MGFN : Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection

Yingxian Chen[1], Zhengzhe Liu[2], Baoheng Zhang[1], Wilton Fok[1], Xiaojuan Qi[1], Yik-Chung Wu[1]

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong
[2]The Chinese University of Hong Kong
{carolcyx, cheungbh}@hku.hk, {wtfok, xjqi, ycwu}@eee.hku.hk, zzliu@cse.cuhk.edu.hk

Abstract

Weakly supervised detection of anomalies in surveillance videos is a challenging task. Going beyond existing works that have deficient capabilities to localize anomalies in long videos, we propose a novel glance and focus network to effectively integrate spatial-temporal information for accurate anomaly detection. In addition, we empirically found that existing approaches that use feature magnitudes to represent the degree of anomalies typically ignore the effects of scene variations, and hence result in sub-optimal performance due to the inconsistency of feature magnitudes across scenes. To address this issue, we propose the Feature Amplification Mechanism and a Magnitude Contrastive Loss to enhance the discriminativeness of feature magnitudes for detecting anomalies. Experimental results on two large-scale benchmarks UCF-Crime and XD-Violence manifest that our method outperforms state-of-the-art approaches.

Introduction

Detection of anomalies in surveillance videos is an important research topic (Chalapathy and Chawla 2019; Guansong et al. 2021). The abnormal events, such as accidents, fighting, arson, and banditry, may cause social harm and even threaten human life (Gopalakrishnan 2012).

Unfortunately, it is extremely challenging to identify and locate anomalies in a long video (Antić and Ommer 2011; Hasan et al. 2016; Kratz and Nishino 2009; Zhao, Li, and Xing 2011). Firstly, "anomaly" is a relative term defined against "normality". Hence, as shown in Fig. 1, it is unreasonable to predict the anomaly based upon only a single or a few nearby frames without the necessary knowledge of the "normality". In addition, "anomalies" include a variety of events mentioned above, so it is challenging to construct a unified representation for all the varying events. Besides, it is tedious and laborious to prepare for the frame-level annotation. Therefore, network training is more feasible under video-level weak supervisions.

To mitigate the above challenges, existing methods can be roughly cast into two branches. One stream learns to detect anomalies by constructing spatial-temporal architectures (Feng, Hong, and Zheng 2021; Zhu and Newsam 2019; Tian et al. 2021; Wu et al. 2020; Wu and Liu 2021).

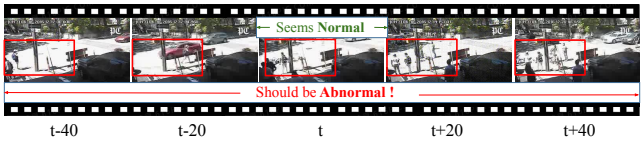

Figure 1: Long-term temporal context information plays a significant role in surveillance video-based anomaly detection. It seems nothing happens in frame t, but the opposite conclusion can be easily derived considering other frames.

Although such models work well for short videos, they typically struggle in tackling long videos where anomalous frames occupy only a small portion, due to the lack of global context awareness and specific focus on the abnormal frames.

The second branch learns to distinguish anomaly from normal by designing loss functions (Tian et al. 2021; Zaheer et al. June 2020; Feng, Hong, and Zheng 2021; Zaheer et al. 2020). For example, the most recent work (Tian et al. 2021) proposes a Robust Temporal Feature Magnitude (RTFM) loss to push abnormal feature magnitudes to be larger and normal ones in the opposite direction. It might be true that within the same video sequence or in similar scenes, abnormal features attain larger magnitudes than normal ones. However, we empirically found that besides the anomaly, the feature magnitude also depends on other attributes of the video such as the object movements, number of objects and persons in the scene, etc. As shown in Fig. 2 (a), the feature magnitudes in the normal video (the below) with substantial object movement are even larger than the abnormal one (the above).

Therefore, it is unreasonable to simply encourage the abnormal features to be larger and normal features to be smaller using the RTFM loss (Tian et al. 2021), since this target is not consistent with the inherent inter-video magnitude distribution and thus harms network training. We further found that even in the same video sequence, as shown in Fig. 2 (b), some normal features learned in (Tian et al. 2021) (green points outside the red bounding box) attain similar or even larger magnitudes compared with abnormal ones (green points in the red bounding box). In addition, the unsatisfying feature separability (Fig. 2 (c)) indicates that

