MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection

Jia-Chang Feng[1,3,4], Fa-Ting Hong[1,3], and Wei-Shi Zheng[1,2,3*]
[1] School of Computer Science and Engineering, Sun Yat-Sen University
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4] Pazhou Lab, Guangzhou, China
fengjch8@mail2.sysu.edu.cn, hongft3@mail2.sysu.edu.cn, wszheng@ieee.org

Abstract

Weakly supervised video anomaly detection (WS-VAD) is to distinguish anomalies from normal events based on discriminative representations. Most existing works are limited in insufficient video representations. In this work, we develop a multiple instance self-training framework (MIST) to efficiently refine task-specific discriminative representations with only video-level annotations. In particular, MIST is composed of 1) a multiple instance pseudo label generator, which adapts a sparse continuous sampling strategy to produce more reliable clip-level pseudo labels, and 2) a self-guided attention boosted feature encoder that aims to automatically focus on anomalous regions in frames while extracting task-specific representations. Moreover, we adopt a self-training scheme to optimize both components and finally obtain a task-specific feature encoder. Extensive experiments on two public datasets demonstrate the efficacy of our method, and our method performs comparably to or even better than existing supervised and weakly supervised methods, specifically obtaining a frame-level AUC 94.83% on ShanghaiTech.

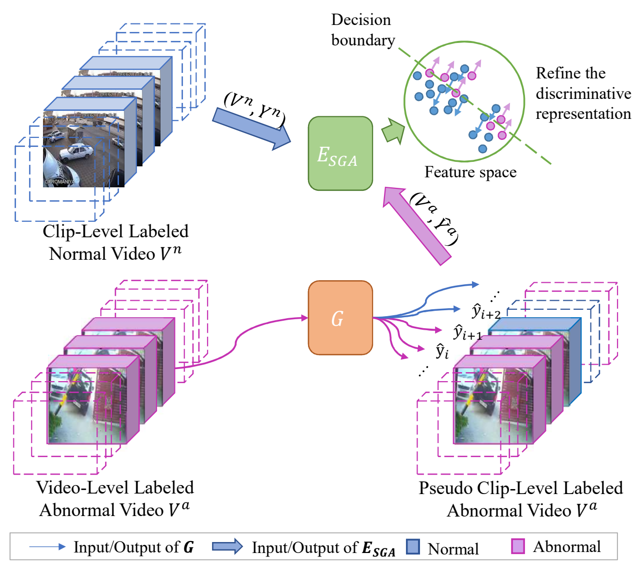

Figure 1: Our proposed MIST first assign clip-level pseudo labels $\hat{Y}^a = \{\hat{y}_i^a\}$ to anomaly videos with the help of a pseudo label generator G. Then, MIST leverages information from all videos to refine a self-guided attention boosted feature encoder E_{SGA}.

1. Introduction

Video anomaly detection (VAD) aims to temporally or spatially localize anomalous events in videos [33]. As increasingly more surveillance cameras are deployed, VAD is playing an increasingly important role in intelligent surveillance systems to reduce the manual work of live monitoring.

Although VAD has been researched for years, developing a model to detect anomalies in videos remains challenging, as it requires the model to understand the inherent differences between normal and abnormal events, especially anomalous events that are rare and vary substantially. Previous works treat VAD as an *unsupervised learning* task [29, 14, 7, 15, 13, 5, 32] , which encodes the usual pattern with only normal training samples, and then detects the distinctive encoded patterns as anomalies. Here, we aim to address the *weakly supervised video anomaly detection* (WS-VAD) problem [20, 31, 28, 34, 24] because obtaining video-level labels is more realistic and can produce more reliable results than unsupervised methods. More specifically, existing methods in WS-VAD can be categorized into two classes, *i.e. encoder-agnostic* and *encoder-based* methods.

The *encoder-agnostic* methods [20, 28, 24] utilize task-agnostic features of videos extracted from a vanilla feature encoder denoted as E (*e.g.* C3D [21] or I3D [2]) to estimate anomaly scores. The *encoder-based* methods [34, 31]

*Corresponding author

train both the feature encoder and classifier simultaneously. The state-of-the-art encoder-based method is Zhong *et al.* [31], which formulates WS-VAD as a label noise learning problem and learns from the noisy labels filtered by a label noise cleaner network. However, label noise results from assigning video-level labels to each clip. Even though the cleaner network corrects some of the noisy labels in the time-consuming iterative optimization, the refinement of representations progresses slowly as these models are mistaught by seriously noisy pseudo labels at the beginning.

We find that the existing methods have not considered training a task-specific feature encoder efficiently, which offers discriminative representations for events under surveillance cameras. To overcome this problem for WS-VAD, we develop a two-stage self-training procedure (Figure 1) that aims to train a task-specific feature encoder with only video-level weak labels. In particular, we propose a Multiple Instance Self-Training framework (MIST) that consists of a multiple instance pseudo label generator and a self-guided attention boosted feature encoder E_{SGA}. 1) *MIL-pseudo label generator.* The MIL framework is well verified in weakly supervised learning. MIL-based methods can generate pseudo labels more accurately than those simply assigning video-level labels to each clip [31]. Moreover, we adopt a sparse continuous sampling strategy that can force the network to pay more attention to context around the most anomalous part. 2) *Self-guided attention boosted feature encoder.* Anomalous events in surveillance videos may occur in any place and with any size [11], while in commonly used action recognition videos, the action usually appears with large motion [3, 4]. Therefore, we utilize the proposed self-guided attention module in our proposed feature encoder to emphasize the anomalous regions without any external annotation [11] but clip-level annotations of normal videos and clip-level pseudo labels of anomalous videos. For our WS-VAD modelling, we introduce a deep MIL ranking loss to effectively train the multiple instance pseudo label generator. In particular, for deep MIL ranking loss, we adopt a sparse-continuous sampling strategy to focus more on the context around the anomalous instance.

