Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection

Shuo Li, Fang Liu*, Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education,
International Research Center for Intelligent Perception and Computation,
Joint International Research Laboratory of Intelligent Perception and Computation,
School of Artificial Intelligent, Xidian University, Xi'an, 710071, P.R. China,
alisure@stu.xidian.edu.cn, f63liu@163.com, lchjiao@mail.xidian.edu.cn

Abstract

Weakly supervised Video Anomaly Detection (VAD) using Multi-Instance Learning (MIL) is usually based on the fact that the anomaly score of an abnormal snippet is higher than that of a normal snippet. In the beginning of training, due to the limited accuracy of the model, it is easy to select the wrong abnormal snippet. In order to reduce the probability of selection errors, we first propose a Multi-Sequence Learning (MSL) method and a hinge-based MSL ranking loss that uses a sequence composed of multiple snippets as an optimization unit. We then design a Transformer-based MSL network to learn both video-level anomaly probability and snippet-level anomaly scores. In the inference stage, we propose to use the video-level anomaly probability to suppress the fluctuation of snippet-level anomaly scores. Finally, since VAD needs to predict the snippet-level anomaly scores, by gradually reducing the length of selected sequence, we propose a self-training strategy to gradually refine the anomaly scores. Experimental results show that our method achieves significant improvements on ShanghaiTech, UCF-Crime, and XD-Violence.

Introduction

Video Anomaly Detection (VAD) aims to detect abnormal events in the video, which has important practical value (Zhang, Qing, and Miao 2019; Guo et al. 2021). Generally, VAD predicts the anomaly score of each snippet in the video. There are three main paradigms: unsupervised VAD (Gong et al. 2019; Cai et al. 2021), weakly supervised VAD (Zhong et al. 2019), and supervised VAD (Liu and Ma 2019; Wan et al. 2021). Unsupervised VAD only learns on normal videos, assuming that unseen abnormal videos have high reconstruction errors. Due to the lack of prior knowledge of abnormality and inability to learn all normal video patterns, the performance of unsupervised VAD is usually poor (Tian et al. 2021). Because fine-grained anomaly label is time-consuming and laborious, it is difficult to collect large-scale datasets for supervised paradigm. With whether the video contains anomalies as video-level label, the weakly supervised paradigm predicts the anomaly score of each frame. The weakly supervised paradigm is verified to be a feasible method because of its competitive performance (Feng, Hong, and Zheng 2021). Recently, many researchers have focused on weakly supervised VAD (Zhong et al. 2019).

Most weakly supervised VADs are based on Multi-Instance Learning (MIL) (Sultani, Chen, and Shah 2018; Zhu and Newsam 2019; Wan et al. 2020; Tian et al. 2021). MIL-based methods treat a video as a bag, which contains multiple instances. Each instance is a snippet. The bag generated from an abnormal video is called a positive bag, and the bag generated from a normal video is called a negative bag. Since the video-level label indicates whether the video contains anomalies, the positive bag contains at least one abnormal snippet and the negative bag contains no abnormal snippet. MIL-based methods learn instance-level anomaly scores through the bag-level labels (Zhong et al. 2019).

In MIL-based methods, at least one instance of the positive bag contains the anomaly, and any instance of the negative bag does not contain the anomaly (Sultani, Chen, and Shah 2018). Generally, MIL-based methods assume that the instance with the highest anomaly score in the positive bag should rank higher than the instance with the highest anomaly score in the negative bag (Zhu and Newsam 2019). Therefore, the important thing for MIL-based methods is to correctly select anomalous instance in the positive bag. Most MIL-based methods regard an instance as an optimization unit (Zhang, Qing, and Miao 2019; Feng, Hong, and Zheng 2021; Tian et al. 2021). However, if the model predicts the anomalous instances incorrectly in the positive bag, this error will be strengthened as the training progresses. That is, if a normal instance is predicted as an abnormal instance, this error will affect subsequent instance selection. In addition, the abnormal event is usually multiple consecutive snippets, but MIL-based methods do not consider this prior.

In order to alleviate the above-mentioned shortcomings, we propose a Multi-Sequence Learning (MSL) method. Our MSL no longer uses an instance as the optimization unit, but a sequence composed of multiple instances as the optimization unit. In other words, instead of choosing the instance with the highest anomaly score, our MSL method chooses the sequence with the highest sum of anomaly scores. This reduces the probability of incorrect selection of anomalous instances. In order to achieve our MSL, we propose a Transformer-based Multi-Sequence Learning Network, which includes a multi-layer Convolutional Transformer Encoder to encode extracted snippet features, a Video Classifier

*Corresponding author

to predict video-level anomaly scores, and a Snippet Regressor to predict snippet-level anomaly scores. In the inference stage, we propose to use video-level anomaly scores to suppress fluctuations in the snippet-level anomaly scores. Since the goal of VAD is to predict fine-grained anomaly scores (Tian et al. 2021), a two stage self-training strategy is used to gradually refine the anomaly scores.

To demonstrate the performance of our MSL, we use VideoSwin (a Transformer-based method) (Liu et al. 2021c) as the backbone to extract snippet-level features and conduct experiments on ShanghaiTech (Luo, Liu, and Gao 2017), UCF-Crime (Sultani, Chen, and Shah 2018), and XD-Violence (Wu et al. 2020). For a fair comparison, we also use C3D (Tran et al. 2015) and I3D (Carreira and Zisserman 2017) as the backbone to extract features. Experiments show that our MSL achieve the state-of-the-art results. In summary, our main contributions are as follows:

- We propose a Multi-Sequence Learning method, which uses a sequence composed of multiple instances as an optimization unit. Based on this, we propose a Multi-Sequence Learning ranking loss, which selects the sequence with the highest sum of anomaly scores.
- Based on Multi-Sequence Learning and its ranking loss, we design a Transformer-based Multi-Sequence Learning network, and propose to use the video-level anomaly classification probability to suppress the fluctuation of the snippet-level anomaly score in the inference stage.
- By gradually reducing the length of selected sequence, we propose a two stage self-training strategy to gradually refine the anomaly scores, because VAD needs to predict fine-grained anomaly scores.
- Experimental results show that our method achieves the state-of-the-art results on ShanghaiTech, UCF-Crime, and XD-Violence. The visualization shows that our method can realize the detection of abnormal snippets.

