



Research Article

A Three-Stage Anomaly Detection Framework for Traffic Videos

Junzhou Chen ^{1,2}, Jiancheng Wang,^{1,2} Jiajun Pu,^{1,2} and Ronghui Zhang ^{1,2}

¹School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, No. 66 Gongchang Road, Guangming District, Shenzhen, Guangdong 518107, China

²Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology, Sun Yat-sen University, Guangzhou 510006, China

Correspondence should be addressed to Ronghui Zhang; zhangrh25@mail.sysu.edu.cn

Received 25 February 2022; Revised 6 April 2022; Accepted 11 June 2022; Published 5 July 2022

Academic Editor: Yong Zhang

Copyright © 2022 Junzhou Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As reported by the United Nations in 2021, road accidents cause 1.3 million deaths and 50 million injuries worldwide each year. Detecting traffic anomalies timely and taking immediate emergency response and rescue measures are essential to reduce casualties, economic losses, and traffic congestion. This paper proposed a three-stage method for video-based traffic anomaly detection. In the first stage, the ViVit network is employed as a feature extractor to capture the spatiotemporal features from the input video. In the second stage, the class and patch tokens are fed separately to the segment-level and video-level traffic anomaly detectors. In the third stage, we finished the construction of the entire composite traffic anomaly detection framework by fusing outputs of two traffic anomaly detectors above with different granularity. Experimental evaluation demonstrates that the proposed method outperforms the SOTA method with 2.07% AUC on the TAD testing overall set and 1.43% AUC on the TAD testing anomaly subset. This work provides a new reference for traffic anomaly detection research.

1. Introduction

With rapid economic development, a leapfrog has been achieved in transportation. Contrary to the wishes of [1], the number of civilian vehicles and the road network density are increasing, and the road network structure is becoming more complex. As a result, traffic management schemes are proposed correspondingly; numerous measures such as CCTV cameras and radars are put on the roadside to regulate the driving behavior of drivers [2–4]. Studies on the vehicle are carried out [5–9]. However, numerous traffic accidents with terrible consequences still happen every year [10]. According to the National Bureau of Statistics, in 2020, there were 244,674 traffic accidents in China, resulting in 61,703 deaths, 250,723 injuries, and a direct property loss of about 206 million dollars [11].

The extent of the damage often depends on when traffic controllers discover the incident and the duration of the traffic incident [12]. The lack of timely accident reporting will result in many deaths due to delays in medical assistance, prolonged traffic jams, and even secondary accidents.

Therefore, real-time detection of traffic incidents is an effective way to reduce their impact significantly. With the development of technology and the advancement of research, various detection technologies and data sources are used in automatic traffic accident detection studies. Traditional traffic data provides rich and relatively available data sources [13, 14], such as traffic data, vehicle speed data, and occupancy data. Numerous machine learning models are also applied to detect traffic incidents with traffic data and have achieved good results [15–18]. Some studies employed online data from mobile phones to detect traffic incidents, such as Twitter and Weibo. Specifically, they used web crawler technology to detect incidents through data processing, filtering, reasoning, and other processes [19, 20]. Moreover, Zhang and He [21] integrated the social media data with traffic data and achieved a better effect.

Another effective solution is to use surveillance video data. On the one hand, surveillance cameras are extensively used on modern roads and help traffic managers obtain rich surveillance video data of road areas. On the other hand, with the rapid development of computer vision and artificial

intelligence, many advancements have been achieved in video analysis and understanding research. Video-based surveillance for traffic incident detection became possible whether in the middle of the night or when the traffic flow is low.

For the research on traffic video anomaly detection, the video anomaly detection method can be divided into two categories according to the model type: the traditional machine learning method and the deep learning method. Traditional machine learning methods are mainly based on the Gaussian mixture model [22], histogram feature [23–25], hidden Markov model [26, 27], appearance feature [28, 29], and Bayesian network model [30]. Deep learning methods are mostly based on appearance features and motion features in specific scenes, and the final anomaly detection is performed by reconstruction error [31–36], prediction error [37–40], or hybrid transfer learning classification [41, 42].

However, the two methods mentioned above are often mixed and cannot be accurately distinguished in recent years. Therefore, we follow [43] and broadly classify video anomaly detection methods into three categories according to the detection granularity: video level, slice level, and frame level. This paper proposes a three-stage anomaly detection framework for traffic video. The main contributions can be summarized as follows:

- (a) We proposed a novel weakly supervised learning method for traffic video anomaly detection. Specifically, in the first stage, the ViVit network is employed as a feature extractor to capture the spatiotemporal features from the input video. In the second stage, the class and patch tokens are fed separately to the segment-level and video-level traffic anomaly detectors. In the third stage, we finished the construction of the entire composite traffic anomaly detection framework by fusing outputs of two traffic anomaly detectors above with different granularity.
- (b) We propose a segment-level traffic anomaly detector based on the global spatiotemporal features (class token), a video-level traffic anomaly detector based on the similarity of patch tokens from different segments, and a composite traffic anomaly detection framework. By entirely using video-level similarity features and all segment-level global spatiotemporal features, the long-tail distribution problem in traffic video anomaly detection tasks can be effectively solved.
- (c) The experimental results demonstrate the effectiveness of the proposed method. Specifically, our proposed architecture achieves 91.71% and 63.09% on the overall set and anomaly subset of the TAD testing set, which are 2.07% and 1.43% higher than the SOTA method, respectively.

The rest of the paper is organized as follows. Section 2 discusses studies related to video anomaly detection in terms of three different detection granularities: video level, segment level, and frame level. The details of our three-stage anomaly detection framework are described in Section 3.

Section 4 shows the implementation details and quantitative results of the experiments. Section 5 gives the conclusions and the focus of future work.

2. Literature Review

Rapid technological progress in computer vision and machine learning has enabled better video understanding. Many studies on traffic anomaly detection via surveillance video have been carried out in recent decades. Following [43], the techniques that could be applied in traffic video anomaly detection can be divided into three categories: video level [44], segment level [45], and frame level [46]. The details of the various method categories are described as follows.

2.1. Video-Level Methods. Popular single-class classification methods directly detect novelty by measuring the gap between the original and reconstructed inputs, such as Support Vector Machine (SVM) [44, 47] and SVDD [48, 49]. In general, video-level methods treat anomaly detection as a novel detection problem. Liu et al. [50] proposed a single-objective generative adversarial active learning method that directly generates information-rich potential outliers based on a mini-max game between the generator and the discriminator. Ngo et al. [51] used a similar approach based on generative adversarial networks (GANs).