To obtain a task-specific feature encoder with smaller domain-gap, we introduce an efficient two-stage self-training scheme to optimize the proposed framework. We use the features extracted from the original feature encoder to produce its corresponding clip-level pseudo labels for anomalous videos by the generator G. Then, we adopt these pseudo labels and their corresponding abnormal videos as well as normal videos to refine our improved feature encoder E_{SGA} (as demonstrated in Figure 1). Therefore, we can acquire a task-specific feature encoder that provides discriminative representations for surveillance videos.

The extensive experiments based on two different feature encoders, *i.e.* C3D [21] and I3D [2] show that our framework MIST is able to produce a task-specific feature encoder. We also compare the proposed framework with other encoder-agnostic methods on two large datasets *i.e.* , UCF-Crime [20] and ShanghaiTech[15]. In addition, we run ablation studies to evaluate our proposed sparse continuous sampling strategy and self-guided attention module. We also illustrate some visualized results to provide a more intuitive understanding of our approach. Our experiments demonstrate the effectiveness and efficiency of MIST.

2. Related Works

Weakly supervised video anomaly detection. VAD aims to detect anomaly events in a given video and has been researched for years[9, 29, 14, 7, 15, 13, 12, 32, 31, 5, 24]. Unsupervised learning methods [9, 29, 7, 30, 15, 13, 32, 5] encode the usual pattern with only normal training samples and then detect the distinctive encoded patterns as anomalies. Weakly supervised learning methods [20, 31, 28, 34, 24] with video-level labels are more applicable to distinguish abnormal events and normal events. Existing weakly supervised VAD methods can be categorized into two classes, *i.e.* , *encoder-agnostic* methods and *encoder-based* methods. 1) *Encoder-agnostic* methods train only the classifier. Sultani *et al.* [20] proposed a deep MIL ranking framework to detect anomalies; Zhang *et al.* [28] further introduced inner-bag score gap regularization; Wan *et al.* [24] introduced dynamic MIL loss and center-guided regularization. 2) *Encoder-based* methods train both a feature encoder and a classifier. Zhu *et al.* [34] proposed an attention based MIL model combined with a optical flow based autoencoder to encode motion-aware features. Zhong *et al.* [31] took weakly supervised VAD as a label noise learning task and proposed GCNs to filter label noise for iterative model training, but the iterative optimization was inefficient and progressed slowly. Some works focus on detecting anomalies in an offline manner [23, 25] or a coarse-grained manner [20, 28, 34, 23, 25], which do not meet the real-time monitoring requirements for real-world applications.

Here, our work is also an encoder-based method and work in an online fine-grained manner, but we use the learned pseudo labels to optimize our feature encoder E_{SGA} rather than using video-level labels as pseudo labels directly. Moreover, we design a two-stage self-training scheme to efficiently optimize our feature encoder and pseudo label generator instead of iterative optimization[31].

Multiple Instance Learning. MIL is a popular method for weakly supervised learning. In video-related tasks, MIL takes a video as a bag and clips in the video as instances [20, 17, 8]. With a specific feature/score aggregation function, video-level labels can be used to indirectly supervise instance-level learning. The aggregation functions vary, *e.g.* max pooling[20, 28, 34] and attention pooling[17, 8]. In

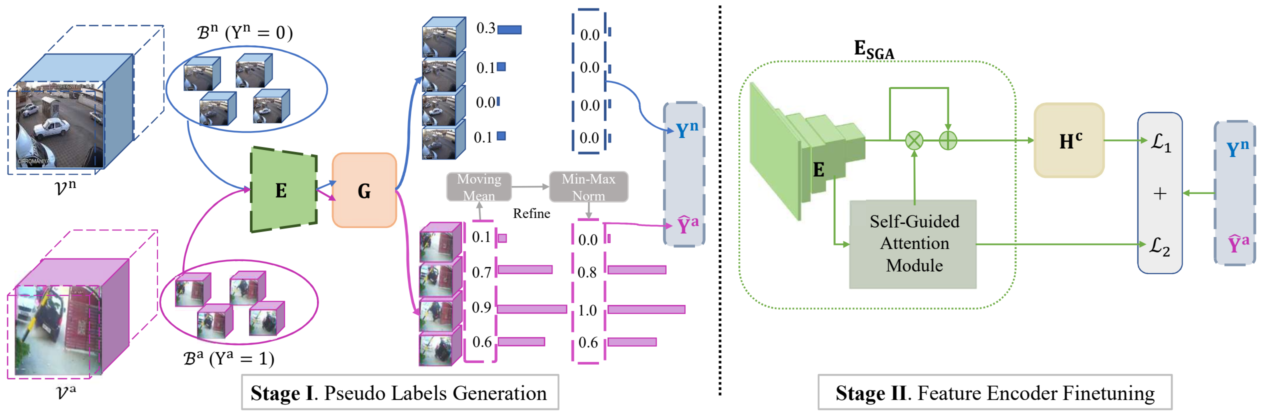

Figure 2: Illustration of our proposed MIST framework. MIST includes a multiple instance pseudo label generator G and self-guided attention boosted feature encoder E_{SGA} followed by a weighted-classification head H_c. We first train a G and then generate pseudo labels for E_{SGA} fine-tuning.

this paper, we adopt a sparse continuous sampling strategy in our multiple instance pseudo label generator to force the network to pay more attention to context around the most anomalous part.

