Scene-Adaptive SVAD Based On Multi-modal Action-based Feature Extraction

Shibo Gao[1,2], Peipei Yang[2,3], and Linlin Huang[1]

[1] Beijing Jiaotong University
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences

Abstract. Due to the lack of anomalous data, most existing semi-supervised video anomaly detection (SVAD) methods rely on designing self-supervised tasks to reconstruct video frames for learning normal patterns from training data, thereby distinguishing anomalous events from normal ones according to the reconstruction quality. However, these methods heavily rely on the frequency of event occurring to judge its abnormality, which often misidentify rare normal events as anomalies. More importantly, they are usually trained to fit a particular scene leading to poor generalization to other scenes. Besides, for all existing methods, the normal/abnormal events are fixed once the training is finished, and cannot conduct test-time adjust without retraining the model. To resolve these problems, we propose a semi-supervised video anomaly detection method based on a multi-modal action-based feature extraction model. Our method exploits a vision-language model pre-trained with an action recognition task for action-based feature extraction, making it robust to scene variations irrelevant to anomalies. A clustering model with learnable prompts is employed for learning the normal patterns and anomaly detection, which does not rely on event frequency and can correctly identify rare normal events. Benefiting from the multi-modal model, our method can conveniently adjust the normal events during test time by text guidance without retraining. We conduct experiments on benchmark datasets and the results demonstrate that our method achieves the start-of-the-art performances. More importantly, our method exhibits obviously better performances in cross-scene experiment and test-time anomalies adjustment experiment.

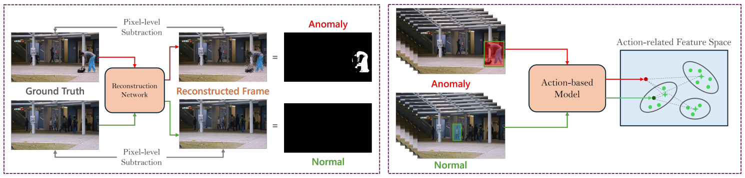
Keywords: Semi-supervised video anomaly detection · Multi-modal model

1 Introduction

Video anomaly detection (VAD) aims to detect anomalous events in video segments that deviate from expected patterns and to pinpoint the time of these anomalies. Due to its broad application prospects, such as in intelligent surveillance systems and video review, VAD has garnered increasing attention from both academia and industry [29, 30, 39, 7, 27, 38, 55, 3].

Semi-supervised video anomaly detection[4] aims to accurately detect anomalies and determine their occurrence in time without using any labeled anomalous data. This approach is more aligned with real-world scenes compared with weakly-supervised VAD (WVAD) [60, 57, 17, 53] or open-set VAD (OSVAD) [1, 6, 65, 64] tasks. Most of existing works resort to designing self-supervised tasks (such as prediction [23, 22], reconstruction [49, 45, 24], jigsaw [46], etc) to learn normal patterns, and exploit them to distinguish anomalous events from normal ones.

However, existing methods based on pixels or images have suffered from the following several drawbacks. First, since there lack explicit representations for normal or abnormal events, existing methods estimate the normality of an event according to only its frequency of occurrence without considering its semantic information, making some rare normal events misidentified as anomalies. Second, while an ideal VAD model is expected to detect abnormal events across different scenes, methods based on self-supervised tasks are easily influenced by appearance differences across scenes, leading to significant performance degradation when the scene changes. Third, in practical applications, we may need to adjust the determination of normal and abnormal events during test phase, which is impossible to achieve without retraining the model using re-annotated data for existing methods.



(a) Existing Pixel-Based Paradigm

(b) Our Method Based on Action-Related Features

Fig. 1. The existing paradigm of SVAD, as shown in (a), aims to design self-supervised tasks and compare them with ground truth to discover anomalies. (b) presents our new paradigm. We directly discriminate anomalies in the feature space by employing a multi-modal action-based feature extraction model.

In recent years, language-vision pre-trained models have achieved remarkable results in many downstream tasks due to their learned cross-modal prior knowledge and robust transfer learning capabilities [16, 62, 41, 43, 18, 19, 21, 34, 37, 40, 50, 51, 58, 63]. This advancement provides us an opportunity to leverage pre-trained models for explicitly constructing semantic representation of the events without labeled data. With these models, we can build a SVAD model to overcome the problems above.

To be specific, our proposed SVAD model exploits an image encoder of CLIP followed by a Temporal Fusion Component (TFC) module to extract action features from video segments. We pre-train the feature extraction module on action

[4] In previous research, many methods have also referred to it as unsupervised VAD.

recognition tasks to make the features insensitive to appearance variations. Then, a clustering model with learnable prompts is employed to learn the distributions of normal patterns from the training set.