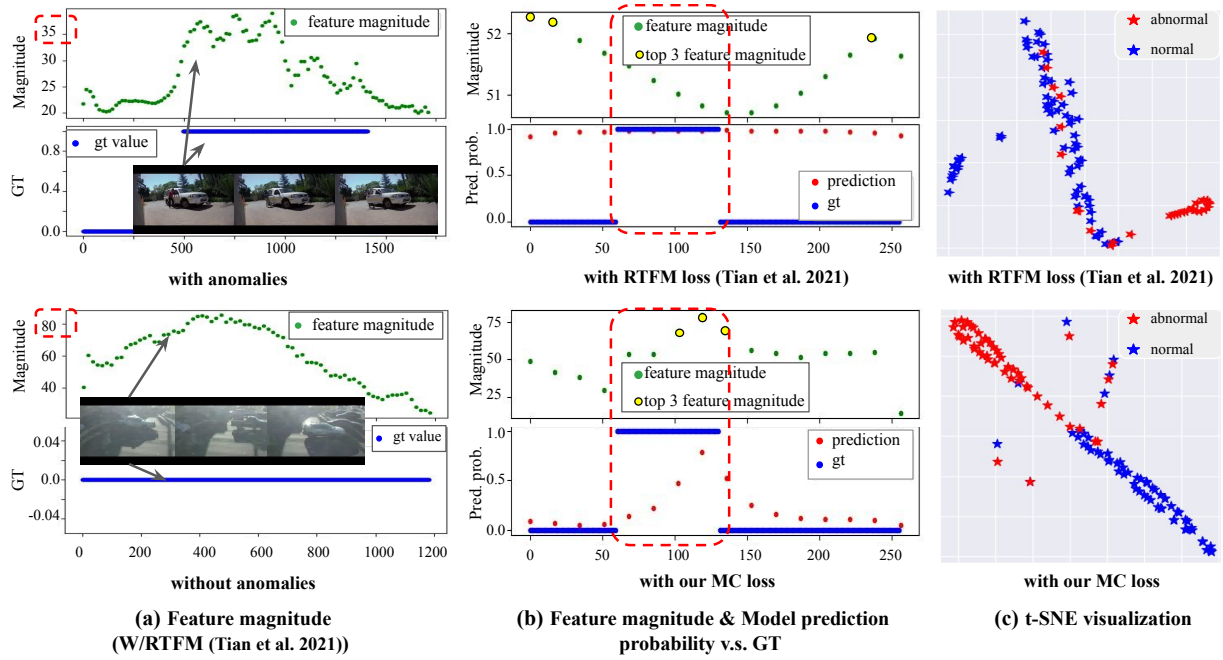

Figure 2: (a) With RTFM loss (Tian et al. 2021), normal feature magnitudes can be larger than abnormal ones due to the various scene attributes. (b) Our MC loss can generate more consistent feature magnitudes with anomalies. (c) Based on 100 random samples, our MC loss achieves the better separability of normal and abnormal features.

RTFM (Tian et al. 2021) cannot effectively separate normal and abnormal features.

To address the aforementioned issues, we propose a Magnitude-Contrastive Glance-and-Focus Network (MGFN) for anomaly detection. Inspired by global-to-local information integration mechanism in the human vision system for detecting anomalies in a long video, MGFN first glances the whole video sequence to capture long-term context information, and then further addresses each specific portion for anomaly detection. Different from existing works that simply fuse the spatial-temporal features, our strategy helps the network to first get an overview of the scene, and then detect scene-adaptive anomaly based on global knowledge as a prior. To the best of our knowledge, this is the first work to explore the glance and focus mechanism for video anomaly detection.

In addition, we propose a simple yet effective Feature Amplification Mechanism (FAM) to enhance the discriminativeness of feature magnitude for anomaly detection. More importantly, unlike the RTFM loss (Tian et al. 2021) that simply pushes normal and abnormal features to the opposite directions without considering different scene attributes, we propose a Magnitude Contrastive (MC) loss to learn a scene-adaptive cross-video magnitude distribution. It encourages the similarity of the feature magnitudes for videos in the same category and the separability of those between normal and abnormal videos. Our MC loss improves the consistency between the feature magnitude and anomaly (Fig. 2(b)), and hence benefits the separability of normal and abnormal features (Fig. 2(c)).

To summarize, our contributions include:

- To the best of our knowledge, we first propose a Magnitude-Contrastive Glance-and-Focus Network (MGFN) to mimic the global-to-local information integration mechanism of the human visual system which first glance the video and then focus on the local portion for video anomaly detection.
- We design the Feature Amplification Mechanism (FAM) to enhance the feature learning and a Magnitude Contrastive Loss to encourage the separability between normal and abnormal features.
- Our proposed MGFN achieves 86.98% (AUC) and 80.11% (AP) on two large-scale datasets UCF-crime (Sultani, Chen, and Shah 2018) and XD-violence (Wu et al. 2020) respectively, outperforming the state of the arts by a large margin. The codes are available in https://github.com/carolchenyx/MGFN.git.

Related Work

Video Anomaly Detection (Tasks)

Due to the huge amount of training data in surveillance video, it is very labor-intensive and time-consuming to annotate every frame of the video. Therefore, researchers tend to focus on the one-class (Zaheer et al. 2022) setting to learn anomaly detection without any annotation, and the weakly supervised setting where only the video-level annotations are available.

In one-class learning, Luo et al. designs a ConvLSTM network to learn the normal segments (Luo, Liu, and Gao 2017). Zhao et al. , Ionescu et al. and Chang et al. adopt

Auto-Encoder (AE) networks to reconstruct features of normal frames (Zhao et al. 2017; Ionescu et al. 2018; Chang et al. 2022). Other works utilize the memory mechanism to memorize normal patterns (Gong et al. 2019; Park, Noh, and Ham 2020; Liu et al. 2021; R. et al. 2021) and use meta-learning (Lu et al. 2020) to enhance model's generalization capability to the unseen normal scenarios.

In weakly supervised learning, the early work (Sultani, Chen, and Shah 2018) uses multiple instance learning to localize anomalous clips in videos. Recently, Zhong et al. adopts graph convolution networks to learn abnormal events (Zhong et al. 2019). However, the generalization capability of such a model is far from satisfactory. To address this issue, Ramachandra et al. builds Siamese network to learn the normal feature simulation builds Siamese network. Afterwards, Wan et al. and Zaheer et al. propose the clustering-based frameworks to distinguish anomalous events (Wan et al. 2020; Zaheer et al. 2020). Many recent works propose some network architectures to learn the spatial-temporal feature ensemble (Wu et al. 2021; Feng, Hong, and Zheng 2021; Tian et al. 2021; Wu and Liu 2021; Li, Liu, and Jiao 2022; Zhang, Qing, and Miao 2019). In this work, we focus on the weakly-supervised anomaly detection due to its good balance between annotation burden and detection performance.

Vision Transformer (Approaches)

Transformers are first utilized in natural language processing (NLP) such as machine translation (Vaswani et al. 2017) and plaintext compression (Brown et al. 2020). They have achieved great improvements on NLP-related tasks and have been proved that they have powerful feature representation capacity (Han et al. 2020).

