 3, the experimental results of our method and the implement details will be introduced. The advantages of our scheme will be summarized in Section 4.

2. Proposed Method

Inspired by [13], we find that the class peak responses typically correspond to strong visual cues residing inside regions of interest. As shown in Figure 1, we first feed the original video into a pretrained CNN with temporal attention (T-CNN) to extract features (Features in Green). This part of features is sent to the spatial attention network to activate class response maps that allows the network to focus on the important part of the video frames. Based on the peaks of class response map each frame of the video is resampled into two new video frames. One of these two video frames focuses on the discriminative region of the image (the orange frame which enlarges the barbell part of the original frame) and the other focuses on the complementary part (the blue frame which enlarges the human body). These two branches then rearrange the video frames into two new videos in the same order as the original video. These two videos will also be fed into the T-CNN as new inputs. Each of these three branches is optimized by a cross-entropy loss function. Finally, the three video branches are jointly optimized to obtain a more accurate classification result.

2.1. Obtaining Class Peak Response Point. We first feed the video into T-CNN (which will be introduced in Section 2.3) that has been trained to extract the feature maps $\mathbf{X} \in \mathrm{R}^{T\times C\times H\times W}$, where T represents the number of frames, $H \times W$ is the size of the feature maps, and C is the number of channels. Then we feed the feature map into a global average pooling (GAP) layer and then go through a fully connected (FC) layer to get the classification score $\mathbf{x} \in \mathrm{R}^S$, where S is the number of categories in the dataset. We expand the feature maps along time dimension into T maps, each with dimension $\mathbf{Y} \in \mathrm{R}^{C\times H\times W}$. Then we let each of these maps go through a GAP and FC layer to get the classification score $\mathbf{y} \in \mathrm{R}^S$ for each frame. With the weight matrix of the FC layer $\mathbf{W^{fc}} \in \mathrm{R}^{C\times S}$, we can compute the class response map $\mathbf{M}_s$ as

$$\mathbf{M}_S = \sum_{c=1}^{C} \mathbf{W}^{\mathbf{fc}}_{c,s} \times \mathbf{Y}_c. \tag{1}$$

The class peak response of class c is defined as a local maximum of the corresponding class response map $\mathbf{M}_c$. The class peak point can be written as $P_c = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{Ns}, y_{Ns})\}$, where Ns is the number of valid peak points in the s-th class. We use these peak points to locate regions that are more important for the classification task and estimate a set of sparse attentions.

Experiments show that peaks in top-1 class response map tend not to cover all discriminative regions, while peaks in top-5 tend to contain the noise points. To seek a balance between these two methods of choosing peak points, we first calculate their entropy to determine the confidence of network predictions. If the confidence is high, we use peaks from the top-1 class response map, and if it is lower, we bring together the top-5 five class response maps to find the peak points. We denote the predicted probability of all S classes as $\mathbf{Prob} = \mathrm{softmax}(\boldsymbol{y}) \in R^S$ and use $\overline{\mathbf{Prob}} \in R^5$ to denote the probability value of the top-5 classes. We compute the entropy as

$$H = -\sum_{i=1}^{5} \mathbf{p}_i \log \mathbf{p}_i, \quad \mathbf{p}_i \in \overline{\mathbf{Prob}}. \tag{2}$$

We construct a response map $\mathbf{R_{map}}$ with the following strategy:

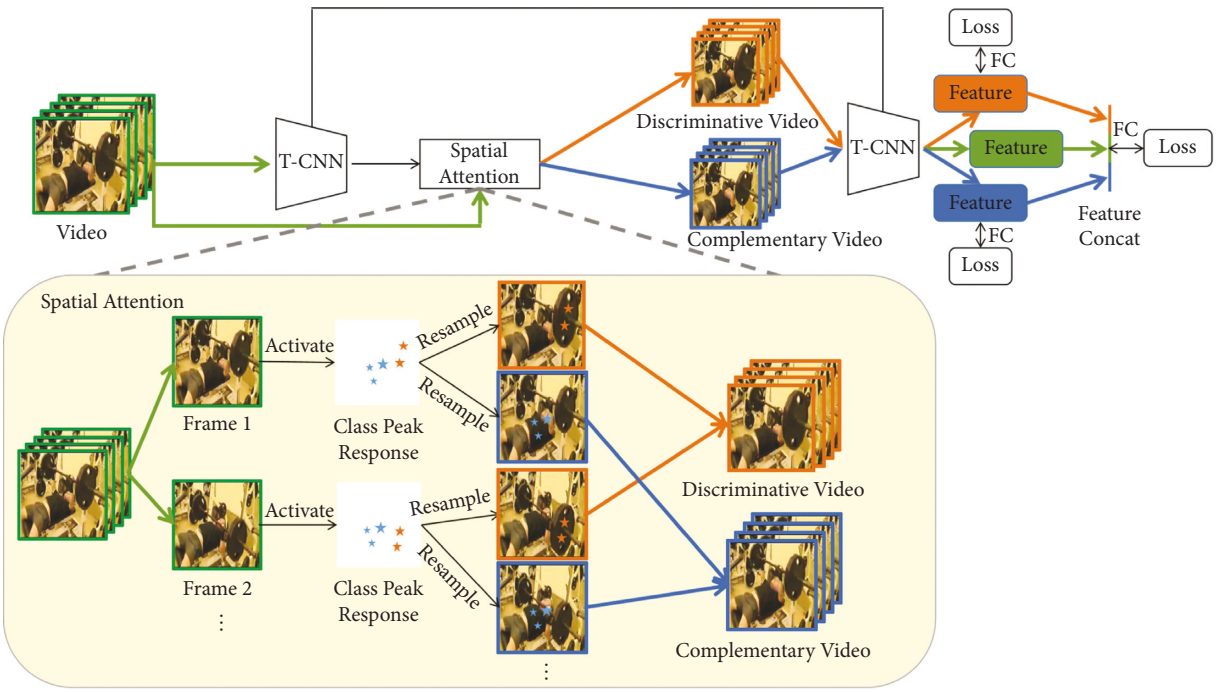


Figure 1: Network structure of our method.

$$\mathbf{R}_{\mathbf{map}} = \begin{cases} \overline{\mathbf{M}_1}, & \text{if } H < \delta, \\ \sum_{k=1}^{5} \overline{\mathbf{M}_k}, & \text{if } H > \delta, \end{cases} \tag{3}$$

where $\overline{M} \in R^{5\times H\times W}$ is the class response maps corresponding to $\overline{\mathbf{Prob}}$. Then we use Min-Max Normalize to map the values of $\mathbf{R}_{\mathbf{map}}$ to [0, 1].

$$\overline{\mathbf{R}_{\mathbf{map}}} = \frac{\mathbf{R}_{\mathbf{map}} - \min\left(\mathbf{R}_{\mathbf{map}}\right)}{\max\left(\mathbf{R}_{\mathbf{map}}\right) - \min\left(\mathbf{R}_{\mathbf{map}}\right)}. \tag{4}$$

We denote their positions as $P = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{Np}, y_{Np})\}$, where N_p is the number of peaks we detected by the above procedure.

2.2. Computing Sparse Attention and Resampling. Reference [14] found that, in fine-grained image classification task, the obtained class peak points can be divided into two sets. One set is the discriminative region and the other is the complementary region, and learning these two sets separately is better than learning all class peak points together directly. Inspired by them, we preset a random number $\varphi_{(x,y)}$ from the uniform distribution between 0 and 1. We compare the response value $\overline{\mathbf{R}_{\mathbf{map}}}$ of the peak point with this random number φ and group all points with response values greater than φ into one set P_{dis} and those less than into another set P_{com}.

$$\begin{cases} P_{\text{dis}} = \left\{(x, y) | (x, y) \in P \quad \text{if} \overline{\mathbf{R}_{\mathbf{map}}}_{(x,y)} \geq \varphi\right\}, \\ P_{\text{com}} = \left\{(x, y) | (x, y) \in P \quad \text{if} \overline{\mathbf{R}_{\mathbf{map}}}_{(x,y)} < \varphi\right\}. \end{cases} \tag{5}$$

The left part of Figure 2 is the original video frame, where the orange dot is the center point of the attention map (middle). As shown in Figure 2(a), points with high response values tend to correspond to discriminative regions, such as bow and arrow, and these peak points are generally grouped into the P_{dis} set. The points with low response values are usually localized at complementary regions as illustrated in Figure 2(b), that is, usually people in the video or the subject of the action, and these peak points will be grouped into the P_{com} set.