2.2. Segment-Level Methods. Segment-level detection is a method between video level and frame level, which divides the input video into multiple segments instead of frames. In recent years, this research has become increasingly popular, and there is a growing body of related work. Some work built memory modules that learn only normal patterns from normal data and determine the presence of anomalies by computing reconstruction errors [33, 35]. In another interesting work, Georgescu et al. [52] proposed joint learning of multiple tasks by self-supervision to produce differential anomaly information: three self-supervised tasks and an ablation study. Moreover, a two-stage framework is also a popular research approach. Waqas et al. [41] applied pre-trained 3D networks to extract spatiotemporal features and trained the classifier with multi-instance learning techniques. Following this work, Zhu and Newsam [45] introduced optical flow; Lin et al. [53] proposed a dual-branch network; Lv et al. [54] replaced the feature extractor with a TSN and proposed an HCE module to capture dynamic changes; Feng et al. [55] applied pseudolabel and self-attentive feature encoders for training; Wu et al. [56] also proposed a dual-branch network but with tubular and temporal branches and so on. This strategy can improve detection accuracy and localize anomalies using a small amount of annotated information.

2.3. Frame-Level Methods. Based on the classical directional optical flow histograms, references [23–25, 29] have developed their own way of extracting frame-level features for

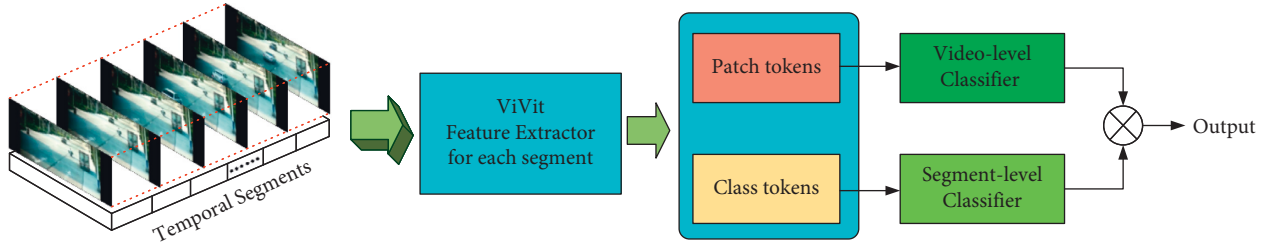


FIGURE 1: Three-stage video-based traffic anomaly detection algorithm framework.

anomaly detection, but they are scene-dependent. More generative models were used to predict future frames and calculate the reconstruction error between predicted and real frames. On this basis, reference [35] used U-Net and memory module; reference [36] used AE and DPU module. Both of them generate “normal” future frames and determine whether they are anomalous. Moreover, after generative adversarial networks (GANs) proved their ability to generate “normal” future frames, many researchers have focused their interest on detecting traffic anomalies at the frame level. In a similar way to determining anomalies by prediction errors [37–40, 46], the frame-level detection method based on GAN networks compares the current frames constructed by GANs with the ground truth current frames [31, 37, 57–59]. Besides GAN, there are other methods to detect traffic incidents at the frame level. Ryan Medel and Svakis [60] built an end-to-end frame-level anomaly detector using a long and short-term memory (Conv-LSTM) network. Zhou et al. [43] first detected boundary frames as potential incident frames and confirmed by encoding spatiotemporal features whether these frames are incident frames.

The following summary can be made from the above review, video-level methods usually aggregate features for single-class prediction, which can take full advantage of fully supervised tasks but cannot identify anomaly locations. Segment-level methods can be trained by weakly supervised learning mechanisms such as multi-instance learning to perform effective anomaly detection and localization while maintaining a few annotations (video-level annotations). Frame-level methods generally perform single-frame detection by calculating the reconstruction error between predicted and real frames, and although the localization is accurate, their application scenarios are limited and have significant errors. Therefore, in this paper, we combine the advantages of video-level methods and fragment-level methods to complement each other and propose a three-stage composite traffic anomaly detection framework to achieve the anomaly detection and localization of anomaly videos.

3. Method

As a carrier of spatiotemporal information, frames in video contain temporal information that is not available in mutually irrelative images. Therefore, understanding and analyzing videos is more complicated and time-consuming than understanding and analyzing images directly. Many current video anomaly detection methods are generally

divided into two steps: the first step is to extract spatiotemporal features from the input video using a pretrained 3D model; the second step is to model the extracted spatiotemporal features and evaluate the anomaly score.

As shown in Figure 1, we propose a three-stage anomaly detection method for traffic videos. Unlike other methods, we use the pretrained ViVit to extract features from video segments and propose a composite framework of video-level and segment-level traffic anomaly detectors. Specifically, we first split the input video into multiple segments and then use the pretrained ViVit to extract spatiotemporal features from those segments. After that, their global spatiotemporal features (class tokens) and local spatiotemporal features (patch tokens) are delivered to the segment-level and video-level traffic anomaly detectors, respectively. Finally, the output results of the above two detectors are compound corrected to complete the final anomaly value evaluation.

In this paper, to avoid ambiguity, class tokens refer to the segment-level global spatiotemporal features extracted by the pretrained ViVit model, and patch tokens refer to the segment-level local spatiotemporal features extracted by the pretrained ViVit model.

3.1. Extract Spatiotemporal Features Based on ViVit.

Unlike the 3D convolution-based feature extractor [61–63], the Transformer-based ViVit model can effectively model the long contextual information of the input video by using its attentional architecture. Therefore, here we use ViVit model 2 (Factorized Encoder) [64], which was pretrained [65] on the Kinetics-400 dataset, as the feature extractor.

The above ViVit model 2 adopts the embedding method of ViVit-B, that is, a tubelet embedding for the input video, whose tubelet size is set to $h \times w \times t = 16 \times 16 \times 2$. The Factorized Encoder consists of two independent transformer encoders. The first is a spatial encoder that models the short spatiotemporal relationships of nonoverlapping adjacent $t = 2$ frames and feeds its output (spatial class token) to the next encoder. The second is the temporal encoder, which uses the spatial class token within the above nonoverlapping adjacent $t = 2$ frames to model the video long spatiotemporal relationship. Finally, the global spatiotemporal features (class token) and the local spatiotemporal feature (patch tokens) are obtained.

