A Novel Two-stream Architecture Fusing Static And Dynamic Features for Human Action Recognition

Fan Yue,Shijian Huang*,Qingming Chen, Siyan Hu,Yong Tan,Suihu Dang, and Derong Du

***Abstract*—Action recognition in real-world videos is difficult because of factors such as scenery muddle, scale alter, dynamic standpoint, and sharp motion.This paper proposes a novel two-stream architecture fusing static and dynamic features to recognize human actions in videos. Firstly, the original image (single frame and optical flow fields) is extracted by Convolutional Neural Network (CNN) to obtain feature maps. Secondly, we extract the obtained feature maps via a 3×3 convolution over all the neighbor features, leading to a static representations of features. Then we concatenate these static features with input feature maps to produce the dynamic attention matrix through two 1×1 convolutions. All of the generated feature maps are then aggregated using the learnt attention matrix, producing a dynamic representation. Thirdly, we take the interaction of the static and dynamic presentations as final outputs. Finally, we utilize Long Short-Term Memory (LSTM) to catch time sequence information among dense optical flow.The experimental results on the three hard datasets UCF101, HMDB51, and Kinetics400 have shown that the method works better than other state-of-the-art methods.**

***Index Terms*—Human action recognition, Two-stream structure, Static features, Dynamic features, Feature fusion**

I. Introduction

THE field of visual computing[1] has recently put increasing emphasis on action recognition, which is a task to infer human actions based on consecutive action executions. Recognizing and distinguishing actions in a video sequence from a surveillance stream is a particularly interesting and challenging field of study in the present day.Monitoring the behavior of old people, surveillance systems, human and computer interaction, video retrieval, public opinion polling, and many more applications are just some of the many ways that action recognition has been put to use.

In recent years, many researchers have developed similar methods for high-level discriminative features learning and making end-to-end systems in video-based action recognition.The effectiveness of adopting Convolutional Neural Networks (ConvNets) on somehow images inspired these researchers. current approaches can be separated into Recurrent Neural Networks (RNN)[2], 3D Convolutional Networks[3], and two-stream Convolutional Networks[4] according to the dimension of the convolution unit that is utilized in the network.LSTM[5]is a special RNN that, because of its memory cells and door control structure, is able to successfully extract information about time sequences information, but it is not sufficient for abstracting pixel-level information.3D Convolution Networks[6] incorporate a temporal dimension. In current data-driven action recognition designs, 3D convolution has been frequently employed. Through 3D convolution, the video's extracted features contain information about both time and space that can be used to generate results for action recognition. 3D Convolutional Networks have a low recognition efficiency,because the accuracy of 3D convolution depends on a large number of network parameters. This obviously impedes the achievement of real-time action recognition.Recent work has established a two-stream architecture to address this issue by utilizing both RGB and optical flow pictures as input data. This model is useful in capturing spatial and motion information. spatial and temporal information plays a crucial role in video recognizing different human actions.

Inspiring by the aforementioned research, we present a novel two-stream architecture for action recognition by fusing static and dynamic information and utilizing ResNet101 as the fundamental framework. The CoT block[7] replaces the 3×3 convolution in ResNet101 architecture called CotNet101. By constructing appropriate representations, the proposed static and dynamic interaction two-stream network is able to overcome the difficulties given by complicated actions.

This paper has two primary contributions. The first is the proposal of novel two-stream architecture. We extract appearance and motion features by CotNet101 architecture on each stream of the two-stream network, and LSTM is used to model time sequence information once motion features have been extracted. The second is that we proposed a data variation augmentation strategy. We randomly rotate and erase the contents of each frame and flip it horizontally. This increases the input variations strengthens the connection

Manuscript received May 1,2022;revised January 25,2023. This work was supported in part by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202001421, KJQN202101446, KJQN202101437), Cooperative Projects between Undergraduate Universities in Chongqing and Institutes affiliated with Chinese Academy of Sciences (Grant No.HZ2021014), University Innovation Research Group of Shale Gas Optical Fiber Intelligent Sensing Technology (Grant No. CXQT20027).

Fan Yue is a postgraduate student of the School of Electronic & Information Engineering of Chongqing Three Gorges University, Chongqing, China. (phone: +8615832232209; email: yuefanpaper@163.com).

Shijian Huang is an associate professor of the Key Laboratory of Micro Nano Optoelectronic Devices and Intelligent Perception Systems, School of Electronic Information and Engineering,Yangtze Normal University, Chongqing, China. (phone:+8615123619526; email: huangshijian@yznu.edu.cn). Shijian Huang is the corresponding author.

Qingming Chen is an undergraduate student of Yangtze Normal University, Chongqing, China. (email: 2476255383@qq.com).

Siyan HU is an undergraduate student of Yangtze Normal University, Chongqing, China. (email: 2377764874@qq.com).

Yong Tan is a professor of Yangtze Normal University, Chongqing, China. (email: tanyong@yznu.edu.cn).

Suihu Dang is a professor of Yangtze Normal University, Chongqing, China. (email: dangsuihu@yznu.edu.cn).

Derong Du is an associate professor of Yangtze Normal University, Chongqing, China. (email: duderong@yznu.edu.cn).

between RGB and optical flow and overcomes the problem to a certain extent. We obtain a competitive performance on the UCF101, HMDB51 and Kinetics400 dataset compared with the other state-of-the-art works. Fusing static and dynamic representations can improve the expression of features, and combing the LSTM can enhance the ability of optical flow field representation verified by experiment.