Related Work

Weakly Supervised Video Anomaly Detection

Most existing weakly supervised VAD methods (He, Shao, and Sun 2018; Zhang, Qing, and Miao 2019) are based on MIL. Since most methods (Li, Mahadevan, and Vasconcelos 2014; Zhao, Fei-Fei, and Xing 2011) earlier than 2017 only used normal training videos, He, Shao, and Sun propose an anomaly-introduced learning method to detect abnormal events, and propose a graph-based MIL model with both normal and abnormal video data (He, Shao, and Sun 2018). Sultani, Chen, and Shah propose a deep MIL ranking loss to predict anomaly scores (Sultani, Chen, and Shah 2018). Zhang, Qing, and Miao further introduces inner-bag score gap regularization by defining an inner bag loss (Zhang, Qing, and Miao 2019). Zhong et al. consider the anomaly detection with weak labels as a supervised learning under noise labels, and design an alternate training procedure to promote the discrimination of action classifiers (Zhong et al. 2019). Zhu and Newsam propose an attention-based temporal MIL ranking loss, which use temporal context to distinguish between abnormal and normal events better (Zhu and Newsam 2019). Wan et al. propose a dynamic MIL loss to enlarge the inter-class distance between anomalous and normal instances, and a center loss to reduce the intra-class distance of normal instances (Wan et al. 2020). Feng, Hong, and Zheng propose a MIL-based pseudo label generator and adopt a self-training scheme to refine pseudo-label by optimizing a self-guided attention encoder and a task-specific encoder (Feng, Hong, and Zheng 2021). Tian et al. propose an robust temporal feature magnitude learning to effectively recognize the anomaly instances (Tian et al. 2021).

Self-Training

Self-training is widely used in semi-supervised learning (Rosenberg, Hebert, and Schneiderman 2005; Tanha, van Someren, and Afsarmanesh 2017; Tao et al. 2018; Li et al. 2019; Jeong, Lee, and Kwak 2020; Tai, Bailis, and Valiant 2021). In self-training, the training data usually contain labeled and unlabeled data (Liu et al. 2011). Self-training includes the following steps (Zheng et al. 2020; Yu et al. 2021): 1) Train model with labeled data; 2) Use the trained model to predict unlabeled data to generate pseudo-labels; 3) Train model with labeled and pseudo-labeled data together; 4) Repeat 2) and 3). In VAD, Pang et al. propose a self-training deep neural network for ordinal regression (Pang et al. 2020). Feng, Hong, and Zheng propose a multi-instance self-training method that assigns snippet-level pseudo-labels to all snippets in abnormal videos (Feng, Hong, and Zheng 2021). Unlike them, our focus is on refining anomaly scores through self-training.

Transformer Combined With Convolution

More and more studies have shown that Transformer has excellent performance (Dosovitskiy et al. 2021; Touvron et al. 2021; Liu et al. 2021b). Dosovitskiy et al. first prove that a pure Transformer architecture can attain state-of-the-art performance (Dosovitskiy et al. 2021). Touvron et al. further explore the data-efficient training strategies for the vision transformer (Dosovitskiy et al. 2021; Touvron et al. 2021). Liu et al. further introduces the inductive biases of locality, hierarchy and translation invariance for various image recognition tasks (Liu et al. 2021b). Because transformer lacks the ability of local perception, many works combine convolution and transformer (d'Ascoli et al. 2021; Wu et al. 2021; Li et al. 2021; Xu et al. 2021; Yan et al. 2021; Zhang and Yang 2021; Liu et al. 2021a). To introduce local inter-frame perception, similar to Wu et al., we turn the linear projection in the Transformer Block into a Depthwise Separable 1D Convolution (Chollet 2017; Howard et al. 2017).

Our Approach

In this section, we first define the notations and problem statement. We then introduce our *Multi-Sequence Learning* (MSL). Finally, we present the pipeline of our approach.

Notations and Problem Statement

In weakly supervised VAD, training videos are only labeled at the video-level. That is, videos containing anomalies are labeled as 1 (positive), and videos without any anomalies