Self-training. Self-training has been widely investigated in semi-supervised learning [1, 10, 6, 27, 22, 35]. Self-training methods increase labeled data via pseudo label generation on unlabeled data to leverage the information on both labeled and unlabeled data. Recent deep self-training involves representation learning of the feature encoder and classifier refinement, mostly adopted in semi-supervised learning [10] and domain adaptation [36, 35]. In unsupervised VAD, Pang *et al.* [18] introduced a self-training framework deployed on the testing video directly, assuming the existence of an anomaly in the given video.

Here, we propose a multiple instance self-training framework that assigns clip-level pseudo labels to all clips in abnormal videos via a multiple instance pseudo label generator. Then, we leverage information from all videos to fine-tune a self-guided attention boosted feature encoder.

3. Approach

VAD depends on discriminative representations that clearly represent the events in a scene, while action recognition datasets pretrained feature encoders are not perfect for surveillance videos because of the existence of a domain gap [11, 3, 4]. To address this problem, we introduce a self-training strategy to refine the proposed improved feature encoder E_{SGA}. An illustration of our method shown in Figure 2 is detailed in the following.

3.1. Overview

Given a video $V = \{v_i\}_{i=1}^N$ with N clips, the annotated video-level label $Y \in \{1, 0\}$ indicates whether an anomalous event exists in this video. We take a video V as a bag and clips v_i in the video as instances. Specifically, a negative bag (*i.e.* $Y = 0$) marked as $B^n = \{v_i^n\}_{i=1}^N$ has no anomalous instance, while a positive bag (*i.e.* $Y = 1$) denoted as $B^a = \{v_i^a\}_{i=1}^N$ has at least one.

In this work, given a pair of bags (*i.e.* a positive bag B^a and a negative bag B^n), we first pre-extract the features (*i.e.* $\{\mathbf{f}_i^a\}_{i=1}^N$ and $\{\mathbf{f}_i^n\}_{i=1}^N$ for B^a and B^n, respectively) for each clip in the video $V = \{v_i\}_{i=1}^N$ using a pretrained vanilla feature encoder, C3D or I3D, forming bags of features $\overline{B}^a$ and $\overline{B}^n$. We then feed the pseudo label generator the extracted features to estimate the anomaly scores of the clips (*i.e.* $\{s_i^a\}_{i=1}^N$, $\{s_i^n\}_{i=1}^N$). Then, we produce pseudo labels $\hat{Y}^a = \{\hat{y}_i^a\}_{i=1}^N$ for anomalous video by performing smoothing and normalization on estimated scores to supervise the learning of the proposed self-guided attention boosted feature encoder, forming as two-stage self-training scheme [10, 36, 35].

Algorithm 1 Multiple instance self-training framework

Input: Clip-level labeled normal videos $V^n = \{v_i^n\}_{i=1}^N$ and corresponding clip-level labels Y^n, video-level labeled abnormal videos $V^a = \{v_i^a\}_{i=1}^N$, pretrained vanilla feature encoder E.

Output: Self-guided attention boosted feature encoder E_{SGA}, multiple instance pseudo label generator G, clip-level pseudo labels $\hat{Y}^a$ for V^a

Stage I. Pseudo Labels Generation.

Extract features of V^a and V^n from E as $\{\mathrm{f}_i^a\}_{i=1}^N$ and $\{\mathrm{f}_i^n\}_{i=1}^N$.

Training G with $\{\mathrm{f}_i^a\}_{i=1}^N$ and $\{\mathrm{f}_i^n\}_{i=1}^N$ and their corresponding video-level labels according to Eq. 7.

Predict clip-level pseudo labels for each clip of V^a via trained G as $\hat{Y}^a$.

Stage II. Feature Encoder Fine-tuning.

Combine E with self-guided attention module as E_{SGA}, then fine-tune E_{SGA} with supervision of $Y^n \cup \hat{Y}^a$.

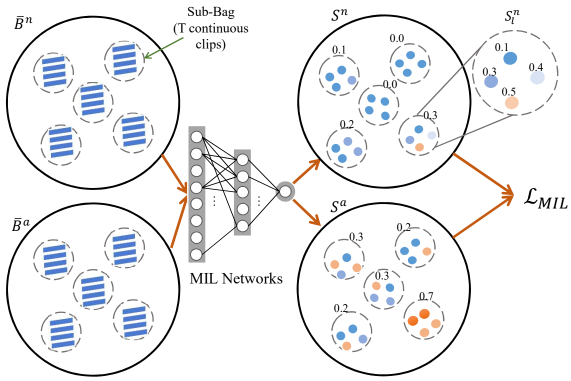

Figure 3: The workflow of our multiple instance pseudo label generator. Each bag contains L sub-bags, and each sub-bag is composed of T continuous clips.

As shown in Figure 2, our proposed feature encoder E_{SGA}, adapted from vanilla feature encoder E (*e.g.* , I3D or C3D) by adding our proposed self-guided attention module, can be optimized with the estimated pseudo labels to eliminate the domain gap and produce task-specific representations. Actually, our proposed approach can be viewed as a two-stage method (see Algorithm 1): 1) we first generate clip-level pseudo labels for anomalous videos that have only video-level labels via the pseudo label generator, while the parameters of the pseudo label generator are updated by means of the deep MIL ranking loss. 2) After obtaining the clip-level pseudo labels of anomalous videos, our feature encoder E_{SGA} can be trained on both normal and anomalous video data. Thus, we form a self-training scheme to optimize both the feature encoder E_{SGA} and pseudo label generator G. The illustration shown in Figure 2 provides an overview of our proposed method.

To better distinguish anomalous clips from normal ones, we introduce a self-guided attention module in the feature encoder, *i.e.* , E_{SGA}, to capture the anomalous regions in videos to help the feature encoder produce more discriminative representations (see Section 3.3). Moreover, we introduce a sparse continuous sampling strategy in the pseudo label generator to enforce the network to pay more attention to the context around the most anomalous part (see Section 3.2). Finally, we introduce the deep MIL ranking loss to optimize the learning of the pseudo label generator, and we use cross entropy loss to train our proposed feature encoder E_{SGA} supervised by pseudo labels of anomalous videos and clip-level annotations of normal videos.