II. RELATED WORK

Self-attention on Classification Task. The research community focuses more on self-attention in visual scenarios, inspired from self-attention in Transformer, which achieves great performance in many NLP tasks. The initial self-attention mechanism in the NLP area[8] intends to get long-range dependency in sequence modeling.It developed for NLP can be easily applied for use in the vision field[9], where it can be used to guide attention across multiple feature vectors located in various parts of an image.Specifically, one of the earliest initiatives to investigate self-attention in Convolutional Networks (ConvNets) is the nonlocal operation,which is a supplementary building component for employing self-attention on convolutional outputs. integrates a global multi-head self-attention method[10] with convolutional operators to optimize picture classification performance. Without using global self-attention across the whole feature map, that really doesn't scale well, should be using local self-attention within a patch. Such a design of local self-attention significantly restricts the number of parameters and computations utilized by the network, and can therefore completely replace convolutions throughout the entire deep architecture.

Deep Learning with Action Recognition. Rapid advances in deep learning have resulted in numerous works in action recognition being investigated in order to gain deep features and devise efficient architectures for video action recognition. We briefly divide the ways to recognize actions into three groups: RNN, 3D Convolutional Networks, and two-stream Convolutional Networks.A recurrent network composed of LSTM cells was used by Donahut et al.[11] to process the returned spatial features at each time step. Unlike conventional models, which have limited spatial and temporal regions and can handle a count of temporal inputs.The research from Karpathy et al.[12] modified the first convolutional way to obtain the spatiotemporal features by stacking successive video frames,and they tested with both early and slow fusion methods.Going this approach further,C3D[13] uses GPU memory inefficiently to switch out all 2D convolutional kernels for 3D kernels.When training 3D convolutional kernels, the high complexity can be avoided by parameterizing the 3D kernels into 2D spatial and 1D temporal kernels,as Sun et al.[14] conducted.Furthermore,another subfield of action recognition research involves the extraction of temporal information from conventional optical flow images.In the field of action recognition, the two-stream ConvNet is now the most prevalent and successful.

Unfortunately, these existing models lack high capacity for modeling, which is not believed to be the interaction between neighboring features. This paper presents a novel two-stream architecture for human action recognition fuses static and dynamic features.

III. Proposed work

A. the proposed two-stream architecture

The overview of the proposed static and dynamic feature interaction two-stream network is shown in Fig.1. The two-stream ConvNet is constructed using two independent spatio-temporal stream ConvNets. The spatial stream network accepts RGB images as input, whereas the temporal-stream network accepts optical flow images that have been stacked. Residual Networks (ResNets) have proven their ability to capture still image features.ResNet uses identity shortcut connections, allowing information to flow between layers without decay, and allows deep networks to build structures of up to lots of layers, which is what is needed to design deeper two-stream Convolutional Networks.A building block in ResNets is defined as:

$$y = F(x, \{w_i\}) + x \tag{1}$$

Here, Layer inputs and outputs are denoted by x and y, w_i presents the weights of i_{th} layer. The residual mapping to be learned is shown by the function F(x,{W_i}). In the left of Fig.2 that has two layers.F+x is computed by performing a shortcut connection followed by an elementwise addition, one on each channel of two feature maps.We replace the 3×3 convolution with CoT block in ResNet-101. CoT block replaces the 3×3 convolution with on the right of Fig.2.

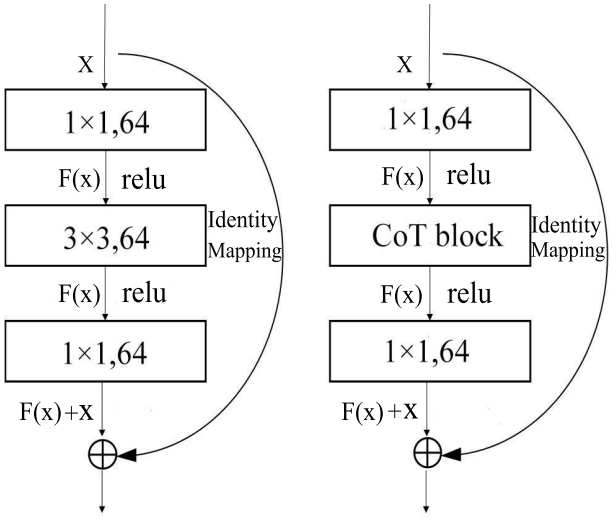


Fig. 2. The residual structure of CoT block.

The CoT block is shown in Fig.3. A 2D feature map $X \in R^{C\times h\times W}$ as input. Keys(K) are defined as X, Queries(Q) as X, and Values(V) as $V = X * W_v$. The representation of each key in CoT block is originally placed in context using k×k group convolution over all adjacent keys in a k×k grid to get the key $K^1 \in R^{C\times h\times W}$. To be sure, K^1 reveals the static relevant information around regional neighbor keys, and it is the static context representation of input X.The attention map is then obtained by following linear 1×1 convolutions, with each convolution being conditioned on the union of contextualized keys k×k and queries Q(W_θ with ReLU activation function and $W_\theta W_\delta$ without activation function).

$$W = [K^1, Q]W_\theta W_\delta \tag{2}$$

Furthermore, the local matrix at each spatial site of W is learned using the query and the learned key feature rather