Inspired by the achievements in NLP, transformer architectures have recently been exploited in computer vision tasks including image classification (Dosovitskiy et al. 2020), image processing (Chen et al. 2021), object detection (Carion et al. 2020), semantic segmentation (Strudel et al. 2021), action classification (Girdhar et al. 2019), and video processing (Bertasius, Wang, and Torresani 2021).

Compared with images, video data has one additional temporal dimension. Recently, many researchers have explored transformer-based architectures to learn the spatial and temporal dependencies for video processing tasks. Fayyaz et al. and Chi et al. use transformer to improve video recognition performance (Fayyaz and Gall 2020; Chi, Wei, and Hu 2020). Yin et al. builds a spatial-temporal transformer to capture the spatial and temporal information for video object detection (Yin et al. 2020). As for untrimmed videos, Seong et al. proposes the multi-task transformer network to reduce the feature redundancy and learn the relationship between different dimensions (Seong, Hyun, and Kim 2019). Inspired by the capabilities of a transformer, in this paper, we propose transformer-based blocks to glance at the whole video globally and then steer attention to each video portion, imitating human beings' global-to-local vision system for anomaly detection.

Our Approach

Overview

The architecture of our framework is shown in Fig. 3. First, feature extractor takes B untrimmed videos V with video-level annotation as input (see Fig. 3 (a)), where $V_i \in \mathbb{R}^{N_i \times H \times W \times 3}$ and N_i, H, W denote the number of frames, height and width of V_i respectively. Then we evenly segment each video sequence into T clips and denote the feature map from feature extractor as $F = \{f^{i,t}$, where $i \in [1, B], t \in [1, T]\} \in \mathbb{R}^{B \times T \times P \times C}$, where P is the number of crops in each video clip (Sultani, Chen, and Shah 2018), and C is the feature dimension. $f^{i,t} \in \mathbb{R}^{P \times C}$ denotes the feature of the t^{th} video clip in V_i.

Taking feature map F as input, Feature Amplification Mechanism (FAM) (Fig. 3 (b)) explicitly calculates the feature norm M to enhance F. Then Glance Block (GB) and Focus Block (FB) (Fig. 3 (c,d)) integrate the global and local features built upon video clip-level transformer (VCT) and self-attentional convolution (SAC) respectively. Instead of simply maximizing the magnitude of abnormal features f_a and minimizing normal ones f_n like that in (Tian et al. 2021), we design a Magnitude Contrastive (MC) loss (Fig. 3 (e)) to maximize the separability of normal features magnitudes and abnormal ones. In the following, we will detail each component of MGFN and the loss functions in the network training.

Feature Amplification Mechanism (FAM)

As shown in Figure. 3 (b), FAM first explicitly calculates the feature norm $M^{i,t}$ of $f^{i,t}$ as Equation (1):

$$M^{i,t} = \left(\sum_{c=1}^{C} \left| f^{i,t,c} \right|^2 \right)^{\frac{1}{2}} \in \mathbb{R}^{1 \times 1 \times P \times 1} \tag{1}$$

where c denotes the feature dimension index.

Afterwards, FAM derives the enhanced features $F_{FAM} = \{f^{i,t}_{FAM}\}$ by adding the 1D-Convolution modulated feature norm, Conv1D($M^{i,t}$) to $f^{i,t}$ as a residue as shown in Equation (2):

$$f^{i,t}_{FAM} = f^{i,t} + \alpha \text{Conv1D}(M^{i,t}) \in \mathbb{R}^{1 \times 1 \times P \times C} \tag{2}$$

where α is a hyper-parameter to control the effect of the norm term, and Conv1D is a single-layer 1D convolutional network to modulate the feature norm for each dimension.

Without affecting the feature map dimension, FAM amplifies the feature map by explicitly incorporating the feature norm, which is a unified abnormal representation, into the network to benefit the Magnitude Contrastive Loss that will be discussed later.

Glance Block

The architecture of the Glance Block is shown in Fig. 4 (a). To reduce the computational burden, we first use a convolution to decrease the feature map dimension from C in F_{FAM} to $C/32$. After a short-cut convolution (SCC) that outputs a feature map $F_{scc\_GB} \in \mathbb{R}^{B \times T \times P \times C/32}$, we construct a video clip-level transformer (VCT) to learn the global correlation among clips. Specifically, we establish an attention

