Look Around for Anomalies:
Weakly-supervised Anomaly Detection via Context-Motion Relational Learning

MyeongAh Cho[1] Minjung Kim[1] Sangwon Hwang[2] Chaewon Park[1] Kyungjae Lee[3] Sangyoun Lee[1]

[1]*Yonsei University*   [2] *Hyundai Motor Company*   [3]*Yong In University*

{maycho0305,mjkima,chaewon28,syleee}@yonsei.ac.kr   sangwonH@hyundai.com   kjlee@yongin.ac.kr

Abstract

Weakly-supervised Video Anomaly Detection is the task of detecting frame-level anomalies using video-level labeled training data. It is difficult to explore class representative features using minimal supervision of weak labels with a single backbone branch. Furthermore, in real-world scenarios, the boundary between normal and abnormal is ambiguous and varies depending on the situation. For example, even for the same motion of running person, the abnormality varies depending on whether the surroundings are a playground or a roadway. Therefore, our aim is to extract discriminative features by widening the relative gap between classes' features from a single branch. In the proposed Class-Activate Feature Learning (CLAV), the features are extracted as per the weights that are implicitly activated depending on the class, and the gap is then enlarged through relative distance learning. Furthermore, as the relationship between context and motion is important in order to identify the anomalies in complex and diverse scenes, we propose a Context–Motion Interrelation Module (CoMo), which models the relationship between the appearance of the surroundings and motion, rather than utilizing only temporal dependencies or motion information. The proposed method shows SOTA performance on four benchmarks including large-scale real-world datasets, and we demonstrate the importance of relational information by analyzing the qualitative results and generalization ability.

1. Introduction

Video anomaly detection (VAD) in surveillance systems refers to the identification of undefined, unusual, or unseen abnormal events (e.g., traffic accidents, robberies, and other unforeseeable events) from amongst normal situations with temporal intervals. Currently, numerous CCTVs installed in public places such as banks, streets, and buildings record
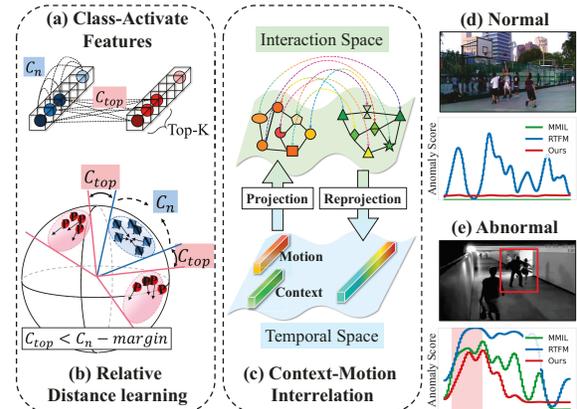
Figure 1. Concept of proposed method. We extract discriminative features that (a) are activated according to normal or abnormal classes, and (b) enlarge their gaps using relative distance learning. Furthermore, by projecting features into an interaction space, we (c) explore relationships between the context and motion information of the scene. For detecting anomalies, the proposed method considers not only motion but also its relationship with the context. For example, (d) shows a normal video with a physical fighting in a basketball game while (e) shows an abnormal fighting video. The red highlighted ranges are ground-truth abnormal frames and ours (red line) accurately detects anomalies without false alarms.

our daily life and play an important role in public safety. However, because it is time-consuming and laborious for humans to pinpoint anomalies in petabytes of surveillance videos or to monitor constantly, the VAD task, which provides automatic and instantaneous responses, is a hot topic in the field of deep learning [5, 26].

Weakly-supervised VAD (WVAD) utilizes minimal knowledge about abnormal events through video-level labeled training data that only has a label stating whether an abnormal event exists in each video clip or not. WVAD faces several challenges. First, it is difficult for the network to learn to classify anomalies at the frame-level through weak labeled training data. Therefore, most WVAD methods [13, 20, 31, 35] learn through a Multiple Instance Learning (MIL)-based approach. When normal and abnormal video clips are divided into multiple snippets and each is

contained in a negative and positive bag, there is at least one abnormal snippet in the positive bag. Therefore, the MIL approach assumes that the highest abnormality score in the positive bag derives from the abnormal snippet, and forces it to be 1 while the highest score in the negative bag is set to 0. However, given that **1) the boundary between normal and abnormal is ambiguous** in the real world, there is a limit to regression learning that forces the predicted score of snippets to a fixed values. Tian *et al*. [33] and Wu *et al*. [37] forced the gap between classes through feature learning by enlarging the feature magnitude and adjusting the distance of the feature with the center feature, respectively. However, **2)** it is **difficult to extract the discrepancy of features from a single-branch model** for enlarging the gap (shown in Fig. 7). Another challenging issue neglected in previous studies is that in real-world scenarios, for a complex and diverse scene, the definition of 'abnormal event' can differ depending on the context and motion relationship. Zhu *et al*. [47] extracted appearance-invariant features by utilizing only optical flow data to focus on moving parts, while [24,33,42] focused on temporal dependencies to consider multi-scale temporal information. However, **3) focusing only on motion or temporal information** and even excluding appearance information leads to an **incomplete understanding of complex scenes.**

