Hindawi
Security and Communication Networks
Volume 2020, Article ID 8876056, 15 pages
https://doi.org/10.1155/2020/8876056

Research Article

Anomaly Event Detection in Security Surveillance Using Two-Stream Based Model

Wangli Hao,[1] Ruixian Zhang,[1] Shancang Li,[2] Junyu Li,[1] Fuzhong Li,[1] Shanshan Zhao,[2] and Wuping Zhang[1]

[1]*School of Software, Shanxi Agricultural University, Taigu District, Jinzhong, Shanxi 030801, China*
[2]*University of the West of England, Bristol BS16 1QY, UK*

Correspondence should be addressed to Fuzhong Li; lifuzhong@sxau.edu.cn

Received 7 March 2020; Revised 10 May 2020; Accepted 23 June 2020; Published 3 August 2020

Academic Editor: Xiaolong Xu

Anomaly event detection has been extensively researched in computer vision in recent years. Most conventional anomaly event detection methods can only leverage the single-modal cues and not deal with the complementary information underlying other modalities in videos. To address this issue, in this work, we propose a novel two-stream convolutional networks model for anomaly detection in surveillance videos. Specifically, the proposed model consists of RGB and Flow two-stream networks, in which the final anomaly event detection score is the fusion of those of two networks. Furthermore, we consider two fusion situations, including the fusion of two streams with the same or different number of layers respectively. The design insight is to leverage the information underlying each stream and the complementary cues of RGB and Flow two-stream sufficiently. Two datasets (UCF-Crime and ShanghaiTech) are used to validate the effectiveness of proposed solution.

1. Introduction

Security surveillance is increasingly utilized at public places such as streets, hospitals, intersections, shopping malls, and banks, to guarantee public safety. However, the law enforcement agencies and monitoring abilities have not been matched. Consequently, the result is that there are obvious defects in the use of surveillance cameras. Anomaly event detection in surveillance videos is an important research topic in computer vision, which has been widely used in many security related scenarios, including traffic accidents investigation, crimes or illegal activities surveillance, forensics investigation, and violence alerting [1]. Because anomalous events rarely appear in real life, behavioral or appearance patterns deviating from normal patterns are often defined as anomalies [1–3].

Anomaly event detection has been effectively performed on the basis of several prevalent theories in the past decade, such as dictionary learning [4–7], probabilistic models [8, 9], and deep learning [10–12]. However, anomaly event detection is still facing a number challenges.

Most existing researches in anomaly event detection mainly focus on the RGB modality when extracting video features in anomaly event detection. In this work, we propose a two-stream-based model to handle the anomaly event detection problem using the RGB and Flow two convolutional neural network (ConvNets) to extract video features. The RGB stream performs anomaly event detection from video frames, whilst the Flow stream is trained to detect anomalies from motion-based on dense optical flow. Moreover, the proposed framework is able to utilize all frames in the video, while almost no additional calculation is introduced in inference when compared to [13]. The main reason is that the final number of features utilized for training the anomaly model is the same. Specifically, one video is divided into several clips, features of all frames in one clip are averaged to obtain the video clip-level feature.

There are noticeable advantages of our two-stream-based anomaly event detection. Instead of only considering RGB features for MIL models, in this work, we propose TAEDM that can leverage information of both RGB and Flow modalities. Specifically, the information from RGB modality is

the static features underlying still images, such as the color, shape, and appearance of objects or people in the event. The information from the Flow modality is the motion features of the event. As a result, TAEDM can capture the complementary information on RGB stream from still images and motion between images in one video sufficiently. We evaluate the developed approach on two different-scale benchmark datasets, including UCF-Crime [13] and ShanghaiTech [14]. The extensive ablation experimental studies demonstrate that our model obtains the state-of-the-art performance.

The main contributions of this work are summarized as follows:

(i) A novel two-stream-based anomaly event detection model is proposed for anomaly detection in surveillance videos. Furthermore, a dense feature extraction method is proposed to obtain video-level feature.

(ii) Proposed models are tested using benchmark datasets UCF-Crime [13] and ShanghaiTech [14], and results from both datasets show good performances than existing works.

The rest of this paper is organized as follows: Section 2 reviews the state-of-the-art research in anomaly event detection. Section 3 proposes a two-stream-based anomaly event detection model. Experimental results are elaborated in Section 4 and further discussion is presented in Section 5. Section 6 concludes this paper.

2. Related Work

In this section, we will discuss the most recent research results in anomaly event detection, and details about anomaly detection, ranking, and two-stream action recognition will be discussed.

2.1. Anomaly Detection. In computer vision, anomaly event detection is one of the most challenging problems and has attracted lots of research efforts in the past decades [15–21], where the commonly used detection methods can be roughly categorized into following three groups.

The first category of anomaly detection methods focuses on the hypothesis that anomalies are rare, and behaviors different from normal patterns seriously are seen as anomalous. In these methods, the regular patterns are encoded through various statistic models, such as Gaussian process based models [22, 23], the model of social force [24], Hidden Markov-based models [15, 25], the spatial-temporal Markov random field based models [26, 27], the combination of dynamic models [21], and treat anomalies as outliers.

The second category of anomaly detection approaches is sparse reconstruction [3, 14, 16, 28], which is utilized for usual pattern learning. Specifically, a dictionary is constructed by employing sparse representation for normal behavior, and the ones with high error are detected as anomalies. Recently, with the promising breakthrough of deep learning, some researchers construct deep neural networks for anomaly detection, including video prediction learning [29], and abstraction feature learning [6, 30, 31].

The third group is the hybrid methods of normal and anomaly behavior for modelling [13, 32, 33], in which, under weakly supervised setting, multi instance learning (MIL) is utilized to model motion patterns [13, 33], e.g., Sultani et al. developed an MIL-based classifier [13], which is employed to detect anomalies. Meanwhile, a deep ranking model is utilized to predict anomaly scores.

Aimed at leveraging the superiority of Sultani's work that considering both normal and anomalous videos, in this work, we rebuilt the model using a weak labelled supervised learning.

2.2. Ranking. Learning to rank is a popular research problem in machine learning and many research efforts have been conducted, including [7, 11, 34–38]. These approaches aimed at boosting relative scores of the pieces rather than individual scores. Rank-SVM [7] was proposed to enhance the retrieval performance of search engines.

The detection algorithm proposed in [34] can solve multiple-instance ranking problems through gradual linear programming. This method has been utilized in computational chemistry to solve hydrogen abstraction problem. More recently, researchers have proposed deep ranking networks for computer vision-related applications and achieved promising success, such as highlight detection [35], person reidentification [11], feature learning [36], Graphics Interchange Format (GIF) generation [37], face detection and verification [38], and metric learning and image retrieval [39]. All the above deep ranking approaches need extensive annotations of both positive and negative samples. Unlikely, in this work, a ranking model is proposed by reformulating anomaly detection problem as a regression problem under the ranking framework based on both normal and anomalous samples. The proposed model utilizes MIL depending on weakly supervised data to train the anomaly model and located anomaly with video segment level during testing. Unlike the conventional multiple instance learning (MIL) setting, the proposed ranking component forces ranking only includes two segments with the highest anomaly score in the negative and positive bags.

2.3. Two-Stream Action Recognition. Video-based action recognition has been extensively researched and achieved comparable attention recently. Among them, the two-stream-based action recognition is superior [40–42]. Inspired by neuroscience, one kind of action recognition methods introduced two-stream neural network architecture [40–42], to perform RGB and Flow feature extraction in parallel. The final score of action classification can be achieved by fusing the results of two paths. In order to further enhance the action recognition performance, Wang et al. developed a novel Temporal Segment Network (TSN) [41], which focuses on modelling the long-range temporal structure in videos. Further, various extensions of two-stream model [40] that explore convolutional fusion [42]

and residual connections [43, 44] were developed. The model in [43] established the residual connections between RGB and Flow streams. The STDDCN [44] integrated the multiscale information into residual connections via dense-connectivity interaction and contained a new knowledge distillation module.

Two-stream-based methods have been widely employed on some other task of video, such as action recognition [40–43, 45, 46]. However, two-stream-based methods are rarely applied to anomaly event detection. Inspired by the two-stream-based action recognition architectures leveraging the complementary information of RGB and Flow modalities underlying actions, we first design a novel two-stream anomaly event detection model. Compared with action recognition, the anomaly event detection can identify the kind of behavior (normal or abnormal) and locate the time range of an exception. That leads this problem more difficult to solve than the others.