As shown in Fig.1, our proposed method detects anomalies by identifying out-of-distribution video segments directly in the action-based feature space, which is significantly different from existing methods. Since the learning process does not rely on fitting normal data, rare normal patterns are also properly represented to avoid misidentification. This novel approach of extracting action features while disregarding appearance features significantly enhances the model's generalization capability across different scenes. Furthermore, benefiting from the vision-language model, we can conveniently achieve text-guided adjustments on the definition of normal or abnormal events without retraining the model.

We evaluate our method on several benchmark datasets including Ped2, Avenue, and Shanghai Tech datasets. The experimental results show that our method achieves the state-of-the-art performance. Particularly, our method obviously outperforms existing methods on cross-scene experiment, exhibiting satisfactory generalization to different scenes. We can also conveniently make text-guided adjustments to the normal or abnormal events during testing.

We summarize our contributions as follows:

- We propose a semi-supervised video anomaly detection approach based on multi-modal action-based feature extraction. To the best of our knowledge, we are the first to use multi-modal information to guide and discriminate anomalies based on action understanding in SVAD.
- By extracting action-based features from video segments and detecting anomalies in this feature space, our method obtains superior generalization across different scenes and better performance on learning rare normal patterns comparing with existing methods.
- Benefiting from the vision-language model, our method can adjust the definition of abnormal events by text guidance during test time without retraining the model.
- Our method achieves state-of-the-art performance on three mainstream datasets including Ped2, Avenue, and Shanghai Tech. More importantly, it obviously outperforms other methods in cross-scene and text-guided anomaly-adjustment experiments.

2 Related Work

2.1 Video Anomaly Detection

Currently, VAD can be roughly categorized into three setting according to the supervision type: semi-supervised anomaly detection(SVAD) [27, 7, 59, 2, 54], weakly supervised anomaly detection(WVAD) [60, 57, 17, 53], and open set anomaly detection(OSVAD) [1, 6, 65, 64]. SVAD aims at detecting abnormal events from videos with only normal samples available during training, which is a more practical technique since the anomalous patterns are quite diverse and difficult to obtain. It's notable that there was a naming confusion that some early works named unsupervised VAD were indeed SVAD, since they also used the normal videos for training.

The mainstream solution for semi-supervised video anomaly detection involves designing a self-supervised task (such as prediction [23, 22], reconstruction [49, 45, 24], jigsaw [46], or rotation) to learn establishing normal patterns, and identifying anomalies by comparing the established video segments with real ones. Some researchers[49, 23, 33, 4] use the reconstruction-based methods to reconstruct normal events and classify events with large reconstruction errors as anomalies. Other researchers[23, 45, 20] focus on predicting future frames using previous video frames and determine whether a frame is an anomaly by calculating the difference between the predicted frame and the actual frame.

Additionally, some researchers [44] combine reconstruction and prediction-based methods to improve detection performance. However, since the self-supervised tasks are trained to fit the training data including the particular scene, it is liable to regard normal patterns in other scenes as anomalies, making these methods poor to generalize to scene variations. Even slight camera rotations can cause the model to fail. Some other methods treat SVAD as a one-class classification task and utilize one-class frameworks [47, 52] for anomaly detection. However, these methods still suffer from interference caused by irrelevant appearance features, resulting in poor performance.

2.2 Action Recognition

The development of video action recognition can be mainly categorized into two types. The first type is based on traditional feature extraction networks, such as two-stream networks, 3D CNNs, and transformer-based networks. Methods based on two-stream networks [9] establish appearance and motion models with two separate networks and fuse them either in the intermediate or final stage. 3D CNNs [5] directly learn spatio-temporal features from RGB frames, adding an additional temporal dimension to the conventional 2D CNNs. Transformer-based networks [8, 36] adopt and modify the latest transformer architecture to jointly encode spatial and temporal features. However, most of these works are single-modal and do not consider the semantic information contained in spatial and temporal features. Recently, there have been some new methods [18, 28, 48] that attempt to introduce multi-modal models into video action recognition.

2.3 Vision-text Multi-modal Model

Frome et al.[10] proposed joint learning of image-text embeddings using category name annotations. Building upon these works, CLIP [41] attempted to explore the relationship between natural images and text, while ALIGN [16] and FILIP [56] further expanded the scale of training data. By utilizing simple noise contrastive learning, the network can learn powerful visual representations from image-text pairs. Based on this foundation, vision-text multi-modal models have been applied to various downstream domains [18, 19, 21, 34, 37] and achieved remarkable results. The rich semantic information brought by the multi-modal model not only improves performance but also empowers the model with strong zero-shot capabilities.

3 Method

3.1 Overview

The semi-supervised video anomaly detection task assumes that during the training phase, only videos containing normal events are available. The goal is to train a detection model that can predict frame-level anomaly probabilities in videos. Given a test video frame F from a video V, the label $y = 1$ if it is an anomalous frame and $y = 0$ otherwise.