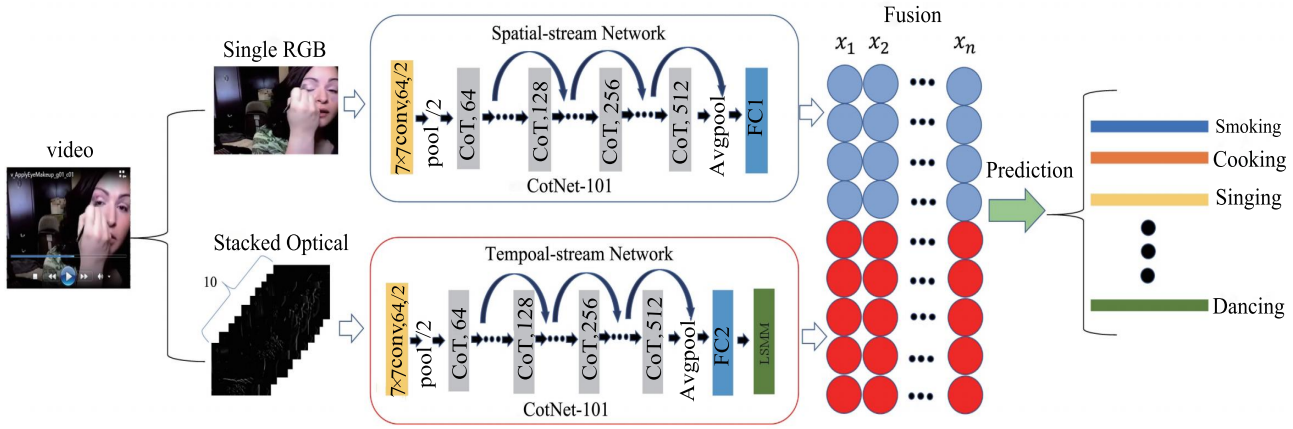


Fig. 1. The overall framework of the proposed action recognition model.

than the separate query-key pairs.Using this method, the efficacy of such miners static context K^1 is added to the enhancement of self-attention learning. Then, the dynamic feature contacts between inputs are obtained by a calculation of the attended feature map K^2. As its ultimate output, the CoT block combines the static context K^1 with the dynamic context K^2.

$$K^2 = V * W \tag{3}$$

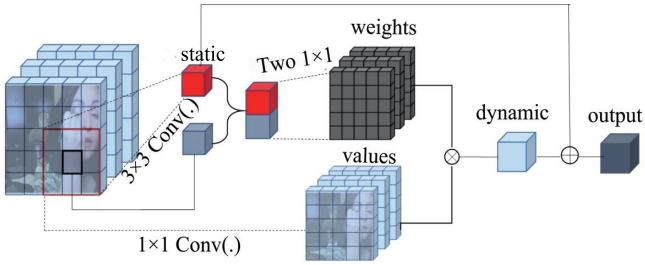


Fig. 3. The structure of CoT block.

B. The Spatial And Temporal Model

Action recognition in video requires associating the visual attributes of objects and motion components for certain action. The proposed architecture for a deep neural network employs the two information stream flows depicted in Fig. 1. The spatial stream accepts an RGB signal image as input, while the temporal stream accepts volumes of optical flow field stacking as input. The temporal stream input is a series of optical flow frames $\{C1, C2...Ct\}$.The frame's temporal order is indicated by where Ct is inserted. The current frame Ct records just the immediate action context, which is insufficient for recognition due to the absence of the temporal component associated with accomplished actions. The performance of a two-stream neural network is contingent on the capacity and resilience of the properties gathered by each information stream.

Spatial stream contains the appearances information existed in the current RGB frame. We use CoTNet-101 network for high-dimensional appearance feature extraction from of each video frame. The present spatial stream is the result of the dense layer FC-1 processing the RGB frame's features.

Temporal stream contains the spatiotemporal synthesis of the input optical flow frames getting the motion and time sequence patterns. We use use CoTNet-101 network to learn the motion flows information from the input optical flow frames sequence as deep features.we utilize LSTM to catch the time sequence information among dense optical flows. LSTM is designed as a resolution to the challenges of typical RNN's long-term dependency.For this reason, LSTM has seen widespread application in capturing the long-term development of actions in video clips. which is illustrated in Fig.4.The fundamental equations are shown below:

$$Z = tanh(W_c * [h_{t-1}, x_t] + b) \tag{4}$$

$$Z^i = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{5}$$

$$Z^o = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{6}$$

$$C^t = Z^f * C^{t-1} + Z^i * Z \tag{7}$$

$$h^t = Z^o * tanh(c^t) \tag{8}$$

where, $x_t$is the current cell input. h_{t-1} is the last cell input. Z is update cell. Z^i is the input gate. Z^f is the forget gate. Z^o is the output gate. C^t is its memory cell. h^t is the final state. σ represents the sigmoid function. $X = \{x_1, x_2...x_t\}$ is the output of CoTNet-101 on the temporal stream , which is the input of LSTM consisting of t time steps of features. $[h_t^l([h_t^0, h_t^1, ..., h_t^i, ...h_t^l])]$ denotes the hidden state of the top LSTM layer, and the input feature x_t at the t_{th} time step. h_t^l represents the l_{th} layer of LSTM at t_{th} time step.After processing the current optical flow frames, the temporal stream's output is the hidden state of the top LSTM layer. This is the final output of the temporal stream.The spatial stream and temporal stream are combined by linear concatenation. The fusion of both streams is carried out in a high-dimensional feature space in order to provide a single representation for the input sequence.Softmax activation is utilized in feature fusion to predict the action category for a given input frame sequence.It's significant that the video stream's spatiotemporal structure is translated into one-dimensional feature streams.