3. Two-Stream Anomaly Event Detection Network

This section will detail the proposed two-stream-based anomaly event detection model as shown in Figure 1. We first will introduce the abnormal video and the normal video, and then divide them into multiple time video clips for extracting the two-stream features (RGB stream and Flow stream) of the video clips. A fully connected neural network will be trained using a ranking loss function, which calculates the highest-scoring instance (shown in blue) and the fusion operation then will be performed.

Video clip can be naturally split into synchronous spatial and temporal parts. The spatial component underlying the individual frame image consists of scenes and object information in the video. The temporal component hidden in the motion across the images carries the movement between the objects and the observer. We designed our anomaly detection model accordingly and decomposed it into two streams, as is illustrated in Figure 1. Each stream is realized via a deep convolutional network (ConvNet), anomaly detection scores of which are fused in the late.

In the proposed model, video segments that obtained high anomaly scores will be marked as anomaly event. Each video will be split into equal number of nonoverlapping segments. The video containing anomaly segment is labelled as positive and a video without any anomaly segment is labelled as negative. A positive/negative video is treated as a positive/negative bag and the segments as instances in the multiple instance learning. Through ranking method, anomaly scores for each video segment can be obtained and the video segments obtained high anomaly scores is seen as anomaly event.

First, given the abnormal video and the normal video, we divided them into multiple time video clips. Secondly, we extracted the two-stream features (RGB stream and Flow stream) of the video clips and then trained a fully connected neural network using a ranking loss function, which calculates the highest-scoring instance (shown in blue) and performed the fusion operation in the last step.

3.1. Problem Formulation. In the past decade, a number of pattern learning methods have been developed [10, 15, 19, 25], most of them assuming that any pattern that violates this common pattern should be abnormal. In fact, it is impossible to propose a method to define a full set of normal patterns, because the normal pattern may contain too many different events and behaviors. To define anomaly events is another challenge, since anomaly events may also contain many similar events and behaviors.

To handle the above issues, the proposed method formulates each anomaly detection task (RGB branch and Flow branch) as a regression problem, which is realized under the ranking framework by leveraging both normal and anomalous data. To achieve more precise segment-level labels, a weakly supervised deep multiple instance learning (MIL) ranking is employed. Specifically, weakly supervised rank indicates that the model only knows whether there is an abnormal event in a video rather than the category of the anomaly event and the corresponding occurrence time during training.

The differences of the proposed pattern learning method from those in [10, 15, 19, 25] is that our model utilizes both normal and anomalous data rather than normal data in previous studies (e.g., [10, 15, 19, 25]). Furthermore, our model is formulated as a regression problem, which means that we consider a certain segment as an abnormal event based on regression prediction score rather than the probability less than a certain threshold.

3.2. Data Formulation. To align the data for deep MIL setting in anomaly detection, the source video is first split into equal number of nonoverlapping segments during training. All segments in the same video are denoted as a bag, and each segment is acted as an instance. All videos formed two different bags, positive bags and negative bags, respectively. The segments of anonymous video are treated as positive bag and those of normal video negative bag. Moreover, as our insight is based on leveraging the complementary information of RGB and Flow streams, video clips in each bag are all decomposed into RGB and Flow components. Each kind of component is fed into the corresponding branch networks separately.

3.3. Network Architecture. The deep MIL framework includes two main branch deep MIL ranking networks: RGB and Flow, as shown in Figure 1. Each branch contains feature extraction and instance scoring parts. Concerning the feature extraction, ResNet [47] is chosen as backbone because of its superiority in both efficiency and effectiveness.

3.3.1. Spatial Branch ConvNet. Spatial Branch ConvNet focuses on single video frame, effectively conducting anomaly event detection from still images. The static RGB stream by itself contains useful information, since some anomaly events are closely associated with specific objects. Actually, as will be reported in the section of experiments,

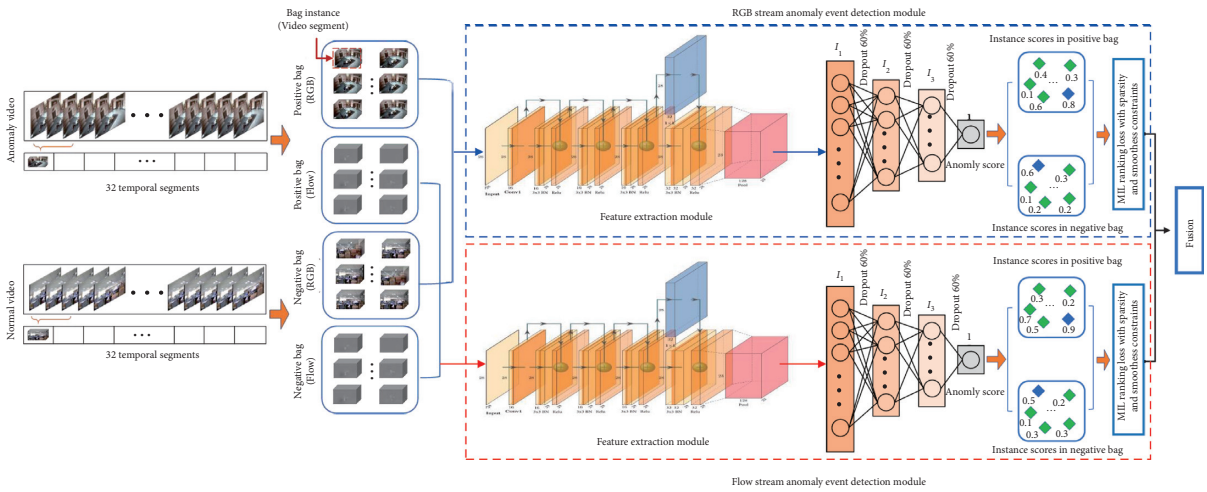

Figure 1: A flowchart of the proposed two-stream based anomaly event detection framework.

anomaly event detection from still images (the RGB anomaly event detection stream) is quite competitive by itself.

3.3.2. Temporal Branch ConvNet. Unlike the conventional ConvNet models, the input of proposed temporal anomaly event detection stream is the stacked optical flow displacement fields among several adjacent frames. This input explicitly models the motion between video clip images, which makes the anomaly event detection easier, as no implicit motion estimation is required.

The dense optical flow is formed using a group of displacement vector fields $\mathbf{v}_t$ between adjacent t and $t+1$ frames. Further, $\mathbf{v}_t(m,n)$ indicates the displacement vector at the corresponding point (m,n) in frame t, which represents the movement of point (m,n) from frame t to frame $t+1$. Moreover, the displacement vector $\mathbf{v}_t$ contains two components, including horizontal and vertical ones, which dubbed as $\mathbf{v}_t^x$ and $\mathbf{v}_t^y$, respectively. $\mathbf{v}_t^x$ and $\mathbf{v}_t^y$ are seen as image channels (as shown in Figure 2) and can be fed into the temporal anomaly event detection stream network.

Figures 2 (a1), (b1) and (a2), (b2) indicate the pair of adjacent video frames with the highlighted moving area outlined with a red rectangle. Figures 2(c1)/(c2) denote the horizontal part $\mathbf{v}_t^x$ of displacement vector field (higher/lower intensity relates to positive/negative values). Figures 2 (d1) and (d2) illustrate the vertical part $\mathbf{v}_t^y$ of displacement vector field.

3.3.3. Loss Function. To pursue better performance, we employ the following loss function (referred from [13]) to train each branch network:

$$\mathscr{L} = l\left(S_p, S_n\right) + \|\mathscr{W}\|_F, \tag{1}$$

where S_p and S_n denote positive and negative bags, respectively. $l(S_p, S_n)$ indicates the loss over these two kind of bags. $\|\mathscr{W}\|_F$ denotes the $F-$ norm regularization on weights of the model, for boosting its generalization. Among them, $l(S_p, S_n)$ is defined as

$$l\left(S_p, S_n\right) = l_{\text{rank}} + \lambda_a l_{\text{smooth}} + \lambda_b l_{\text{sparsity}}, \tag{2}$$

in which l_{rank} denotes the ranking loss, l_{smooth} denotes the temporal smooth restrict and l_{sparsity} represents the sparsity constraint. λ_a and λ_b are two hyper-parameters which balance the strengths of corresponding terms. Among them, l_{rank} is formulated as

$$l_{\text{rank}} = \max\left(0, 1 - \max_{i \in S_p} f\left(C_a^i\right) + \max_{i \in S_n} f\left(C_n^i\right)\right), \tag{3}$$

where S_p and S_n share the same meanings with those of equation (1). C_n and C_a indicate normal and anomalous video instances. $f(C_a^i)$ and $f(C_n^i)$ denote the predicted scores for the corresponding video instances.

The l_{rank} forces rank only on two segments with the highest anomaly score in the negative and positive bags separately, rather than every segment of the bag. Thus, the max operation is performed over all instances in each bag. The reason for this different setting is the absence of video segment-level annotations in anomaly event detection task.