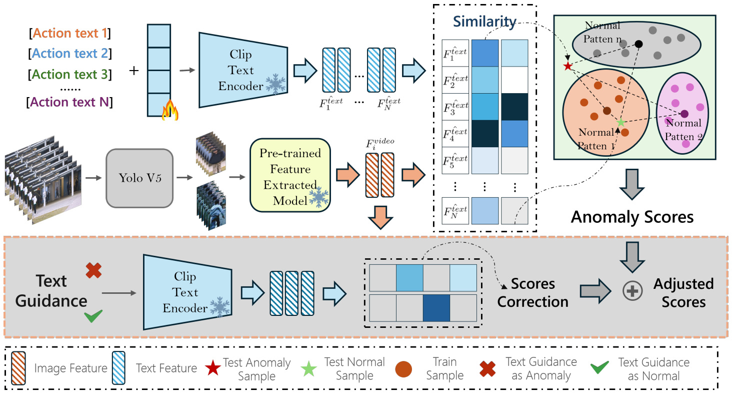


Fig. 2. The structure of our Model. Different from existing SVAD methods resorting to self-supervised tasks, our approach distinguishes between normal and abnormal events based on their response to action prompts. The process within the gray box represents an optional step that allows for the redefinition of normal and abnormal events during testing based on text guidance, without the need for retraining the model or re-annotating the dataset. The green check-mark indicates that the events conform to the current text guidance should be seen as normal, thereby reducing the anomaly scores for such events, and vice versa.

The overall model structure is illustrated in Fig.2. Firstly, we utilize a pre-processed, annotated action recognition dataset to align video action features with semantic information. Subsequently, we employ learnable prompts and a clustering model to obtain more robust descriptions of the normal patterns. Finally, we use the cosine distance to the nearest normal pattern as the anomaly score. Additionally, we can choose to leverage the zero-shot capability of the CLIP model to make text-guided anomalies adjustments during testing.

In Section 3.2, we will introduce the construction of the pre-training dataset and model. In Section 3.3, we will detail how to guide the pretrained model towards the VAD task. In Section 3.4, we will describe the details about inference. In Section 3.5, we will describe how to leverage the zero-shot capability of CLIP to achieve text-guidance anomalies adjustments.

3.2 Multi-modal Pre-training for Action Feature Extraction

In SVAD, most anomalies are closely related to human activities(cycling on sidewalks, chasing and playing, skateboarding and others) and not related on appearance. However, existing methods tend to overfit to specific scenes, where irrelevant features beyond the human behaviors become noise and significantly degrade model performance. Therefore, we focus on extracting behavior-related features to accurately detect anomalies in videos, free from the interference of appearance factors, thereby achieving greater generalization across different scenes.

Recently, outstanding works [18, 48, 28] have successfully introduced CLIP into the field of action recognition and achieved significant effects. Inspired by these works, we propose a vision-text multi-modal model to extract action-related features and pre-train it using action recognition datasets to address the issue of lacking anomalous data. In addition to accurately detecting anomalies, such a model can generalize across different scenes and flexibly adjust the anomalous events without retraining, paving the way for semi-supervised video anomaly detection.

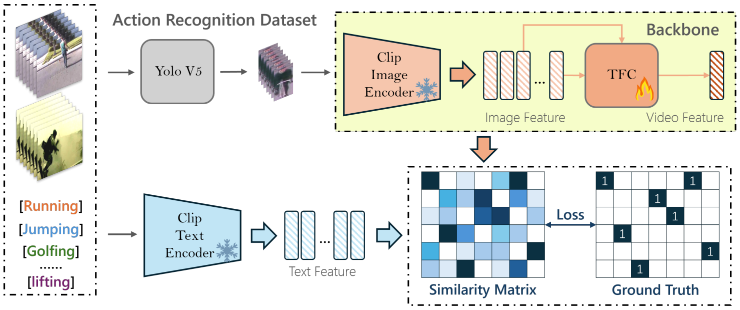


Fig. 3. During the pre-training stage, we leverage action recognition datasets and CLIP to establish a connection between video segments and their semantic meanings.

Due to the fact that SVAD applications often involve multiple objects in the same frame and require fine-grained detection for each one, while action recognition task mainly focuses on the behavior of the main object in the scene, directly using mainstream action recognition pre-trained models for extracting action-based features often yields unsatisfactory results. To overcome this challenge, we use an object detector to simultaneously process VAD datasets and action recognition datasets. Specifically, we utilize existing object detectors to process popular action recognition datasets(HMDB51, UCFSports, Weizmann, K400, and KTH), while excluding data containing viewpoint changes. This allows us to construct a pre-training dataset consisting of 30 labels and over 50000 samples that are relevant to common actions such as walking, running, golf swinging, diving, cycling, pull-ups, and others.