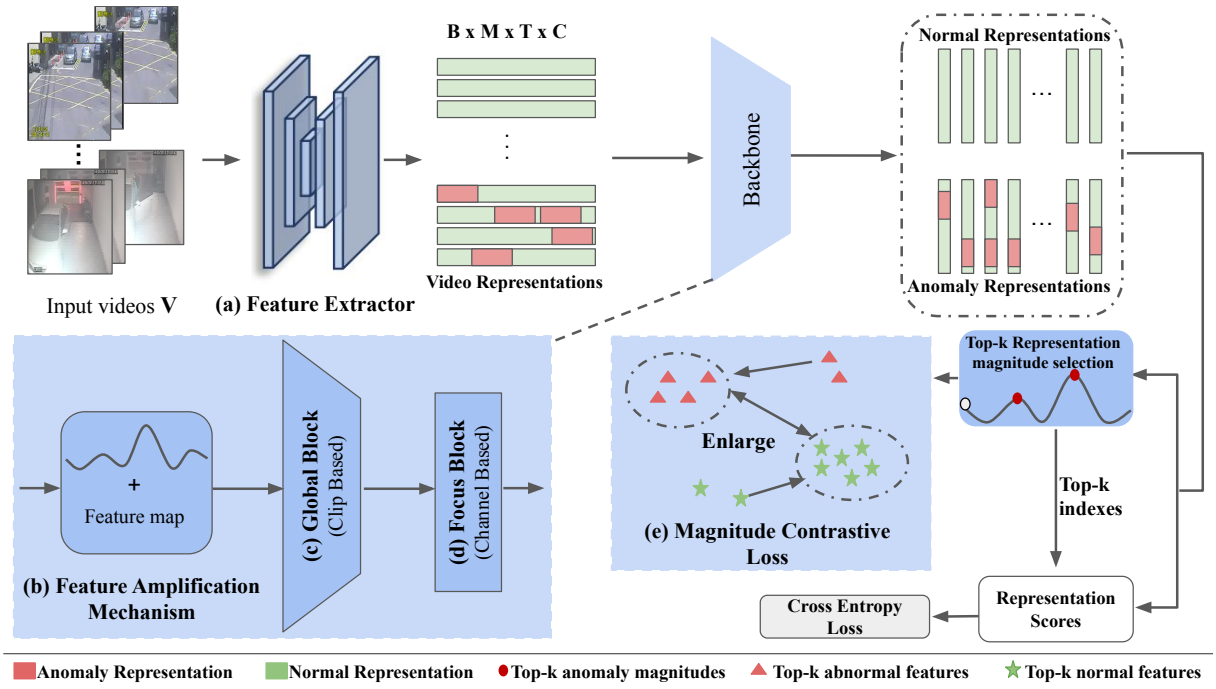

Figure 3: The overview of our MGFN network architecture. The framework takes V videos as input. After (a) Feature extractor, (b) Feature Amplification Mechanism (FAM) calculates the feature magnitude and incorporates it as a residue explicitly. Then (c) Glance Block (GB) and (d) Focus Block (FB) extract the global context information and enhance the local feature respectively. (e) Magnitude Contrastive (MC) loss encourages the separability of normal and abnormal features by shrinking the intra-category feature magnitude distances and enlarge the inter-category differences using the top-k normal and abnormal feature magnitudes.

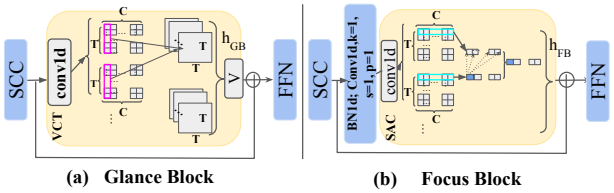

Figure 4: Architectures of Glance Block and Focus Block.

map $A \in \mathbb{R}^{1\times T\times T\times P}$ to explicitly correlate the different temporal clips.

$$A^{i,t_1,t_2} = \sum_{c=1}^{C} Q(F^{i,t_1,c}_{scc\_GB}) K(F^{i,t_2,c}_{scc\_GB}) \tag{3}$$

where $t_1, t_2 \in [1, T]$, Q, K are 1D "query" and "key" convolutions of transformer.

Next, we use the softmax normalization to generate $a \in \mathbb{R}^{1\times T\times T\times P}$ where $a^{i,t_1,:}$ represents how other clips associate with clip t_1.

$$a^{i,t_1,t_2} = \frac{e^{A^{i,t_1,t_2}}}{\sum_{t_2=1}^{T} e^{A^{i,t_1,t_2}}} \tag{4}$$

The output of VCT $F_{att\_GB} \in \mathbb{R}^{1\times T\times P\times C/32}$ is the weighted average of all clips in the long video containing both normal and abnormal (if exists) ones:

$$F^{i,t_1,c}_{att\_GB} = \sum_{t_2=1}^{T} a^{i,t_1,t_2} V(F^{i,t_2,c}_{scc\_GB}) \tag{5}$$

where V is the 1D "value" convolution of transformer. Therefore, Glance Block provides the network with the knowledge of "what the normal cases are like" to better detect the abnormal events. In addition, it helps the network to better utilize the long-term temporal contexts.

Glance Block contains an additional Feed-Forward Network (FFN) including two fully-connected layers and a GeLU non-linear function to further improve the model's representation capability. The output feature map F_{GB} is fed to the following Focus Block.

Focus Block

As shown in Fig. 4 (b), the Focus Block (FB) consists of a short-cut convolution (SCC), a self-attentional convolution (SAC), and a Feed-Forward-Network (FFN). With F_{GB} as input, we first increase the channel number to $C/16$ with a convolution. Then SSC generates the feature map F_{scc_FB}.

Inspired by the self-attention mechanism, we propose a self-attentional convolution (SAC) to enhance the feature learning in each video clip. Specifically, we exploit F_{scc_FB} as both the feature map and convolution kernel, and formulate this step to be a convolution with kernel_size=5 as Equa-

tion (6) and (7).

$$F_{sac\_FB} = F_{scc\_FB} \circledast F_{scc\_FB} \in \mathbb{R}^{B\times T\times P\times C/16} \quad (6)$$

where

$$F^{i,t,k_1}_{sac\_FB} = \sum_{k_1,k_2=0}^{C/16} F^{i,t,k_1}_{scc\_FB} F^{i,t,k_2}_{scc\_FB} \in \mathbb{R}^{1\times 1\times P\times 1} \quad (7)$$

Like self-attention, our self-attentional convolution allows each channel to get access to the nearby channels to learn the channel-wise correlation without any learnable weight. After a two-layered FFN, FB outputs the feature map F_{FB}.

Loss Functions

In this section, we introduce our loss functions. Since anomaly detection is a binary-class classification problem, a natural loss function is the sigmoid cross-entropy loss for the predicted score like existing works (Wan et al. 2020; Zhong et al. 2019; Wu et al. 2020; Tian et al. 2021): $L_{sce} = -y\log(s^{i,j}) - (1-y)\log(1-s^{i,j})$, where y is the video-level ground truth ($y = 1$ indicates abnormal) and $s^{i,j}$ is the predicted abnormal probabilities of clip j.