Before extracting the temporal features, we perform a preprocessing operation on the input video. Specifically, we resize each frame in the video to 224×224 and normalize it. Like Waqas et al. [41], we slice the processed video into

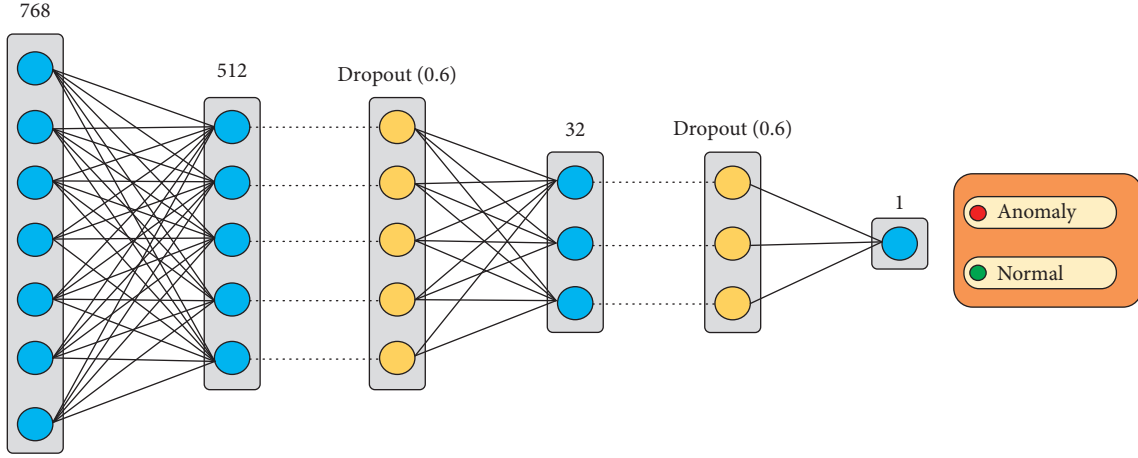


FIGURE 2: Segment-level classifier.

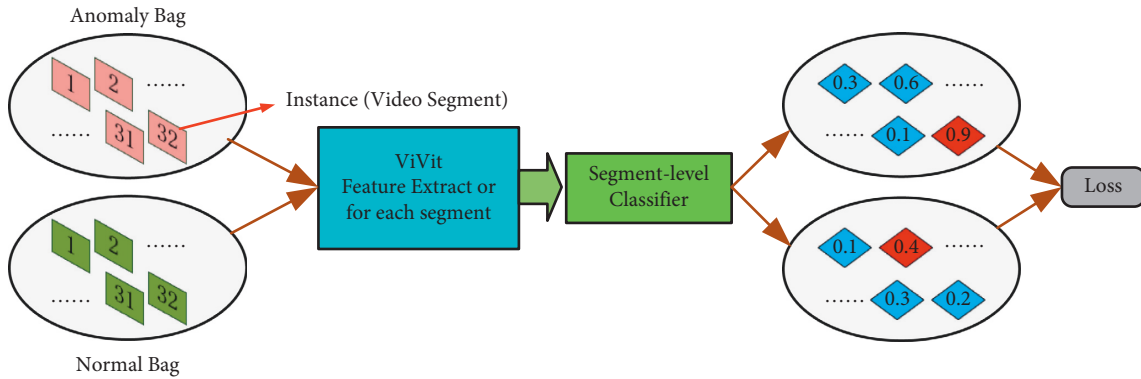


FIGURE 3: Multi-instance learning-based segment-level classification.

multiple video subunits, which are then distributed into 32 segments, where each video subunit is 16 frames. However, unlike the reference, we perform the averaging operation for each video subunit in the segment directly rather than after feature extraction. Then, each segment is subjected to spatiotemporal feature extraction using the pretrained ViViT model to obtain 1 class token and 8 patch tokens. The class token aggregates all the spatiotemporal features of the whole segment and represents the whole spatiotemporal segment. The patch token aggregates the certain local spatiotemporal features in the segment and represents the local spatiotemporal segment and its local contextual spatiotemporal segment. Finally, the class tokens of all segments are delivered to the segment-level anomaly detector for segment-level detection; the patch tokens of all segments are delivered to the video-level anomaly detector for video-level detection.

3.2. Segment-Level Traffic Anomaly Detector. As shown in Figure 2, we propose a segment-level classifier based on class token (768 dimensions). Our segment-level classifier is made up of five layers, detailed in Figure 2. Its last layer outputs an anomaly score, and the closer the score to 0, the greater the probability that the input segment is normal. Conversely, the closer the score to 1, the greater the probability that the input segment is abnormal.

Here, we use the multi-instance learning mechanism to train our segment-level traffic anomaly detector, a weakly supervised learning method, following [41]. As shown in Figure 3, it is the training framework of our segment-level traffic anomaly detector based on multi-instance learning:

- (a) Input both positive bag (anomaly video) and negative bag (normal video) into 32 segments, and then compile those segments as a positive bag \mathcal{B}_a and a negative bag \mathcal{B}_n . Each segment in its bag is called the instance, so the positive bag and the negative bag can be described as follow:

$$\begin{aligned} \mathcal{B}_a &= \{a_i, i = 1, \dots, m\}, \\ \mathcal{B}_n &= \{n_i, i = 1, \dots, m\}, \end{aligned} \quad (1)$$

where a_i is the instance in the positive pack and n_i is the instance in the negative pack. Our use case has $m = 32$.

- (b) Under the basic assumption of multi-instance learning, there are only bag-level labels. Besides, each positive bag contains at least one positive example, while each negative bag contains no positive examples:

$$\begin{aligned} y_i^a &= 1, \quad \exists a_i \in \mathcal{B}_a, \\ y_i^n &= 0, \quad \forall n_i \in \mathcal{B}_n, \end{aligned} \quad (2)$$

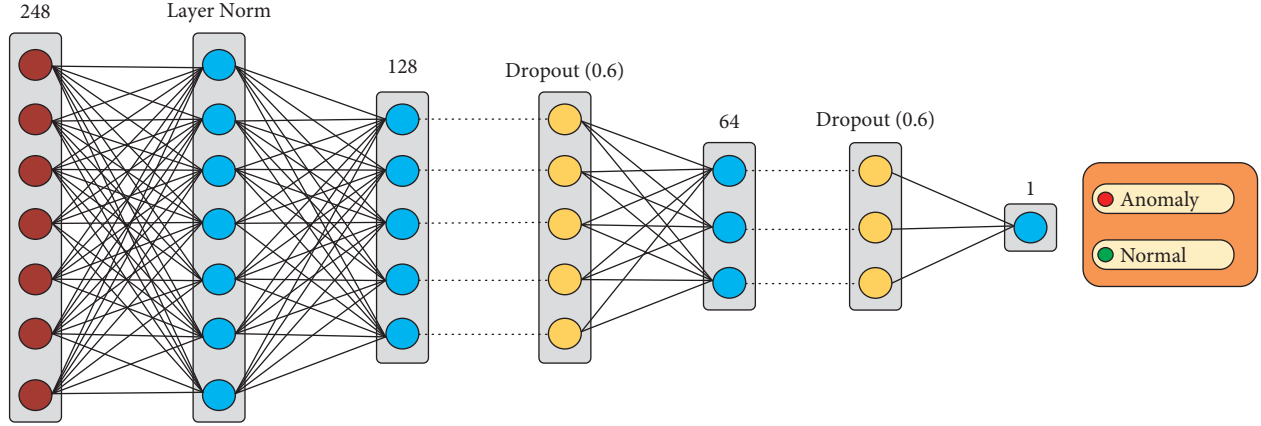


FIGURE 4: Video-level classifier.

where y_{a_i} is the label of instance a_i and y_{n_i} is the label of instance n_i . The instance is a positive instance when its label is 1, but a negative instance when its label is 0.