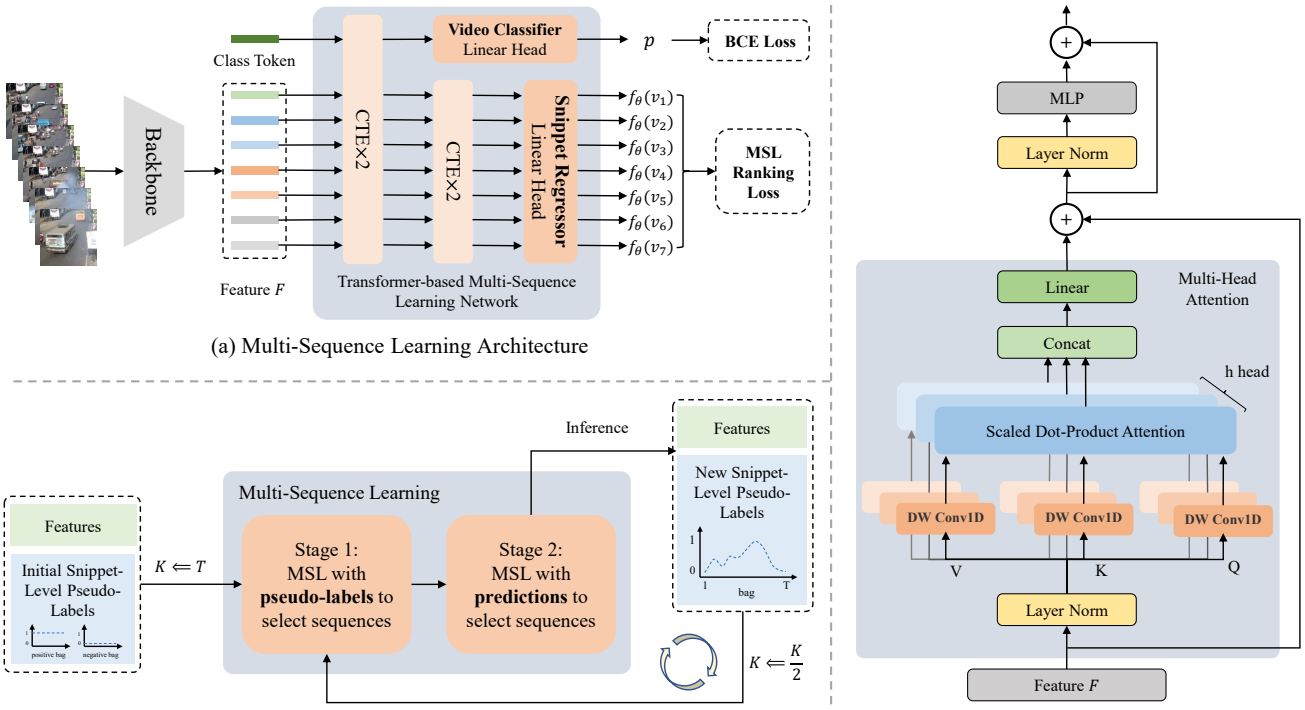


Figure 1: Overall framework. (a) The architecture of our Multi-Sequence Learning (MSL), which includes a Backbone and a Transformer-based MSL Network (MSLNet). The feature $F \in \mathcal{R}^T \times D$ extracted by the Backbone is input into MSLNet to predict the anomaly scores, where T is the number of snippets and D is the feature dimension of each snippet. MSLNet contains a video classifier to predict the probability p of the video containing anomalies and a snippet regressor to predict the snippet anomaly score $f_\theta(v_i)$ of the i-th snippet. BCE is the Binary Cross Entropy loss. (b) The pipeline of self-training MSL, where K gradually changes from T to 1 through a self-training mechanism. According to the way of selecting sequences, the optimization of MSL includes two stages: the first stage uses pseudo-labels to select sequences and the second stage uses predictions to select sequences. (c) Convolutional Transformer Encoder (CTE), which is similar to (Dosovitskiy et al. 2021), except that the linear projection is replaced with DW Conv1D (Depthwise Separable 1D Convolution) (Howard et al. 2017).

are labeled as 0 (negative). Given a video $V = \{v_i\}_{i=1}^T$ with T snippets and its video-level label $Y \in \{0, 1\}$. MIL-based methods treat video V as a bag and each snippet as an instance. A positive video is regarded as a positive bag $\mathcal{B}_a = (a_1, a_2, ..., a_T)$, and a negative video is regarded as a negative bag $\mathcal{B}_n = (n_1, n_2, ..., n_T)$. The goal of VAD is to learn a function f_θ maps snippets to their anomaly scores, ranging from 0 to 1. Generally, MIL-based VAD assumes that abnormal snippets have higher abnormal scores than normal snippets. Sultani, Chen, and Shah formulate VAD as an anomaly score regression problem and propose a MIL ranking objective function and a MIL ranking loss (Sultani, Chen, and Shah 2018):

$$\max_{i \in \mathcal{B}_a} f_\theta(a_i) > \max_{i \in \mathcal{B}_n} f_\theta(n_i). \quad (1)$$

$$\mathcal{L}(\mathcal{B}_a, \mathcal{B}_n) = \max(0, \max_{i \in \mathcal{B}_n} f_\theta(n_i) - \max_{i \in \mathcal{B}_a} f_\theta(a_i)). \quad (2)$$

The intuition behind Eq.1 and Eq.2 that the snippet with highest anomaly score in the positive bag should rank higher than the snippet with highest anomaly score in the negative bag (Zhu and Newsam 2019). In order to keep a large margin between the positive and negative instances, Sultani, Chen, and Shah give a hinge-based ranking loss:

$$\mathcal{L}(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f_\theta(a_i) + \max_{i \in \mathcal{B}_n} f_\theta(n_i)). \quad (3)$$

At the beginning of the optimization, f_θ needs to have a certain ability to predict abnormalities. Otherwise, it will be possible to select a normal instance as an abnormal instance. If f_θ predicts the instances in the positive bag incorrectly, *e.g.* predicting normal instances as abnormal instances, this error will be strengthened as the training progresses. In addition, the abnormal event is usually multiple consecutive snippets, but MIL-based methods do not consider this prior.