The l_{rank} loss here is superior for anomaly detection task due to several appealing reasons. First, it can enforce the anomalous video segments to achieve higher anomaly scores compared to normal ones. Furthermore, it can separate the positive instances and negative instances based on anomaly score.

On the other hand, l_{smooth} and l_{sparsity} are defined as

$$\begin{aligned} l_{\text{smooth}} &= \sum_{i=1}^{n-1} \left(f\left(C_a^i\right) - f\left(C_a^{i+1}\right)\right)^2, \\ l_{\text{sparsity}} &= \sum_{i=1}^{n} f\left(C_a^i\right), \end{aligned} \tag{4}$$

where n is the number of instances in the specific bag. l_{smooth} is utilized to guarantee the temporal smoothness via minimizing the difference of anomaly scores between neighboring video instances in a bag. l_{sparsity} is employed to enforce the sparsity of scores in the anomalous bag. The

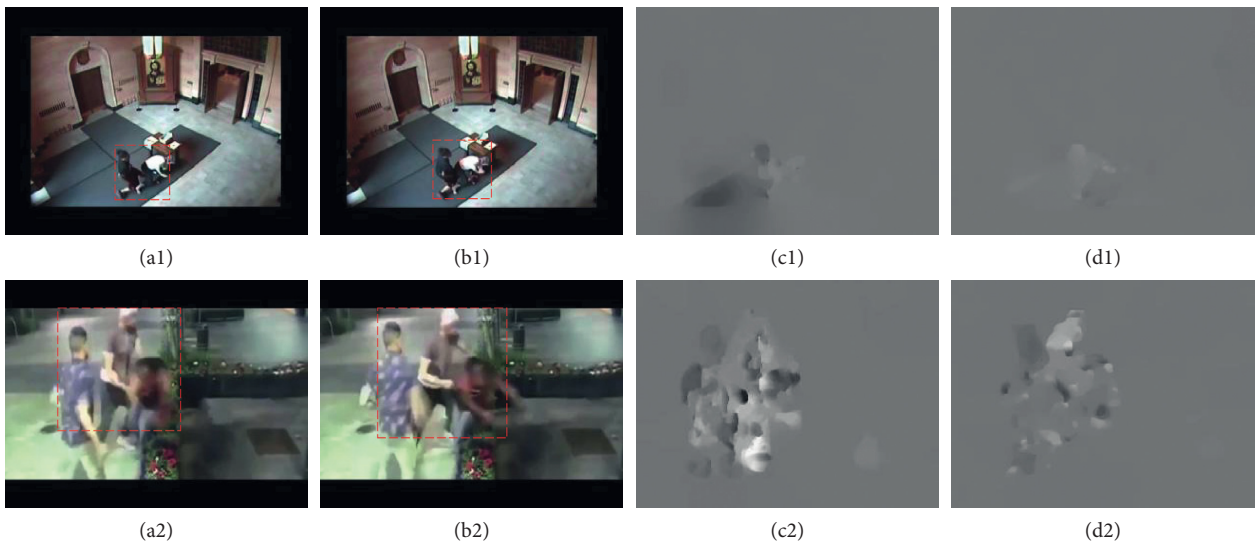

Figure 2: Optical flow examples. (a1, b1) and (a2, b2): the pair of adjacent video frames with the highlighted moving area outlined with a red rectangle. (c1, c2) horizontal part $\mathbf{v}_t^x$ of displacement vector field (higher/lower intensity relates to positive/negative values). (d1, d2): vertical part $\mathbf{v}_t^y$ of displacement vector field.

reason for introducing l_{sparsity} loss function is that few segments may involve the anomaly event.

4. Experiments

In this section, we will illustrate our experiments in detail from aspects including datasets, implementation details, evaluation metric, and sufficient quantitative and qualitative experiments, respectively.

4.1. Datasets. UCF-Crime [13] and ShanghaiTech [14] are two popular benchmark datasets commonly used in anomaly detection task. In this work, we use both datasets to validate the superiority of our proposed anomaly event detection model. With following steps, the proposed model can also support other datasets: (1) extract the RGB image and optical flow images of each video in the dataset; (2) extract their corresponding features; and (3) feed both the RGB and Flow stream features into corresponding branch subnetworks in the proposed model for training and obtaining expected test results. Before introducing the details, we first briefly introduce the two benchmark datasets as follows.

4.1.1. UFC-Crime. Reference [13] is a large-scale dataset that contains a total of 13 anomaly events and 1900 real-word surveillance videos. Among them, 950 videos include clear anomalies and the other videos are treated as normal video. Further, concerning the dataset partitioning, the training set contains 1610 videos (800 normal videos, 810 anomalous videos), and the test set contains 290 videos (150 normal, 140 anomalous videos).

4.1.2. ShanghaiTech. Reference [14] is a medium-scale dataset with a total of 437 videos, which contains 130 abnormal events of 13 scenes. This dataset cannot be utilized directly to perform anomaly event detection because the training set has no abnormal video. To tackle this problem, Zhong et al. [48] rebuilt the dataset via randomly choosing abnormal test videos and putting them into the training data and vice versa. Simultaneously, both training and test dataset contain 13 scenes. This new organization of dataset made it suitable for anomaly event detection task. Thus, we perform the same operation as that in [48], before executing the experiments.

4.2. Implementation Details and Evaluation Metrics. To implement the proposed model, we first extract features of RGB and Flow images from the last fully connected (FC) layer of the ResNet network [47]. Concerning the RGB stream, ResNet features for every frame are computed. The video segment-level feature can be obtained by averaging all frame features in the corresponding video segment. Similarly, for the Flow stream, features can be extracted using the same way of RGB stream. The only difference between these two streams is that each frame in Flow stream contains two directional flow images, namely, vertical ($\mathbf{v}_t^x$) and and horizontal ($\mathbf{v}_t^y$) images as stated above, which makes the ResNet infeasible to extract their features. Specifically, $\mathbf{v}_t^x$ and $\mathbf{v}_t^y$ are all grayscale images with only one channel (the concatenation of them only has two channels), while the input sample of feature extraction network (ResNet) needs three ones. To handle this problem, we concatenate the two directional flow images and their average variant to form the input flow sample with three channels for the feature extraction network.

After obtaining segment-level RGB and Flow ResNet features, we feed them (2048D) into a three-layer FC neural network as that of [13]. Further, the Adagrad optimizer is utilized, which initial learning rate is 0.001. To perform a fair comparison, the smoothness constraint, the sparsity restriction, and the segment number of each video are the

same with those of [13]. We stop our training at 20, 000 iterations.

The following commonly used evaluation metrics are adopted to validate the performance our model. They are receiver operating characteristic (ROC) curve and the area under the curve (AUC), respectively. The reason we utilize ROC and AUC is that they are two popular metrics for anomaly event detection tasks [13, 21, 48]. For fair comparison with other works and to verify the effectiveness of our model, ROC and AUC are employed.

4.3. Experimental Results

4.3.1. Evaluation of the Proposed Model. To validate the performance of the proposed method, we compare the results with those of state-of-the-art models, based on UCF-Crime [13] and ShanghaiTech [14]. Comparison ROC curves are shown in Figure 3. In Figure 3, RGB, Flow and Two denote the anomaly event detection results of different models based on RGB stream network, Flow stream network and the fusion of them separately.

Figure 3 illustrates that RGB, Flow, and Two obtain better results than the other models, validating the dense feature extraction is effective. Further, Two yields better results than those of RGB and Flow, which verifies the superiority of the proposed model.

The AUC results from different models on UCF-Crime [13] and ShanghaiTech [14] are displayed in Tables 1 and 2, respectively. It can be seen that the results are the same with those of Figure 3, which further validates the effectiveness of our model.

4.3.2. Ablation Studies. In this section, several ablation studies are designed to demonstrate the effectiveness of the proposed model.

(1) Evaluation of the Generalization Capacity of the Model. To validate the generalization of the proposed method, we present the results of proposed method based on models with different depths and architectures, including ResNet50, ResNet100, ResNet150, and VGG16, respectively, as shown in Tables 3 and 4. The results in Tables 3 and 4 illustrate that model Two achieves better results than those of the corresponding RGB and Flow models in all cases, which verifies the generalization capacity of the proposed model in terms of model depth and architecture.

Additionally, the ROC curves of ResNet50, ResNet100, ResNet150, and VGG16 are exhibited in Figures 4 and 5, respectively. Among them, Figures 4(a)–4(c) and 5(a)–5(c) report the ROC curves of RGB, Flow, and Two networks from ResNet50, ResNet100, ResNet150, and VGG16 models, respectively. Figures 4 and 5 further validate the generalization capacity of our method on model depth and architecture.