- (c) Using pretrained ViVit mentioned in Section 3.1, we can extract the feature from all instances in both positive bag and negative bag to obtain their corresponding class token vector as follow:

$$\begin{aligned} \mathcal{C}_a &= \{c_i^a, i = 1, \dots, m\}, \\ \mathcal{C}_n &= \{c_i^n, i = 1, \dots, m\} \end{aligned} \quad (3)$$

where \mathcal{C}_a is the class tokens feature set, extracted from the positive bag \mathcal{B}_a with pretrained ViVit model, the same as \mathcal{C}_n .

- (d) Put extracted feature (class token) of each instance into the segment-level classifier and acquire an anomaly score:

$$\begin{aligned} \mathcal{S}_a &= \{s_i^a = \mathcal{F}_s(c_i^a), i = 1, \dots, m\}, \\ \mathcal{S}_n &= \{s_i^n = \mathcal{F}_s(c_i^n), i = 1, \dots, m\}. \end{aligned} \quad (4)$$

Each training sample \mathcal{X} should include one positive bag and one negative bag together, namely, $\mathcal{X} = \{\mathcal{B}_a, \mathcal{B}_n\}$. We use a combination of the following three loss functions to train the segment-level classifier \mathcal{F}_s . The first loss function is margin ranking loss. Choose the biggest instance anomaly score in positive and negative packets as their bag-level anomaly score for metric ranking loss calculation, where the metric parameter margin is set to 1.

$$l_{\text{margin}} = \max\left(0, \max_{n_i \in \mathcal{B}_n} \mathcal{F}_s(c_i^n) - \max_{a_i \in \mathcal{B}_a} \mathcal{F}_s(c_i^a) + \text{margin}\right). \quad (5)$$

The second loss function is the temporal smoothness term. Since video is a sequence of continuous frames combined, we split it into segments. In theory, the output anomaly score should be relatively smooth between segments. The temporal smoothness term is designed as

$$l_{\text{smooth}} = \sum_i^{(m-1)} (\mathcal{F}_s(c_i^a) - \mathcal{F}_s(c_{i+1}^a))^2. \quad (6)$$

The third one is the sparsity term. For anomalies only take a small part of the entire video, the anomaly instance should be sparse in the positive bag:

$$l_{\text{sparsity}} = \sum_i^m \mathcal{F}_s(c_i^a). \quad (7)$$

Therefore, our final loss function becomes

$$\mathcal{L}_s = l_{\text{margin}} + \eta_1 l_{\text{smooth}} + \eta_2 l_{\text{sparsity}}. \quad (8)$$

Here, the η_1 and η_2 coefficients weight time smooth loss and sparse term loss separately.

3.3. Video-Level Traffic Anomaly Detector. As shown in Figure 4, we presented a novel video-level classifier. The input layer of the classifier is the similarity of patch tokens from adjacent segments of the same video. Our video-level classifier is made up of six layers, detailed in Figure 4. Its last layer outputs an anomaly score, and the closer the score to 0, the greater the probability that the input video is normal. Conversely, the closer the score to 1, the greater the probability that the input video is abnormal.

Here, we choose the cosine similarity to measure the degree of difference between two feature vectors. For example, given \mathcal{P}_i and \mathcal{P}_j , let the cosine similarity calculation function be \mathcal{F}_{cos} , and then, the similarity Sim_{ij} between two vectors is calculated as follows:

$$\text{Sim}_{ij} = \mathcal{F}_{\text{cos}}(\mathcal{P}_i, \mathcal{P}_j),$$

$$\mathcal{F}_{\text{cos}}(\mathcal{P}_i, \mathcal{P}_j) = \frac{\mathcal{P}_i \cdot \mathcal{P}_j}{\|\mathcal{P}_i\| \|\mathcal{P}_j\|} = \frac{\sum_{k=1}^n \mathcal{P}_k^i \times \mathcal{P}_k^j}{\sqrt{\sum_{k=1}^n (\mathcal{P}_k^i)^2} \times \sqrt{\sum_{k=1}^n (\mathcal{P}_k^j)^2}}. \quad (9)$$

A normal video should remain continuous in its timeline, even after it is segmented. The continuity between adjacent segments can be reflected in their similarity. Therefore, a normal video should maintain a relatively high similarity between adjacent segments. In contrast, an abnormal video would be discontinuous in its timeline due to the presence of abnormal clips. So, the similarity between

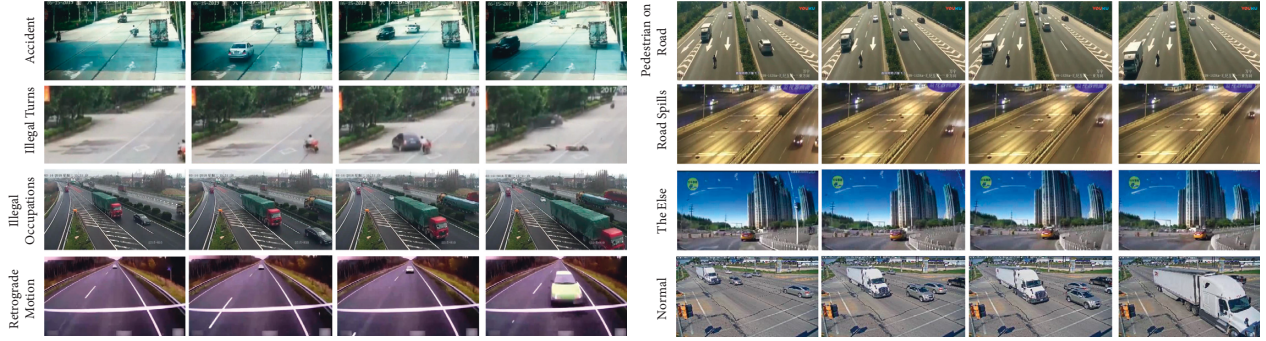


FIGURE 5: Examples of anomaly frames in the TAD dataset [54].

adjacent segments in anomaly video should dramatically decrease and unstable similarity between adjacent segments in which anomaly occurs.

Based on the above observation and analysis, we proposed a video-level traffic anomaly detector to focus on the feature similarity between segments and output the video-level anomaly score. Specifically, after the feature extraction in Section 3.1, an input video could get 32 groups of patch tokens (8 in each group). Then, we calculate the cosine similarity between each corresponding pair of patch token features in adjacent groups and finally get $8 \times 31 = 248$ patch token cosine similarity. Therefore, an entire video can be represented by a 248-dimensional similarity space feature vector, which is fed into a video-level traffic anomaly discriminator for forwarding derivation to obtain its video-level anomaly score.