As shown in Fig.3, we first input the video segment $S_i = I_1, I_2, \ldots, I_n$ from the samples into the image encoder of CLIP to extract visual features $F_1, F_2, \ldots F_n$. Then we input the corresponding action annotation into the text encoder of CLIP to extract text feature F_i^{text}. Since CLIP is trained to connect the text with static images, overlooking the temporal information among successive frames that is crucial to activity recognition, we introduce a simple and lightweight temporal feature confusion(TFC) [18] module to integrate image features of multiple frames for extracting the feature of the video segment F_i^{video}. The TFC module is a transformer encoder following the image encoder for extracting temporal features from several independent image embeddings. Additionally, we use a adapter structure to fuse the original image features and the action features to preserve their zero-shot capability.

The feature of a video segment is extracted as:

$$F_k = \mathrm{CLIP}_{image}(I_k), k \in [1, 2, \ldots n] \tag{1}$$

$$F_i^{video} = \alpha \cdot \mathrm{TFC}(F_1, F_2, \ldots F_n) + \beta \cdot \mathrm{CLIP}_{image}(I_{middle}), \tag{2}$$

where i is the object index in the corresponding image, α and β are employed as "residual ratio" to preserve the original image features of CLIP. I_{middle} represents the middle image in S_i.

Then, the action-based features and their corresponding text features can be aligned using a simple NCE Loss:

$$L_{nce} = -\sum_i (\log \frac{\exp(< F_i^{video} \cdot F_{g(i)}^{text} > / \sigma)}{\sum_j \exp(< F_i^{video} \cdot F_j^{text} > / \sigma)}), \tag{3}$$

where σ is the temperature hyper-parameter, $F_{g(i)}^{text}$ is the ground truth of S_i in pre-train datasets. In this way, we effectively convert the powerful image adaptation capability of CLIP into videos.

3.3 Multi-modal Action-based Feature Extracted Model for SVAD

As mentioned in Section 3.2, the pre-trained model can establish the connection between semantic information and action segments, enabling us to extract action-based features. Based on this, an intuitive way is to directly apply it to the SVAD task for action classification on both the training and testing sets. In this way, action segments that do not appear in the training set are considered as anomalies.

However, this seemingly simple approach often fails in practice. There are two main reasons for this:

First, the pre-training dataset and the SVAD datasets have inconsistent distributions. Directly using the pre-trained model to detect anomalies will result in a significant amount of mis-classification.

Second, the majority of samples in the pre-training datasets are mutually exclusive. However, it is difficult to describe the samples in the SVAD dataset with a simple action category. For example, a person may be simultaneously waving and walking, but such samples may be lacking in the pre-training datasets.

Inspired by [35], we describe events in SVAD datasets based on their responses to a series of action prompts. Normal samples are expected to cluster around several response patterns, while abnormal samples should exhibit distinct response patterns. Specifically, we describe the samples by utilizing the response of many action prompts and obtain the centers of normal modes through clustering with cosine similarity. The cosine distance is then used as the anomaly score for each sample.

In the pre-training dataset, action text is often presented in the form of words or phrases, which may not comprehensively describe the complex and diverse action patterns in videos. To ensure robustness in action prompts and better adaptability to downstream tasks, we draw inspiration from [61] and introduce learnable prompts into the action text. It is worth noting that these action-related text have no specific restrictions. Therefore, any distinctive batch of action-related text can be used to describe the normal patterns. In this paper, we utilize all the annotations from the K400 action recognition dataset as action text set $T = [T_1, T_2, \ldots, T_N]$.

The process of obtaining the text features $\hat{F}_j^{text}$ after embedding the learnable prompts can be described as:

$$p_i = \underset{j=1,\ldots,N}{\arg\max} < F_i^{video} \cdot F_j^{text} >, \tag{4}$$

$$Prompt_j = \{C_1, C_2, \ldots, Tokenizer(T_{p_i}), \ldots, C_k\}, \tag{5}$$

$$\hat{F}_j^{text} = \mathrm{CLIP}_{text}(Prompt_j). \tag{6}$$

where $\{C_1, \ldots, C_k\}$ are learnable prompts, and N represents the number of action prompts. At this stage, we are still using the method mentioned in Section 3.2 for training.

For the same dataset, the normal samples in the test set have a similar distribution to the normal samples in the training set, while the distribution of abnormal samples in the test set differs from that of the normal samples. By introducing learnable prompts to enhance the sample's maximal response to action prompts, the model can continuously optimize the way it describes normal samples, thereby capturing the characteristics of normal samples more accurately.

3.4 SVAD Inference

After transferring the pre-trained model to the SVAD task and introducing learnable prompts, we extract features from the training set using this model. In this paper, we use the $Sim(x, y)$ function to calculate the similarity between x and y:

$$Sim(x, y) = \frac{< x \cdot y >}{||x|| \cdot ||y||}. \tag{7}$$

For a given sample, we utilize its response to the action prompts with learnable prompts to provide a detailed description:

$$Response_i[j] = Sim(F_i^{video}, \hat{F}_j^{text}), j \in [1, 2, \dots N]. \tag{8}$$

We then attempt to group the normal samples into several clusters and use these clusters to describe the entire set of normal samples as comprehensively as possible. Since the training set lacks annotations, a simple yet effective approach is to use clustering model to obtain several cluster centers $Center_m, m \in [1, 2, \dots M]$ representing normal modes. It is notable that the same intensity of responses at different positions in the features represents different actions. Therefore, using existing Euclidean distance is not suitable. In this paper, we use K-means clustering based on cosine similarity to obtain the centers of normal modes. The number of cluster centers is determined by the size of the dataset, and the specific settings can be found in Section 4.2.

