

# Real-world Video Anomaly Detection by Extracting Salient Features in Videos

Yudai Watanabe\*

Makoto Okabe

Yasunori Harada

Naoji Kashima

Shizuoka University

Chubu Electric Power Co., Inc.

watanabe.yudai@jp.panasonic.com, m.o@acm.org, {Harada.Yasunori,kashima.naoji}@chuden.co.jp

## Abstract

*We propose a lightweight and accurate method for detecting anomalies in videos. Existing methods used multiple-instance learning (MIL) to determine the normal/abnormal status of each segment of the video. Recent successful researches argue that it is important to learn the temporal relationships among segments to achieve high accuracy, instead of focusing on only a single segment. Therefore we analyzed the existing methods that have been successful in recent years, and found that while it is indeed important to learn all segments together, the temporal orders among them are irrelevant to achieving high accuracy. Based on this finding, we do not use the MIL framework, but instead propose a lightweight model with a self-attention mechanism to automatically extract features that are important for determining normal/abnormal from all input segments. As a result, our neural network model has 1.3% of the number of parameters of the existing method. We evaluated the frame-level detection accuracy of our method on three benchmark datasets (UCF-Crime, ShanghaiTech, and XD-Violence) and demonstrate that our method can achieve the comparable or better accuracy than state-of-the-art methods.*

## 1. Introduction

The number of surveillance cameras in the world is increasing every year, and they are used for crime prevention in cities and for safety confirmation in factories, power plants, and other large-scale facilities. However, since it is difficult for humans to see and confirm all of these videos, there is an urgent need to develop technology that enables artificial intelligence to analyze the videos and automatically detect abnormal events on behalf of humans. Since abnormal events are rarely observed, many methods have been proposed that use only the normal state as training data, and judge whether the input video is normal or abnormal based on the criterion of how much it deviates from the

learned normal state when inferring [13, 2, 7, 19, 8]. However, these methods can only detect abnormalities based on low-level features such as differences in the appearance and velocity in the video. Therefore, recently, a method has been proposed to train an anomaly detector using a weakly supervised dataset that contains both normal and abnormal videos [27, 28].

In the weakly supervised dataset, each video is labeled as normal or abnormal. That is, the videos labeled as normal contain only the normal state throughout all frames. On the other hand, the videos labeled as abnormal contain a mixture of normal and abnormal frames. By using such a dataset, it is no longer necessary to label each frame in the video as normal or abnormal, thus reducing the labeling effort.

Many existing methods treated multiple consecutive frames as a single short-term segment and used multiple instance learning (MIL) to determine normal/abnormal for each segment [27]. However, recent successful methods have argued that it is important to learn the temporal relationships among segments by taking all segments from the video as input together, rather than focusing on only a single segment [28]. To confirm this, we analyzed these methods on a dataset of video segments randomly sorted in time. We then found that while it is indeed important to train all segments together, the temporal order among them is irrelevant for high accuracy.

Based on this finding, we propose a novel model that does not rely on MIL, but instead takes all segments as input and has a self-attention mechanism to automatically extract features important for determining normal/abnormal from them. Despite the fact that the proposed neural network has 1.3% of the number of parameters of the existing method [28], the proposed method can achieve the comparable or better accuracy than state-of-the-art methods. We report the frame-level detection accuracy on three benchmark datasets (UCF-Crime [27], ShanghaiTech [18], and XD-Violence [33]).

In summary, this paper has three contributions:

- We analyzed the existing methods and found that while it is important to learn

\*Current affiliation: Panasonic System Networks R&D Lab. Co., Ltd.

all segments together, the temporal orders among them are irrelevant to achieving high accuracy.

- Based on this finding, we propose a lightweight and accurate method for video anomaly detection. Our method outperforms existing methods for the UCF-Crime dataset despite its lightweight model, a neural network with 1.3% trainable parameters. We also show that our method outperforms existing methods for the XD-Violence dataset when the model is extended by adding a bi-directional LSTM.
- A detailed analysis of our method was performed: visualization and observation of the attention map show that our self-attention mechanism works for extract salient features similarly to the top- $k$  strategy in existing methods. The relationship between hyperparameters and accuracy was also investigated in detail.

Our method is a simple model with a self-attention mechanism and appears to be a commonplace model. However, our technical contribution is that we arrived at such a lightweight model based on our observations and analysis of existing methods and found that this model is sufficient to achieve the comparable or better accuracy than state-of-the-art methods.