Multi-Sequence Learning

In order to alleviate the above shortcomings in MIL-based methods, we propose a novel Multi-Sequence Learning (MSL) method. As shown in Figure 2, given a video $V = \{v_i\}_{i=1}^T$ with T snippets, the anomaly score curve is predicted through a mapping function f_θ. Let us assume that

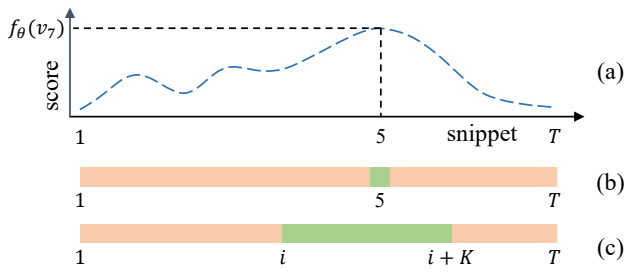


Figure 2: Comparison of instance selection method between MIL and our MSL. (a) Anomaly score curve of a video containing T snippets, assuming that the 5-th snippet has the largest anomaly score $f_\theta(v_5)$. (b) Instance selection method of MIL, which selects the 5-th snippet. (c) Instance selection method of our MSL, which selects a sequence consisting of K consecutive snippets starting from the i-th snippet.

the 5-th snippet v_5 has the largest anomaly score $f_\theta(v_5)$. In MIL-based methods, the 5-th snippet will be selected to optimize the network (Zhu and Newsam 2019). In our MSL, given a hyperparameter K, we propose a sequence selection method, which selects a sequence that contains K consecutive snippets. In detail, we calculate the mean of anomaly scores of all possible sequences of K consecutive snippets:

$$S = \{s_i\}_{i=1}^{T-K}, \quad s_i = \frac{1}{K}\sum_{k=0}^{K-1} f_\theta(v_{i+k}), \tag{4}$$

where s_i represents the mean of anomaly scores of the sequence of K consecutive snippets starting from the i-th snippet. Then, the sequence with the largest mean of abnormal scores can be selected by $\max_{s_i \in S} s_i$.

Based on the above sequence selection method, we can simply use an MSL ranking objective function as:

$$\max_{s_{a,i}\in S_a} s_{a,i} > \max_{s_{n,i}\in S_n} s_{n,i},$$
$$s_{a,i} = \frac{1}{K}\sum_{k=0}^{K-1} f_\theta(a_{i+k}), \quad s_{n,i} = \frac{1}{K}\sum_{k=0}^{K-1} f_\theta(n_{i+k}). \tag{5}$$

where $s_{a,i}$ and $s_{n,i}$ represent the mean of abnormal scores of K consecutive snippets starting from the i-th snippet in abnormal video and normal video, respectively. The intuition of our MSL ranking objective function is that the mean of abnormal scores of K consecutive snippets in abnormal videos should be greater than the mean of abnormal scores of K consecutive snippets in normal videos. To keep a large margin between the positive and negative instances, similar to Eq. 3, our hinge-based MSL ranking loss is defined as:

$$\mathcal{L}(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{s_{a,i}\in S_a} s_{a,i} + \max_{s_{n,i}\in S_n} s_{n,i}). \tag{6}$$

It can be seen that MIL is a case of our MSL. When $K = 1$, MIL and our MSL are equivalent. When $K = T$, our MSL treats every snippet in the abnormal video as abnormal.

Transformer-based MSL Network

Convolutional Transformer Encoder Before introducing our Transformer-based MSL architecture, we first introduce the basic layer. Transformer (Vaswani et al. 2017) uses sequence data as input to model long-range relationships, and has made great progress in many tasks. We adopt Transformer as our basic layer. The representation between the local frames or snippets of the video is also very important. However, Transformer is not good at learning local representations of adjacent frames or snippets (Yan et al. 2021). Motivated by this, as shown in Figure 1(c), we replace the linear projection in the original Transformer with a DW Conv1D (Depthwise Separable 1D Convolution) (Howard et al. 2017) projection. The new Transformer is named Convolutional Transformer Encoder (CTE). In this way, our CTE can inherit the advantages of Transformer and Convolutional Neural Network.

Transformer-based MSL Network As shown in Figure 1 (a), our architecture includes a Backbone and a MSLNet. Any action recognition method can be used as the Backbone, such as C3D (Tran et al. 2015), I3D (Carreira and Zisserman 2017), and VideoSwin (Liu et al. 2021c). Similar to (Tian et al. 2021), the Backbone uses pre-trained weights on the action recognition datasets (Karpathy et al. 2014; Kay et al. 2017). Through the Backbone, a feature $F \in \mathcal{R}^{T\times D}$ extracts from a video containing T snippets, where D is the feature dimension of each snippet. Our MSLNet will use F as the input to predict anomalies.

Our MSLNet includes a video classifier and a snippet regressor. The video classifier is used to predict whether the video contains anomalies. Specifically, the video classifier contains two layers of CTE and a linear head for predicting the probability of whether the video contains anomalies:

$$p = \sigma(\mathcal{W}^c \cdot E^c[0]), \quad E^c = CTE_{\times 2}(class\ token||F), \tag{7}$$

where $\mathcal{W}^c$ is the parameter of the linear head, p is the probability that the video contains anomalies, and *class token* is used to predict the probability by aggregated features in CTE. Since whether the video contains anomalies is a binary classification problem, σ chooses the sigmoid function.

The snippet regressor is used to predict the anomaly score of each snippet. Specifically, the snippet regressor contains two layers of CTE and a linear head for predicting the anomaly score of each snippet:

$$f_\theta(v_i) = \sigma(\mathcal{W}^r \cdot E^r[i]), \quad E^r = CTE_{\times 2}(E^c), \tag{8}$$

where $\mathcal{W}^r$ is the parameter of the linear head, $f_\theta(v_i)$ is the abnormal score of the i-th snippet, and $E^r[i]$ is the feature of the i-th snippet. Since predicting the anomaly score is treated as a regression problem, σ chooses the sigmoid function.

We regard the optimization of the video classifier and snippet regressor as a multi-task learning problem. The total loss to optimize the parameters of MSLNet is the sum of our hinge-based MSL ranking loss and the classification loss:

$$\mathcal{L} = \mathcal{L}(\mathcal{B}_a, \mathcal{B}_n) + BCE(p, Y), \tag{9}$$

where $\mathcal{L}(\mathcal{B}_a, \mathcal{B}_n)$ is the Eq. 6, and *BCE* is the Binary Cross Entropy loss between the output p and the target Y.