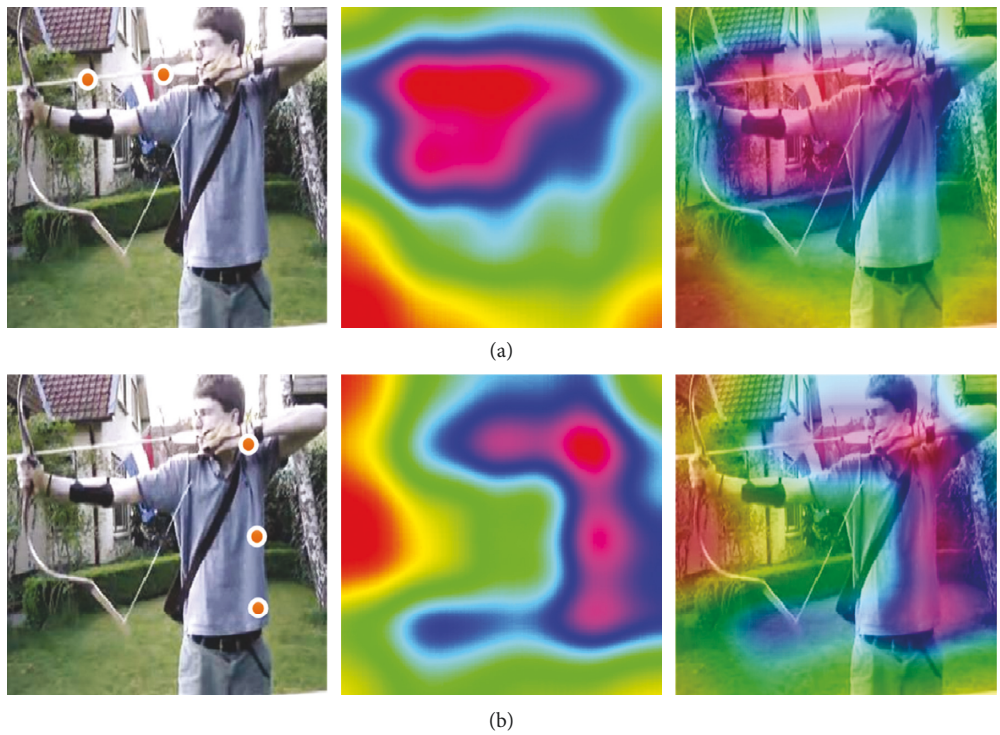
For each peak set, we compute a set of sparse attention $\mathbf{A} \in R^{Np\times H\times W}$ using Gaussian kernel.

$$\mathbf{A}_{i,x,y} = \begin{cases} R_{x_i,y_i} e^{-(x-x_i)^2+(y-y_i)^2/R_{x_i,y_i}\beta_1^2}, & \text{if } (x_i, y_i) \in P_{\text{dis}}, \\ \frac{1}{R_{x_i,y_i}} e^{-(x-x_i)^2+(y-y_i)^2/R_{x_i,y_i}\beta_2^2}, & \text{if } (x_i, y_i) \in P_{\text{com}}. \end{cases} \tag{6}$$

Both β_1 and β_2 are learnable parameters.

With the previously obtained sparse attention, we can resample the discriminative regions from the original video frames while also preserving the contextual information around the image regions. After the above series of operations, each video frame can be resampled to obtain two new frames, and we use $\mathbf{Q}_{\text{dis}}$ to denote the feature map of the extracted discriminative branch and $\mathbf{Q}_{\text{com}}$ to correspond to the feature map of the complementary branch.

$$\begin{cases} \mathbf{Q}_{\text{dis}} = \sum \mathbf{A}_i, & \text{if } (x_i, y_i) \in P_{\text{dis}}, \\ \mathbf{Q}_{\text{com}} = \sum \mathbf{A}_i, & \text{if } (x_i, y_i) \in P_{\text{com}}. \end{cases} \tag{7}$$



(a)

(b)

Figure 2: Visualization of discriminative and complementary branches: (a) discriminative branch; (b) complementary branch.

The resampling process is implemented using convolution following the method of [15] and can be embedded into the end-to-end training. So both β_1 and β_2 can be updated by the classification loss function. The input video has multiple frames and each frame of video can produce two frames according to the above method; we then line up all the discriminative and complementary branch images in the order of the original video input to form two new videos $\mathbf{V_{com}}$ and $\mathbf{V_{dis}}$.