## 2. Related work

In the real world, most of what we can observe are normal states, and abnormal events are rarely observed. For this reason, many anomaly detection methods have been developed using unsupervised learning approaches that learn only the normal state. In inference, the input video is judged to be normal or abnormal based on how much it deviates from the learned normal state. The normal state can be learned using a set of mixture of dynamic textures models [13], a space-time Markov random field (MRF) model [12], Gaussian mixture models [2], and sparse dictionary learning [4, 17], etc. Deep learning approaches have been also proposed, such as a method for analyzing the temporal changes in the CNN features [23], autoencoder-based methods [34, 7, 19, 8, 11] and GAN-based methods [24, 26], etc. A method that uses multi-task learning has also been proposed [6]. However, these methods are basically only able to detect anomalies based on low-level features such as differences in the appearance and velocities of the video.

Recently, in order to develop higher-level anomaly detectors, a number of methods have been proposed to learn

anomaly detectors using weakly supervised datasets containing both normal and abnormal videos [27, 28, 33, 39, 40, 36, 5, 22]. Usually, to train a frame-by-frame anomaly detector, we need to label and train every frame of the video, which is expensive to label. Therefore, Sultani et al. proposed a weakly supervised dataset where not each frame but each video is labeled as normal or abnormal [27]. They also developed an anomaly detector by introducing MIL, which considers a video as a bag and selects the segment with the highest anomaly score from the bag for training.

Most of the anomaly detection methods using weakly supervised datasets are based on MIL. MIL-based methods have a problem that they have a negative impact on learning when they select normal parts of anomalous videos for learning. To cope with this problem, Zhong et al. proposed an approach that considers the weakly supervised dataset as a dataset containing incorrect labels, and uses graph convolutional neural networks to correct the incorrect labels and learn from them [40]. Tian et al. proposed a top- $k$  strategy that calculates the difference in anomaly scores between segments and selects the top  $k$  segments with the highest scores for training [28]. Sapkota et al. proposed a model that does not require the selection of the parameter  $k$  by introducing distributionally robust optimization [25]. Zaheer et al. does not rely on MIL-based approach but proposed a method for detecting anomalies using global features of the entire video and local features of each segment using the attention mechanism and per-video clustering loss [36, 37]. These recent successful methods achieve high detection accuracy by efficiently using the features of multiple segments in the video instead of just a single segment.

Recently, Zaheer et al. proposed a method for unsupervised video anomaly detection and reported significant improvements over existing unsupervised methods through experiments on the benchmark datasets (UCF-Crime and ShanghaiTech) [38]. Acsintoae et al. proposed UBnormal, a new benchmark dataset for supervised video anomaly detection consisting of virtual scenes and annotated at the pixel level [1].

## 3. Method

We propose a lightweight and accurate learning method for detecting anomalies in videos. The proposed method analyzes the entire video and automatically extracts and learns the features that are important for determining normal/abnormal.

Let  $\mathcal{D} = \{(\mathbf{V}_i, y_i)\}$  be the dataset. where  $\mathbf{V}_i$  is the  $i$ th video in the training dataset and  $y_i$  is the label attached to  $\mathbf{V}_i$ .  $y_i = \{0, 1\}$ , 0 indicates normal, and 1 indicates abnormal. In a video labeled as normal, only the normal state is recorded in all frames. On the other hand, the video labeled as abnormal contains a mixture of frames with normal and abnormal states.  $\mathbf{V}_i$  is divided into  $T$  segments, and each

segment is converted into a  $D$ -dimensional feature vector  $\mathbf{F}_{i,j}$  by the feature extractor:  $\mathbf{F}_{i,j}$  represents the  $j$ th feature vector of  $\mathbf{V}_i$ ; the feature extractor used throughout all experiments was I3D [3], which has been trained on the Kinetics dataset.

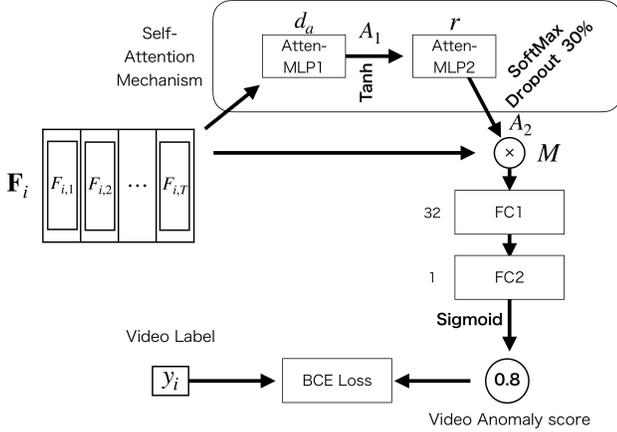


Figure 1. An overview diagram of our method

Our method is a simple model that consists of a self-attention mechanism and two fully connected layers (Figure 1). In the self-attention mechanism, the input  $\mathbf{F}_i$  is transformed by Atten-MLP1 multilayer perceptron into a  $d_a \times T$  matrix  $\mathbf{A}_1$ . At this time, the Tanh function is used for activation.  $\mathbf{A}_1$  is converted to  $\mathbf{A}_2$ , a matrix of  $r \times T$ , by Atten-MLP2 multilayer perceptron. At this time, the softmax function is used for activation. In addition, a dropout regularization of 30% is also performed.