C. Data Variation Augmentation strategy

Video, unlike an image, is three-dimensional data with varying lengths of time.In the initial two-stream network,Videos were broken up into frames based on a time interval, and the optical flow fields between those frames were used to model the motion information. But because of

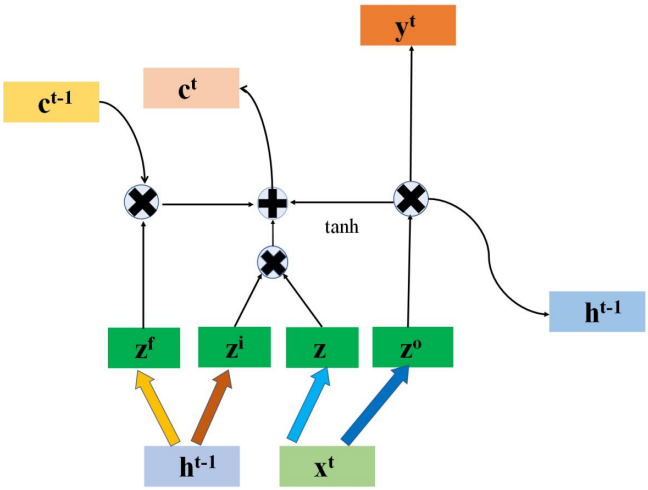


Fig. 4. The structure of LSTM.

data duplication across successive frames, action recognition would lack the ability to discriminate between similar actions.To boost data diversity, we not only crop the prominent portions of the image during training in the proposed work, but we also offer a strategy of data variation augmentation. With a fixed frame size of 256×256, all frames are randomly rotated and erased the contents of the frame,which was meant to make use of representations of diversity.After resizing the cropped areas to 224×224 and flipping them horizontally. Figure.5 is an illustration of several RGB and optical flow frame samples, along with an example of data augmentation that goes along with it.This manner of augmentation strategy significantly enhances the variability of inputs, which also helps to eliminate the problem of overfitting.

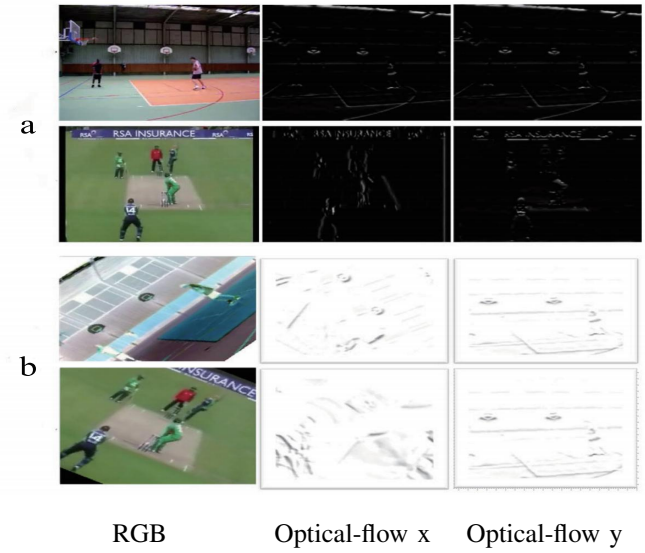


Fig. 5. Examples of data augmentation.In here, a represents the original RGB images and corresponding optical flow frames in x and y directions , b represents corresponding data augmentation.

D. Implementation considerations

In order to implement the proposed work some critical details must be considered, and we will discuss them below.

Network Input: In the proposed two-stream structure, a single RGB image serves as the input for the spatial network. TVL1 optical flow is used for transitions between frames. dense optical flow frames (10 in our experiment) are used to record motion information for the temporal network. Figure.5 shows some frame samples and the accompanying optical flow fields. It has been discovered that the background contains a large degree of horizontal or vertical movement. In addition, more training samples can be produced by means of the data variation augmentation strategy.

Network Training: The suggested two-stream Network is trained to utilize the PyTorch toolkit. Momentum is set to 0.9, and stochastic gradient descent (SGD) is performed using a 64-mini-batch size. An initial learning rate of 0.001 is used in the model and a maximum iteration of 500. At iteration number 450, the learning rate is lowered to 0.0001.

IV. EXPERIMENTS

We conducted our experiments on UCF101, HMDB51 and Kinetics400 dataset.Under **UCF101**, 13,000 video clips are sorted into 101 distinct action genres. The vast majority of these films are from unscripted, impresses YouTube videos with wildly varying lighting, camera angles, levels of occlusion, etc. The duration of the clip is between 1 and 71 seconds, while the frame rate is 25 and the resolution is 320 x 240. Training and testing data are separated into three equal parts in this dataset. There are the usual three divisions for training and testing in the dataset.**HDMB51** is a collection of films found on the internet that show individuals carrying out a variety of activities, including moving their faces with and without touching anything, moving their bodies with and without touching anything, and conversing with one another. There are a total of 6,800 films included in the collection, and these videos have been categorized into 51 distinct sorts of activities. Each action category has at least 101 footage, with the shortest clip clocking in at only one second in duration. The action categories also have at least 101 clips overall. The HMDB51, in contrast to the UCF101, maintains 30 frames per second even while scaling video to maintain a height of 240 pixels across the frame. In addition, there are three train/test splits of 70/30 for each action type inside the dataset.**Kinetics400** is a massive video dataset that includes 240 thousand training clips, twenty thousand validation movies, and forty thousand testing clips.