In essence, our feature-similarity-based video-level traffic anomaly detector is a binary classification task whose parameters can be optimized with Binary Cross-Entropy Loss. After training on a large set of video-level labeled data, it is capable of performing high-performance anomaly traffic video discrimination.

$$\mathcal{L}_v = \mathcal{Y} * \log(\hat{\mathcal{Y}}) + (1 - \mathcal{Y}) * \log(1 - \hat{\mathcal{Y}}). \quad (10)$$

Here, \mathcal{Y} is the label of input video, and $\hat{\mathcal{Y}}$ is the output of the video-level traffic anomaly detector.

3.4. Composite Traffic Anomaly Detection. As mentioned earlier, video-level traffic anomaly detectors focus on feature similarity between adjacent video segments, while segment-level traffic anomaly detectors pay attention to modeling global spatiotemporal features within video segments. Theoretically, feature similarity between segments has stronger integrity and stability compared to global spatiotemporal features within segments. Therefore, the video-level anomaly traffic detector can provide a more reliable output and assist the segment-level detector in anomaly identification. Inspired by [33, 35], we design the following composite operation (equation (11)). When the anomaly score of the video-level traffic anomaly detector exceeds the threshold value, we normalize the output of the segment-level traffic anomaly detector by a min-max normalization [37]:

TABLE 1: Statistic of TAD dataset.

Dataset	Videos	Frames	Label level
Training set	400	452,220	Video level
Testing overall set	100	88,052	Frame level
Testing anomaly subset	60	18,900	Frame level

$$\mathcal{S}_C = \begin{cases} \frac{\mathcal{S} - \min_{i \in \{1, \dots, m\}} \mathcal{S}}{\max_{i \in \{1, \dots, m\}} \mathcal{S} - \min_{i \in \{1, \dots, m\}} \mathcal{S}}, & \text{if } \hat{\mathcal{Y}} > \lambda, \\ \mathcal{S}, & \text{otherwise,} \end{cases} \quad (11)$$

where \mathcal{S}_C is the composite traffic anomaly score, \mathcal{S} is the output of segment-level traffic anomaly detector, $\hat{\mathcal{Y}}$ is the output of video-level traffic anomaly detector, and λ is the preset threshold.

4. Experiment

4.1. Dataset and Training Details. We conducted the experiments on the TAD dataset built by Lv et al. [54], a total of 500 traffic surveillance videos with 250 normal and anomaly videos, respectively. The average frames in each clip of the TAD dataset are 1075. The anomalies randomly occur in the anomaly clips and take about 80 frames on average. The anomalies, including vehicle accidents, illegal turns, illegal occupations, retrograde motion, pedestrians on the road, and road spills, take place in various scenarios, weather conditions, and daytime periods.

Some examples of anomaly videos in the TAD dataset are shown in Figure 5. While training and testing, we followed [54] to split the TAD dataset into two parts, with a training set of 400 videos and a test set of 100 videos. Other statistics are shown in Table 1.

All experiments were carried out on PyTorch and hardware configuration of NVIDIA GeForce RTX 2070 GPU, 16 G RAM, CPU i7-10700k @3.80 GHz machine. We jointly use margin ranking loss, time smooth loss, and sparse term loss to train our segment-level anomaly traffic detector as mentioned in Section 3.2, where we set margin = 1, $\eta_1 = 8 \times 10^{-5}$, and $\eta_2 = 8 \times 10^{-5}$. It was trained of 1000 epochs with batch size 4. Binary Cross-Entropy Loss was

TABLE 2: Result of TAD dataset

Class	Method	Overall set AUC (%)	Anomaly subset AUC (%)
Unsupervised	Luo et al. [32]	57.89	55.84
	Liu et al. [37]	69.13	55.38
Weakly supervised	Sultani et al. [41]	81.42	55.97
	Zhu et al. [45]	83.08	56.89
	Lv et al. [54]	89.64	61.66
	Ours	91.71	63.09

applied to train the video-level traffic anomaly detector, which was 1000 epochs with batch size 8.

Both detectors were SGD Optimizer paired with Cosine Annealing LR; we both set their Optimizer parameters $lr = 0.001$, momentum = 0.9, and weight_decay = 1×10^{-4} and kept the best performed model parameters as the optimal model. In our experiment, by comparing different preset thresholds, it is proved that $\lambda = 0.6$ works best.

4.2. Evaluation Metrics. For the evaluation metrics of anomaly detection, we first defined “true positive (TP),” “false positive (FP),” “true negative (TN),” and “false negative (FN),” which represent the difference between the predicted and actual classes.

TP: the predicted class is “anomaly,” and so is the actual class.

TN: the predicted class is “normal,” and so is the actual class.

FN: the predicted class is “normal,” but the actual class is “anomaly.”

FP: the predicted class is “anomaly,” but the actual class is “normal.”

The true positive rate (TPR) is the probability that an actual positive will test positive, and the false positive rate (FPR) is defined as the probability of falsely rejecting the null hypothesis. TPR and FPR are calculated as follows:

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{aligned} \quad (12)$$

We choose the area under the frame-level ROC curve (AUC) as the primary evaluation metric for traffic video anomaly detection. The frame-level AUC is insensitive to the imbalance of sample classification and, therefore, suitable as our primary evaluation metric. Meanwhile, as an evaluation metric, the frame-level AUC reflects the detection performance of a method in locating traffic video anomalies. The closer the AUC value is to 1, the better the detection performance is.

The receiver operating characteristic curve (ROC) mentioned above is a graph showing the performance of the classification model at all classification thresholds, and the plotted curve represents the relationship between TPR and FPR.

We also used some other evaluation metrics to evaluate the ablation study of our proposed method. Precision and

recall are two important evaluation metrics for detection evaluation. The precision (equation (13)) of a class reflects the proportion of the number of TP among the total number of elements that are predicted and labeled as the positive class. Recall (equation (14)) is defined as the proportion of the number of TP among the total number of the positive classes. Recall and precision are contradictory measures, and the $F1$ -score (equation (15)) is defined as a combination of recall and precision.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (13)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (14)$$

$$F1 - \text{score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (15)$$

4.3. Comparison with SOTA Method. In this paper, we compare the performance of the proposed method with several other SOTA methods, and their quantitative results on TAD are shown in Table 2. Among all the methods, the work by Luo et al. [32] and Liu et al. [37] uses an unsupervised approach and trains with only the normal video training set. Otherwise, Sultani et al. [41], Zhu et al. [45], Lv et al. [54], and our work use weakly supervised learning methods with the video-level labeled training set for training. The above SOTA results on TAD refer to [54].