Using the weight matrix  $\mathbf{A}_2$  obtained from the self-attention mechanism, we calculate  $\mathbf{M} = \mathbf{F}_i \mathbf{A}_2^t$ . After reshaping  $\mathbf{M}$  into a vector of  $D \times r$  dimensions, it is transformed into the anomaly score through two fully connected layers FC1 (32 units) and FC2 (1 unit). The activation function of FC1 is an identity function, and that of FC2 is a sigmoid function. The binary cross entropy (BCE) function was used as the loss function.

Note that our self-attention mechanism does not deal with temporal orders. As we see in the operations through Atten-MLP1 and Atten-MLP2, the vector in column  $j$  of the attention map  $\mathbf{A}_2$  depends only on the feature vector  $\mathbf{F}_{i,j}$  in column  $j$ . In other words, shuffling the input feature vectors  $\mathbf{F}_{i,j}$  in the temporal direction (column direction) does not change the video anomaly score, which is the output of the model.

### 3.1. Motivation of our method

Several existing methods take into account the temporal relationships between segments in a video [27, 39, 28, 36, 37]. Sultani et al. introduced a term in the loss function to impose continuity of the anomaly score, as the

anomaly score should vary continuously in the video [27]. Zaheer et al. also used the similar term in their loss function [36, 37]. Tian et al. introduced a multi-scale temporal network (MTN) to capture local and global temporal features [28]. All of these methods achieve high accuracy in anomaly detection. Therefore, we investigated how the mechanism for capturing the temporal relationship between segments contributes to the high accuracy.

Specifically, in  $\mathbf{F}_i$ , the set of feature vectors obtained from the training video  $\mathbf{V}_i$ , the feature vectors are ordered by default as  $\{\mathbf{F}_{i,1}, \mathbf{F}_{i,2}, \dots, \mathbf{F}_{i,T}\}$ , but we randomly rearranged this order to create a new dataset and used it for training. The results of the experiments using the UCF-Crime dataset [27] are shown in Table 1. In both methods [27, 28], there was no degradation in accuracy due to random reordering of feature vectors.

Table 1. Comparison of AUC performance using the UCF-Crime dataset [27], where the feature vectors of each video are randomly reordered. Here, we used not C3D but I3D with 10-crop augmentation for Sultani et al.’s method [27]

Method	Reorder	AUC(%)
Sultani et al. [27]		81.39
	✓	81.54
RTFM [28]		84.30
	✓	84.26

This result indicates that capturing the temporal order between segments does not contribute to the accuracy of the anomaly detector.

The temporal smoothness term introduced by Sultani et al. minimizes the difference in anomaly scores between temporally adjacent segments [27]. However, from the above results, we infer that this term has the regularization effect of ensuring that all segments in a video have similar anomaly scores, rather than constraining the temporal order between segments. Zaheer et al. reported a decrease in accuracy without this term [36, 37], which may indicate the importance of the regularization effect that the temporal smoothness term brings. Although the MTN and top- $k$  strategies used in Tian et al.’s method [28] are more complex, and we would expect them to analyze more complex temporal relationships than just temporal order, the above results indicate that temporal order is not important. Based on the above observations, we hypothesized that the high anomaly detection accuracy achieved by these methods is due to the fact that they have a mechanism that allows all segments from the video to be trained together and extracts salient features that are important for determining normal/abnormal.

Our method (Figure 1) is designed based on the above insights. Our method is not an MIL framework but it is a

model that takes all segments in a video as input and determines whether the video is normal or abnormal. Since we do not use MIL, the extraction of salient features can be achieved with a simple self-attention mechanism. For the self-attention mechanism, we introduce a model inspired by Lin et al.’s method [14]: their method targets sentence classification and can deal with variable length input. We adopted this mechanism because we divide the video  $\mathbf{V}_i$  into  $T$  segments during training, but the number of segments during inference should be different dependent on the length of the input video. Our method is accurate and also lightweight because it does not have any mechanism to capture temporal orders.

Our self-attention mechanism is similar in concept to that of Claws [36, 37], but there are some differences between them. Claws generates the attention map twice but our method generates it once. While Claws’ paper states that the purpose of introducing the self-attention mechanism is to suppress normalcy, our purpose is to extract salient features that are important in determining normality/abnormality.