We calculate the cosine distance between a test sample and the nearest cluster center as the sample's anomaly score. Normal samples should have higher cosine similarity and shorter cosine distance to one of the centers, while abnormal samples would exhibit lower cosine similarity and longer cosine distance to all cluster centers. The process of assigning an anomaly score to a sample can be described as follows:

$$Score_i = 1 - \max_m [Sim(Response_i, Center_m)], m \in [1, 2, \dots M], \tag{9}$$

where M represents the number of cluster centers obtained from the normal modes in the training set.

3.5 Test-time anomaly redefine

Existing SVAD methods define events that did not appear in the training set as anomalies. However, in the real world, the definition of anomalies is often subjective and situational. For example, running rapidly inside a shopping mall or riding a bicycle on a sidewalk is usually considered abnormal behavior, whereas running in the playground or cycling on a highway is considered normal. For existing SVAD methods, each anomalies adjustment requires adjusting the dataset and retraining the model.

In this paper, we aim to leverage the rich semantic information and zero-shot capabilities of CLIP to adjustment normal and abnormal events during testing based on the text guidance, without incurring any additional computational cost. Specifically, we start by providing a set of text as text guidance T_g, which can be single words, phrases, or sentences, and explicitly specify which ones represent normal and which ones represent anomalies.

Then, text guidance T_g are passed through the text encoder, and their similarity with visual features is computed by $Sim(F_i^{video}, \text{CLIP}_{text}(T_g))$.If the similarity value between a sample and the text guidance exceeds a predefined threshold, we consider it as a match with the text guidance and adjust its score accordingly:

$$S_{New} = \begin{cases} S + S_{min} \cdot \theta \cdot (Similarity - threshold) & (if\ T_g = Anomaly) \\ S - S_{max} \cdot \gamma \cdot (Similarity - threshold) & (if\ T_g = Normal) \end{cases} \tag{10}$$

In this equation, S represents the anomaly score of the current frame. S_{max} and S_{min} denote the maximum and minimum anomaly scores in the current video, respectively. $Similarity$ refers to the similarity between the sample and the text guidance. θ and γ represent the constant terms.

4 Experiment

4.1 SVAD Datasets

- ***UCSD Ped2*** The UCSD dataset [32] contains 16 training videos and 12 test videos of pedestrians walking on sidewalk. The abnormal events are about cyclists, carts, cars, or people walking across the surrounding grass.
- ***CUHK Avenue*** The CUHK dataset [26] contains 16 training videos and 21 test videos about sidewalk. The anomalies include anomalies of pedestrians, wrong direction of movement, appearance of anomalous objects, etc.
- ***Shanghai Tech*** The Shanghai Tech dataset [30] is a challenging anomaly detection dataset. It contains 330 training videos and 107 testing ones with 130 abnormal events. Totally, it consists of 13 scenes and various anomaly types.

4.2 Implementation Details

During pre-training of the action feature extraction module, video clips corresponding to each object are resized to $224 \times 224 \times 9$ and fed into the model, while both the image encoder and text encoder utilize the pre-trained CLIP (ViT-B/16) and are kept frozen. The only trainable component is the TFC module, which fuses the image features of consecutive frames for temporal feature extraction. The model is optimized using AdamW, where we set the learning rate to 0.0007, the batch-size 512, and the temperature hyper-parameter σ to 0.1.

For training the SVAD model by learning normal patterns, we set the number of cluster centers to 10 for the Ped2 dataset and Avenue dataset while 20 for the Shanghai Tech dataset. During this process, image encoder, text encoder, and

TFC module are all kept frozen, while only the learnable prompts are trained. We also use AdamW for optimization and set the learning rate to 0.0002, the batch size to 128, and the residual ratio α to 0.7.

We evaluate the model based on recent research, considering both micro and macro AUC metrics. For micro AUC, videos are concatenated before AUC calculation. For macro AUC, AUC is computed per video, then averaged, yielding a single value.

4.3 Comparison with State-of-the-art Methods on Single Dataset

Comparison of our method with various representative methods is shown in Table 1. In the evaluation of metric Mic AUC, our method achieves sort-of-the-art performance across three mainstream datasets: Shanghai Tech, Avenue, and Ped2. For metric Mac AUC, our result on the Ped2 dataset is consistent with state-of-the-art method, while on the more challenging Shanghai Tech and Avenue datasets, our performance is the second-best. Undoubtedly, our approach demonstrates the best overall performance in single-dataset evaluations.