The comparative results of the performance on TAD are given in Table 2. They represent that the weakly supervised learning methods outperform the unsupervised learning methods. For example, the relatively inefficient weakly supervised learning method [41] reaches 81.42% AUC on the overall set and 55.97% AUC on the anomaly subset, yet still about 12.29% and 0.13% higher than the best unsupervised learning method [37]. Besides, among the current SOTA methods, Lv et al. perform best on both the overall set and anomaly subset, with 89.64% AUC on the overall set and 61.66% AUC on the anomaly subset. However, the proposed method outperforms the optimal SOTA with 2.07% and 1.43% higher AUC on the overall set and anomaly subset, separately. The results show that our work has been the SOTA on the TAD dataset.

The above quantitative analysis proves the following points: (1) Unsupervised learning methods have limited performance in complex scenarios and when data anomalies are not significant. (2) Weakly supervised learning methods

TABLE 3: Ablation studies on TAD dataset.

Dataset	Methods	Recall (%)	Precision (%)	F1-score	AUC (%)
Overall set	T-SAD	92.16	90.17	0.9088	91.05
	T-CAD	92.00	90.48	0.9109	91.71
Anomaly subset	T-SAD	66.68	62.68	0.6154	62.04
	T-CAD	66.62	63.15	0.6279	63.09

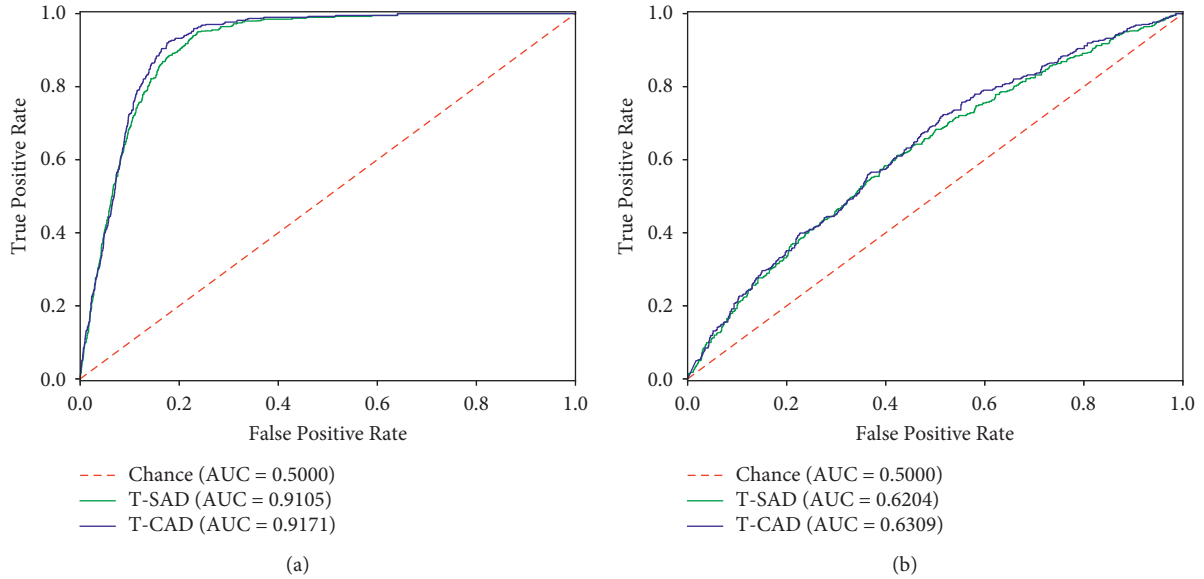


FIGURE 6: ROC curves on TAD dataset.

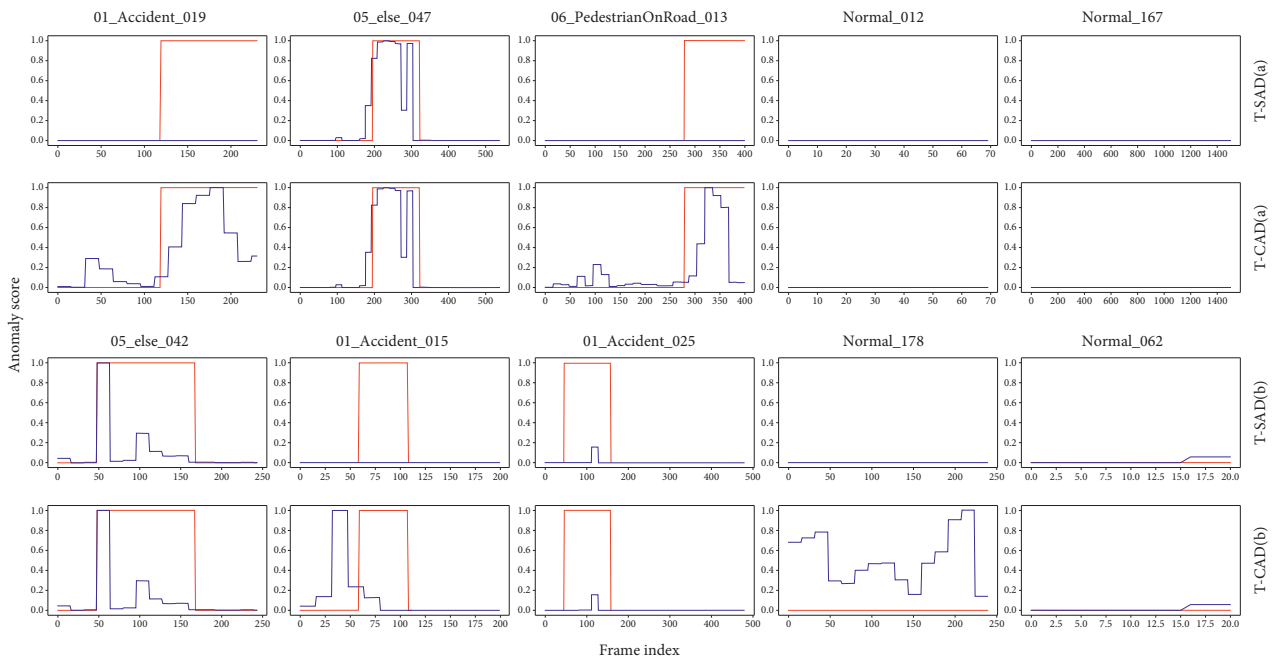


FIGURE 7: Video detection result.

can significantly improve the learning and representational ability of neuronal networks on training data while maintaining a small number of annotations. (3) The proposed method is more advanced in anomaly detection and

localization, where the ViVit-based feature extractor can effectively characterize the pattern features of video data, and the ViVit-based composite traffic anomaly detection method can more accurately capture the anomalous features

in video data, making the method in this paper significantly better than the existing SOTA method [54].