Magnitude Contrastive Loss To better encourage the feature separability, we propose a Magnitude Contrastive (MC) Loss as Equation (8).

$$\begin{aligned} L_{mc} = &\sum_{p,q=0}^{B/2} (1-l)(D(M^p_n, M^q_n)) + \sum_{u,v=B/2}^{B} (1-l)(D \\ &(M^u_a, M^v_a)) + \sum_{p=0}^{B/2}\sum_{u=B/2}^{B} l(\text{Margin} - D(M^p_n, M^u_a)) \end{aligned} \quad (8)$$

Note that we sample $B/2$ normal videos and $B/2$ abnormal videos in a training batch B. p, q are the indexes of normal clips, and u, v are for abnormal clips. M_a indicates the top-k feature magnitudes of abnormal clips and M_n means those of normal clips. $D(\cdot,\cdot)$ is a distance metric that will be described below. l is an indicator function where $l = 1$ denotes a pair of normal and abnormal clips p, u is sampled. In this case, L_{mc} increases the feature magnitude distance of them. $l = 0$ denotes that the two sampled clips p, q or u, v are both normal or abnormal, where L_{mc} groups them together.

Instead of simply increasing the feature magnitudes of abnormal clips and reducing those of normal clips like Tian et al. (Tian et al. 2021), our MC loss learns a scene-adaptive cross-video feature magnitude distribution. Specifically, MC loss does not force all abnormal features in different scenes to be larger than normal ones. Instead, it encourages the model to separate them with the proper distribution. For example, we allow a normal video with substantial movement to have larger feature magnitudes than an abnormal one. Accompanying with other features learned by the network, such as the scene- and movement-related features, the model can still correctly predict the anomaly. In addition, we further found that in the same video and similar scenes, the abnormal feature magnitudes are typically larger than normal ones (see Fig. 2 (c)), which is consistent with the aim of RTFM loss (Tian et al. 2021)].

$D(M^p_n, M^q_n)$ is defined in Equation (9), and $D(M^u_a, M^v_a)$ is defined similarly.

$$D(M^p_n, M^q_n) = \min_{t=0,\dots,T, r=0,\dots,T} (\mathbb{1}(\left\|f^{p,t}_{FB}\right\|_2) - \mathbb{1}(\left\|f^{q,r}_{FB}\right\|_2)) \quad (9)$$

$\mathbb{1}$ is a top-k mean function where $\mathbb{1}(\left\|f^{p,t}_{FB}\right\|_2)$ is one of the top-k feature magnitudes among $f^{p,\{1,\dots,T\}}_{FB}$. $D(\cdot,\cdot)$ derives the feature distance based on the top-k-largest-feature-magnitude clips throughout the T clips of each video. Choosing top-k clips instead of the whole video sequence benefits the model training under the video-level weak supervision. Due to the absence of the clip-level abnormality ground truth, top-k selection helps L_{mc} to focus on the clips that are most likely to be abnormal in the abnormal videos, as well as the hardest cases in normal videos. At the same time, we also select the maximum-distance pair among the top-k normal and top-k abnormal clips and encourage their similarity using L_{mc}.

Similarly, the distance $D(M^p, M^u)$ of normal feature p and abnormal u is defined as Equation (10).

$$D(M^p_n, M^u_a) = \max_{t=0,\dots,T, r=0,\dots,T} (\mathbb{1}(\left\|f^{p,t}_{FB}\right\|_2) - \mathbb{1}(\left\|f^{u,r}_{FB}\right\|_2)) \quad (10)$$

Overall Loss Functions Following Sultani et al. (Sultani, Chen, and Shah 2018), we adopt temporal smoothness loss $L_{ts} = \sum_{j=1}^{T} s^{i,j}_a$, and sparsity loss $L_{sp} = \sum_{j=1}^{T} (s^{i,j}_a - s^{i,j+1}_a)^2$, where s_a denotes the prediction score of abnormal clips. They work as regularizations to smooth the predicted scores of adjacent video clips. Thus, the total loss in the model training is: $L = L_{sce} + \lambda_1 L_{ts} + \lambda_2 L_{sp} + \lambda_3 L_{mc}$, where $\lambda_1, \lambda_2, \lambda_3$ are loss weights to balance the loss terms.

Experiment

Benchmarks and Evaluation Metrics

We consider two benchmarks in our analysis, UCF-Crime (Sultani, Chen, and Shah 2018) and XD-Violence (Wu et al. 2020). For these two benchmarks, only video-level annotations are provided. The abnormal videos contain both normal and abnormal frames, and the normal videos only contain normal frames. Following (Zaheer et al. 2020; Wu et al. 2020; Tian et al. 2021; Wu and Liu 2021; Zhu and Newsam 2019), we adopt Area Under the Curve (AUC) under the Receiver Operating Characteristic (ROC) curve as an evaluation metric for the UCF-Crime, and utilize Average Precision (AP) as the metirc for the XD-Violence. Larger AUC and AP indicate the better performance of the model.

Implementation Details

Our proposed method is implemented in PyTorch (Paszke et al. 2019). The feature extractors are I3D (Carreira and Zisserman 2017) and VideoSwin (Liu et al. 2022). The hyperparameters are set as $T = 32$, $P = 10$, $\alpha = 0.1$, $k = 3$, $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = 0.001$. To train the network, we used Adam optimiser (Kingma and Ba 2015) with a weight decay of 0.0005 and a learning rate of 0.001. The batch size B in the training is 16.