Table 1. Micro and macro AUC scores of several state-of-the-art methods on the single dataset. We mark the first, second, and third places in the results with red, orange, and blue respectively.

Method	Ped2		Avenue		Shanghai Tech	
	Mic AUC	Mac AUC	Mic AUC	Mac AUC	Mic AUC	Mac AUC
Ristea *et al.*[42]	-	-	91.6%	92.5%	83.8%	90.5%
Georgescu *et al.*[11]	97.5%	99.8%	91.5%	92.8%	82.4%	90.2%
BA Framework[12]	98.7%	99.7%	92.3%	90.4%	82.7%	89.3%
Hirschorn *et al.*[14]	-	-	-	-	85.9%	-
OCAE[15]	94.3%	97.8%	87.4%	90.4%	78.7%	84.9%
Liu *et al.*[24]	99.3%	-	89.9%	93.5%	74.2%	83.2%
Zheng *et al.*[25]	-	-	91.8%	92.3%	83.8%	87.8%
Madan *et al.*[31]	-	-	93.2%	91.8%	83.3%	89.3%
Wang *et al.*[46]	99.0%	-	92.2%	-	84.3%	-
Park *et al.*[39]	97.0%	-	82.8%	86.8%	68.3%	79.7%
Ours	99.3%	99.8%	93.6%	93.1%	86.1%	90.3%

* For more details, please refer to the supplement.

Fig.4 illustrates the visualization of our method's scores on two videos, namely 01_0051 and 01_0071. By focusing on action-related feature instead of pixels, our method achieves precise detection with high anomaly scores right at the onset of anomalies, rather than exhibiting a slow rise as observed in existing self-supervised tasks, which strongly demonstrates the reliability of our approach.

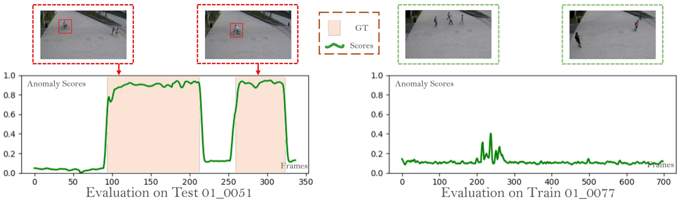


Fig. 4. Visualizations of abnormal and normal videos from the Shanghai Tech dataset by our model. Our model can accurately identify anomalies with high anomaly scores.

4.4 Experiment on Cross-Scene

As mentioned previously, a significant advantage of our method lies in its ability to generalize to different scenes without retraining. By leveraging action recognition pre-training datasets and multi-modal models, our method can disregard visual differences between samples and focus on extracting action-related features, thereby achieving strong generalization across different scenes.

For the three mainstream datasets, we design four distinct experimental schemes to explore the generalization capability of our proposed method across different scenes: Shanghai Tech→Avenue, Avenue→Shanghai Tech, Avenue→Ped2, and Shanghai Tech→Ped2. The results are shown in Table 2.

Table 2. Experimental results on generalization to different scenes. The metric in this figure is Micro AUC.

Method	Avenue→SHT	SHT→Avenue	SHT→Ped2	Avenue→Ped2
ZS CLIP[41]	60.9%	62.3%	52.7%	51.9%
ZS CLIP IB[13]	61.3%	64.5%	53.6%	52.8%
Astrid *et al.*[2]	51.7%	54.3%	65.9%	62.7%
Wang *et al.*[46]	59.3%	62.9%	75.6%	73.1%
Ours	78.7%	86.2%	97.8%	95.2%

It is evident that when evaluating between different scenes, both CLIP and CLIP IMAGE BIND, as well as existing SVAD methods, fail to maintain their performance, whereas our method continues to achieve high detection accuracy. This clearly demonstrates the strong generalization capability of our method across different scenes.

4.5 Experiment on Test-time redefine

Compared to existing SVAD methods, another significant advantage of our approach is its ability to adjust normal and abnormal events during testing based on text guidance, without re-annotation the dataset and retraining the model.

To illustrate this point specifically, we conduct additional evaluations on the challenging Shanghai Tech dataset.

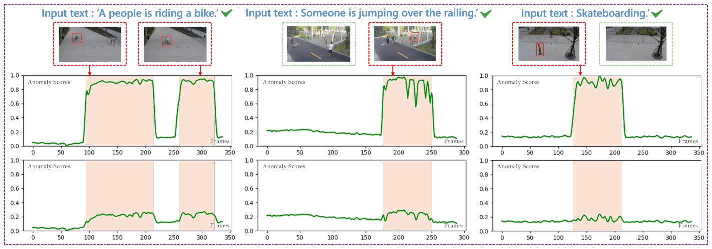


Fig. 5. Visualizations of the experiment of test-time redefinition. We specify three types of anomalies present in the dataset and asked the model to classify them as normal. It can be observed that the anomaly scores for the abnormal frames are significantly reduced to levels close to those normal ones.