To reduce the fluctuation of the abnormal scores predicted by the snippet regressor, we propose a score correction method in the inference stage. Specifically, the score correction method corrects the abnormal scores by using the

probability of whether the video contains anomalies:

$$\hat{f}_\theta(v_i) = f_\theta(v_i) \times p. \tag{10}$$

The intuition of this method is that to keep the anomaly scores when the video classifier predicts that the video contains anomalies with a higher probability, and weaken the anomaly scores when the video classifier predicts that the video contains anomalies with a lower probability.

Self-Training MSL

As shown in Figure 1 (b), we propose a self-training mechanism to achieve the training from coarse to fine. The training process of our MSLNet includes two training stages. Before introducing our self-training mechanism, we first get the pseudo-labels $\hat{\mathcal{Y}}$ of the training videos. By taking the known video-level labels $\mathcal{Y}$ in weakly supervised VAD as the anomaly scores of snippets, we can immediately get the initial snippet-level pseudo-labels $\hat{\mathcal{Y}}$. That is, for an abnormal video, the pseudo label of each snippet is 1, and for a normal video, the pseudo label of each snippet is 0.

In the initial stage of training, the function f_θ has a poor ability to predict abnormalities. Therefore, if the sequence is selected directly through the prediction of f_θ, there is a probability of selecting the wrong sequence. Based on this motivation, we propose a transitional stage (stage one): MSL with pseudo-labels to select sequences. Specifically, by replacing the predicted anomaly score $f_\theta(v_i)$ in Eq. 4 with the pseudo-label $\hat{y}_i$ of each snippet v_i, we select the sequence with the largest mean of pseudo labels by $\max_{s_i \in S} s_i$. Based on this sequence, we can calculate $s_{a,i}$ and $s_{n,i}$, and then optimize MSLNet through the hinge-based MSL ranking loss:

$$\mathcal{L}(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - s_{a,i} + s_{n,i}), \tag{11}$$

where $s_{a,i}$ and $s_{n,i}$ are the sequence with the largest mean of pseudo labels starting from the i-th snippet in the abnormal and normal video, respectively. After E_1 epochs training, f_θ has a preliminary ability to predict the anomaly scores.

In stage two, MSLNet is optimized with predictions to select sequences. This stage uses Eq. 5 and Eq. 6 to calculate the ranking loss. After E_2 epochs training, the new snippet-level pseudo-labels $\hat{\mathcal{Y}}$ of training videos are inferenced. By halving the sequence length K and repeating the above two stages, the predicted anomaly scores are gradually refined.

The role of the transitional stage is to establish a connection between MSL and different self-training rounds. By introducing a self-training mechanism, we achieve the prediction of anomaly scores from coarse to fine. For better understanding, we show our self-training MSL in Algorithm 1.

Algorithm 1: Our Self-Training Multi-Sequence Learning.

Input: A set of features $\mathcal{F}$ and its video-level labels $\mathcal{Y}$.
Parameter: The number T of snippets.
Output: MSLNet.

```
Set K ← T.
Get the initial snippet-level pseudo-labels Ŷ by Y.
while K ≥ 1 do
    Initialize the parameters of MSLNet.
    // Stage one: with pseudo-labels to select sequences.
    Optimize MSLNet with K by F, Ŷ, and Eq. 11.
    // Stage two: with predictions to select sequences.
    Optimize MSLNet with K by F and Eq. 6.
    Inference the new snippet-level pseudo-labels Ŷ.
    Set K ← K/2.
end while
return MSLNet.
```

Experiments

Datasets and Evaluation Metrics

We conduct sufficient experiments on the ShanghaiTech, UCF-Crime, and XD-Violence datasets.

ShanghaiTech is a medium-scale dataset that contains 437 campus surveillance videos with 130 abnormal events in 13 scenes (Luo, Liu, and Gao 2017). However, all the training videos of this dataset are normal. In line with the weakly supervised setting, we adopt the split proposed by (Zhong et al. 2019): 238 training videos and 199 testing videos.

UCF-Crime is a large-scale dataset that contains 1,900 untrimmed real-world street and indoor surveillance videos with 13 classes of anomalous events and a total duration of 128 hours (Sultani, Chen, and Shah 2018). The training set contains 1,610 videos with video-level labels, and the test set contains 290 videos with frame-level labels.

XD-Violence is a large-scale dataset that contains 4,754 untrimmed videos with a total duration of 217 hours and collect from multiple sources, such as movies, sports, surveillances, and CCTVs (Wu et al. 2020). The training set contains 3,954 videos with video-level labels, and the test set contains 800 videos with frame-level labels.

Following previous works (Zhong et al. 2019; Wan et al. 2020), we use the AUC (Area Under Curve) of frame-level ROC (Receiver Operating Characteristic) as our metric for ShanghaiTech and UCF-Crime. Following previous works (Wu et al. 2020; Tian et al. 2021), we use the AP (Average Precision) as our metric for XD-Violence. Note that the larger the value of AUC and AP, the better the performance.