### 3.2. Inference

Let  $\mathbf{V}^e$  be the video to which we want to apply our method for anomaly detection. First, we divide  $\mathbf{V}^e$  into  $N$  segments  $\{\mathbf{V}_1^e, \mathbf{V}_2^e, \dots, \mathbf{V}_N^e\}$  using 16 consecutive frames as one segment. Each segment  $\mathbf{V}_i^e$  is converted by I3D [3] into a  $D$ -dimensional feature vector  $\mathbf{F}_i^e$ . Let the set of feature vectors be  $\mathbf{F}^e = \{\mathbf{F}_1^e, \mathbf{F}_2^e, \dots, \mathbf{F}_N^e\}$ . Let  $l$  be the split size and we divide  $\mathbf{F}^e$  into  $m = N/l$  bags for every  $l$  segments:

$$\mathbf{F}^e = \{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_m\} = \{\{\mathbf{F}_1^e, \dots, \mathbf{F}_l^e\}, \{\mathbf{F}_{l+1}^e, \dots, \mathbf{F}_{2l}^e\} \dots \{\mathbf{F}_{N-l+1}^e, \dots, \mathbf{F}_N^e\}\}.$$

Next, we input  $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_m$  one by one into our anomaly detector, and we obtain the inference result  $\mathbf{S}^v = \{s_1^v, s_2^v, \dots, s_m^v\}$ .  $s_i^v$  is used as the anomaly score for each segment of  $\{\mathbf{V}_{(i-1)l+1}^e, \dots, \mathbf{V}_{il}^e\}$ . When it is necessary to produce the results of frame-level anomaly detection, the obtained scores are assigned to all frames in each segment.

## 4. Experiments

To evaluate the proposed method, we conducted experiments on three weakly supervised datasets for anomaly detection: UCF-Crime dataset [27], ShanghaiTech dataset [18], and XD-Violence dataset [33]. We compare the detection accuracy with several existing methods.

### 4.1. Dataset and evaluation measure

**UCF-Crime dataset** is a large dataset of real-world surveillance video [27]. This dataset contains 13 types of

anomalies. The total number of videos is 1900, and the total duration of the videos is 128 hours. 1610 of the 1900 videos are for training and 290 are for testing. Each training video is labeled as normal/abnormal on a per-video basis. In each test video, each frame is labeled as normal or abnormal.

**ShanghaiTech dataset** is a medium-sized dataset of videos captured by fixed cameras installed in the university. The dataset contains 437 videos captured by 13 fixed cameras. Of the 437 videos, 307 are normal videos and 130 are videos with anomalies. The original dataset was intended for the development of an anomaly detector based on unsupervised learning [18]. However, Zhong et al. labeled each video so that it could be used as a dataset for weakly supervised learning [40]. We constructed a weakly supervised dataset using the same procedure as Zhong et al. [40] and conducted experiments.

**XD-Violence dataset** is a large dataset containing various types of videos, such as movies and videos from video sharing sites. The total number of videos is 4754, and the total duration of the videos is 217 hours. 2405 of the 4754 videos contain 6 types of anomalies: *fighting*, *shooting*, *riot*, *abuse*, *explosion*, and *car accident*. 2349 of the 4754 videos are normal. As in UCF-Crime dataset, each training video is labeled with a per-video label, and each test video is labeled with a per-frame label.

**Evaluation measure:** for UCF-Crime and ShanghaiTech datasets, we used the Area Under the Curve (AUC) with the Receiver Operating Characteristic (ROC) curve, which is calculated based on the frame-level anomaly detection accuracy, as in the existing studies [27, 28, 39, 41, 40, 36, 33, 5, 8, 7]; a larger AUC value indicates a more accurate anomaly detector. For XD-Violence dataset, we used Average Precision (AP) as the evaluation measure, as in the existing study [33, 28]; a larger AP value indicates a more accurate anomaly detector.

### 4.2. Implementation detail

Our method was developed and evaluated using PyTorch [21]. Radam [16] was used as the optimization algorithm, and the learning rate was set to 0.001. The batch size was set to 64. As in the existing method [27], mini-batches were created so that each mini-batch contained an equal number of normal and abnormal videos, i.e., 32 normal videos and 32 abnormal videos in our case. The hyperparameter  $T$  was set to 32. For UCF-Crime and ShanghaiTech datasets, as in the existing method [28, 5], 10-crop augmentation was performed for each video. For XD-Violence dataset, we performed 5-crop augmentation for each video as in the existing method [33].

### 4.3. Analysis on Our Model

We have experimented with adding the bi-directional long short-term memory (LSTM) to our model and re-

moving the self-attention mechanism from our model. In the context of sentence classification, the model of Lin et al. [14] achieves high classification accuracy, where the self-attention mechanism is used after the bi-directional LSTM. Inspired by this, we also extended our model by adding the bi-directional LSTM and evaluated the accuracy of anomaly detection. Specifically, our model proposed in Section 3 directly inputs  $F_i$  into the Atten-MLP1 (Figure 1). In the model with the bi-directional LSTM,  $F_i$  is first input to the bi-directional LSTM and the output from it is input to the Atten-MLP1. The dimension of the hidden layer of the LSTM was set to 256.