3.2. Pseudo Label Generation via Multiple Instance Learning

In contrast to [31], which simply assigns video-level labels to each clip and then trains the vanilla feature encoder at the very beginning, we introduce a MLP-based structure as the pseudo label generator trained under the MIL paradigm to generate pseudo labels, which are utilized in the refinement process of our feature encoder E_{SGA}.

Even though recent MIL-based methods [20, 28] have made considerable progress, the process of slicing a video into fixed segments in an coarse-grained manner regardless of its duration is prone to bury abnormal patterns as normal frames that usually constitute the majority, even in abnormal videos [24]. However, by sampling with a smaller temporal scale in a fine-grained manner, the network may overemphasize on the most intense part of an anomaly but ignore the context around it. In reality, anomalous events often last for a while. With the assumption of minimum duration of anomalies, the MIL network is forced to pay more attention to the context around the most anomalous part.

Moreover, to adapt to the variation in duration of untrimmed videos and class imbalance in amount, we introduce a sparse continuous sampling strategy: given the features for each clip extracted by a vanilla feature encoder E from a video $\{\mathbf{f}_i\}_{i=1}^N$, we uniformly sample L subsets from these video clips, and each subset contains T consecutive clips, forming L sub-bags $\overline{\mathcal{B}} = \{\mathbf{f}_{l,t}\}_{l=1,t=1}^{L,T}$, as shown in Figure 3. Remarkably, T, a hyperparameter to be tuned, also plays as the assumption of minimum duration of anomalies, as discussed in the previous paragraph. Here, we combine the MIL model with our continuous sampling strategy, as shown in Figure 3. We feed extracted features into our pseudo label generator to produce corresponding anomalous scores $\{s_{l,t}\}_{l=1,t=1}^{L,T}$. Next, we perform average pooling of the predicted instance-level scores $s_{l,t}$ of each sub-bag score as $\mathcal{S}_l$ below, which can be utilized in Eq. 7.

$$\mathcal{S}_l = \frac{1}{T}\sum_{t=1}^{T} s_{l,t}. \tag{1}$$

After training, the trained multiple instance pseudo label generator predicts clip-level scores for all abnormal videos marked as $S^a = \{s_i^a\}_{i=1}^N$. By performing temporal smoothing with a moving average filter to relieve the jitter of anomaly scores with kernel size of k,

$$\tilde{s}_i^a = \frac{1}{2k}\sum_{j=i-k}^{i+k} s_j^a, \tag{2}$$

and min-max normalization,

$$\hat{y}_i^a = \left(\tilde{s}_i^a - \min \tilde{S}^a\right) / (\max \tilde{S}^a - \min \tilde{S}^a)), i \in [1, N], \tag{3}$$

we refine the anomaly scores into $\hat{Y} = \{\hat{y}_i^a\}_{i=1}^N$. Specifically, $\hat{y}_i^a$ is in $[0, 1]$ and acts as a soft pseudo label. Then, the pseudo labeled data $\{V^a, \hat{Y}^a\}$ are combined with clip-level labeled data $\{V^n, Y^n\}$ as $\{V, Y\}$ to fine-tune the proposed feature encoder E_{SGA}.

3.3. Self-Guided Attention in Feature Encoder

In contrast to vanilla feature encoder E, which provides only task-agnostic representations for the down-stream task,

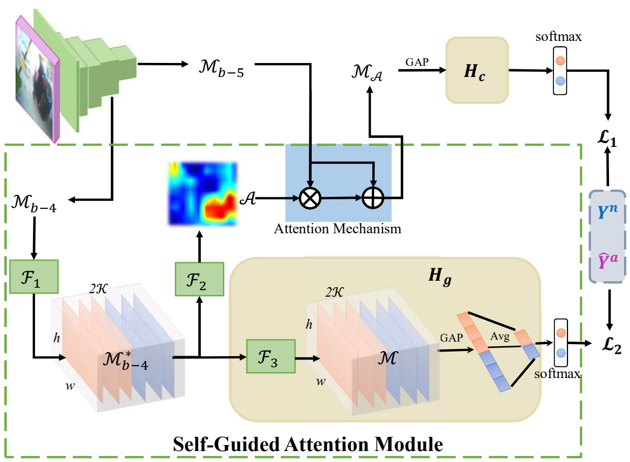

Figure 4: The structure of self-guided attention boosted feature encoder E_{SGA}. GAP means the global average pooling operation, while Avg means $\mathcal{K}$ channel-wise average pooling in producing guided anomaly scores in guided classification head H_g. $\mathcal{A}$ is the attention map. $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ are three encoding units constructed by convolutional layers.

we propose a self-guided attention boosted feature encoder E_{SGA} adapted from E, which optimizes attention map generation via pseudo labels supervision to enhance the learning of task-specific representations.

As Figure 4 shows, the self-guided attention module (SGA) takes feature maps $\mathcal{M}_{b-4}$ and $\mathcal{M}_{b-5}$ as input, which are produced by the 4^{th} and 5^{th} blocks of vanilla feature encoder E, respectively. SGA includes three encoding units, namely $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$, which are all constructed by convolutional layers. $\mathcal{M}_{b-4}$ is encoded as $\mathcal{M}^*_{b-4}$ and then applied to attention map $\mathcal{A}$ generation, denoted as

$$\mathcal{A} = \mathcal{F}_1(\mathcal{F}_2(\mathcal{M}_{b-4})). \tag{4}$$

Finally, we obtain $\mathcal{M}_A$ via the attention mechanism below:

$$\mathcal{M}_{\mathcal{A}} = \mathcal{M}_{b-5} + \mathcal{A} \circ \mathcal{M}_{b-5}, \tag{5}$$

where $\circ$ is element-wise multiplication, and $\mathcal{M}_{\mathcal{A}}$ is applied for final anomaly scores prediction via weighted-classification head $\mathcal{H}_c$, a fully connected layer.