Evaluation indicators: Top-1 and top-5 accuracy are commonly employed as measures of action recognition precision.Top-1 accuracy indicates that the action class that has the greatest possibility of being classifiable based on the prediction results has been correctly classified.It is believed that the sample is authentic.Top- 5 accuracy indicates that the sample is deemed accurate if the correct classification is among the five action groups that have the greatest likelihood of categorizing in the prediction result. This suggests that the sample is accurate.We use the top-1 accuracy on the UCF101 and HMDB51 dataset, and the top-1 and top-5 accuracy on the Kinetics400 dataset to verify the performance.

A. Experiments with Dropout

We assess the proposed method on the UCF101, HMDB51, and Kinetics400 datasets. The results are summarized

TABLE I: Recognition accuracy on UCF101,HMDB51 and Kinetics400 datasets

Dataset	train	test
UCF101	94.2%	92.6%
HMDB51	75.5%	73.7%
Kinetics400-top-1	76.5%	74.7%

in TABLE I. To evaluate the proposed network architecture's learning and generalizability, we record the training set's accuracy across all datasets.TABLEAU I illustrates that the proposed network is able to utilize spatial and temporal connection for successful fusion and categorization. The existence of many dense layers in the network needs an analysis of the dropout pattern on the FC-1 and FC-2 layers. Dropout pattern implementation attempts in the proposed network architecture are detailed in TABLE II.(dropout rate relates to retention rate). In the UCF101 dataset, we have a 2.4% increase in test case accuracy, whereas in the Kinetics400 dataset, we achieve a 1.2% increase in test case accuracy. This reveals that the network is overfitting for the two datasets because of the high dimension of the medial features. By applying a dropout pattern on FC-1 and FC-2 layers, we gain a minor increase of 0.3% for the HMDB51 dataset. Although the improvement is minor, we retain dropout for training the completed network since it reduces generalization errors. Separate streams of spatial and temporal features, together with their respective identification accuracies, are also provided.The finding shows that fusion effectively takes up the multimodal connection, resulting in improved performance in video recognition.

TABLE II: Dropout pattern:evaluation on UCF101,HMDB51 and Kinetics400 dataset.

Dataset	Dropout layer and rate	Accuracy (%)		
		spatial	temporal	*fusion*
UCF101	{FC-1, 0.5}	88.5	87.3	93.6
	{FC-2, 0.5}	86.5	87.8	93.6
	{FC-1, 0.5},{FC-2, 0.5}	87.1	89.5	**95.0**
HMDB51	{FC-1, 0.5}	67.2	66.4	72.4
	{FC-2, 0.5}	66.8	68.9	73.8
	{FC-1, 0.5},{FC-2, 0.5}	68.3	71.2	**75.0**
Kinetics400-top-1	{FC-1, 0.5}	68.3	67.2	73.2
	{FC-2, 0.5}	68.7	69.5	72.7
	{FC-1, 0.5},{FC-2, 0.5}	69.6	70.5	**74.9**

B. Exploration study

We compare two training methods: (1) baseline setting in initial two-stream ConvNets[15],where a crop of a fixed size is flipped haphazardly from the entire frame. (2) data variation augmentation strategy, which is discussed in III.C. The outcomes are shown in Fig.6.It is evident to us that the effectiveness of the data variation augmentation strategy is significantly higher than that of the baseline setting(97.3%vs 95% on UCF101 dataset,78.5% vs 75% on HMDB51 dataset,75.8% vs 74.9% top-1 and 95.2% vs 93.5% top-5 accuracy on the Kinetics400 dataset.), indicating that the data variation augmentation strategy can effectively improve the action recognition accuracy.

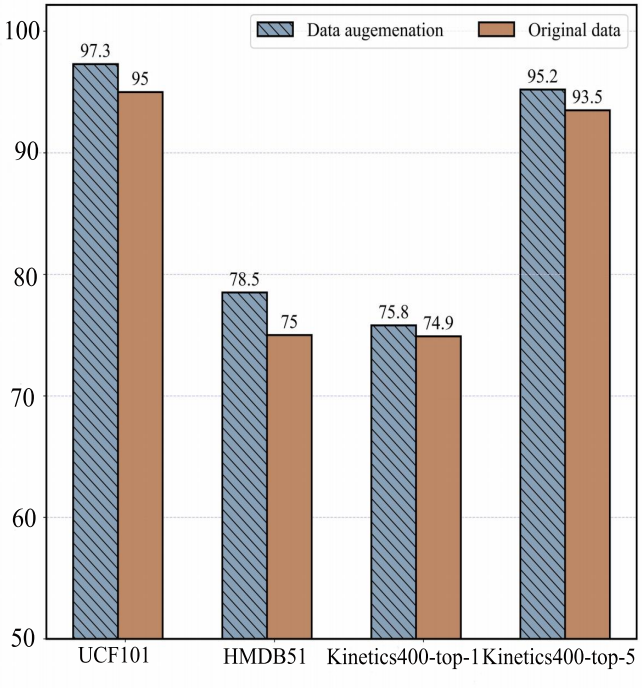


Fig. 6. Comparison of effectiveness of the data variation augmentation on UCF101 dataset and HMDB51 dataset.