Implementation Details

We extract the 4,096D features from the fc_6 layer of the pre-trained C3D (Tran et al. 2015) on Sports-1M (Karpathy et al. 2014), the 1,024D features from the $mixed_{5c}$ layer of the pre-trained I3D (Carreira and Zisserman 2017) on Kinetics-400 (Kay et al. 2017), and the 1,024D features from the $Stage_4$ layer of the pre-trained VideoSwin (Liu et al. 2021c) on Kinetics-400. Following previous works (Tian et al. 2021), we divide each video into 32 snippets, that is, $T = 32$ and $K \in \{32, 16, 8, 4, 2, 1\}$. The length of each snippet is 16. Our MSLNet is trained using the SGD optimizer with a learning rate of 0.001, a weight decay of 0.0005 and a batch size of 64. We set E_1 to 100 and $E2$ to 400. Following (Tian et al. 2021), each mini-batch is composed of 32 randomly selected normal and abnormal videos. In abnormal videos, we randomly select one of the top 10% snippets as the abnormal snippet. In CTE, we set the number of headers to 12 and use DW Conv1D with kernel size is 3.

Method	Feature	Crop	AUC(%) ↑
MIL-Rank†	I3D RGB	one	85.33
GCN	C3D-RGB	ten	76.44
GCN	TSN-Flow	ten	84.13
GCN	TSN-RGB	ten	84.44
IBL	I3D-RGB	one	82.50
AR-Net†	C3D RGB	one	85.01
AR-Net	I3D Flow	one	82.32
AR-Net	I3D RGB	one	85.38
AR-Net	I3D-RGB+Flow	one	91.24
CLAWS	C3D-RGB	one	89.67
MIST	C3D-RGB	one	93.13
MIST	I3D-RGB	one	94.83
RTFM	C3D-RGB	ten	91.51
RTFM	I3D-RGB	ten	**97.21**
RTFM*	VideoSwin-RGB	ten	96.76
Ours	C3D-RGB	one	**94.23**
Ours	I3D-RGB	one	**95.45**
Ours	VideoSwin-RGB	one	**96.93**
Ours	C3D-RGB	ten	**94.81**
Ours	I3D-RGB	ten	96.08
Ours	VideoSwin-RGB	ten	**97.32**

Table 1: Compared with related methods on ShanghaiTech. The methods with † are reported by (Feng, Hong, and Zheng 2021) or (Tian et al. 2021). * indicates we re-train the method. Under the same feature, the highest result is bolded.

Results on ShanghaiTech

We report the results on ShanghaiTech (Zhong et al. 2019) in Table 1. For a fair comparison, we use two features: one-crop and ten-crop. One-crop means cropping snippets into the center. Ten-crop means cropping snippets into the center, four corners, and their flipped version (Zhong et al. 2019). Under the same backbone and crop, compared with the previous weakly supervised methods, our methods achieve the superior performance on AUC. For example, with the one-crop I3D-RGB feature, our model achieves an AUC of 95.45% and outperforms all other methods with the same crop, and with the ten-crop VideoSwin-RGB feature, our model achieves the best AUC of 97.32%.

Results on UCF-Crime

We report our experimental results on UCF-Crime (Sultani, Chen, and Shah 2018) in Table 2. Under I3D and VideoSwin as the backbone, our method outperforms all previous weakly supervised methods on the frame-level AUC metric. Under C3D as the backbone, our method has also achieved competitive result. For example, with the one-crop I3D-RGB feature, our model achieves an AUC of 85.30% and outperforms all other methods, and with the one-crop VideoSwin-RGB feature, our model achieves the best AUC of 85.62% which is higher than RTFM by 2.31%.

Results on XD-Violence

We report our results on XD-Violence (Wu et al. 2020) in Table 3. For a fair comparison, we use the same five-crop features with other methods. Five-crop means cropping snippets into the center and four corners. Under the same backbone, our method outperforms all previous weakly supervised VAD methods on the AP metric. For example, with five-crop I3D-RGB features, our model achieves an AP of 78.28% and outperforms all other methods, and with five-crop VideoSwin-RGB features, our model achieves an AP of 78.59% which higher than RTFM by 0.64%.

Method	Feature	Crop	AUC(%) ↑
MIL-Rank	C3D RGB	one	75.41
MIL-Rank†	I3D RGB	one	77.92
Motion-Aware	PWC-Flow	one	79.00
GCN	C3D-RGB	ten	81.08
GCN	TSN-Flow	ten	78.08
GCN	TSN-RGB	ten	82.12
IBL	C3D-RGB	one	78.66
CLAWS	C3D-RGB	ten	83.03
MIST	C3D-RGB	one	81.40
MIST	I3D-RGB	one	82.30
RTFM	C3D-RGB	ten	**83.28**
RTFM	I3D-RGB	ten	84.03
RTFM*	VideoSwin-RGB	one	83.31
Ours	C3D-RGB	one	82.85
Ours	I3D-RGB	one	**85.30**
Ours	VideoSwin-RGB	one	**85.62**

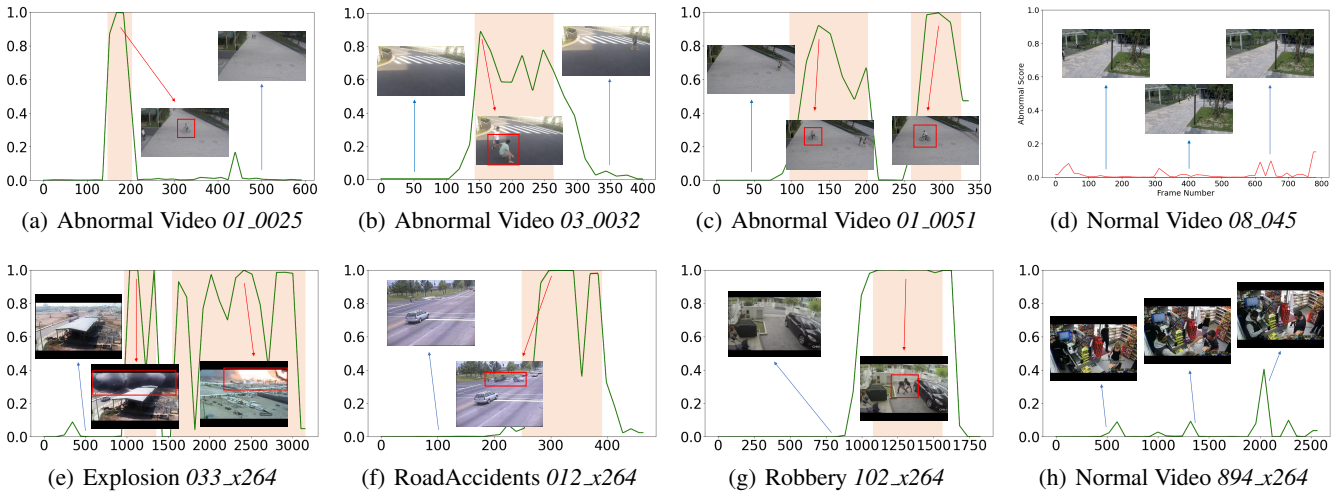
Table 2: Compared with other methods on UCF-Crime. The method with † is reported by (Tian et al. 2021). * indicates we re-train the method. Bold represents the best results.