4.4. Ablation Studies. We conducted ablation experiments to analyze the performance advantages of the Transformer-based Segment-level traffic Anomaly Detector (T-SAD) itself and the performance advantages of the Transformer-based Composite traffic Anomaly Detection method (T-CAD). As shown in Table 3, the AUC of T-SAD reached 91.05% and 62.04% on the overall set and anomaly subset, respectively, exceeding the current SOTA method [54] by 1.14% and 0.38%, respectively. In addition, the AUC values of T-CAD were 0.66% and 1.05% higher than those of T-SAD on the overall set and anomaly subset, respectively, demonstrating the better performance of T-CAD compared with T-SAD in anomaly localization.

Figure 6 visualizes the ROC curves of T-SAD and T-CAD on the overall set and anomaly subset and vividly demonstrates the superiority of the proposed method. As seen from Figure 6, the ROC curve of T-CAD clearly wraps around the ROC curve of T-SAD, proving that the T-CAD outperforms the T-SAD in all aspects of the overall set and anomaly subset.

We further visualized the detection results of T-SAD and T-CAD on the overall set separately. In the visualized results in Figure 7, row T-CAD (a) shows some reliable outputs from T-CAD on the test set, where T-SAD (a) is the corresponding outputs of T-SAD. It shows an improvement that T-CAD did compare to T-SAD. Still, in Figure 7, T-CAD (b) is some failure outputs from T-CAD on the overall testing set and its corresponding T-SAD. Enhancing detection ability could cause a higher probability of mis-detection to catch abnormal features distributed sparsely in anomalous videos. The exaggeration of the failure outputs is in keeping with the trait of T-CAD, widening the gap in T-SAD results. Nonetheless, no matter the overall set or anomaly subset, performance enhancement proved the effectiveness of our T-CAD structure.

5. Conclusion

In this work, we propose a three-stage anomaly detection for traffic videos. First, we utilize a pretrained ViVit model as the feature extractor to capture the spatiotemporal features of the input video. Then, we put the class tokens into the segment-level traffic anomaly detector for segment-level detection, pretrained with a multi-instance learning strategy. We similarly put the patch tokens into the video-level traffic anomaly detector for video-level detection. Finally, we fuse the video-level and segment-level detection outputs as our final output. From the experimental results, our proposed architecture achieves 91.71% AUC and 63.09% AUC on testing overall set and testing anomaly subset, which outperforms the SOTA method with 2.07% and 1.43%, respectively. Overall, the quantitative results demonstrate the effectiveness of using a spatiotemporal feature extractor and our composite traffic anomaly detection framework on the traffic video anomaly detection problem.

The feature extraction, fusion of foreground and background information, and modeling of relationships between foreground objects may be helpful for anomaly feature extraction, which is worth doing in the future. In addition, the spatial location detection of anomalies and the specific classification of anomalies are also worthy topics for research.

Data Availability

The data used to support the findings of this study are available from the first author and the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the Shenzhen Fundamental Research Program (no. JCYJ20200109142217397), the Guangdong Natural Science Foundation (nos. 2021A1515011794 and 2021B1515120032), the National Natural Science Foundation of China (no. 52172350), and the Guangzhou Science and Technology Plan Project (nos. 202007050004 and 202206030005).

References

- [1] J. d D. Ortúzar, "Future transportation: sustainability, complexity and individualization of choices," *Communications in Transportation Research*, vol. 1, Article ID 100010, 2021.
- [2] A. Franklin, "The future of cctv in road monitoring," *IEE Seminar on CCTV and Road Surveillance*, vol. 10, 1999.
- [3] Ki Yong-Kul, J.-W. Choi, Ho-J. Joun, G.-H. Ahn, and K.-C. Cho, "Real-time estimation of travel speed using urban traffic information system and cctv," in *Proceedings of the 2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 22-24 May 2017.
- [4] F. Baselice, G. Ferraioli, G. Matuozzo, V. Pascasio, and G. Schirinzi, "3d automotive imaging radar for transportation systems monitoring," in *Proceedings of the 2014 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems Proceedings*, 17-18 September 2014.
- [5] R.-H. Zhang, Z.-C. He, H.-W. Wang, F. You, and Ke-N. Li, "Study on self-tuning tyre friction control for developing main-servo loop integrated chassis control system," *IEEE Access*, vol. 5, pp. 6649–6660, 2017.
- [6] Q. Yang, G. Shen, C. Liu, Z. Wang, K. Zheng, and R. Zheng, "Active fault-tolerant control of rotation angle sensor in steer-by-wire system based on multi-objective constraint fault estimator," *Journal of Intelligent and Connected Vehicles*, vol. 12, 2020.
- [7] Y. Cai, T. Luan, H. Gao et al., "Yolov4-5d: an effective and efficient object detector for autonomous driving," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, no. 1–13, pp. 1–13, 2021.
- [8] X. Zhao, X. Li, Y. Chen, H. Li, and Y. Ding, "Evaluation of fog warning system on driving under heavy fog condition based