In complex scenes, the boundary between normal and abnormal is ambiguous, and the distinction sometimes differs depending on the situation. That is, rather than having a fixed explicit prior to the abnormal class, it is necessary to implicitly learn class representative features by relatively comparing each class. Furthermore, abnormal events occurring in the real world vary depending on the relationship between context and motion. For example, in Fig. 1, (d) a physical skirmish during a basketball game is a normal and acceptable event; but (e) a physical fight on the street is an abnormal event. Thus, the same motion has a different class depending on the relationship between motion and surrounding or appearance. Therefore, our motivation is to extract *class-activated features* by considering the *relative boundary* between classes and to understand the *reciprocal relationship* between context and motion information.

To overcome the aforementioned challenges, we propose distance learning that adjusts the interval between normal and abnormal through **1) relative feature distance rather than individual values** such as magnitude or score. This adjusts the relative distance between the hard-negative normal sample and the abnormal sample based on the intra-class variance of normal samples. In addition, **2) Class-Activate Feature Learning (CLAV)** is proposed with an add-on Implicit Class-Activate (ICA) module to **implicitly activate representative features from a single branch** for each class with Class-Specific (CS) loss function as an auxiliary task to explore each normal or abnormal pattern. Fur-

thermore, for the first time in WVAD, we address the importance of the relationship between static and dynamic information for WVAD and propose **3) a Context-Motion Interrelation Module (CoMo)** that has a dynamic path and a context path **focusing on motion and appearance**, respectively, in the scene, for **modeling the relationship** between these two information. Then, each feature is projected from the temporal space to the interaction space and correlate propagation is performed by the graph convolution module. As shown in Fig. 1, (a) the CLAV feature enlarged the gap by (b) distance learning and explored relational information through (c) CoMo, and has no false alarm in (d) the basketball game scene with physical fighting, and shows accurate temporal localization in (e) the abnormal scene with fighting. We evaluate and discuss the effectiveness of the proposed method on four weak-labeled benchmarks, including large-scale real-world dataset UCF-Crimes [31] and XD-Violence [38], and it showed SOTA results.

2. Related Works

Weakly-supervised Video Anomaly Detection. As abnormal data are difficult to acquire and annotate for training owing to their rarity in the real world, many studies have been conducted in an unsupervised manner using reconstructive autoencoders [1, 12, 14, 28, 46] and frame predicting networks [22, 32] that model normal patterns through a training set consisting of a large number of normal videos without labels and estimate the abnormal regions through out-of-distribution. Although this approach has the advantages of easy data acquisition and no labeling cost, it detects patterns other than the training data as anomalies, which results in high false-positives and a severe bias in normal training data. To alleviate this issue, WVAD methods using weakly-labeled training data (annotated normal or abnormal in video-level) aim to differentiate between normal and anomalous through minimal supervision of abnormal events in order to avoid overfitting on prior information. WVAD approaches [9, 19, 30, 31, 43, 45] have shown substantially improved performance compared to labeling cost. Zhong *et al*. [45] proposed a label correcting method by propagating supervision signals from high-confidence video snippets to the low-confidence ones based on the feature similarity and temporal consistency. Zhang *et al*. [41] proposed a robust method on the unseen patterns, which learns to determine unseen open data. Sapkota *et al*. [30] used the dynamic non-parametric hierarchical clustering technique to efficiently group temporally and semantically similar segments.