We compare two network structures:(1)The essential structure-ResNet101, where the backbone has 3×3 convolution.(2)The structure-CotNet101, where the 3×3 convolution in the backbone is replaced by Cotblock.The results can be seen in Fig. 7. It turns out that the CotNet101 performs somewhat better than the ResNet101 (97.3%vs 96.8% on the UCF101 dataset,78.5% vs 77.8% on the HMDB51dataset, 75.8% vs 74.9% top-1 and 95.2% vs 94.8% top-5 accuracy on the Kinetics400 dataset),indicating that the CotNet101 network is more effective for action recognition.

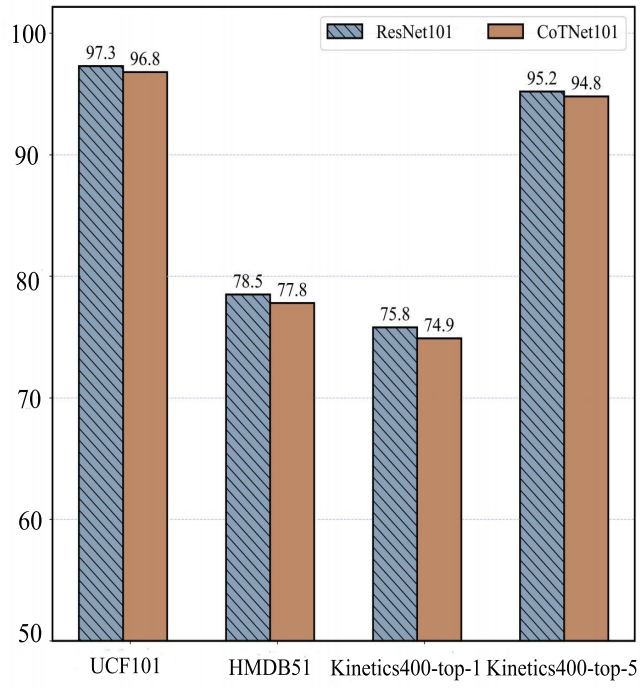


Fig. 7. Comparison of effectiveness of different network structures on the UCF101,HMDB51 and Kinetics400 dataset.

C. Comparison with those hand-designed descriptors

We compare the proposed two-stream network on the UCF101 and HMDB51 datasets using a variety of manually crafted descriptors. We cannot compare the experimental results to those of the Kinetics400 dataset since there are no manually defined features for conducting tests on the Kinetics400 dataset. According to TABLE III, traditional approaches have a performance rate of approximately 80% in UCF101. Since the HMDB51 dataset is more complicated than the UCF101 dataset, the overall performance in HMDB51 is pretty poor, below 70%.The hand-crafted features, such as iDT, are shallow. Even though these features are handled to reduce noise and capture hidden relationships, they are not distinctive enough for precise action recognition. With shallow features, the handcrafted features can't generalize well because they can't model the spatiotemporal information under complex backgrounds and quickly changing action dynamics. TABLE III demonstrates that our suggested method performs significantly better than those manually-designed descriptors, demonstrating that the interaction between static and dynamic features has a greater capacity for discrimination in the action recognition.

TABLE III: We compare our proposed model with hand-designed descriptors on the UCF101 and HMDB51 dataset. Experiments verify the effectiveness of our model.

Method	UCF101	HMDB51
HOG [16]	72.4%	40.2%
HOF[17]	76.0%	48.9%
MBH [17]	80.8%	52.1%
HOF+MBH [17]	82.2%	57.2%
iDT[18]	84.7%	72.4%
Ours	**97.3%**	**78.5%**