(2) Evaluation of the Fusion of Two Streams. As stated above, different backbone feature extraction models are employed to assess the proposed method. This naturally raises the following evaluations, including the fusion of two streams with the same number of layers and the fusion of two streams with different number of layers separately:

(1) Fusion of two streams with the same number of layers: To validate the effectiveness of the fusion of two streams with the same number of layers, we utilize the identical network (including ResNet50, ResNet100, ResNet150, and VGG16, respectively) to perform both RGB and Flow stream feature extractions. Comparison results are presented in Tables 1 and 2. Tables 1 and 2 show model Two obtains uniformly better results than those of the corresponding RGB and Flow models, which illustrates the effectiveness of the proposed method under the same layer fusion setting (dubbes as $\text{Fusion}_{\text{same}}$ setting).

(2) Fusion of two streams with different number of layers: To verify the effectiveness of the fusion of two streams with different number of layers, we employ different networks to perform RGB and Flow stream feature extractions, respectively.

Tables 5 and 6 illustrate that the performance of model Two is consistently superior to those of corresponding RGB and Flow models, which validates the superiority of the proposed method under the different layers fusion setting (dubbed as $\text{Fusion}_{\text{dif}}$ setting). Further, an appealing conclusion can be drawn that $\text{Fusion}_{\text{dif}}$ surpasses $\text{Fusion}_{\text{same}}$ in most cases. In addition, the case that RGB stream with ResNet50 and Flow stream with VGG16 yields our best anomaly event detection results, which again verifies the effectiveness of fusion under different model architectures and depths.

(3) Evaluation of Fusion Proportion. This paper obtains the final anomaly detection scores via fusing two streams via the following equation: $Score = \beta * Score_{Flow} + (1-\beta) * Score_{RGB}$. To validate the effects of fusion proportion β between two streams, we perform anomaly event detection with various fusion proportion ranges from 0.1 to 0.9 with step size 0.1.

(1) Results of fusion proportion with same number of layers: Figure 6 reports the results of $\text{Fusion}_{\text{same}}$ with different fusion proportions. From Figure 6, we can see that each ResNet backbone (including RenNet50, RenNet100, and RenNet150) obtains similar fusion anomaly detection results respectively under different fusion proportions, with differences range from 0.2 to 1.23 (UCF-Crime) and 0.05 to 0.6 (ShanghaiTech). Further, the best fusion results are seemed obtained at $\beta = 0.7$ in most ResNet backbone cases. On the other hand, VGG backbone's fusion varies a lot, about 4 points (UCF-Crime) and 1 point (ShanghaiTech). Moreover, the best fusion result is achieved at $\beta = 0.9$.

(2) Results of fusion proportion with different number of layers: Figure 7 shows the results of $\text{Fusion}_{\text{dif}}$ with different fusion proportions on UCF-Crime. Figures 7(a)–7(d) denote the results of ResNet50 (Flow), ResNet100 (Flow), ResNet150 (Flow), and VGG16 (Flow) fusing with different RGB networks,

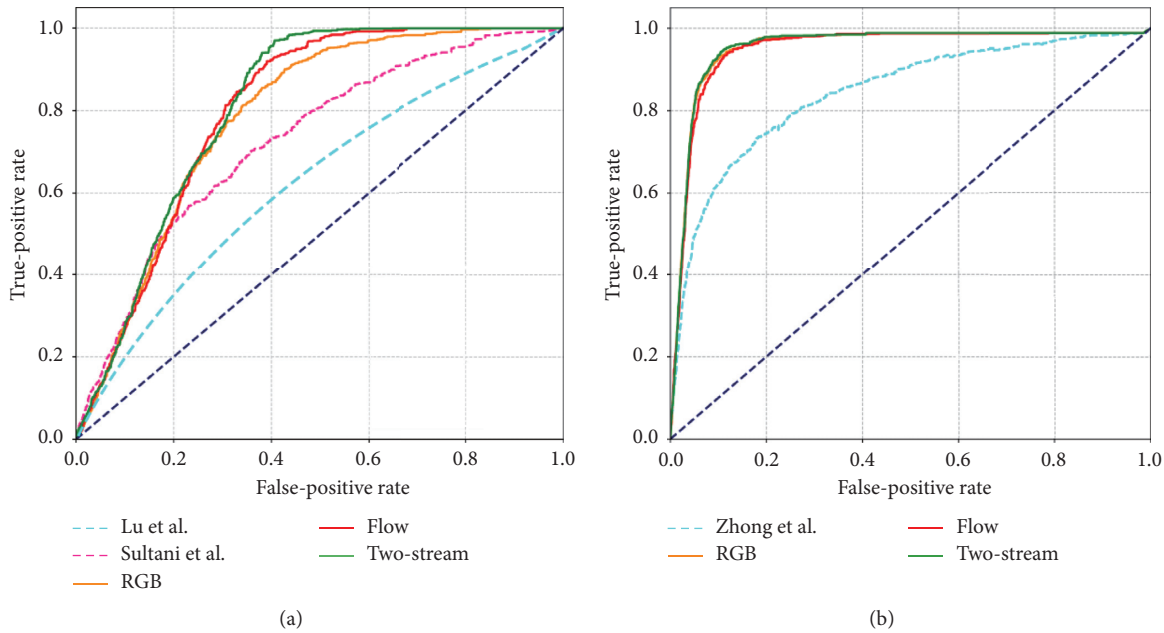

Figure 3: ROC curves of different models on UCF-Crime [13] and ShanghaiTech [14]. (a) Results of UCF-Crime [13] and (b) that of ShanghaiTech [14].

Table 1: Quantitative comparison on UCF-Crime.

Method	AUC (%)
Hasan et al. [6]	50.6
Lu et al. [28]	65.51
Sultani et al. [13] with w/o constraints	74.44
Sultani et al. [13] with w constraints	75.41
Ours (RGB)	78.51
Ours (flow)	80.29
Ours (two)	81.22

Table 2: Quantitative comparison on ShanghaiTech.

Method	AUC (%)
Zhong et al. [48]	84.44
Ours (RGB)	94.53
Ours (flow)	95.42
Ours (two)	96.74

respectively. Figure 7 shows that as fusion proportion value β increases, the trends of the AUC curves of Figures 7(a) and 7(d) are monotonically increasing and those of Figures 7(b) and 7(c) are monotonically decreasing. Reasons are that anomaly scores of flow streams of Figures 7(a) and 7(d) are superior to those of their corresponding fused RGB streams, and anomaly scores of flow streams of Figures 7(a) and 7(d) are worse than or comparable to those of their corresponding fused RGB streams. In other words, which stream has better performance, the fusion result will be better when its proportion is higher. The general proportion value for that stream with better results is 0.8 in most cases.

Table 3: Comparison results of different models on UCF-Crime.

Model	ResNet50	ResNet100	ResNet150	VGG16
RGB AUC (%)	78.51	76.84	77.44	71.21
Flow AUC (%)	79.92	77.51	77.93	80.29
Two AUC (%)	80.87	79.54	79.19	80.58

Table 4: Comparison results of different models on ShanghaiTech.

Model	ResNet50	ResNet100	ResNet150	VGG16
RGB AUC (%)	94.53	95.11	95.28	83.07
Flow AUC (%)	95.42	94.81	95.28	95.49
Two AUC (%)	95.53	95.23	95.38	95.85

Figure 8 presents the results of $\text{Fusion}_{\text{dif}}$ with different fusion proportions on ShanghaiTech. Figures 8(a)–8(d) show the results of ResNet50 (Flow), ResNet100 (Flow), ResNet150 (Flow), and VGG16 (Flow) fusing with different RGB networks respectively. Figure 8 shows that as fusion proportion value β increases, the trends of the almost all AUC curves of four subfigures are monotonically increasing, and also three curves are almost monotonically decreasing. Reason is that flow streams of these curves have higher anomaly scores than those of their corresponding fused RGB streams, and anomaly scores of flow streams are worse than or comparable to those of their corresponding fused RGB streams. In this dataset, we also obtain the similar conclusion that for the stream with a better result, the fusion result will be better when its proportion is higher. The general proportion value for that stream with better results is 0.9 in most cases.