To assist the learning of the attention map, we introduce a guided-classification head $\mathcal{H}_g$ that uses the pseudo labels as supervision. In $\mathcal{H}_g$, $\mathcal{F}_3$ transforms $\mathcal{M}^*_{b-4}$ into $\mathcal{M}$. Specifically, $\mathcal{M}^*_{b-4}$ and $\mathcal{M}$ hav $2\mathcal{K}$ channels as $\mathcal{K}$ multiple detectors for each class, *i.e.* , normal and abnormal, to enhance the guided supervision [26]. Then, we deploy spatiotemporal average pooling, $\mathcal{K}$ channel-wise average pooling on $\mathcal{M}$ and Softmax activation to obtain the guided anomaly scores for each class.

Remarkably, there are two classification heads in E_{SGA}, *i.e.* , weighted-classification head $\mathcal{H}_c$ and guided classification head $\mathcal{H}_g$, which are both supervised by pseudo labels via $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. That is, we optimize E_{SGA} with the pseudo labels (see Section 3.2). Therefore, the feature encoder E_{SGA} can update its parameters on video anomaly datasets and eliminate the domain gap from the pretrained parameters.

3.4. Optimization Process

- Deep MIL Ranking Loss: Considering that the positive bag contains at least one anomalous clip, we assume that the clip from a positive bag with the highest anomalous score is the most likely to be an anomaly [8].To adapt our sparse continuous sampling in 3.2, we treat a sub-bag as an instance and acquire a reliable relative comparison between the mostly likely anomalous sub-bag and the most likely normal sub-bag:

$$\max_{1 \le l \le L} \mathcal{S}^n_l < \max_{1 \le l \le L} \mathcal{S}^a_l \tag{6}$$

Specifically, to avoid too many false positive instances in positive bags, we introduce a sparse constraint on positive bags, which instantiates Eq. 6 as a deep MIL ranking loss with sparse regularization:

$$\mathcal{L}_{MIL} = \left(\epsilon - \max_{1 \le l \le L} \mathcal{S}^a_l + \max_{1 \le l \le L} \mathcal{S}^n_l\right)_+ + \frac{\lambda}{L}\sum_{l=1}^{L} \mathcal{S}^a_l. \tag{7}$$

where $(\cdot)_+$ means $\max(0, \cdot)$, and the first term in Eq. 7 ensures that $\max_{1 \le l \le L} \mathcal{S}^a_l$ is larger than $\max_{1 \le l \le L} \mathcal{S}^n_l$ with a margin of ϵ. ϵ is a hyperparameter that is equal to 1 in this work. The last term in Eq. 7 is the sparse regularization indicating that only a few sub-bags may contain the anomaly, while λ is another hyperparameter used to balance the ranking loss with sparsity regularization.

- Classification Loss: After obtaining the pseudo labels for an abnormal video in Eq. 3, we obtain the training pair $\{V^a, \hat{Y}^a\}$ that is further combined with $\{V^n, Y^n\}$ to train our feature encoder E_{SGA}. For this purpose, we apply the cross entropy loss function to the two classification heads ($\mathcal{H}_c$ and $\mathcal{H}_g$) in E_{SGA}, *i.e.* $\mathcal{L}_1$ and $\mathcal{L}_2$ in Figure 4.

Finally, we train a task-specific feature encoder E_{SGA} with the combination of $\mathcal{L}_1$ and $\mathcal{L}_2$. In the inference stage, we use E_{SGA} to predict clip-level scores for videos via weighted-classification head $\mathcal{H}_c$.

4. Experiments

4.1. Datasets and Metrics

We conduct experiments on two large datasets, *i.e.* , UCF-Crime [20] and ShanghaiTech [15], with two feature encoders, *i.e.* C3D [21] or I3D [2].

UCF-Crime is a large-scale dataset of real-world surveillance videos, including 13 types of anomalous events with 1900 long untrimmed videos, where 1610 videos are training videos and the others are test videos. Liu *et al.* [11]

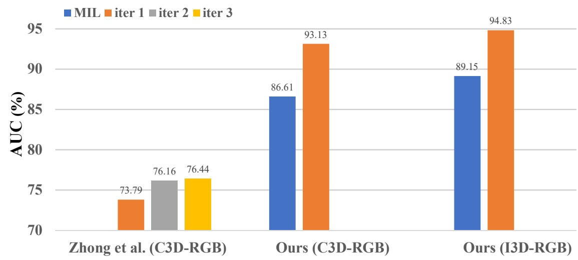

Figure 5: Comparisons with the state-of-the-art encoder-based method Zhong *et al.* [31] on ShanghaiTech.

manually annotated bounding boxes of anomalous regions in one image per 16 frames for each abnormal video, and we use their annotation of test videos only to evaluate our model's capacity to identify anomalous regions.

ShanghaiTech is a dataset of 437 campus surveillance videos. It has 130 abnormal events in 13 scenes, but all abnormal videos are in the test set, as the dataset is proposed for unsupervised learning. To adapt to the weakly supervised setting, Zhong *et al.* [31] re-organized the videos into 238 training videos and 199 testing videos.

Evaluation Metrics. Following previous works [13, 11, 20, 24], we compute the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) as the main metric, where a larger AUC implies higher distinguishing ability. We also follow [20, 24] to evaluate robustness by the false alarm rate (FAR) of anomaly videos.