D. Comparison with other state-of-the-art methods

On the UCF101, HMDB51, and Kinetics400 datasets, we compare the objective outcomes of our method to those of deep learning methods. Based on the two streams, we select comparable approaches for the UCF101 and HMDB51 datasets, including Two-stream CNN, Temporal Seg.Net, Two Stream+LSTM, and L2LSTM. In addition, we also select some state-of-the-art methods based on 3D ConvNets, including C3D+IDT and Temporal 3D CNN. The performance of deep learning-based methods far outperforms that of more conventional approaches. Since the deep features considerably enhance the spatial and temporal representation, the accuracy is approximately over 90% on UCF101 and 60% on HMDB51.From TABLE IV, we can see that Two-stream CNN's accuracy obtains 88.0% on the UCF101 dataset, 59.4% on the HMDB51dataset. Temporal Seg.Net's accuracy achieves 94.2% on UCF101 datasets, 60.4% on HMDB51dataset. Two Stream+LSTM achieves 88.6%the accuracy on the UCF101 dataset. L2LSTM achieves an accuracy of 93.6% on the UCF101 dataset and 66.2% on the HMDB51 dataset. C3D+ IDT achieves 90.4% accuracy on the UCF101 dataset. The accuracy of a temporal 3D CNN is 93.2% on the UCF101 dataset and 63.5% on the HMDB51 dataset. In the meanwhile, we can see that our approach outperforms the competition on both datasets. The accuracy of our method is 97.3% on the UCF101 dataset, which is 9.3%, 3.1%, 8.7% ,3.7%, 6.9% and 4.1% greater than Two-stream CNN, Temporal Seg.Net, Two Stream+LSTM, L2LSTM,C3D+IDT and Temporal 3D CNN respectively. Our method achieves 78.5% accuracy on the HMDB51 dataset, outperforming Two-stream CNN, Temporal Seg.Net, L2LSTM, and Temporal 3D CNN by 21.1%, 18.1%, 12.3%, and 5.0%, respectively. In the HMDB51 dataset, the improvement brought about by our method is more evident. This demonstrates that fusing static and dynamic features is crucial for action recognition.From TABLE V, it can be seen that for the Kinetics400 dataset, STM accuracy is 84.34% top-1 and 96.23% top-5. MSNet achieves a top-1 accuracy of 76.4%. R(2+1) D-Two-Stream achieves 75.4% top-1 and 91.9% top-5 accuracy. Two-stream+CMA yields top-1 accuracy of 76.4%. The top-1 and top-5 accuracy rates for I3D non-local are 75.5% and 92.4%, respectively. S3GD achieves 74.7% top-1 and 93.4% top-5 accuracy. MARS achieves a top-1 accuracy of 74.9%. On the Kinetics400 dataset, the accuracy of our proposed method is 75.8% top-1 and 95.2% top-5. Our method has the best top-1 and top-5 accuracy among these state-of-the-art methods, as shown in TABLE V. On the one hand, the results presented in TABLE IV and TABLE V demonstrate that the propsoed two-stream architecture can significantly improve action recognition by adopting the proposed data variation augmentation strategy. On the other hand, it may suggest that most deep convolutional networks' poor performance is due to their disregard for the relationship between static and dynamic features.

TABLE IV: comparison with state-of-the-arts methods on the UCF101 and HMDB51 dataset.

Method	UCF101	HMDB51
Two-stream CNN[15]	88.0%	59.4%
Temporal Seg,Net[19]	94.2%	60.4&
Two Stream+LSTM[20]	88.6%	–
L2LSTM[21]	93.6%	66.2%
C3D+IDT[13]	90.4%	–
Temporal 3D CNN [22]	93.2%	63.5%
Ours	**97.3%**	**78.5%**

TABLE V: comparison with state-of-the-arts methods on the Kinetics400 dataset.

Method	top-1	top-5
STM[23]	73.7%	91.6%
MSNet[24]	76.4%	–
R(2+1)D-Two-Stream [25]	75.4%	91.9%
Two-stream+CMA[26]	75.98%	–
I3D non-local[27]	75.5%	92.4%
S3GD [28]	74.7%	93.4%
MARS[29]	74.9%	–
Ours	**75.8%**	**95.2%**

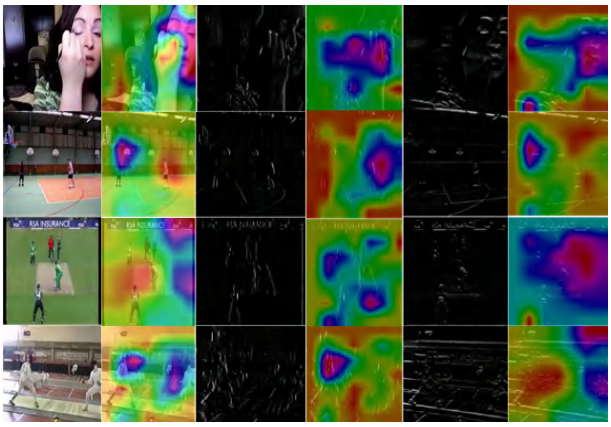
E. Visualization

To understand the static and dynamic feature interplay of the trained two-stream network. From the UCF101 dataset, certain RGB and optical flow frames are chosen to input into

the spatial and temporal stream, each representing a different activity category like “ApplyEyeMakeup”,“Basketball”,“ CricketShot” and“Fencin”. The Class Activates Mapping (CAM)[30] of the last layer is then visualized.Fig.8 displays the outputs of the visualization process. From the visualized samples, we can see that the CAM of the last layer is highly concentrated on appearance and motion area. It suggests that the proposed two-stream network integrating static and dynamic variables has a greater modeling ability to produce more contextual power for video action recognition.

V. CONCLUSIONS AND FUTURE WORK

Throughout this work, we introduced an innovative two-stream architecture that learns high-dimensional representations for action cognition in videos by combining static and dynamic feature.According to the results of the experiments, the proposed static and dynamic feature interaction two-stream network can achieve a higher recognition accuracy than the state-of-the-art, specifically 97.3% on the UCF101 dataset, 78.5% on the HMDB51 dataset, 75.8% top-1, and 95.2% top-5 on the Kinetics400 dataset..Given the rapid growth of live video streaming,it is important that human action be detected in real-time. Our network needs to compute optical flow, Which is hard to achieve the real-time video recognition requirements. Our future work will focus on optimizing our model using transformer structures to improve the feature representation, allowing for more accurate action recognition in real-time.



a b c d e f

Fig. 8. We visualize their corresponding CAM of the last layer. here, a is RGB,b is spatial CAM, c is Flow-x, d is flow-x CAM, e is Flow-y, f is Flow-y CAM.