4.3.3. Qualitative Results. To provide a more intuitive perception of the proposed model, we introduce the scores of anomalies per segment in a video obtained via our

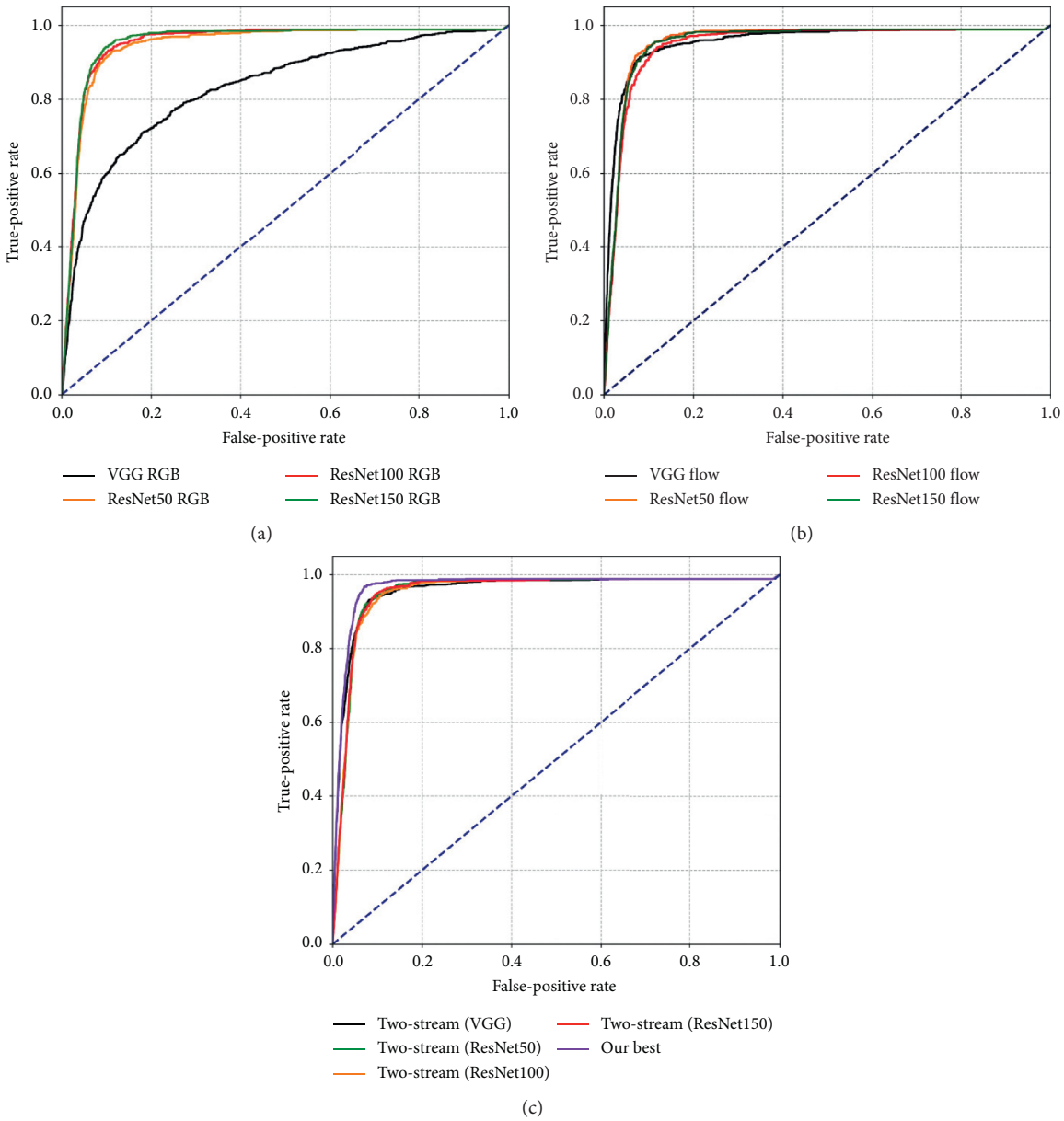

Figure 4: ROC curves of different models on ShanghaiTech. (a) Results of RGB stream. (b) Results of Flow stream. (c) Results of the fusion of RGB and Flow two streams.

method. Meanwhile, corresponding events with the highest or lower abnormal event scores in the video are also presented, with results presented in Figure 9 for UCF-Crime and Figure 10 for ShanghaiTech. Specifically, three example events are displayed in Figures 9 and 10, respectively. The first row of Figures 9 and 10 show the visualization results obtained by our best model variant, and the green blocks in the gray rectangle in Figure 9 or purple rectangles in Figure 10 represent the ground truth time period in which the anomaly event occurred. The second row of Figures 9 and 10 present the visualization results of different variants of our model, including results of ResNet50, ResNet50, ResNet150, and the best model variants, respectively. Simultaneously, several frames at the corresponding time are exhibited, including corresponding frames with the highest or lower abnormal event scores in the video. The area marked by the red circle in the image is the corresponding abnormal event. From Figures 9 and 10, we can see that our model can effectively predict the time period of anomalous events.

5. Discussion

It is noted that the RGB stream focuses on the appearance information and Flow stream concentrates on motion clues underlying a certain video. The fusion of these two streams with the same number of layers boosts the anomaly event detection performance effectively, as Table 1 and Table 2 show. Reason is that this fusion can

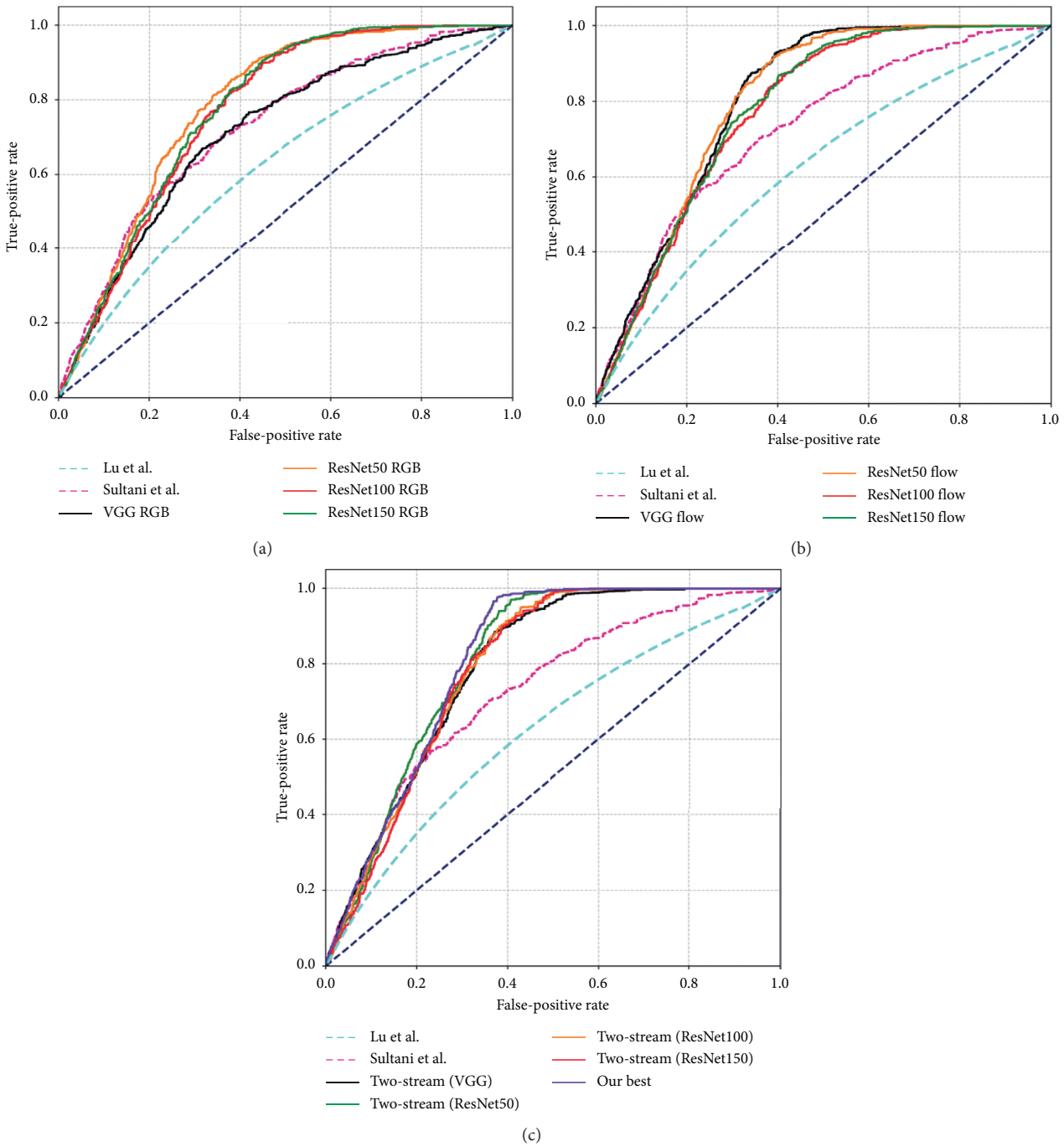

Figure 5: ROC curves of different models on ShanghaiTech. (a) Results of RGB stream. (b) Results of Flow stream. (c) Results of the fusion of RGB and Flow two streams.