- on driving simulator,” *Journal of intelligent and connected vehicles*, vol. 4, no. 2, pp. 41–51, 2021.
- [9] K. Li Lim, J. Whitehead, D. Jia, and Z. Zheng, “State of data platforms for connected vehicles and infrastructures,” *Communications in Transportation Research*, vol. 1, Article ID 100013, 2021.
- [10] G. R. Gang and Z. Zhuping, “Traffic safety forecasting method by particle swarm optimization and support vector machine,” *Expert Systems with Applications*, vol. 38, no. 8, pp. 10420–10424, 2011.
- [11] Prc Bureau of Statistics, “China Statistical Yearbook in 2021,” 2022, <http://www.stats.gov.cn/tjsj/ndsj/2021/indexch.htm>.
- [12] W. Zhu, J. Wu, T. Fu, J. Wang, J. Zhang, and Q. Shangguan, “Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on lstm and mlp,” *Journal of intelligent and connected vehicles*, vol. 4, no. 2, pp. 80–91, 2021.
- [13] D. Ma, X. Song, and P. Li, “Daily traffic flow forecasting through a contextual convolutional recurrent neural network modeling inter- and intra-day traffic patterns,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2627–2636, 2021.
- [14] Y. Liu, C. Lyu, Y. Zhang, Z. Liu, W. Yu, and X. Qu, “Deeptsp: deep traffic state prediction model based on large-scale empirical data,” *Communications in Transportation Research*, vol. 1, Article ID 100012, 2021.
- [15] Ho-C. Kho and S. Kho, “Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data,” *Accident Analysis & Prevention*, vol. 88, no. 9–19, pp. 9–19, 2016.
- [16] J. Wang, W. Xie, B. Liu, S. Fang, and D. R. Ragland, “Identification of freeway secondary accidents with traffic shock wave detected by loop detectors,” *Safety Science*, vol. 87, pp. 195–201, 2016.
- [17] Y. W. Liyanage, D.-S. Zois, and C. Chelmiss, “Near Real-Time Freeway Accident Detection,” vol. 1, 2020 *IEEE Transactions on Intelligent Transportation Systems*.
- [18] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, “Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis,” *Accident Analysis & Prevention*, vol. 136, Article ID 105405, 2020.
- [19] Z. Zhang, Q. He, J. Gao, and M. Ni, “A deep learning approach for detecting traffic accidents from social media data,” *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 580–596, 2018.
- [20] E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, “Real-time detection of traffic from twitter stream analysis,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2269–2283, 2015.
- [21] Z. Zhang and Q. He, “On-site traffic accident detection with both social media and traffic data,” in *Proceedings of the 9th Triennial Symp. Transp. Anal.*, TRISTAN), China, June 19–26, 2022.
- [22] Y. Li, W. Liu, and Q. Huang, “Traffic anomaly detection based on image descriptor in videos,” *Multimedia Tools and Applications*, vol. 75, no. 5, pp. 2487–2505, 2016.
- [23] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, “Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions,” in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1 20–25 June 2009.
- [24] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, “Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 673–682, 2017.
- [25] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, “Combining motion and appearance cues for anomaly detection,” *Pattern Recognition*, vol. 51, pp. 443–452, 2016.
- [26] L. Kratz and N. Ko, “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*, pp. 1446–1453, IEEE, Miami, FL, USA, 1 20–25 June 2009.
- [27] T. Hospedales, S. Gong, and T. Xiang, “A Markov clustering topic model for mining behaviour in video,” in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 1165–1172, IEEE, Kyoto, Japan, 29 September 2009 - 02 October 2009.
- [28] C. Yang, J. Yuan, and Ji Liu, “Sparse reconstruction cost for abnormal event detection,” in *Proceedings of the CVPR 2011*, pp. 3449–3456, IEEE, Colorado Springs, CO, USA, 20–25 June 2011.
- [29] W. Li, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.
- [30] C. G. Blair and N. M. Robertson, “Event-driven dynamic platform selection for power-aware real-time anomaly detection in video,” in *Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 54–63, IEEE, Lisbon, Portugal, 05–08 January 2014.
- [31] M. Hasan, J. Choi, J. Neumann, K. Amit, and L. S. Davis, “Learning temporal regularity in video sequences,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–742, Las Vegas, June 2016.
- [32] W. Luo, L. Wen, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE international conference on computer vision*, p. 341, Venice, Italy, 22–29 October 2017.
- [33] D. Gong, L. Liu, V. Le et al., “Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, Seoul Korea, October 2019.
- [34] T.-N. Nguyen and J. Meunier, “Anomaly detection in video sequence with appearance-motion correspondence,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1273–1283, Seoul Korea, October 2019.
- [35] H. Park, J. Noh, and B. Ham, “Learning memory-guided normality for anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14372–14381, Seoul Korea, 2020.
- [36] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, “Learning normal dynamics in videos with meta prototype network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages, pp. 15425–15434, Seoul Korea, 2021.
- [37] L. Wen, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—a new baseline,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, Salt Lake City, June 2018.
- [38] S. Lee, H. G. Kim, and Y. M. Ro, “Bman: bidirectional multi-scale aggregation networks for abnormal event detection,” *IEEE Transactions on Image Processing*, vol. 29, pp. 2395–2408, 2020.

- [39] F. Dong, Yu Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," *IEEE Access*, vol. 8, pp. 88170–88176, 2020.
- [40] K. Yilmaz and Y. Yilmaz, "Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate," *Pattern Recognition*, vol. 114, Article ID 107865, 2021.
- [41] S. Waqas, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," pp. 6479–6488 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, June 2018.
- [42] K. Doshi and Y. Yilmaz, "Continual learning for anomaly detection in surveillance videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 254–255, New Orleans, Louisiana, June 19th – 23rd.
- [43] Z. Zhou, X. Dong, Z. Li, K. Yu, C. Ding, and Y. Yang, "Spatio-temporal feature encoding for traffic accident detection in vanet environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 110 pages, 2022.
- [44] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, Salt Lake City, 2018.
- [45] Yi Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," 2019, <https://arxiv.org/abs/1907.10211>.
- [46] H. Zenati, C. S. Foo, L. Bruno, G. Manek, and V. R. Chandrasekhar, "Efficient gan-based anomaly detection," 2018, <https://arxiv.org/abs/1802.06222>.
- [47] C. Vapnik and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [48] D. M. J. Duin and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.
- [49] K. T. Chui, R. W. Liu, M. Zhao, and P. O. De Pablos, "Predicting performance with school and family tutoring using generative adversarial network-based deep support vector machine," *IEEE Access*, vol. 8, pp. 86745–86752, 2020.
- [50] Y. Liu, Z. Li, C. Zhou et al., "Generative adversarial active learning for unsupervised outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, p. 1, 2019.
- [51] P. Cuong Ngo, A. Aristo Winarto, S. Park, F. Akram, and H. Kuan Lee, "Fence gan: towards better anomaly detection," in *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 141–148, IEEE, Portland, OR, USA, 04–06 November 2019.
- [52] M.-I. Georgescu, A. Barbalau, R. Tudor Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 12742, Salt Lake City, October 2021.
- [53] S. Lin, H. Yang, X. Tang, T. Shi, and L. Chen, "Social mil: interaction-aware for crowd anomaly detection," in *Proceedings of the 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS)*, 18–21 September 2019.
- [54] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 4505–4515, 2021.
- [55] J.-C. Feng, Fa-T. Hong, and W.-S. Zheng, "Mist: multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14009–14018, Seoul Korea, June 2021.
- [56] J. Wu, W. Zhang, G. Li et al., "Weakly-supervised spatio-temporal anomaly detection in surveillance video," 2021, <https://arxiv.org/abs/2108.03825>.
- [57] J. Donahue, P. Krahenbuhl, and Trevor Darrell, "Adversarial Feature Learning," 2016, <https://arxiv.org/abs/1605.09782>.
- [58] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Lecture Notes in Computer Science* vol. 146, Springer 157 pages, Springer, 2017.
- [59] T. Schlegl, P. Seeböck, S. M. Waldstein, and G. Langs, "Schmidt-Erfurth: f-AnoGAN: f," *Medical Image Analysis*, vol. 54, pp. 30–44, 2019.
- [60] J. Ryan Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, <https://arxiv.org/abs/1612.00390>.
- [61] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, Salt Lake City, June 2017.
- [62] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Salt Lake City, June 2019.
- [63] C. Feichtenhofer, "X3d: expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Soeul Korea, 2020.
- [64] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: a video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Soeul Korea, 2021.
- [65] 2022, <https://github.com/mx-markdsludslu/VideoTransformer-pytorch>.