Method	Feature	Crop	AP(%) ↑
MIL-Rank†	C3D RGB	five	73.20
MIL-Rank†	I3D RGB	five	75.68
Multimodal-VD	I3D-RGB	five	75.41
RTFM	C3D-RGB	five	**75.89**
RTFM	I3D-RGB	five	77.81
RTFM*	VideoSwin-RGB	five	77.95
Ours	C3D-RGB	five	75.53
Ours	I3D-RGB	five	**78.28**
Ours	VideoSwin-RGB	five	**78.59**

Table 3: Compared with related methods on XD-Violence. The methods with † are reported by (Wu et al. 2020) or (Tian et al. 2021). * indicates we re-train the method.

Complexity Analysis

Generally, Transformer has been often computationally expensive, but our method can achieve real-time surveillance. On an NVIDIA 2080 GPU, with VideoSwin (Liu et al. 2021c) as the backbone processes 3.6 snippets per second (a snippet has 16 frames), which is 57.6 frames per second (FPS); with I3D (Carreira and Zisserman 2017) as the backbone processes 6.5 snippets per second, which is 104 FPS. Our MSL Network can reach 156.4 forwards per second. Overall, the speed with VideoSwin as the backbone is 42 FPS, and the speed with I3D as the backbone is 63 FPS.



(a) Abnormal Video *01_0025* (b) Abnormal Video *03_0032* (c) Abnormal Video *01_0051* (d) Normal Video *08_045*

(e) Explosion *033_x264* (f) RoadAccidents *012_x264* (g) Robbery *102_x264* (h) Normal Video *894_x264*

Figure 3: Visualization of abnormal score curves. The horizontal axis represents the number of frames, and the vertical axis represents the abnormal scores. Videos of (a), (b), (c), and (d) are from the ShanghaiTech dataset, and videos of (e), (f), (g), and (h) are from the UCF-Crime dataset. The curves indicate the abnormal scores of the video frames, pink areas indicate that the interval contains an abnormal event, and the red rectangles indicate the location of abnormal events. Best viewed in color.

Basic Layer	ShanghaiTech	UCF-Crime
Transformer	96.51	85.41
CTE	**96.93 (+0.42)**	**85.62 (+0.21)**

Table 4: Compared with Transformer (Dosovitskiy et al. 2021), AUC(%) improvement brought by CTE on the ShanghaiTech and UCF-Crime datasets.

Score correction	ShanghaiTech	UCF-Crime
×	95.98	84.94
✓	**96.93 (+0.95)**	**85.62 (+0.68)**

Table 5: Performance improvement brought by the score correction method in the inference stage measured by AUC(%) on the ShanghaiTech and UCF-Crime datasets.

Qualitative Analysis

In order to further demonstrate the effect of our method, as shown in Figure 3, we visualize the anomaly score curves. The first row shows the ground truth and prediction anomaly scores of three abnormal videos and one normal video from the ShanghaiTech dataset. From the first row of Figure 3, we can see that our method can detect abnormal events in surveillance videos. Our method successfully predicts short-term abnormal events (Figure 3 (a)) and long-term abnormal events (Figure 3 (b)). Furthermore, our method can also detect multiple abnormal events in a video (Figure 3 (c)). The second row shows the ground truth and predicted anomaly scores of three abnormal videos and one normal video from the UCF-Crime dataset. From the second row of Figure 3, we can see that our proposed method can also detect abnormal events in complex surveillance scenes.

Ablation Analysis

In order to further evaluate our method, we perform ablation studies on the ShanghaiTech and UCF-Crime datasets with one-crop VideoSwin-RGB features.

Improvement brought by CTE. To evaluate the effect of our CTE, we replace CTE with the standard Transformer (Dosovitskiy et al. 2021). The dimension of the standard Transformer is the same as our CTE. Table 4 reports the results of this ablation experiment. Compared with the result using the standard Transformer as the basic layer, the result with CTE as the basic layer increases an AUC by 0.42% and 0.21% on the ShanghaiTech and UCF-Crime datasets.

Impact of score correction in the inference stage. As shown in Table 5, we conduct an experiment to report the performance improvement brought by the score correction method in the inference stage. From Table 5 we can observe that score correction can bring an AUC improvement of 0.95% and 0.68% with the one-crop features on the ShanghaiTech and UCF-Crime datasets, respectively.