Table 5: Comparison results of different models of two streams with different number of layers on UCF-Crime.

Flow	RGB			
	ResNet50	ResNet100	ResNet150	VGG16
ResNet50	80.87	80.48	80.19	80.08
ResNet100	80.22	79.54	79.16	78.28
ResNet150	80.11	79.66	79.19	78.75
VGG16	81.21	80.8	80.54	80.58

Table 6: Comparison results of different models of two streams with different number of layers on ShanghaiTech.

Flow	RGB			
	ResNet50	ResNet100	ResNet150	VGG16
ResNet50	95.53	95.54	95.55	95.85
ResNet100	94.93	95.23	95.13	95.33
ResNet150	94.93	95.13	95.38	95.33
VGG16	96.66	96.74	96.61	95.85

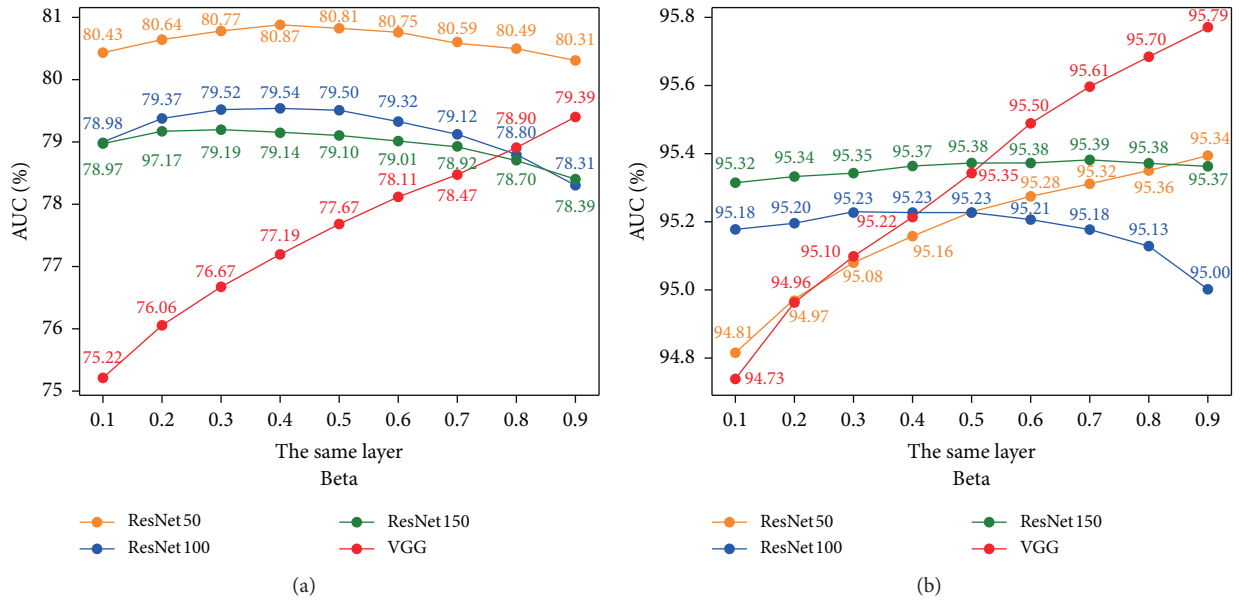

Figure 6: Results of different fusion proportions under $\text{Fusion}_{\text{same}}$ setting on UCF-Crime and ShanghaiTech. (a) Results of UCF-Crime. (b) Results of ShanghaiTech.

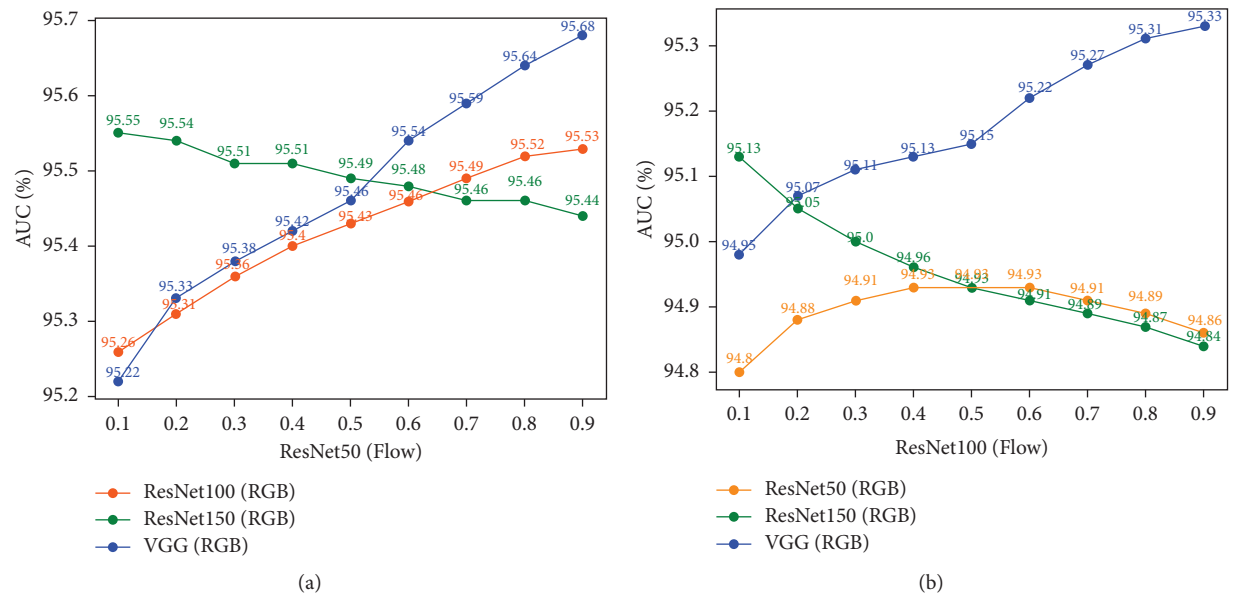

Figure 7: Continued.

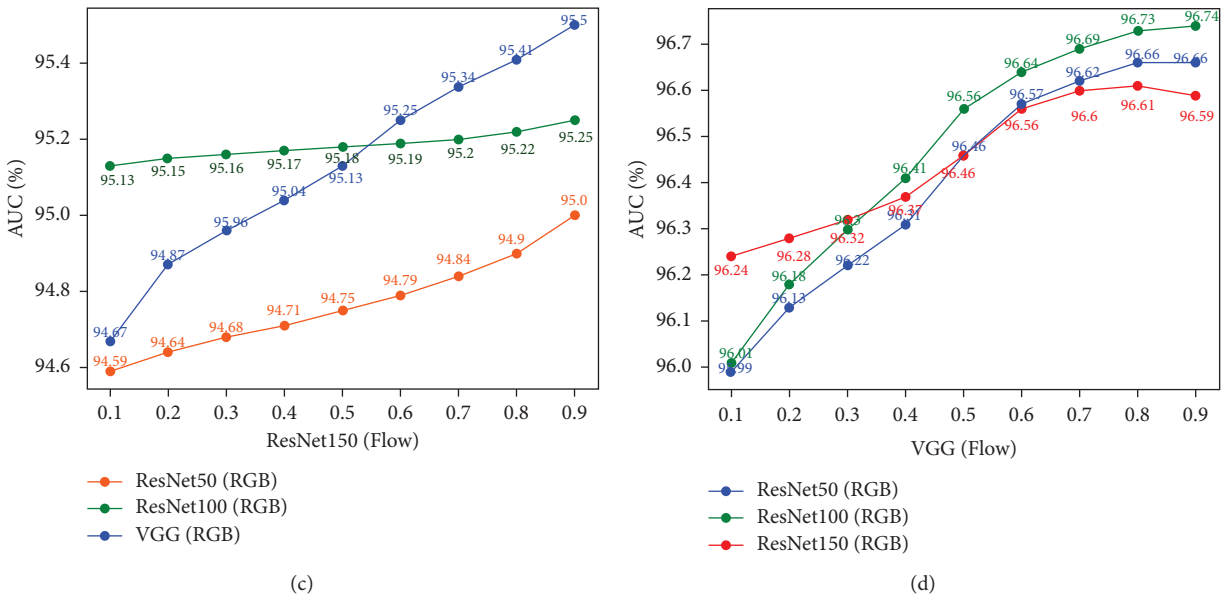

Figure 7: Results of different fusion proportions under $\text{Fusion}_{\text{dif}}$ setting on UCF-Crime.