2.3. Network Structure and Loss Function. Essentially, the sparse attention we propose is for spatial attention of video images. Our input is an original video $\mathbf{V_o}$ and the output is two videos $\mathbf{V_{com}}$ and $\mathbf{V_{dis}}$. The video frames of $\mathbf{V_{dis}}$ focus more on discriminative regions, while $\mathbf{V_{com}}$ focuses on regions that are complementary.

As we mentioned before, in order to obtain the two new videos, we first feed the original video into a pretrained CNN with temporal attention (T-CNN) to extract features. Table 1 shows the network structure of T-CNN; "Dilation Conv(4)" means a dilation convolution with a dilation of 4 is used in the temporal dimension. T-CNN comes from a network structure that we obtained previously using neural architecture search [12]. In this work we explored how many frames are needed in each stage of the network. In fact, it is about allowing the network to focus on only the appropriate number of video frames to complete the final classification task. After we get the two new videos, $\mathbf{V_{com}}$ and $\mathbf{V_{dis}}$, we will refeed them to T-CNN as new data to learn. These two new videos are equivalent to a data augmentation of our original input. So our method does not significantly improve the number of model parameters, although the computational complexity increases.

Our loss function is a cross-entropy loss. Each input video will produce three predictions, which we denote as $\mathbf{F_o}$, $\mathbf{F_{com}}$, and $\mathbf{F_{dis}}$. These three predictions come from the original video $\mathbf{V_o}$, the discriminative video $\mathbf{V_{dis}}$, and the complementary video $\mathbf{V_{com}}$, respectively. Comparing them with the classification labels will produce three losses. We will also concatenate the features of the three videos together and pass them through a FC layer to obtain the fourth prediction $\mathbf{F_{total}}$. So our final loss function consists of four components, which can be written as

$$L(\mathbf{X}) = \sum_{i \in \{O,C,D\}} L_{cls}(\mathbf{F}_i, \mathbf{F}^*) + L_{cls}(\mathbf{F_{total}}, \mathbf{F}^*), \tag{8}$$

where L_{cls} denotes the cross-entropy loss and $\mathbf{F}^*$ is the ground-truth label vector.

3. Experimental Results and Discussions

3.1. Datasets and Implementation Details. To evaluate the effectiveness of our proposed method, we have done experiments on two common datasets for video action recognition, UCF101 and HMDB51.

The UCF101 dataset [16] has 13,320 videos from 101 action categories. Each of these categories can be divided into 25 groups, each containing 4–7 action videos. This dataset is highly diverse in terms of motion and varies greatly in terms of camera movement, object appearance and pose, object scale, point of view, cluttered backgrounds, lighting conditions, etc.

Table 1: Network structure of T-CNN.

Input: $3 \times 16 \times 224 \times 224$	
Stage 1	Conv 3–32 + BN + ReLU
	Conv 32–32 + BN + ReLU
Stage 2	Conv 32–64 + BN + ReLU
	Conv 64–64 + BN + ReLU
Stage 3	Conv 64–96 + BN + ReLU
	Conv 96–96 + BN + ReLU
Stage 4	Conv 96–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Dilation Conv(4) 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Dilation Conv(4) 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Dilation Conv(4) 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Conv 160–160 + BN + ReLU
	Dilation Conv(4) 160–160 + BN + ReLU
Stage 5	Conv 160–224 + BN + ReLU
	Conv 224–224 + BN + ReLU
	Conv 224–224 + BN + ReLU
	Dilation Conv(2) 224–224 + BN + ReLU
	Conv 224–224 + BN + ReLU
	Dilation Conv(4) 224–224 + BN + ReLU
	Conv 224–224 + BN + ReLU
	Dilation Conv(4) 224–224 + BN + ReLU
Stage 6	Conv 224–288 + BN + ReLU
	Conv 288–288 + BN + ReLU
	Conv 288–288 + BN + ReLU
	Dilation Conv(2) 288–288 + BN + ReLU
	Conv 288–288 + BN + ReLU
	Dilation Conv(2) 288–288 + BN + ReLU
	Conv 288–288 + BN + ReLU
	Dilation Conv(2) 288–288 + BN + ReLU
	Conv 288–288 + BN + ReLU
	Conv 288–288 + BN + ReLU
Stage 7	Conv 288–512 + BN + ReLU
	Conv 512–512 + BN + ReLU
	Conv 512–512 + BN + ReLU
	Dilation Conv(2) 512–512 + BN + ReLU
Global average pooling	
Fully connected layer	
Softmax	
Classification result	