4.2. Implementation Details

The multiple instance pseudo label generator, is a 3-layer MLP, where the number of units is 512, 32 and 1, respectively, regularized by dropout with probability of 0.6 between each layer. ReLU and Sigmoid functions are deployed after the first and last layer, respectively. Here, We adopt hyperparameters $L = 32$, $T = 3$, and $\lambda = 0.01$ and train the generator with the *Adagrad* optimizer with a learning rate of 0.01. While *fine-tuning*, we adopt the *Adam* optimizer with a learning rate of $1e-4$ and a weight decay of 0.0005 and train 300 epochs. More details about implementation are reported in Supplementary Material.

4.3. Comparisons with Related Methods

In Table 1, we present the **AUC**, **FAR** to compare our MIST with related state-of-the-art online methods in terms of *accuracy and robustness*. We can find that MIST outperforms or performs similarly to all other methods in terms of all evaluation metrics from Table 1, which confirms the efficacy of MIST. Specifically, the results of Zhong *et al.* [31], marked with $*$, are re-tested from the official released models[1] without deploying *10-crop*[2] for fair comparison, while the results in brackets are reported on [31] using 10-crop augmentation. However, 10-crop augmentation may improve the performance but requires 10 times the computation. Notably, the result of our MIST still slightly overtakes that of Zhong *et al.* [31] using 10-crop augmentation (81.08% vs. 81.40% in terms of AUC and 2.2% vs. 2.19% for FAR). Moreover, our method outperforms the supervised method of Liu *et al.* [11], which trains $C3D^{RGB}$ with external temporal annotations and NLN^{RGB} with external spatiotemporal annotations. These results verify that our proposed MIST is more effective than previous works.

[1] https://github.com/jx-zhong-for-academic-purpose/GCN-Anomaly-Detection.

[2] 10-crop is a test-time augmentation of cropping images into the center, four corners and their mirrored counterparts.

Method	Supervised	Grained	Encoder	AUC (%)	FAR (%)
Hasan et al. [7]	Un	Coarse	AE^{RGB}	50.6	27.2
Lu et al. [14]	Un	Coarse	Dictionary	65.51	3.1
SVM	Weak	Coarse	$C3D^{RGB}$	50	-
Sultani et al. [20]	Weak	Coarse	$C3D^{RGB}$	75.4	1.9
Zhang et al. [28]	Weak	Coarse	$C3D^{RGB}$	78.7	-
Zhu et al. [34]	Weak	Coarse	AE^{Flow}	79.0	-
Zhong et al. [31]	Weak	Fine	$C3D^{RGB}$	80.67*(81.08)	3.3*(2.2)
Liu et al. [11]	Full(T)	Fine	$C3D^{RGB}$	70.1	-
Liu et al. [11]	Full(S+T)	Fine	NLN^{RGB}	82.0	-
MIST	Weak	Fine	$C3D^{RGB}$	**81.40**	2.19
MIST	Weak	Fine	$I3D^{RGB}$	**82.30**	**0.13**

Table 1: Quantitative comparisons with existing online methods on UCF-Crime under different levels of supervision and fineness of prediction. The results in $(\cdot)$ are tested with *10-crop*, while those marked by * are tested without.

Method	Feature Encoder	Grained	AUC (%)	FAR (%)
Sultani *et al.* [20]	$C3D^{RGB}$	Coarse	86.30	0.15
Zhang *et al.* [28]	$C3D^{RGB}$	Coarse	82.50	0.10
Zhong *et al.* [31]	$C3D^{RGB}$	Fine	76.44	-
AR-Net [24]	$C3D^{RGB}$	Fine	85.01*	0.57*
AR-Net [24]	$I3D^{RGB}$	Fine	85.38	0.27
AR-Net [24]	$I3D^{RGB+Flow}$	Fine	91.24	0.10
MIST	$C3D^{RGB}$	Fine	**93.13**	1.71
MIST	$I3D^{RGB}$	Fine	**94.83**	**0.05**

Table 2: Quantitative comparisons with existing methods on ShanghaiTech. The results with * are re-implemented.

For the ShanghaiTech dataset results in Table 2, our MIST far outperforms other RGB-based methods [20, 28, 31, 24], which validates the capacity of MIST. Remarkably, MIST also surpasses the multi-model method of AR-Net [24] ($I3D^{RGB+Flow}$) on AUC by more than 4% to 94.83% and gains a much lower FAR of 0.05%.

We detail the comparison with the state-of-the-art encoder-based method [31] on ShanghaiTech in Figure 5. The multiple instance pseudo label generator performs much better than Zhong *et al.* [31], which indicates the drawback of utilizing video-level labels as clip-level labels. Even though Zhong *et al.* [31] optimizes for three iterations, it falls far behind our MIST with 16.69% AUC on C3D, which solidly verifies the efficiency and efficacy of MIST. Moreover, our MIST is much faster in the inference stage, as Zhong *et al.* [31] applies *10-crop* augmentation.

Encoder-Agnostic Methods	AUC (%) UCF-Crime pretrained	fine-tuned	ShanghaiTech pretrained	fine-tuned
Sultani *et al.* [20]	78.43	81.42	86.92	92.63
Zhang *et al.* [28]	78.11	81.58	88.87	92.50
AR-Net [24]	78.96	82.62	85.38	92.27
Our MIL generator	79.37	81.55	89.15	92.24

Table 3: Quantitative comparisons between the features from the pretrained vanilla feature encoder and those from MIST on UCF-Crime and ShanghaiTech datasets by adopting encoder-agnostic methods.