Results on UCF-Crime

As shown in Table 1, our result outperforms all the existing one-class baselines (Sohrab et al. 2018; Sun et al. 2020; Wang and Cherian 2019), unsupervised work (Zaheer et al. 2022) and weakly-supervised works (Sultani, Chen, and Shah 2018; Feng, Hong, and Zheng 2021; Wu et al. 2020; Tian et al. 2021; Zaheer et al. 2022) by a large margin. Specifically, our approach achieves better performance consistently compared with (Wu et al. 2020; Tian et al. 2021; Wu and Liu 2021) that utilize spatial-temporal feature learning, indicating the effectiveness of our Glance and Focus Blocks. Besides, with I3D features, our method outperforms RTFM (Tian et al. 2021) by 2.85% which utilizes the RTFM loss, manifesting the superiority of our proposed MC loss. With VideoSwin features, our approach even surpasses the SOTA approach MSL (Li, Liu, and Jiao 2022) by 1.05% AUC, which is already a significant improvement in the weakly-supervised video anomaly detection field.

Results on XD-Violence

Table 2 shows the results on XD-Violence dataset. Again, we achieve the superior performance over all the existing works. Specifically, our approach outperforms the spatial-temporal feature ensemble approaches (Wu et al. 2020; Wu and Liu 2021; Tian et al. 2021) consistently, demonstrating the high performance of our Glance-Focus mechanism. Thanks to our MC loss, we outperform SOTA approaches RTFM (Tian et al. 2021) by 1.38% with I3D features and MSL (Li, Liu, and Jiao 2022) by more than 1.53% AP with VideoSwin features, indicating the effectiveness of MC loss over their RTFM loss. The consistent superiority of our approach demonstrates the effectiveness of our proposed model for weakly-supervised video anomaly detection.

Ablation Studies

We conduct extensive ablation studies to validate the effectiveness of the key components in our MGFN: Glance-Focus mechanism, FAM, and MC loss.

Glance-Focus Mechanism To verify the effectiveness of the Glance-Focus mechanism to first glance the whole video and then focus on local video portions, we construct four baselines including three different organizations of our Glance Block (GB) and Focus Block (FB) as shown in Fig. 5 (a to d), and the state-of-the-art work (Tian et al. 2021).

As shown in Fig. 6, FF, FG, GF-Fusion and the state-of-the-art work (Tian et al. 2021) cannot make the consistent prediction throughout a normal video (see Fig. 6 (a)). In addition, FF, FG and GF-Fusion fail to spot out the abnormal frames in Fig. 6 (b), and (Tian et al. 2021) creates some false-positive predictions (see the points just outside the red bounding box in Fig. 6 (b) RTFM (Tian et al. 2021)). On the contrary, our MGFN with the Glance-Focus mechanism can not only produce the stable predictions throughout a normal video, but also successfully detect the abnormal events in the abnormal video (see Fig. 6 (a,b) GF (Ours)).

The results in Table 3 further manifest the effectiveness of our GF mechanism. FF achieves only 82.66% AUC on UCF datasets and 72.14% AP on XD datasets due to the lack of global context information. When adding G in the network after learning the local information with F, the per-

Supervision	Method	Features	T=32	AUC(%)
One-class	SVM Baseline	-	-	50.00
	SSV (2018)	-	-	58.50
	BODS (2019)	I3D-RGB	-	68.26
	GODS (2019)	I3D-RGB	-	70.46
	SACR (2020)	-	-	72.70
	Zaheer et al. (2022)	ResNext	-	74.20
Un-supervised	Zaheer et al. (2022)	ResNext	-	71.04
Weakly-supervised	Sultani et al. (2018)	C3D-RGB	✓	75.41
	Sultani et al. (2018)	I3D-RGB	✓	77.92
	IBL et al. (2019)	C3D-RGB	✓	78.66
	Zaheer et al. (2022)	ResNext	-	79.84
	GCN (2021)	TSN-RGB	-	82.12
	MIST (2021)	I3D-RGB	-	82.30
	Wu et al. (2020)	I3D-RGB	✓	82.44
	CLAWS (2021)	C3D-RGB	-	82.30
	RTFM (2021) ∗	VideoSwin-RGB	✓	83.31
	RTFM (2021)	I3D-RGB	✓	84.03
	Wu and Liu (2021)	I3D-RGB	✓	84.89
	MSL (2022)	I3D-RGB	✓	85.30
	MSL (2022)	VideoSwin-RGB	✓	85.62
	MGFN(Ours)	**I3D-RGB**	✓	**86.98**
	MGFN(Ours)	**VideoSwin-RGB**	✓	**86.67**

Table 1: Comparison with existing works on UCF-Crime dataset. (T=32 means a video is divided into 32 non-overlap clips. ∗ means the result is reported by (Li, Liu, and Jiao 2022))

Supervision	Method	Features	T=32	AP(%)
Weakly-supervised	Wu et al. (2020)	C3D-RGB	–	67.19
	Sultani et al. (2020) ⋄	I3D-RGB	–	73.20
	MSL (2022)	C3D-RGB	–	75.53
	Wu et al. (2020) ∗	I3D-RGB	✓	75.68
	Wu and Liu (2021)	I3D-RGB	✓	75.90
	RTFM (2021)	I3D-RGB	✓	77.81
	MSL (2022)	I3D-RGB	✓	78.28
	MSL (2022)	VideoSwin-RGB	✓	78.58
	MGFN(Ours)	**I3D-RGB**	✓	**79.19**
	MGFN(Ours)	**VideoSwin-RGB**	✓	**80.11**

Table 2: Frame-level AP on XD-Violence dataset. ∗ means the result is derived using I3D-RGB features from (Tian et al. 2021). ⋄ means result was reported by (Wu et al. 2020).

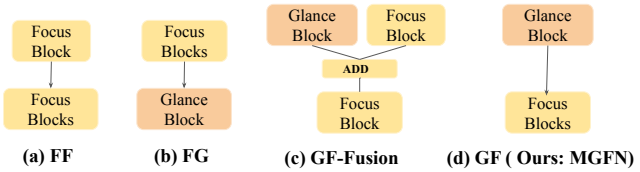

Figure 5: Glance-Focus structure (GF) analysis. (G: Glance Block, F: Focus Block) (a)-(d): Different designed network structures.