The experimental results are shown in Table 2. For UCF-Crime and ShanghaiTech datasets, we used the features obtained by I3D, while for XD-Violence dataset, since it contains audio information, we used the features obtained by VGGish [10] in addition to I3D. For UCF-Crime and ShanghaiTech datasets, the model using only the self-attention mechanism without the bi-directional LSTM (the model proposed in Section 3) achieved the highest detection accuracy. For XD-Violence dataset, the model with both bi-directional LSTM and self-attention mechanism achieved the highest detection accuracy. This may be due to the fact that each video in XD-Violence dataset contains audio information, and the bi-directional LSTM may have worked effectively for audio information.

Table 2. Results of experiments on the addition of the bi-directional LSTM (BL) to our model and the removal of the self-attention mechanism (SA) from our model

BL	SA	UCF-Crime	ShanghaiTech	XD-Violence
✓		81.72	92.90	75.92
	✓	<b>84.91</b>	<b>95.72</b>	75.46
✓	✓	83.28	94.06	<b>82.89</b>

#### 4.4. Results on UCF-Crime

Table 3 shows the frame-level AUC performance on the UCF-Crime dataset. The split size  $l$  is set to 28. Compared to MIST [5] and Wu et al. [33], which use the same I3D RGB features, our method achieves higher detection accuracy. Although our method is a simple model, it achieves the highest detection accuracy 84.91% and this is 0.61% higher than RTFM [28], which has the highest accuracy among the existing methods.

#### 4.5. Results on ShanghaiTech

Table 4 shows the frame-level AUC performance on the ShanghaiTech dataset. The split size  $l$  is set to 21. Our detection accuracy is 95.72% and this is 1.49% inferior to RTFM [28], which has the highest accuracy among the existing methods. However, our method achieves the second

Table 3. Comparison of frame-level AUC performance on the UCF-Crime dataset. **Blue** is the highest value and **red** is the second highest value

Supervision	Method	Feature Type	AUC(%)
One-class classifier	SVM Baseline	-	50.00
	Conv-AE [8]	-	50.60
	Lu et al. [17]	-	65.51
	BODS [30]	-	68.26
	GODS [30]	-	70.46
Supervised	NLN [31]	NLN RGB	78.9
	Lin et al. [15]	C3D RGB	70.1
	Lin et al. [15]	NLN RGB	82.0
Weakly Supervised	Sultani et al. [27]	C3D RGB	75.41
	Zhang et al. [39]	C3D RGB	78.66
	Motion-Aware [41]	PWC Flow	79.00
	GCN-Anomaly [40]	TSN RGB	82.12
	CLAWS Net [36]	C3D RGB	83.03
	CLAWS Net+ [37]	C3D RGB	83.37
	CLAWS Net+ [37]	3DResNext	84.16
	Wu et al. [33]	I3D RGB	82.44
	MIST [5]	I3D RGB (Fine)	82.30
	RTFM [28]	C3D RGB	83.28
	RTFM [28]	I3D RGB	84.30
	Our ( $d_a=64, r=3$ )	I3D RGB	<b>84.74</b>
	Our ( $d_a=128, r=7$ )	I3D RGB	<b>84.91</b>

highest value and outperforms the other existing methods. In addition, the accuracy of more than 95% is achieved, indicating that a sufficiently practical anomaly detector can be trained.

Table 4. Comparison of frame-level AUC performance on the ShanghaiTech dataset. **Blue** is the highest value and **red** is the second highest value

Supervision	Method	Feature Type	AUC(%)
One-class classifier	Conv-AE [8]	-	60.85
	Frame-Pred [32]	-	73.40
	Mem-AE [7]	-	71.20
	VEC [35]	-	74.80
Weakly Supervised	GCN-Anomaly [40]	TSN RGB	84.44
	Zhang et al. [39]	I3D RGB	82.50
	CLAWS Net [36]	C3D RGB	89.67
	CLAWS Net+ [37]	C3D RGB	90.12
	CLAWS Net+ [37]	3DResNext	91.46
	AR-Net [29]	I3D RGB&Flow	91.24
	MIST [5]	I3D RGB (Fine)	94.83
	RTFM [28]	C3D RGB	91.51
	RTFM [28]	I3D RGB	<b>97.21</b>
	Our ( $d_a=64, r=3$ )	I3D RGB	<b>95.72</b>

#### 4.6. Results on XD-Violence

Table 5 shows the frame-level AP performance on the XD-Violence dataset. The split size  $l$  is set to 9. When using only I3D RGB features, our method was inferior to Wu et al. [33] and RTFM [28]. Since XD-Violence dataset contains audio information, we can also use audio features (VGGish) [10] in addition to I3D RGB features as Wu et al. [33] and Pang et al. [20] did. When using the audio features, as mentioned in Section 4.3, we extend our model by adding the bi-directional LSTM in front of the self-attention mechanism, which is represented as “Ours†” in Table 5. This extended model achieves the highest detection accuracy 82.89%: this is 1.2% higher than that of Pang et al. [20], which is a method dedicated to anomaly detection based on multimodal information.