As described in Section 3.5, we use the phrases "someone is jumping over the railing." "a people is riding a bike.," and "fast running." as text guidance, and test them on the test videos with IDs 01_0051, 03_0061, and 04_0050, respectively. Fig.5 shows the visualization of the scores before and after adjustment. It can be observed that when provided with prior text guidance specifying the event type as normal, our model leverages CLIP's zero-shot capability to adjust the scores, thereby no longer recognizing events not included in the training set as anomalies. This ability to adjust the classification of normal and anomalous events during the testing phase without additional computational cost further enhances the applicability of our method in real-world scenes.

4.6 Ablation Study

To further elucidate our method, we conduct a study on the selection of backbones. We aim to demonstrate that any CLIP-based action recognition model can be applied within our method and achieve outstanding results. Table 3 presents the performance of our method when three different state-of-the-art multi-modal action recognition models are used as backbones. It is evident that our method maintains a high level of performance across different backbones.

Table 3. Experimental results on different backbones.

Backbone	Action CLIP[48]		CLIP4CLIP[28]		Prompt VLM[18]	
Pre-train	×	68.3%	×	69.1%	×	70.9%
	✓	85.2%	✓	84.3%	✓	86.1%

Notably, failing to pre-train the model on datasets preprocessed can lead to catastrophic performance degradation. The specific reasons have already been analyzed in Section 3.2 and Section 3.3. Additionally, we provide a further detailed analysis and a visual sample in Supplement.

Table 4. Experimental results on the number of input images. The metric in this Table is Micro AUC.

Input Number	5	7	9	11
Avenue results	84.7%	91.5%	93.6%	82.7%
Shanghai Tech results	78.1%	85.7%	86.1%	79.2%

Furthermore, we investigate the influence of the number of images included in the input video sequence on model performance, with experimental results shown in Table 4. It can be observed that there is no significant difference in model performance when the number of inputs is 7 or 9. This is primarily due to the presence of a large amount of video material with varying frame rates in the training dataset.

5 Conclusion

This paper proposes a multi-modal action-based feature extraction model for semi-supervised video anomaly detection task. Without explicit semantic representations for video segments, existing SVAD methods struggle to correctly learn rare normal patterns and demonstrate poor generalization to different scenes. Considering that the majority of anomalies are closely related to human behaviors, we integrate the TFC module with CLIP image encoder for extract the action-based features from video segments, which is pre-trained using action recognition task. We then guide this model to SVAD training data for learning normal patterns, where the feature representations are generated based on their responses to different action prompts. The anomaly score of a video segment is obtained according to the similarity between the features. Experimental results on mainstream public datasets demonstrate outstanding performance of our proposed model in SVAD task. Moreover, our method benefits from superior generalization to different scenes, and can conveniently adjust the anomaly events by text guidance during test phase without retraining. Admittedly, our method has some limitations. Firstly, our method requires pre-training on preprocessed action recognition datasets to achieve best performance. Secondly, the clustering model we use is relatively simple.

Acknowledgments. This work has been supported by "Scientific and Technological Innovation 2030" Program of China Ministry of Science and Technology (2021ZD0113803) and the National Natural Science Foundation of China (NSFC) grant 62276258.

References

1. Acsintoae, A., Florescu, A., Georgescu, M.I., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: New benchmark for supervised open-set video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20143–20153 (2022)
2. Astrid, M., Zaheer, M.Z., Lee, S.I.: Synthetic temporal anomaly guided end-to-end video anomaly detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 207–214 (2021)
3. Bergamin, L., Carraro, T., Polato, M., Aiolli, F.: Novel applications for vae-based anomaly detection systems. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2022)
4. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: CVPR 2011. pp. 3449–3456. IEEE (2011)
5. Diba, A., Fayyaz, M., Sharma, V., Arzani, M.M., Yousefzadeh, R., Gall, J., Van Gool, L.: Spatio-temporal channel correlation networks for action classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 284–299 (2018)
6. Ding, C., Pang, G., Shen, C.: Catching both gray and black swans: Open-set supervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7388–7398 (2022)
7. Dong, F., Zhang, Y., Nie, X.: Dual discriminator generative adversarial network for video anomaly detection. IEEE Access **8**, 88170–88176 (2020)
8. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6824–6835 (2021)
9. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. Advances in neural information processing systems **26** (2013)
11. Georgescu, M.I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12742–12752 (2021)
12. Georgescu, M.I., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: A background-agnostic framework with adversarial training for abnormal event detection in video. IEEE transactions on pattern analysis and machine intelligence **44**(9), 4505–4523 (2021)
13. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
14. Hirschorn, O., Avidan, S.: Normalizing flows for human pose anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13545–13554 (2023)
15. Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L.: Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7842–7851 (2019)