REFERENCES

[1] A. Gershon, M. Rabindranath, and J. Yao, “Compound computer vision workflow for efficient and automated immunohistochemical analysis of whole slide images,” *Journal of Clinical Pathology*, vol. 5, no. N, pp. 9–20, 2022.

[2] G. A. O. e. a. Gimeno P, Ribas D, “Convolutional recurrent neural networks for speech activity detection in naturalistic audio from apollo missions,” *IberSPEECH*, vol. 45, no. J, p. 5, 2021.

[3] W. Zhu, Y. S. Vang, Y. Huang, and X. Xie, “Deepem: Deep 3d convnets with em for weakly supervised pulmonary nodule detection,” *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 11071, no. C, pp. 812–820, 2018.

[4] Y. Liu, R. Ma, H. Li, C. Wang, and Y. Tao, “Rgb-d human action recognition of deep feature enhancement and fusion using two-stream convnet,” *Journal of Sensors*, vol. 2021, no. 1, pp. 1–10, 2021.

[5] I. Priyadarshini and C. Cotton, “A novel lstm–cnn–grid search-based deep neural network for sentiment analysis,” *The Journal of Super Computing*, vol. 77, no. 12, pp. 13 911–13 932, 2021.

[6] A. Zhu, “Pose-guided inflated 3d convnet for action recognition in videos,” *Signal Processing Image Communication*, vol. 91, no. 13, p. 116098, 2020.

[7] Y. Li, T. Yao, Y. Pan, and T. Mei, “Contextual transformer networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 323, no. 7, pp. 123–127, 2022.

[8] S. H. Maronikolakis A, Dufter P, “Wine is not vi n.–on the compatibility of tokenizations across languages,” *ArXiv Preprint arXiv*, vol. 2109, no. J, p. 05772, 2021.

[9] Q. Yang, Y. Lian, Y. Liu, W. Xie, and Y. Yang, “Multi-agv tracking system based on global vision and apriltag in smart warehouse,” *Journal of Intelligent and Robotic Systems*, vol. 104, no. 3, pp. 1–16, 2022.

[10] H. Zang, R. Xu, L. Cheng, T. Ding, and G. Sun, “Residential load forecasting based on lstm fusing self-attention mechanism with pooling,” *Energy*, vol. 229, no. J, p. 120682, 2021.

[11] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, and Venugopalan, “Long-term recurrent convolutional networks for visual recognition and description,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 45, no. J, pp. 2625–2634, 2015.

[12] B. SravyaPranati, D. Suma, C. ManjuLatha, and S. Putheti, “Large-scale video classification with convolutional neural networks,” *International Conference on Information and Communication Technology for Intelligent Systems*, vol. 124, no. C, pp. 689–695, 2020.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d fonvolutional networks,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

[14] J. K. Y. D.-Y. Sun, Lin and B. E. Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 78, no. C, pp. 4597–4605, 2015.

[15] F. N., “Two-stream convolutional networks for end-to-end learning of self-driving cars,” *ArXiv Preprint arXiv:1811.05785*, vol. 29, no. J, p. 8, 2018.

[16] H. Wang, Schmid, and Cordelia, “Action recognition with improved trajectories,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 3, no. C, pp. 3551–3558, 2013.

[17] INRIA, H. Wang, and C. Schmid, “Lear-inria submission for the thumos workshop,” *ICCV Workshop on Action Recognition With a Large Number of Classes*, vol. 2, no. 7, p. 8, 2013.

[18] S. C. Wang H, “Action recognition with improved trajectories,” *IEEE International Conference on Computer Vision*, vol. 78, no. C, pp. 3551–3558, 2013.

[19] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks for action recognition in videos,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[20] C. Dai, X. Liu, and J. Lai, “Human action recognition using two-stream attention based lstm networks,” *Applied Soft Computing*, vol. 86, no. J, p. 105820, 2020.

[21] E. Hassan, “Learning video actions in two stream recurrent neural network,” *Pattern Recognition Letters*, vol. 151, no. J, pp. 200–208, 2021.

[22] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, “Temporal 3d convnets: New architecture and transfer learning for video classification,” *ArXiv Preprint arXiv:1711.08200*, no. J, 2017.

[23] G. W. e. a. Jiang B, Wang M M, “Stm: Spatiotemporal and motion encoding for action recognition,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 9, no. C, pp. 2000–2009, 2019.

[24] K. S. e. a. Kwon H, Kim M, “Motionsqueeze: Neural motion feature learning for video understanding,” *European Conference on Computer Vision. Springer, Cham*, vol. 56, no. C, pp. 345–362, 2020.

[25] T. L. e. a. Tran D, Wang H, “A closer look at spatiotemporal convolutions for action recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 56, no. C, pp. 6450–6459, 2018.

[26] L. Chi, G. Tian, Y. Mu, and Q. Tian, “Two-stream video classification with cross-modality attention,” *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, vol. 34, no. J, pp. 0–0, 2019.

[27] H. K. e. a. Wu C Y, Girshick R, “A multigrid method for efficiently training video models,” *Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition*, vol. 66, no. C, pp. 153–162, 2020.

[28] H. J. e. a. Xie S, Sun C, "Rethinking spatiotemporal feature learning for video understanding," *ArXiv Preprint arXiv:1712.04851*, no. C, p. 5, 2017.
[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, no. C, pp. 2921–2929, 2016.
[30] K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun, "Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images," *Remote Sensing*, vol. 11, no. 5, p. 544, 2019.