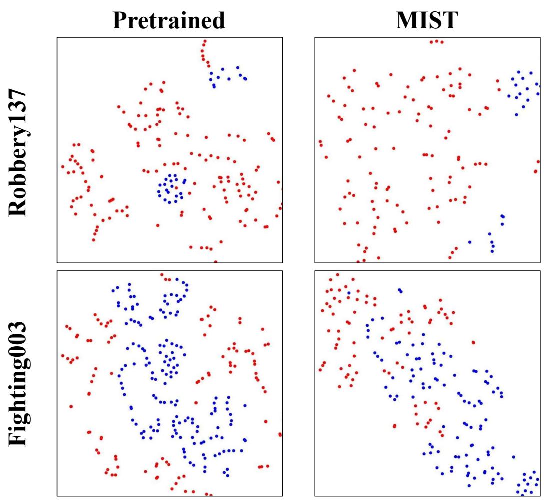

Figure 6: Feature space visualization of pretrained vanilla feature encoder **I3D** and the MIST fine-tuned encoder via t-SNE [16] on UCF-Crime testing videos. The red dots denote anomalous regions while the blue ones are normal.

4.4. Task-Specific Feature Encoder

To verify that our feature encoder can produce task-specific representations that facilitate the other encoder-agnostic methods, we also conduct related experiments with **I3D** as presented in Table 3. It is noticeable that all results of encoder-agnostic methods are boosted after using our MIST fine-tuned features, showing a reduction in the domain gap. For example, AR-Net [24] increases from 85.38% to 92.27% on the UCF-Crime dataset and achieves an improvement of 6.89% on the ShanghaiTech dataset. Therefore, our MIST can produce a more powerful task-specific feature encoder that can be utilized in other approaches. We visualize the feature space of the pretrained I3D vanilla feature encoder and the MIST-fine-tuned encoder via t-SNE[16] in Figure 6, which also indicates the refinement of feature representations.

4.5. Ablation Study

At first, we introduce another evaluation metric, *i.e. score gap*, which is the gap between the average scores of abnormal clips and normal clips. Larger score gap

Dataset	Feature	AUC (%) Uniform	Sparse Continuous	ΔAUC (%)
UCF-Crime	$C3D^{RGB}$	74.29	75.51	+1.22
	$I3D^{RGB}$	78.72	79.37	+0.65
ShanghaiTech	$C3D^{RGB}$	83.68	86.61	+2.93
	$I3D^{RGB}$	83.10	89.15	+6.05

Table 4: Performance comparisons of sparse continuous sampling and uniform sampling for MIL generator training.

indicates the network is more capable of distinguishing anomalies from normal events [13]. We conduct ablation studies on UCF-Crime to analyze the impact of generated pseudo labels (PLs), the self-guided attention module (SGA), and classifier head H_g in SGA of proposed feature encoder E_{SGA} in Table 5. Compared with the baseline and MIST$^{\text{w/o PLs}}$, our MIST achievees a significant improvement when the generated pseudo labels are utilized. In particular, we observe 8.17% improvement in AUC and an approximately 17% score gap, which shows the efficacy of our multiple instance pseudo label generator with the sparse continuous sampling strategy. Pseudo labels also plays an important role. Compared with MIST, the performance of MIST$^{\text{w/o PLs}}$ drops seriously, even worse than the baseline for the low-quality supervision that influences the attention map $\mathcal{A}$ generation from SGA.

Moreover, SGA enhances the feature encoder on emphasizing the informative regions and distinguishing abnormal events from normal ones. Compared with $MIST^{w/oSGA}$, MIST increases by 2% in AUC and 5% in the score gap. Specifically, the guided-classification branch in SGA plays an important role in guiding the attention map generation, and there is a drop of more than 2% if such a branch is removed.

Ablation studies are also conducted on a sparse continuous sampling strategy on UCF-Crime and ShanghaiTech with $C3D^{RGB}$ and $I3D^{RGB}$ features. As shown in Table 4, when sampling the same number of clips for a bag and selecting the same number of top clips to represent the bag, our sparse continuous sampling strategy pays more attention to the context and does better than uniform sampling. Especially in ShanghaiTech, sparse continuous sampling gains 2.93% and 6.05% on two kinds of features.

4.6. Visual Results

To further evaluate the performance of our model, we visualize the temporal predictions of the models. As presented in Figure 7, our model exactly localizes the anomalous events and predicts anomaly scores very close to zero on normal videos, showing the effectiveness and robustness of our model. We collect some failed samples in the right row of Figure 7. In addition, our model predicts the highest score at the end of *Arrest001*, where a man walks across the scene with his arm pointing forward as if brandishing a

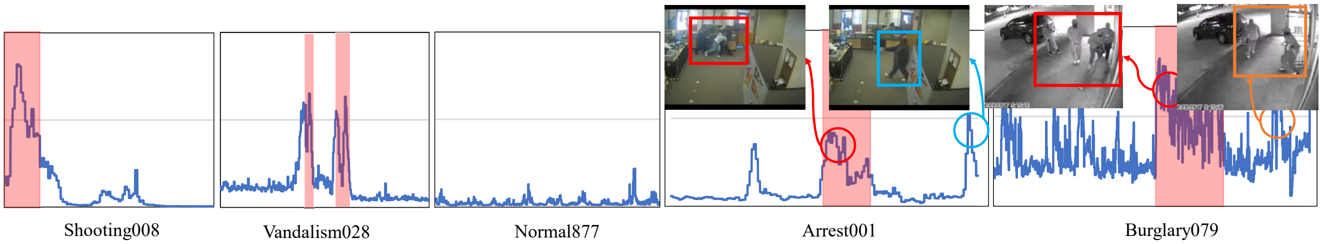

Figure 7: Visualization of the testing results on UCF-Crime (better viewed in color). The red blocks in the graphs are temporal ground truths of anomalous events. The orange circle shows the wrongly labeled ground truth, the blue circle indicates the wrongly predicted clip, and the red cricle indicates the correctly predicted clip.