Conclusion

In this work, we first propose an MSL method and a hinge-based MSL ranking loss. We then design a Transformer-based network to learn both video-level anomaly probability and snippet-level anomaly scores. In the inference stage, we propose to use the video-level anomaly probability to suppress the fluctuation of snippet-level anomaly scores. Finally, since VAD needs to predict the instance-level anomaly scores, by gradually reducing the length of selected sequence, we propose a self-training strategy to refine the anomaly scores. Experimental results show that our method achieves significant improvements on three public datasets.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62076192), Key Research and Development Program in Shaanxi Province of China (No.2019ZDLGY03-06), the State Key Program of National Natural Science of China (No.61836009), in part by the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT_15R53), in part by The Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No.B07048), in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the National Key Research and Development Program of China.

References

Cai, R.; Zhang, H.; Liu, W.; Gao, S.; and Hao, Z. 2021. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In *AAAI*, 938–946.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 4724–4733.

Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*, 1800–1807.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In *ICML*, volume 139, 2286–2296.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 14009–14018.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and van den Hengel, A. 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *ICCV*, 1705–1714.

Guo, Z.; Zhao, J.; Jiao, L.; Liu, X.; and Liu, F. 2021. A Universal Quaternion Hypergraph Network for Multimodal Video Question Answering. *IEEE Transactions on Multimedia*, 1–1.

He, C.; Shao, J.; and Sun, J. 2018. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools and Applications*, 77(22): 29573–29588.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, abs/1704.04861.

Jeong, J.; Lee, S.; and Kwak, N. 2020. Self-Training using Selection Network for Semi-supervised Learning. In *ICPRAM*, 23–32.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Li, F. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*, 1725–1732.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950.

Li, W.; Mahadevan, V.; and Vasconcelos, N. 2014. Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1): 18–32.

Li, Y.; Xing, R.; Jiao, L.; Chen, Y.; Chai, Y.; Marturi, N.; and Shang, R. 2019. Semi-Supervised PolSAR Image Classification Based on Self-Training and Superpixels. *Remote. Sens.*, 11(16): 1933.

Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; and Gool, L. V. 2021. LocalViT: Bringing Locality to Vision Transformers. *CoRR*, abs/2104.05707.

Liu, K.; and Ma, H. 2019. Exploring Background-Bias for Anomaly Detection in Surveillance Videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, 14901499.

Liu, X.; Li, K.; Zhou, M.; and Xiong, Z. 2011. Enhancing Semantic Role Labeling for Tweets Using Self-Training. In *AAAI*.

Liu, Y.; Sun, G.; Qiu, Y.; Zhang, L.; Chhatkuli, A.; and Gool, L. V. 2021a. Transformer in Convolutional Neural Networks. *CoRR*, abs/2106.03180.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *CoRR*, abs/2103.14030.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2021c. Video Swin Transformer. *CoRR*, abs/2106.13230.

Luo, W.; Liu, W.; and Gao, S. 2017. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In *ICCV*, 341–349.

Pang, G.; Yan, C.; Shen, C.; van den Hengel, A.; and Bai, X. 2020. Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection. In *CVPR*, 12170–12179.

Rosenberg, C.; Hebert, M.; and Schneiderman, H. 2005. Semi-Supervised Self-Training of Object Detection Models. In *WACV/MOTION*, 29–36.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*, 6479–6488.

Tai, K. S.; Bailis, P.; and Valiant, G. 2021. Sinkhorn Label Allocation: Semi-Supervised Classification via Annealed Self-Training. In *ICML*, volume 139, 10065–10075.

Tanha, J.; van Someren, M.; and Afsarmanesh, H. 2017. Semi-supervised self-training for decision tree classifiers. *Int. J. Mach. Learn. Cybern.*, 8(1): 355–370.

Tao, Y.; Zhang, D.; Cheng, S.; and Tang, X. 2018. Improving semi-supervised self-training with embedded manifold transduction. *Trans. Inst. Meas. Control*, 40(2): 363–374.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. *CoRR*, abs/2101.10030.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, 10347–10357.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wan, B.; Fang, Y.; Xia, X.; and Mei, J. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning. In *ICME*, 1–6. IEEE.

Wan, B.; Jiang, W.; Fang, Y.; Luo, Z.; and Ding, G. 2021. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. CvT: Introducing Convolutions to Vision Transformers. *CoRR*, abs/2103.15808.

Wu, P.; Liu, j.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *ECCV*.

Xu, W.; Xu, Y.; Chang, T. A.; and Tu, Z. 2021. Co-Scale Conv-Attentional Image Transformers. *CoRR*, abs/2104.06399.

Yan, H.; Li, Z.; Li, W.; Wang, C.; Wu, M.; and Zhang, C. 2021. ConTNet: Why not use convolution and transformer at the same time? *CoRR*, abs/2104.13497.

Yu, F.; Zhang, M.; Dong, H.; Hu, S.; Dong, B.; and Zhang, L. 2021. DAST: Unsupervised Domain Adaptation in Semantic Segmentation Based on Discriminator Attention and Self-Training. In *AAAI*, 10754–10762.

Zhang, J.; Qing, L.; and Miao, J. 2019. Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection. In *ICIP*, 4030–4034.

Zhang, Q.; and Yang, Y. 2021. ResT: An Efficient Transformer for Visual Recognition. *CoRR*, abs/2105.13677.

Zhao, B.; Fei-Fei, L.; and Xing, E. P. 2011. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 3313–3320.

Zheng, H.; Zhang, Y.; Yang, L.; Wang, C.; and Chen, D. Z. 2020. An Annotation Sparsification Strategy for 3D Medical Image Segmentation via Representative Selection and Self-Training. In *AAAI*, 6925–6932.

Zhong, J.; Li, N.; Kong, W.; Liu, S.; Li, T. H.; and Li, G. 2019. Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection. In *CVPR*, 1237–1246.

Zhu, Y.; and Newsam, S. D. 2019. Motion-Aware Feature for Improved Video Anomaly Detection. In *BMVC*, 270.