Table 5. Comparison of frame-level AP performance on XD-Violence dataset. **Blue** is the highest value and **red** is the second highest value. “Ours†” represents the model that adds bi-directional LSTM to our original model

Supervision	Method	Feature Type	AP(%)
One-class classifier	SVM baseline	-	50.78
	Hasan et al. [9]	-	30.77
Weakly Supervised	Sultani et al. [27]	C3D RGB	73.20
	Wu et al. [33]	I3D RGB	75.68
	RTFM [28]		77.81
	Ours ( $d_a=64, r=3$ )		73.25
	Wu et al. [33]	I3D RGB +VGGish	78.64
	Pang et al. [20]		<b>81.69</b>
	Ours ( $d_a=64, r=3$ )		75.46
	Ours† ( $d_a=64, r=3$ )		79.92
Ours† ( $d_a=128, r=1$ )	<b>82.89</b>		

#### 4.7. Comparison of the number of trainable parameters

Table 6 shows the number of parameters that can be trained in the model for each method. Our method is an extremely lightweight model with a much smaller number of parameters than the existing methods. Even though the number of parameters is only 1.3% of RTFM [28] when  $d_a = 64$  and  $r = 3$ , our method achieves higher accuracy than RTFM [28] in Table 3, and achieves a comparable high accuracy in Table 4. It is also the lightest model among the existing methods even when  $d_a = 128$  and  $r = 7$ , which achieves even higher accuracy.

#### 4.8. Visualization of Attention Map

It was described in Section 3 that our method does not use the MIL framework and so the self-attention mechanism can be used for extracting salient features. Therefore, we visualize  $A_2$ , the attention map, to investigate whether

Table 6. Comparison of the number of trainable parameters

Method	Number of Parameters
Sultani et al. [27]	2,114,113
Wu et al. [33]	769,155
RTFM [28]	24,718,849
Ours ( $d_a = 64, r = 3$ )	328,004
Ours ( $d_a = 128, r = 7$ )	721,992

such salient features are really extracted. Here we visualize  $A_2$  for “Burglary030”, which is one of the training videos in the UCF-Crime dataset. Figure 2 shows the heatmap of  $A_2$  for the 50-th iteration, the 500-th iteration, and the highest detection accuracy. Note that one iteration here means training on one mini-batch. As the training progresses, the weights are concentrated on specific segments, indicating that the self-attention mechanism is automatically learning the segments of interest. The fact that the weights are concentrated on multiple segments instead of one indicates that the self-attention mechanism has an effect similar to the top- $k$  strategy of RTFM [28].

#### 4.9. Performance on each anomaly class

Figure 3 shows the AUC performance of the proposed method for each anomaly class on the UCF-Crime dataset. Our method achieves higher or comparable detection accuracy compared to RTFM [28] for six classes of anomalies: Abuse, Arrest, Assault, Explosion, RoadAccidents, and Stealing. In particular, for the four classes of anomalies, Abuse, Assault, Explosion, and Stealing, the detection accuracy was improved by more than 6%. Since these four classes of anomalies cannot be detected without long-term analysis of the motion of objects and people, we suspect that these results indicate that the proposed method successfully captures the long-term features of the videos. For the four classes of Arrest, Arson, RoadAccidents, and Shooting, our method performed comparably to RTFM. For the four classes of Burglary, Fighting, Robbery, and Shoplifting, our method was inferior to RTFM. Our method may be still difficult to classify instantaneous anomalies. In addition, RTFM seems to be a method that is good at capturing features on human movement.

#### 4.10. Analysis on the hyperparameters $d_a$ and $r$

We defined  $d_a$  and  $r$  as hyperparameters in Section 3. By changing the hyperparameters, the number of trainable parameters changes, and the detection accuracy also changes. Therefore, using the UCF-Crime dataset, we investigated how changes in hyperparameters affect the detection accuracy. The split size  $l$  was set to 32. The results are shown in Figure 4. The points surrounded in red represent the highest detection accuracy for each  $d_a$ . Focusing only on  $r$ , if  $r$  is larger than 3, detection accuracy is almost stable and its