Method	AUC (%)	Score Gap (%)
Baseline	74.13	0.375
MIST$^{\text{w/o PLs}}$	73.33	0.443
MIST$^{\text{w/o }H_g}$	81.97	15.37
MIST$^{\text{w/o SGA}}$	80.28	12.74
MIST	**82.30**	**17.71**

Table 5: Ablation Studies on UCF-Crime with $I3D^{RGB}$. Baseline is the original **I3D** trained with video-level labels [31]. MIST is our whole model. MIST$^{\text{w/o PLs}}$ is trained without pseudo labels but with video-level labels. MIST$^{\text{w/o }H_g}$ is MIST trained without H_g. MIST$^{\text{w/o SGA}}$ is trained without the self-guided attention module).

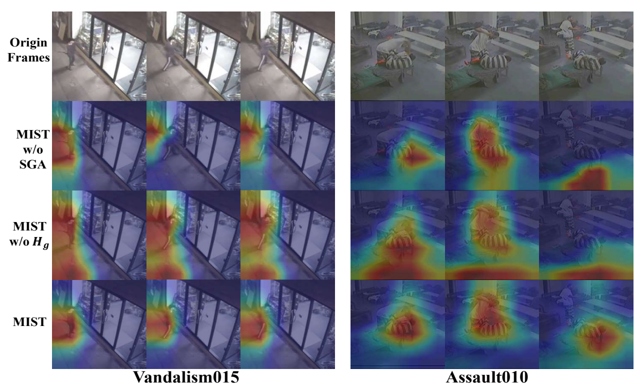

Figure 8: Visualization results of anomaly activation maps (better viewed in color).

gun. As the videos in UCF-Crime are low-resolution, it is difficult to judge such a confusing action without any other context information. Furthermore, the bottom-right part of Figure 7 shows another failed case; *i.e.* , our model successfully localizes the major part of the anomalous burglary event and raises an alarm when the thieves are rushing out of the house, which should be treated as an anomaly but is wrongly labeled as a normal event in the ground truth. We also visualize the spatial activation map via Grad-CAM on $\mathcal{M}_{\mathcal{A}}$ [19] for spatial explanation. As Figure 8 shows, our model is able to sensitively focus on informative regions that help decide whether the scene is anomalous . This verifies that our self-guided attention module can boost the feature encoder to focus on anomalous regions. Additionally, compared with the activation maps generated from the MIST without guided-classification head H_g and the MIST without the SGA module, the results of MIST are concentrated on the anomalous regions, which shows the rationality and effectiveness of our self-guided attention module.

4.7. Discussions

The key of our MIST is to design a two stage self-training strategy to train a task-specific feature encoder for video anomaly detection. Each component of our framework can be replaced by any other advanced module, *e.g.* , replacing C3D with I3D, or a stronger pseudo label generator to take the place of the multiple instance pseudo label generator. Additionally, the scheme of our framework can be adapted to other tasks, such as weakly supervised video action localization and video highlight detection.

5. Conclusions

In this work, we propose a multiple instance self-training framework (MIST) to fine-tune a task-specific feature encoder efficiently. We adopt a sparse continuous sampling strategy in the multiple instance pseudo label generator to produce more reliable pseudo labels. With the estimated pseudo labels, our proposed feature encoder learns to focus on the most probable anomalous regions in frames facilitated by the proposed self-guided attention module. Finally, after a two-stage self-training process, we train a task-feature encoder with discriminative representations that can also boost other existing methods. Remarkably, our MIST makes significant improvements on two public datasets.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC (U1911401, U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong NSF Project (Nos.2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), Research Projects of Zhejiang Lab (No.2019KD0AB03), and the Key-Area Research and Development Program of Guangzhou (202007030004).

References

[1] Massih-Reza Amini and Patrick Gallinari. Semi-supervised logistic regression. In *ECAI*, 2002.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[3] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Adv. Neural Inform. Process. Syst.*, 2019.

[4] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. 2020.

[5] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Int. Conf. Comput. Vis.*, 2019.

[6] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Adv. Neural Inform. Process. Syst.*, 2005.

[7] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[8] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. *arXiv preprint arXiv:2007.09833*, 2020.

[9] Timothy Hospedales, Shaogang Gong, and Tao Xiang. A markov clustering topic model for mining behaviour in video. In *Int. Conf. Comput. Vis.*, 2009.

[10] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

[11] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *ACM Int. Conf. Multimedia*, 2019.

[12] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, 2019.

[13] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[14] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Int. Conf. Comput. Vis.*, 2013.

[15] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Int. Conf. Comput. Vis.*, 2017.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[17] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[18] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12173–12182, 2020.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[20] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. Comput. Vis.*, 2015.

[22] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284, 2015.

[23] Waseem Ullah, Amin Ullah, Ijaz Ul Haq, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks. *Multimedia Tools and Applications*, pages 1–17, 2020.

[24] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *Int. Conf. Multimedia and Expo*, 2020.

[25] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[26] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[27] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[28] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *IEEE Int. Conf. Image Process.*, 2019.

[29] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2011.

[30] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *ACM Int. Conf. Multimedia*, 2017.

[31] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[32] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly

detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550, 2019.

[33] Sijie Zhu, Chen Chen, and Waqas Sultani. Video anomaly detection for smart surveillance. *arXiv preprint arXiv:2004.00222*, 2020.

[34] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *Brit. Mach. Vis. Conf.*, 2019.

[35] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Int. Conf. Comput. Vis.*, 2019.

[36] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Eur. Conf. Comput. Vis.*, 2018.