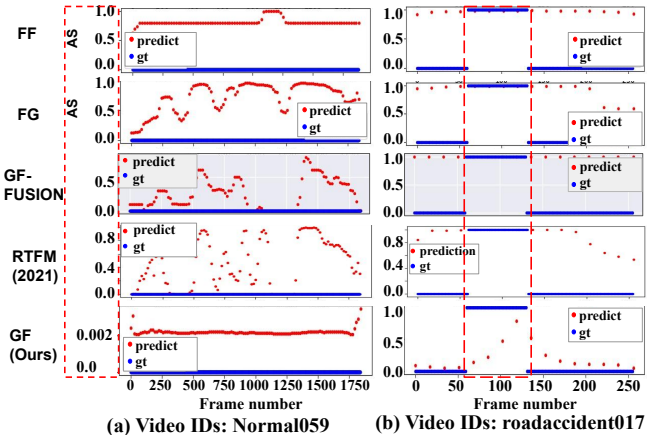

Figure 6: Glance-Focus structure (GF) analysis. (AS: anomaly score) (a) and (b) are the predicted anomaly scores (red points) of different designed structures (FF, FG, GF-FUSION and GF) and baseline (RTFM) v.s. ground truth (blue lines) on the UCF-crime dataset ("normal059" and "roadaccident017").

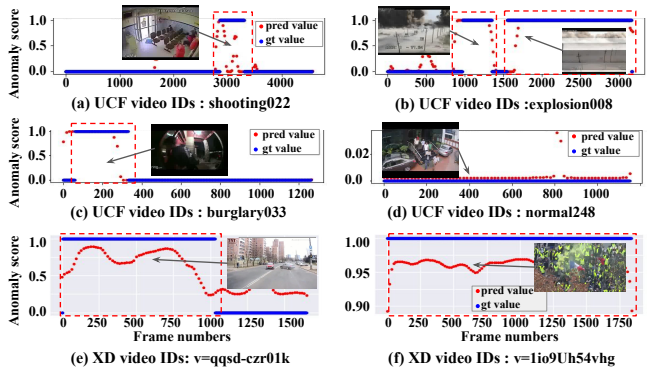

Figure 7: Anomaly scores (red points) predicted by our MGFN on UCF-crime (a - d), and XD-violence (e - f). Red boxes indicate the ground truth anomalies.

formance is boosted to 83.1% AUC on UCF and 74.30% AP on XD, indicating the importance of the global feature. Tian et al. (Tian et al. 2021) and GF-Fusion share the similar architecture of the parallel global and local feature extractor. They further improve the AUC to 84.3% and 85.05% respectively, indicating the complementarity of global and local features. At last, our MGFN further improves the performance to 85.80% AUC on UCF and 78.50% AP on XD, manifesting the effectiveness of our GF mechanism.

FAM and MC Loss In this section, we conduct ablation studies on FAM and MC loss on UCF-crime and XD-violence datasets. As shown in Table 4, FAM enhance the model's awareness of feature magnitude, thus improving AUC by more than 1.65% AUC on the UCF dataset and 1.02% AP on XD dataset. Furthermore, MC loss encourages the magnitude feature learning and thus introduces nearly 2.85% AUC improvement on UCF dataset and 3.69% AP

Network structure	AUC(%)-UCF	AP(%)-XD
FF	82.66	72.14
FG	83.10	74.30
RTFM (2021)	84.30	77.81
GF-Fusion	85.05	78.03
MGFN (GF)	**85.80**	**78.50**

Table 3: Ablation studies of Glance-Focus mechanism (G: Glance Block, F: Focus Block). All methods use RTFM loss (Tian et al. 2021) for the fair comparison.

Baseline	FAM	L_{mc}	AUC(%)-UCF	AP(%)-XD
✓	-	-	83.20	75.11
✓	✓	-	84.85	76.13
✓	-	✓	86.05	78.80
✓	✓	✓	**86.98**	**80.11**

Table 4: Ablation studies of FAM and MC loss on UCF-crime and XD-Violence. "Baseline" means MGFN without FAM and trained only sigmoid cross-entropy loss.

increment on XD dataset. In addition, combining FAM and MC Loss can further improve the performance, indicating their compatibility in video anomaly detection.

Qualitative Analysis In this section, we show the qualitative results of our MGFN. As shown in Figure 7, the distributions of red points match the ground truth very well in most cases. Specifically, our approach is able to precisely detect the abnormal events (see a,c,e), even if the video contains multiple anomalous clips (see b). For the normal video (d), our method has stable prediction throughout the video. Note that due to a sudden big movement in the video scene, the predicted abnormal probability is changed from 0 to 0.04 on the rising edge of (d), which is small enough and will not produce an alarm. It also manifests the robustness of our approach against the noise like the object movements in the video. For video (f) which is only composed of abnormal frames, our MGFN keeps predicting the high anomaly probability, indicating that our approach does not overfit the common scenes where most of the frames are normal.

Conclusion

This paper introduces a novel framework MGFN with a Glance-and-Focus module and a Magnitude Contrastive loss for anomaly detection. Imitating human beings' global-to-local vision system, the proposed MGFN contains a Glance and Focus mechanism to effectively integrate the global context and local features. In addition, a Feature Amplification Mechanism (FAM) is proposed to enhance the model's awareness of feature magnitudes. To learn a scene-adaptive cross-video feature magnitude distribution, a Magnitude Contrastive loss was introduced for encouraging the separability of normal and abnormal feature magnitudes. Experimental results on two large-scale datasets UCF-Crime and XD-Voilence demonstrate that the approach outperforms the state-of-the-art works by a large margin.

Acknowledgments

This work was supported by the Smart Traffic Fund (PSRI/27/2201/PR) funded by the Hong Kong Productivity Council and the Transport Department of HKSAR.