The HMDB51 dataset [17] contains 51 action categories, a total of 6849 videos, and each action contains at least 51 videos. The action categories can be divided into four major categories: (1) general facial actions (laughing, chewing); (2) facial and object actions (smoking, eating); (3) human body actions (hugging, inversion); (4) interactive actions with objects (horse riding, archery).

For both video datasets, during training, we sample 16 consecutive frames from each video, and each frame is converted to 256×342 resolution by preprocessing. And then we randomly crop 224×224 pixels from the frame and feed them into the network. To make a fair comparison with other methods, we follow the common reference method [18]. We divide each video into 10 clips equally, with each clip including 16 video frames, resize the short edge of each image to 224 pixels, and cut three 224×224 crops from the left, middle, and right of the image. Each crop of each clip is called a "view," so we have 30 views, and the final prediction result of each video is obtained by averaging the softmax scores of these 30 views.

The whole model is trained for 150 epochs with a batch size of 16. We use SGD optimizer with 0.9 momentum and 4×10^{-5} weight decay. The learning rate strategy uses cosine annealing learning rate schedule [19]. The initial learning rate was 0.1 and the lowest was 1×10^{-4}. The dropout probability is 0.5 after the final GAP layer. Finally, it is sent to the linear layer to classify according to the number of classes of each dataset.

3.2. Comparison with SOTA. On the two commonly used datasets, UCF101 and HMDB51, we compare the proposed method with the SOTA methods. Since our method uses only RGB images as input, when comparing with other methods that have multiple input modalities like I3D, we only compare with their results obtained with RGB modality. From the results of the comparisons in Tables 2 and 3, we can see that the classification accuracy of our method on both datasets exceeds the best available methods.

The advantage of our method comes first from our spatial attention. From the results, the network of T-CNN with only temporal attention has lower accuracy than TSM and I3D RGB on both datasets. In particular, for I3D RGB, T-CNN is 0.3% lower than it on UCF101 and 1.5% lower than it on HMDB51. When the spatial attention proposed in this paper is added, our method achieves a reversal on both datasets. This effect is related to these two datasets, which are more sensitive to spatial information, so the increase of attention to spatial information will produce such a huge improvement (1.4% for UCF101 and 1.9% for HMDB51). Then there is the fact that the spatial attention in this paper is finally externalized to two new data inputs, which actually has the effect of data augmentation. This is very important because both UCF101 and HMDB51 are easy to overfit. Data augmentation helps to improve generalization ability and reduce the occurrence of overfitting.

The second advantage of our approach comes from the fact that we integrate spatial and temporal attention, allowing them to complement each other and improve the final classification accuracy. Not all video frames have positive implications for classification. As shown in Figure 3, this image is difficult to classify based on the original picture and spatial attention. It is easy to be classified as "holding something" rather than "shooting an arrow." At this point, we can rely on temporal attention in the frame selection strategy to reduce our chances of picking this image frame, thus reducing the number of misleading cases.

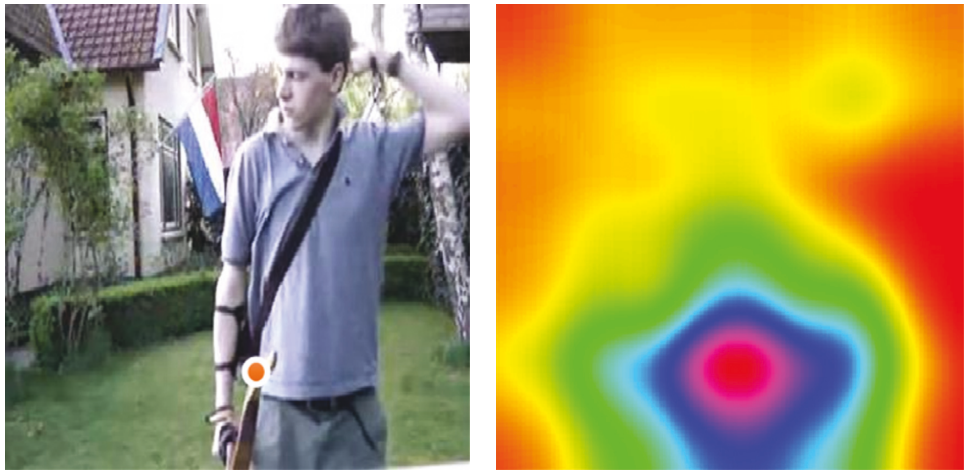
There are also shortcomings in our method. From the last column in Tables 2 and 3, we can see that the computational complexity of our method has increased several