16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916 (2021)
17. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3230–3234 (2023)
18. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: European Conference on Computer Vision. pp. 105–124 (2022)
19. Ju, C., Li, Z., Zhao, P., Zhang, Y., Zhang, X., Tian, Q., Wang, Y., Xie, W.: Multi-modal prompting for low-shot temporal action localization. arXiv preprint arXiv:2303.11732 (2023)
20. Kanu-Asiegbu, A.M., Vasudevan, R., Du, X.: Bipoco: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection. arXiv preprint arXiv:2207.02281 (2022)
21. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11144–11154 (2023)
22. Lee, S., Kim, H.G., Ro, Y.M.: Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. Transactions on Image Processing **29**, 2395–2408 (2019)
23. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6536–6545 (2018)
24. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13588–13597 (2021)
25. Liu, Z., Wu, X.M., Zheng, D., Lin, K.Y., Zheng, W.S.: Generating anomalies for video anomaly detection with prompt-based feature mapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24500–24510 (2023)
26. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2720–2727 (2013)
27. Lu, Y., Yu, F., Reddy, M.K.K., Wang, Y.: Few-shot scene-adaptive anomaly detection. In: European Conference on Computer Vision. pp. 125–141 (2020)
28. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
29. Luo, W., Wen, L., Gao, S.: Remembering history with convolutional lstm for anomaly detection. In: 2017 IEEE International Conference on Multimedia and Expo (ICME) (2017)
30. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE international conference on computer vision. pp. 341–349 (2017)
31. Madan, N., Ristea, N.C., Ionescu, R.T., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M.: Self-supervised masked convolutional transformer block for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

32. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1975–1981 (2010)
33. Munawar, A., Vinayavekhin, P., De Magistris, G.: Limiting the reconstruction capability of generative neural network using negative learning. In: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2017)
34. Nag, S., Zhu, X., Song, Y.Z., Xiang, T.: Zero-shot temporal action detection via vision-language prompting. In: European Conference on Computer Vision. pp. 681–697 (2022)
35. Nam, G., Heo, B., Lee, J.: Lipsum-ft: Robust fine-tuning of zero-shot models using random text guidance. arXiv preprint arXiv:2404.00860 (2024)
36. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3163–3172 (2021)
37. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision. pp. 1–18 (2022)
38. Park, C., Cho, M., Lee, M., Lee, S.: Fastano: Fast anomaly detection via spatio-temporal patch transformation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2249–2259 (2022)
39. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14372–14381 (2020)
40. Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., et al.: Freeseg: Unified, universal and open-vocabulary image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19446–19455 (2023)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
42. Ristea, N.C., Croitoru, F.A., Ionescu, R.T., Popescu, M., Khan, F.S., Shah, M.: Self-distilled masked auto-encoders are efficient video anomaly detectors. arXiv preprint arXiv:2306.12041 (2023)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Shi, C., Sun, C., Wu, Y., Jia, Y.: Video anomaly detection via sequentially learning multiple pretext tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10330–10340 (2023)
45. Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., Yang, J.: Integrating prediction and reconstruction for anomaly detection. Pattern Recognition Letters **129**, 123–130 (2020)
46. Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., Huang, D.: Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In: European Conference on Computer Vision. pp. 494–511 (2022)
47. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8201–8211 (2019)

48. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)
49. Wang, Y., Qin, C., Bai, Y., Xu, Y., Ma, X., Fu, Y.: Making reconstruction-based method great again for video anomaly detection. In: 2022 IEEE International Conference on Data Mining (ICDM). pp. 1215–1220 (2022)
50. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y.G.: Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In: International Conference on Machine Learning. pp. 36978–36989 (2023)
51. Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., et al.: Towards open vocabulary learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
52. Wu, P., Liu, J., Shen, F.: A deep one-class neural network for anomalous event detection in complex scenes. IEEE transactions on neural networks and learning systems **31**(7), 2609–2622 (2019)
53. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. arXiv preprint arXiv:2308.11681 (2023)
54. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE computer society conference on computer vision and pattern recognition. pp. 2054–2060. IEEE (2010)
55. Xu, Z., Zeng, X., Ji, G., Sheng, B.: Improved anomaly detection in surveillance videos with multiple probabilistic models inference. Intelligent Automation & Soft Computing **31**, 1703–1717 (2022)
56. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
57. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14744–14754 (2022)
58. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
59. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 1933–1941 (2017)
60. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. arXiv preprint arXiv:2302.05160 (2023)
61. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
62. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv preprint arXiv:2310.18961 (2023)
63. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023)
64. Zhu, J., Ding, C., Tian, Y., Pang, G.: Anomaly heterogeneity learning for open-set supervised anomaly detection. arXiv preprint arXiv:2310.12790 (2023)
65. Zhu, Y., Bao, W., Yu, Q.: Towards open set video anomaly detection. In: European Conference on Computer Vision. pp. 395–412 (2022)