References

Antić, B.; and Ommer, B. 2011. Video parsing for abnormality detection. In *2011 International Conference on Computer Vision*, 2415–2422.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *CoRR*, abs/2005.14165.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chalapathy, R.; and Chawla, S. 2019. Deep Learning for Anomaly Detection: A Survey. *CoRR*, abs/1901.03407.

Chang, Y.; Tu, Z.; Xie, W.; and Yuan, J. 2022. Clustering Driven Deep Autoencoder for Video Anomaly Detection. In *Computer Vision – ECCV 2020, 16th European Conference*, 329–345.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-Trained Image Processing Transformer. arXiv:2012.00364.

Chi, C.; Wei, F.; and Hu, H. 2020. RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder. *CoRR*, abs/2010.15831.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fayyaz, M.; and Gall, J. 2020. SCT: Set Constrained Temporal Transformer for Set Supervised Action Segmentation. *CoRR*, abs/2003.14266.

Feng, J.; Hong, F.; and Zheng, W. 2021. MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection. In *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video Action Transformer Network. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *IEEE International Conference on Computer Vision (ICCV)*.

Gopalakrishnan, S. 2012. A public health perspective of road traffic accidents. *Family medicine and primary care*.

Guansong, P.; Chunhua, S.; Longbing, C.; and Den, H. A. V. 2021. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.*, 54(2).

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; Yang, Z.; Zhang, Y.; and Tao, D. 2020. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1–1.

Hasan, M.; Choi, J.; Neumann, j.; Roy-Chowdhury, A. K.; and Davis, L. 2016. Learning Temporal Regularity in Video Sequences. In *Proceedings of IEEE Computer Vision and Pattern Recognition*.

Ionescu, R. T.; Khan, F. S.; Georgescu, M.; and Shao, L. 2018. Object-centric Auto-encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *the 3rd International Conference for Learning Representations*.

Kratz, L.; and Nishino, K. 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1446–1453.

Li, S.; Liu, F.; and Jiao, L. 2022. Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2): 1395–1403.

Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3202–3211.

Lu, Y.; Yu, F.; Reddy, M. K. K.; and Wang, Y. 2020. Few-shot Scene-adaptive Anomaly Detection. In *European Conference on Computer Vision*.

Luo, W.; Liu, W.; and Gao, S. 2017. Remembering history with convolutional LSTM for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 439–444.

Park, H.; Noh, J.; and Ham, B. 2020. Learning Memory-guided Normality for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14372–14381.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Chilamkurthy, A. T. S.; Steiner, B.; Fang, L.; Baiand, J.; and Chintala, S. 2019. Automatic differentiation in pytorch. In *Advances in Neural Informa- tion Processing Systems*, 8024–8035.

R., C.; H., Z.; S., L. W. G.; and Z., H. 2021. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In *(2021). Appearance-Motion Memory Consistency Network for Video Anomaly Detection. Proceedings of the AAAI Conference on Artificial Intelligence*, 938–946.

Seong, H.; Hyun, J.; and Kim, E. 2019. Video Multitask Transformer Network. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 1553–1561.

Sohrab, F.; Raitoharju, J.; Gabbouj, M.; and Iosifidis, A. 2018. Subspace Support Vector Data Description. *CoRR*, abs/1802.03989.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for Semantic Segmentation. *arXiv preprint arXiv:2105.05633*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, C.; Jia, Y.; Hu, Y.; and Wu, Y. 2020. *Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos*, 184–192. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4975–4986.

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 *Advances in Neural Information Processing Systems*.

Wan, B.; Fang, Y.; Xia, X.; and Mei, J. 2020. Weakly Supervised Video Anomaly Detection via Center-Guided Discriminative Learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Wang, J.; and Cherian, A. 2019. GODS: Generalized One-class Discriminative Subspaces for Anomaly Detection. *CoRR*, abs/1908.05884.

Wu, J.; Zhang, W.; Li, G.; Wu, W.; Tan, X.; Li, Y.; Ding, E.; and Lin, L. 2021. Weakly-Supervised Spatio-Temporal Anomaly Detection in Surveillance Video. In *he Thirtieth International Joint Conference on Artificial Intelligence*.

Wu, P.; and Liu, J. 2021. Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection. *IEEE Transactions on Image Processing*, 30: 3513–3527.

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 *European Conference on Computer Vision (ECCV)*.

Yin, J.; Shen, J.; Guan, C.; Zhou, D.; and Yang, R. 2020. LiDAR-based Online 3D Video Object Detection with Graph-based Message Passing and Spatiotemporal Transformer Attention. *CoRR*, abs/2004.01389.

Zaheer, M. Z.; Lee, J.-H.; Astrid, M.; Mahmood, A.; and Lee, S.-I. June 2020. Cleaning label noise with clusters for minimally supervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Zaheer, M. Z.; Mahmood, A.; Astrid, M.; and Lee, S.-I. 2020. CLAWS: Clustering Assisted Weakly Supervised Learning with Normalcy Suppression for Anomalous Event Detection. In *European Conference on Computer Vision (ECCV)*, 358–376. Springer.

Zaheer, M. Z.; Mahmood, A.; Khan, M. H.; Segu, M.; Yu, F.; and Lee, S.-I. 2022. Generative Cooperative Learning for Unsupervised Video Anomaly Detection.

Zhang, J.; Qing, L.; and Miao, J. 2019. Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, 4030–4034.

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

Zhao, Y.; Deng, B.; Shen, C.; Liu, Y.; Lu, H.; and Hua, X.-S. 2017. Spatio-Temporal AutoEncoder for Video Anomaly Detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, 1933–1941. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349062.

Zhong, J.-X.; 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, Y.; and Newsam, S. D. 2019. Motion-Aware Feature for Improved Video Anomaly Detection. In *British Machine Vision Conference (BMVC)*.