Table 2: Comparisons with other methods on UCF101 dataset.

Model	Pretraining dataset	Accuracy (%)	GFLOPs
C3D [20]	Sports-1M	82.3	38.57
TRN [21]	—	83.5	83.83
Res3D [22]	Sports-1M	85.8	—
P3D [23]	Imagenet + Sports-1M	88.6	18.51
T3D [24]	Kinetics-400	90.3	—
TSN [8]	Imagenet + Kinetics-400	91.1	80
R(2 + 1)D [25]	Sports-1M	93.6	41.69
TSM [26]	Kinetics-400	95.5	32.88
I3D RGB [27]	Imagenet + Kinetics-400	95.6	108
T-CNN [12]	Kinetics-400	95.3	15.78
T-CNN + spatial	Kinetics-400	96.7	52.3

Table 3: Comparisons with other methods on HMDB51 dataset.

Model	Pretraining dataset	Accuracy (%)	GFLOPs
Res3D [22]	Sports-1M	54.9	—
T3D [24]	Kinetics-400	59.2	—
R(2 + 1)D [25]	Sports-1M	66.6	41.69
TSM [26]	Kinetics-400	73.6	32.88
I3D RGB [27]	Imagenet + Kinetics-400	74.8	108
T-CNN [12]	Kinetics-400	73.3	15.78
T-CNN + spatial	Kinetics-400	75.2	52.3



(a) (b)

Figure 3: An example of classification error based only on spatial attention. (a) original frame. (b) Spatial attention.

times compared to T-CNN. The T-CNN has the lowest computational complexity among existing methods (15.78 GFLOPs), but with the addition of the spatial attention part, the computational complexity comes directly to the back half of the list. This is mainly due to the fact that our spatial attention resamples two new videos into the network, which equates to one video input that needs to be computed 3 times through the network, plus the fact that we need to compute the class response maps and sparse attentions for each frame and resample them. All of them add to the computational complexity.

3.3. Effects of Different Extraction Branches. To verify the effects of each branch, we tried to omit one or more branches and observe their effects on the final classification results. From Table 4, we can draw the following conclusions. (1) Both O + D mode and O + C mode are improved for the final classification accuracy. It indicates that both complement and discriminative regions are helpful for classification, and it also verifies that the spatial attention extraction method in this paper is effective. (2) In the absence of the complementary branch, our overall accuracy decreases the least (from 96.7 to 96.3), indicating that the complementary

Table 4: Ablation study on UCF1051 dataset based on different branches.

Branches	Original branch	Discriminative branch	Complement branch	Total accuracy
O	95.3	—	—	95.3
O + D	96.0	95.8	—	96.3
O + C	95.5	—	95.0	95.8
C + D	—	95.9	95.0	96.2
O + C + D	96.2	95.6	94.8	96.7

branch is indeed the region containing the least discriminative information compared to the other branches. However, the accuracy of the classification still decreases when this part is missing, suggesting that sometimes the subject of the action can also play a crucial role in the classification task.

4. Conclusions

In this paper we integrate temporal and spatial attention to construct a network structure. We learn a set of sparse attention by computing class response maps. It selectively collects visual evidence of dynamic information areas based on image content and surrounding context. Based on these regions obtained by spatial attention we resampled two new videos. These new videos are fed into the network as completely new data, enhancing the generalization ability of our network structure. We then feed these two videos with the spatial attention together with the original video into a temporal attention network. So our network learns the most discriminative regions in these most useful frames, resulting in a more accurate result. And the network where we extract spatial attention is the same as the network that finally completes the classification task. So our extraction of discriminative regions in image frames is also getting accurate as well as the final classification accuracy. A positive cycle is formed to continuously improve the classification results. Integrating attention in temporal and spatial is actually consistent with human vision. We also recognize the other person's action by observing key object information in consecutive actions. Extensive experimental results on some benchmark datasets illustrate the promising performance of the proposed scheme.