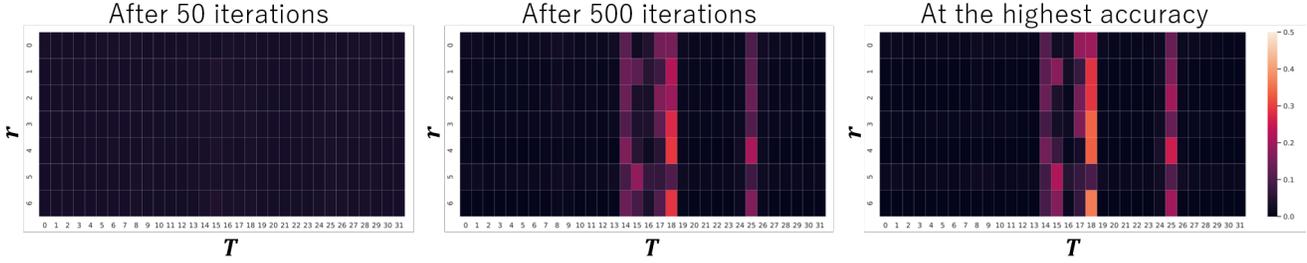


Figure 2. Visualization of the attention map  $A_2$

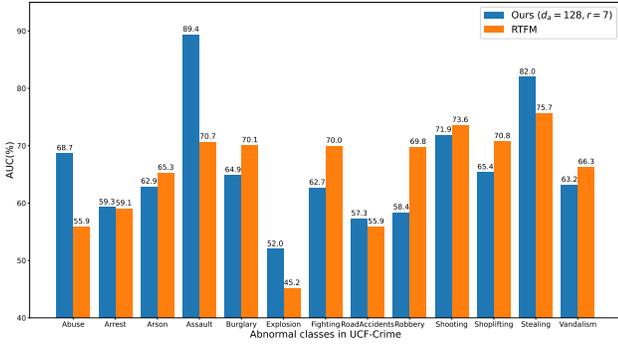


Figure 3. AUC performance by anomaly classes in UCF-Crime dataset

variation is small. Focusing on the relationship between  $d_a$  and  $r$ , if  $d_a$  is increased,  $r$  should also be increased to some extent to obtain better detection accuracy. However, if  $r$  is made too large, the detection accuracy will decrease. Up to 256, better detection accuracy could be obtained by increasing  $d_a$ , but when  $d_a$  was increased to 512, the overall detection accuracy decreased regardless of the value of  $r$ . This may be due to overfitting caused by making the model too large.

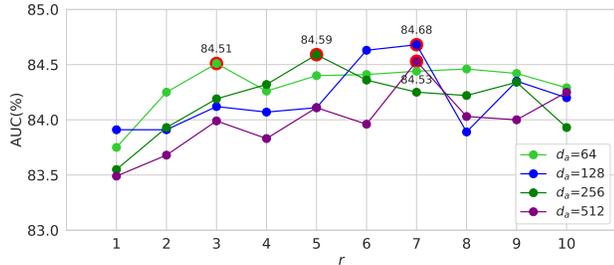


Figure 4. Relationship between hyperparameters ( $d_a$  and  $r$ ) and AUC performance on the UCF-Crime dataset

We also investigated the influence of hyperparameters on the detection accuracy using the XD-Violence dataset. Here we use the extended model with the bi-directional LSTM and set the split size  $l$  to 32. The results are shown in Figure 5. It can be seen that good detection accuracy is

achieved when  $r = 1$ , regardless of the value of  $d_a$ . In the case of the model with the bi-directional LSTM, usage of a model that is too large for the self-attention mechanism may cause overfitting.

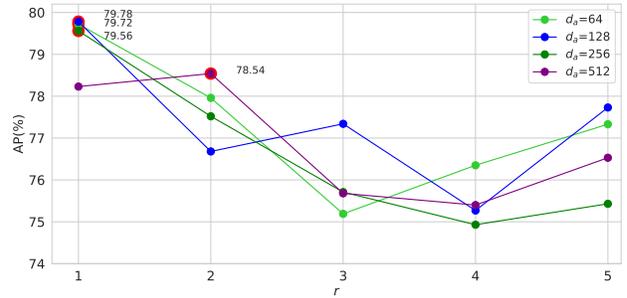


Figure 5. Relationship between hyperparameters ( $d_a$  and  $r$ ) and AP performance on the XD-Violence dataset

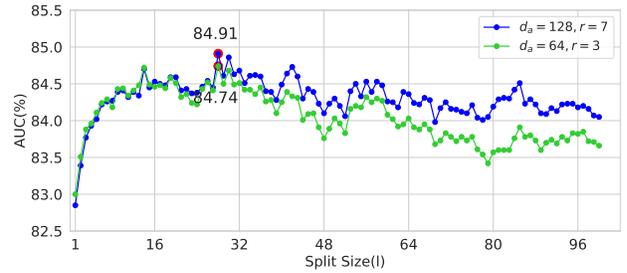


Figure 6. Relationship between the split size  $l$  and AUC performance during inference on the UCF-Crime dataset

#### 4.11. Analysis on the split size $l$

We defined the split size  $l$  in Section 3.2. Since our method divides  $N$  feature vectors into  $m = N/l$  bags during inference, each bag contains  $l$  feature vectors. Since the detection accuracy varies depending on the split size  $l$ , we used the UCF-Crime dataset to investigate how the change in  $l$  affects the detection accuracy. The result for  $d_a = 64$  and  $r = 3$  and the result for  $d_a = 128$  and  $r = 7$  are shown in Figure 6. The points surrounded in red represent the highest detection accuracy for each hyperparameter set. The de-