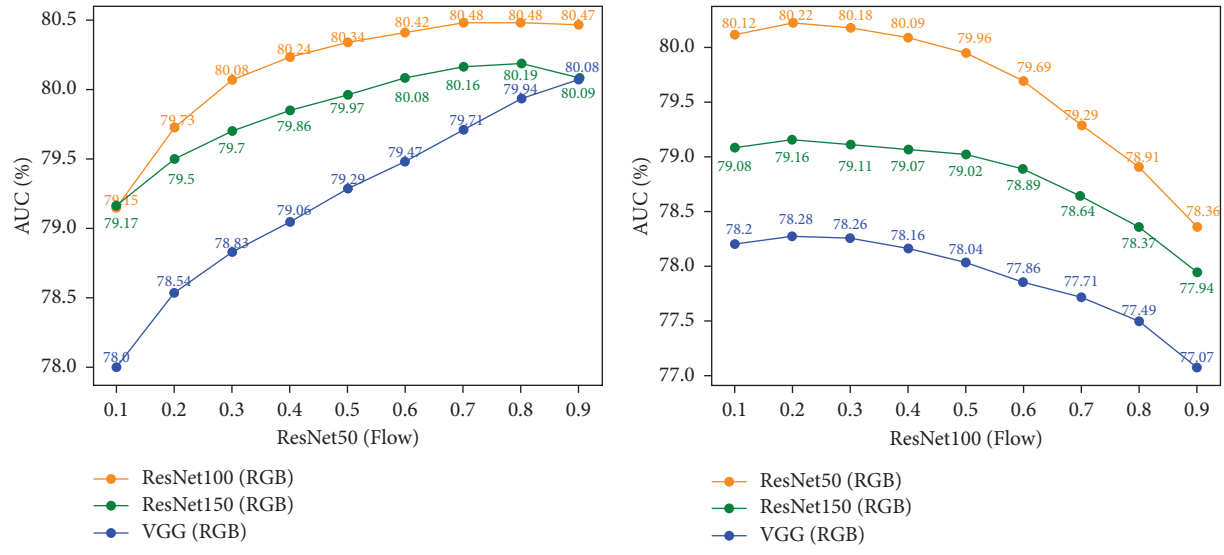

Figure 8: Continued.

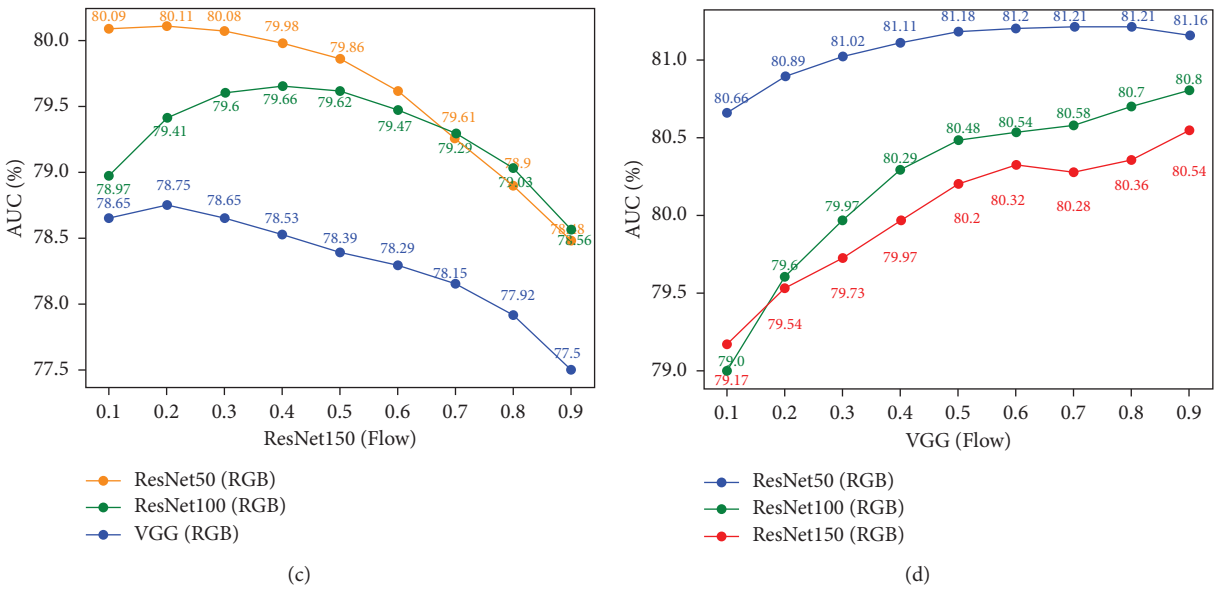

Figure 8: Results of different fusion proportions under $\text{Fusion}_{\text{dif}}$ setting on ShanghaiTech.

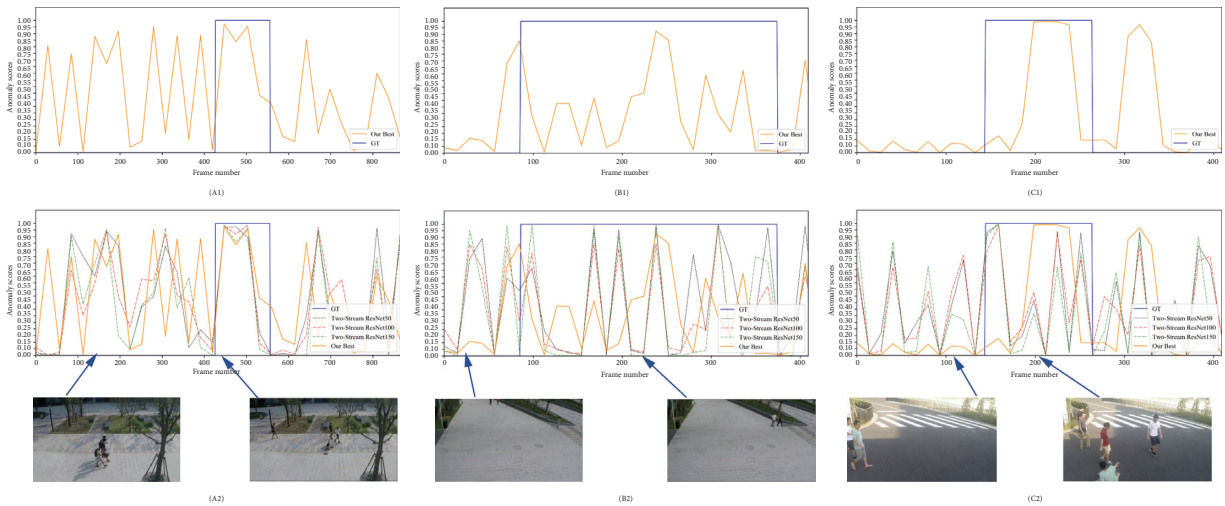

Figure 9: The visualization results of our method on testing videos on UCF-Crime. The first row shows the visualization results obtained by our best model variant, and the green block in the gray rectangle represent the ground truth time period in which the anomaly event occurred. The second row presents the visualization results of different variants of our model, including results of ResNet50, ResNet50, ResNet150, and the best model variants, respectively.

leverage the complementary spatiotemporal information on the same scale underlying videos. In addition, the fusion of two streams with different number of layers achieves better results than those of the same layer fusion. Reason is this different layer fusion not only utilizes the complementary information between two streams, but also leverages the multiscale information at different layers, as Tables 5 and 6 show. Thus, fusion of RGB and Flow two streams is optimal in anomaly event detection task.

The benefits of our proposed solution are that it can further improve the performance of anomaly event detection significantly by leveraging the complementary information of RGB.