Data Availability

The datasets used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] F. Wang, G. Wang, Y. Du, Z. He, and Y. Jiang, "A two-stage temporal proposal network for precise action localization in untrimmed video," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 8, pp. 2199–2211, 2021.

[2] P. Li, P. Zhang, and X. Xu, "Graph convolutional network meta-learning with multi-granularity POS guidance for video captioning," *Neurocomputing*, vol. 472, pp. 294–305, 2022.

[3] J. Zhang, J. Shao, R. Cao, L. Gao, X. Xu, and H. T. Shen, "Action-centric relation transformer network for video question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 63–74, 2022.

[4] M. Bruno, B. Lavi, Z. Dias, and R. Anderson, "Harnessing high-level concepts, visual, and auditory features for violence detection in videos," *Journal of Visual Communication and Image Representation*, vol. 78, Article ID 103174, 2021.

[5] K. Simonyan and A. Zisserman, "Two-Stream convolutional networks for action recognition in videos,," in *Proceedings of the 27th International Conference on Neural Information Processing Systems NIPS*, pp. 568–576, Montreal Canada, December 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[7] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 1–9, Boston, MA, June 2015.

[8] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision ECCV*, no. 8, pp. 20–36, Amsterdam, The Netherlands, October 2016.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proceedings of the 2019 International Conference on Computer Vision ICCV*, pp. 6201–6210, Seoul, Korea, November 2019.

[10] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning ICML*, pp. 813–824, July 2021.

[11] Y. Wang, Z. Chen, H. Jiang, S. Song, Y. Han, and G. Huang, "Adaptive focus for efficient video recognition," in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, October 2021.

[12] Y. Zhou, B. Li, Z. Wang, and H. Li, "Video action recognition with neural architecture search," in *Proceedings of the 13th Asian Conference on Machine Learning ACML*, pp. 1675–1690, November 2021.

[13] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *Proceedings of the CVPR*, pp. 3791–3800, Salt Lake City, Utah, June 2018.

[14] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision ICCV*, pp. 6598–6607, Seoul, Korea (South), November 2019.

[15] A. Recasens, P. Kellnhofer, S. Simon, W. Matusik, and A. Torralba, "Learning to zoom: a saliency-based sampling layer for neural networks," in *Proceedings of the ECCV*, no. 9, pp. 52–67, Munich, Germany, September 2018.

[16] K. Soomro, Amir Roshan Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," 2012, https://arxiv.org/abs/1212.0402.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the 2011 International Conference on Computer Vision ICCV*, pp. 2556–2563, Barcelona, Spain, November 2011.

[18] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: temporal excitation and aggregation for action recognition," in *Proceedings of the CVPR*, pp. 906–915, Seattle, Washington, June 2020.

[19] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *Proceedings of the ICLR*, Palais des Congrès Neptune, Toulon, Fr, April 2017.

[20] Du Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision ICCV*, pp. 4489–4497, Santiago, Chile, 2015.

[21] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision ECCV*, no. 1, pp. 831–846, Munich, Germany, September 2018.

[22] D. Tran, J. Ray, S. Zheng, S. F. Chang, and M. Paluri, "ConvNet architecture search for spatiotemporal feature learning,," 2017, https://arxiv.org/abs/1708.05038.

[23] J. Wei, H. Wang, Y. Yi, Q. Li, and D. Huang, "P3D-Ctn: Pseudo-3D convolutional tube network for spatio-temporal action detection in videos," in *Proceedings of the 2019 IEEE International Conference on Image Processing ICIP*, pp. 300–304, Taipei, Taiwan, September 2019.

[24] D. Ali, M. Fayyaz, V. Sharma et al., "Temporal 3D ConvNets: new architecture and transfer learning for video classification," 2017, https://arxiv.org/abs/1711.08200.

[25] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, pp. 6450–6459, Salt Lake City, UT, USA, June 2018.

[26] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision ICCV*, pp. 7082–7092, 2019.

[27] J. Carreira, A. Zisserman, and Q. Vadis, "Action recognition? A new model and the kinetics dataset," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 4724–4733, Honolulu HI, USA, July 2017.