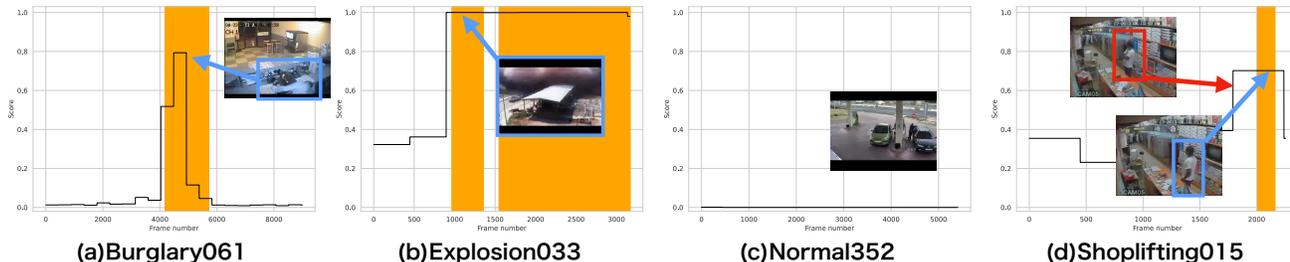


Figure 7. Visualization of the anomaly scores of our method. Black lines show the transition of anomaly scores. Orange blocks indicate ground truth. Blue arrows indicate correctly detected anomalies. Red arrows indicate incorrectly detected anomalies

tection accuracy is increased until  $l$  is around 16. Since our method analyzes the entire video, it requires a certain length of the video, which means that a certain split size is necessary. On the other hand, when the split size  $l$  is larger than 16, the detection accuracy is stable, indicating that our method can analyze the whole video efficiently if the video is long enough.

#### 4.12. Qualitative Analysis

Figure 7 shows the transition of anomaly scores predicted by our method for the videos of “Burglary061”, “Explosion033”, “Normal352”, and “Shoplifting015” in the UCF-Crime datasets. The hyperparameters  $d_a$  and  $r$  are set to 128 and 7, respectively, and the split size  $l$  is set to 16. Our method successfully detects abnormal frames when a burglar is breaking the window glass and stealing items as shown in Figure 7-a, or abnormal frames when an explosion accident happens and then dust is flying as shown in Figure 7-b. Our method correctly detects anomalies that are occurring for a long period of time. Figure 7-c shows a video at a gas station that does not contain any anomalies, and our method does not cause any false positives. Figure 7-d is an failure case. This is a video of a man shoplifting items placed on a counter. Although our method is able to detect anomalies, there are moments when false positives are detected. As described in Section 4.9, our method is not good at detecting short-term anomalies. For a instantaneous anomaly such as shoplifting, the frames before and after the anomaly may be identified as anomalies in addition to the moment when the anomaly actually occurred.

### 5. Conclusion

We have proposed a lightweight and accurate weakly supervised learning method for anomaly detection from video. Since MIL is not used, the extraction of salient features can be achieved with a simple self-attention mechanism. We show that the proposed model is simple and lightweight, yet achieves the comparable or better accuracy than state-of-the-art methods.

### References

- [1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnorm: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20143–20153, June 2022.
- [2] Arslan Basharat, Alexei Gritai, and Mubarak Shah. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [4] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456, 2011.
- [5] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14009–14018, 2021.
- [6] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12742–12752, 2021.
- [7] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.
- [8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.
- [9] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal reg-

- ularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.
- [10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.
- [11] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.
- [12] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, 2009.
- [13] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014.
- [14] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
- [15] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, page 1490–1499, New York, NY, USA, 2019. Association for Computing Machinery.
- [16] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.
- [17] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.
- [18] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.
- [19] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1273–1283, 2019.
- [20] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2260–2264, 2021.
- [21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.
- [22] Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 173–183, October 2021.
- [23] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698, 2018.
- [24] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581, 2017.
- [25] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2188–2196. PMLR, 13–15 Apr 2021.
- [26] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.
- [27] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.
- [28] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.
- [29] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
- [30] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019.
- [31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [32] Dongze Lian Wen Liu, Weixin Luo and Shenghua Gao. Future frame prediction for anomaly detection - - a new baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2018*, pages 6536–6545, 2018.
- [33] 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.
- [34] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017.
- [35] Guang Yu, Siqu Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020.
- [36] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*, pages 358–376. Springer, 2020.
- [37] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Clustering aided weakly supervised training to detect anomalous events in surveillance videos. *arXiv preprint arXiv:2203.13704*, 2022.
- [38] M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14744–14754, June 2022.
- [39] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019.
- [40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019.
- [41] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.