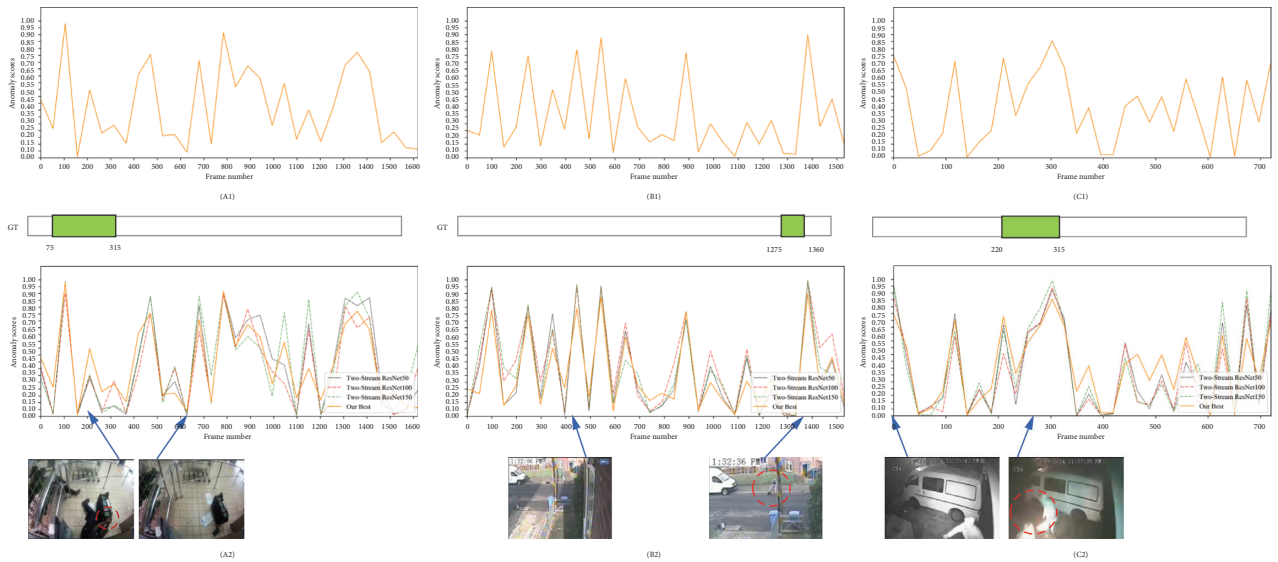

FIGURE 10: The visualization results of our method on testing videos on ShanghaiTech. The first row exhibits the visualization results obtained by our best model variant, and the purple rectangles represent the ground truth time period in which the anomaly event occurred. The second row shows the visualization results of different variants of our model, including results of ResNet50, ResNet50, ResNet150, and the best model variants, respectively, and Flow modalities in the video. Moreover, our proposed solution can provide inspiration for other video-related tasks, including video classification, video segmentation, video tracking and video detection, through bistream setting to obtain the improved.

6. Conclusion

This paper proposes a novel two-stream-based model for anomaly event detection. Specifically, this model consists of RGB and Flow two branch networks, and the final anomaly detection score is the fusion of two networks. Meanwhile, we consider two fusion strategies, including the fusion of two streams with the same of different number of layers, respectively. The proposed model can utilize the complementary information of the two streams hidden in the video, which can improve the performance of anomaly event detection. Ablative studies based on two benchmark datasets UCF-Crime and ShanghaiTech have validated the effectiveness of the proposed model. Future work should focus more on effective feature extraction methods for improved anomaly event detection using new inputs [49] in edge computing environment [50–52].

Data Availability

The datasets used to support the findings of this study are available at https://webpages.uncc.edu/cchen62/dataset.html (UFC-Crime) and https://svip-lab.github.io/datasets.html (ShanghaiTech).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was partially funded by Shanxi Agricultural University Young Science and Technology Innovation Programme (41257914) and Shanxi Key Research and Development Program (201703D221033-3).

References

[1] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, FL, USA, June 2009.

[2] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, and M. Intelligence, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[3] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2011.

[4] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 2008.

[5] J. C. Duchi and E. Hazan, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[6] M. Hasan, J. Choi, J. Neumann, A. K. Roychowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," 2016, https://arxiv.org/abs/1604.04574.

[7] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Edmonton*, Alberta, Canada, July 2002.

[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[9] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27 th International Conference on Machine Learning*, Haifa, Israel, 2010.

[10] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proceedings of the Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2011.

[11] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[12] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 108–118, 2000.

[13] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," 2018, https://arxiv.org/abs/1801.04264.

[14] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proceedings of the International Conference on Computer Vision*, Venice, Italy, October 2017.

[15] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009.

[16] B. Zhao, L. Feifei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proceedings of the Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2011.

[17] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010.

[18] N. Li, H. Guo, D. Xu, and X. Wu, "Multi-scale analysis of contextual information within spatio-temporal video volumes for anomaly detection," in *Proceedings of the International Conference on Image Processing*, Paris, France, October 2014.

[19] B. Antic and B. Ommer, "Video parsing for abnormality detection," in *Proceedings of the International Conference on Computer Vision*, Barcelona, Spain, November 2011.

[20] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proceedings of the 26 th International Conference on Machine Learning*, Montreal, Canada, 2009.

[21] W. Li, V. Mahadevan, V. NJIToPA, and M. Intelligence, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 18–32, 2014.

[22] N. Li, X. Wu, H. Guo et al., "Anomaly detection in video surveillance via Gaussian process," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, no. 6, Article ID 1555011, 2015.

[23] K. Cheng, Y. Chen, and W. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015.

[24] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009.

[25] T. M. Hospedales, S. Gong, and T. Xiang, "A Markov clustering topic model for mining behaviour in video," in *Proceedings of the International Conference on Computer Vision*, Kyoto, Japan, October 2009.

[26] C. Wang, Z. Chen, K. Shang, and H. Wu, "Label-removed generative adversarial networks incorporating with *K*-Means," *Neurocomputing*, vol. 361, pp. 126–136, 2019.

[27] T. Meng, K. Wolter, H. Wu, Q. Wang, and M. Computing, "A secure and cost-efficient offloading policy for Mobile Cloud Computing against timing attacks," *Pervasive and Mobile Computing*, vol. 45, pp. 4–18, 2018.

[28] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proceedings of the International Conference on Computer Vision*, Sydney, Australia, December 2013.

[29] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—a new baseline," 2018, https://arxiv.org/abs/1712.09867.

[30] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," 2017, https://arxiv.org/abs/1701.01546.

[31] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proceedings of the International Conference on Multimedia and Expo*, Hong Kong, China, July 2017.

[32] K. P. Adhiya, S. R. Kolhe, and S. S. Patil, "Tracking and identification of suspicious and abnormal behaviors using supervised machine learning technique," in *Proceedings of the International Conference on Advances in Computing, Communication and Control*, Mumbai India, January 2009.

[33] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29573–29588, 2018.

[34] C. Bergeron, J. Zaretzki, C. M. Breneman, and K. Bennett, "Multiple instance ranking," in *Proceedings of the 25 th International Conference on Machine Learning*, Helsinki, Finland, 2008.

[35] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proceedings of the 016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[36] J. Wang, Y. Song, T. Leung et al., "Learning fine-grained image similarity with deep ranking," 2014, https://arxiv.org/abs/1404.4661.

[37] M. Gygli, Y. Song, and L. Cao, "Video2GIF: automatic generation of animated GIFs from video," 2016, https://arxiv.org/abs/1605.04850.

[38] S. Sankaranarayanan, A. Alavi, and R Chellappa, *Triplet Similarity Embedding for Face Verification*, https://arxiv.org/abs/1602.03418, 2016.

[39] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "Deep image retrieval: learning global representations for image search," 2016, https://arxiv.org/abs/1604.01325.

[40] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," 2014, https://arxiv.org/abs/1406.2199.

[41] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," 2016, https://arxiv.org/abs/1608.00859.

[42] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," 2016, https://arxiv.org/abs/1604.06573.

[43] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal residual networks for video action recognition," 2016, https://arxiv.org/abs/1611.02155.

[44] H. Kwon, Y. Kim, J. S. Lee, and M. Cho, "First person action recognition via two-stream ConvNet with long-term fusion pooling," *Pattern Recognition Letters*, vol. 112, pp. 161–167, 2018.

[45] L. Sevillalara, Y. Liao, F. Guney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," 2018, https://arxiv.org/abs/1712.08416.

[46] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," 2017, https://arxiv.org/abs/1711.10305.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, https://arxiv.org/abs/1512.03385.

[48] J. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection," 2019, https://arxiv.org/abs/1903.07256.

[49] X. Xu, X. Liu, Z. Xu, F. Dai, X. Zhang, and L. Qi, "JIIoTJ. Trust-oriented IoT service placement for smart cities in edge computing," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4084–4091, 2019.

[50] X. Xu, X. Zhang, H. Gao, Y. Xue, L. Qi, and W. Dou, "BeCome: blockchain-enabled computation offloading for IoT in mobile edge computing," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4187–4195, 2020.

[51] X. Xu, C. He, Z. Xu, L. Qi, S. Wan, and Z. A. Bhuiyan, "JIIoTJ. Joint optimization of offloading utility and privacy for edge computing enabled IoT," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2622–2629, 2019.

[52] S. Li, "Zero trust based internet of things," *EAI Endorsed Transactions on Internet of Things*, vol. 5, no. 20, p. 6, 2020.