MIL-based Methods on WVAD. Many WVAD studies [31, 33, 35, 37, 44] have attempted to detect anomalies based on the MIL framework. Maximum score-based MIL method [31] have shown promising result by maximizing the highest score gap between two classes. Furthermore, score distance learning approaches [35,44] have been proposed, which leverage the highest and lowest anomaly
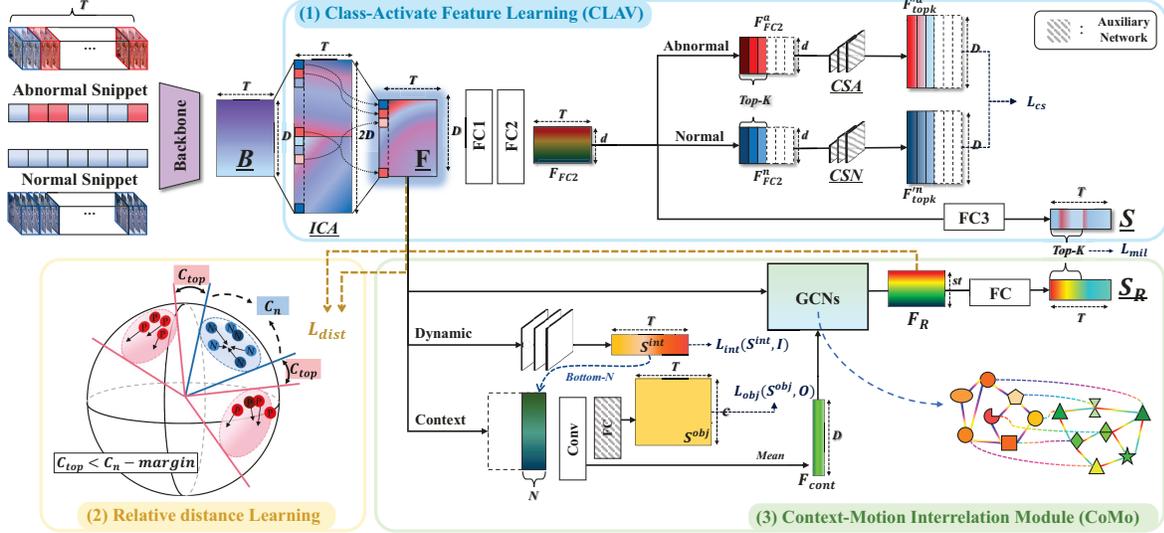
Figure 2. Overall proposed framework. Weakly-labeled training videos are split into snippets and input into the backbone. In (1) CLAV, feature F from the ICA is used for predicting abnormal score S through FC layers and class-specific auxiliary learning with CSA and CSN. F is adjusted by (2) relative distance learning, and is input to (3) CoMo for relational feature F_R and relational score S_R.

scores to reduce the intra-class and enlarge the inter-class scores. Although these methods have shown noteworthy performance, MIL approaches still have limitations in learning using few or a single highest snippet and relying only on the regression output, without feature-based decision. For feature learning, Tian *et al.* [33] enlarged the feature magnitude value between classes, and Wu *et al.* [37] adjusted the feature distance through the center feature. However, it is difficult to enlarge the gap of features with weak-labels through a single-branched backbone.

Feature Aggregation on WVAD. The feature extraction procedures of the existing WVAD methods [3,24,43,47] are mostly focused on information regarding the motion, temporal relation, and temporal dependency. Li *et al.* [19] used sequences of multiple video instances as units in order to consider the temporal relations. Wu *et al.* [37] developed the MIL method by capturing the temporal cues between video snippets, and Zhang *et al.* [42] focused on multi-scale temporal dependencies. However, these methods neglect the relationship between the surroundings and the motion, which is crucial in real-world scenarios.

3. Proposed Methods

Overview. In WVAD, as training data for detecting anomaly event in frame-level, there are the normal clip V^n and abnormal clip V^a with a video-level label $Y = \{0, 1\}$. While training, each normal and abnormal input clip is divided into T snippets, which are included in the negative bag and positive bag. As shown in Fig. 2, first, the input snippets become a backbone feature B focused on temporal dependency and motion information through a pretrained backbone, then CLAV is performed through the ICA to create a class-representative feature F through CS fea-

ture learning. To consider complex real-world scenarios, F passes through a CoMo and becomes a relational feature F_R that focuses on the interrelation between context and motion information. In CoMo, by predicting the motion information of the snippet through the dynamic path, a static feature with low motion intensity is selected and passed through a context path to extract F_{cont} containing context information. After that, to consider the interrelation between the context feature F_{cont} and the class representative feature F containing motion information, features are projected into an interaction space, and the final feature F_R is output by propagating relations through a graph-based reasoning network. Each feature F and F_R is adjusted using the inter- and intra-class gap with proposed relative distance loss, and scores S and S_R predicted through features are trained using top-K MIL loss.

3.1. Class-Activate Feature Learning (CLAV)

3.1.1 Implicit class-activate (ICA) module

To find the anomalous event among all snippets through the weak label, it is important to learn the discriminant characteristic for the normal/abnormal class. Similar to previous methods [9, 19, 30, 31], we extract the features of each snippet through the backbone, which is pretrained on the large-scale action recognition dataset. Therefore, the backbone feature B contains motion information, but it is difficult to capture representative normal and abnormal information. In addition, features from the single backbone have a limit with regard to having a distinguishing inter-class difference, and it is difficult to explicitly divide the network into two streams in a class-specific manner. Therefore, inspired by [39], which performed effective crossdomain face recognition by differentially activating weights depending on the domain, we suggest the Implicit Class-

Activation module to make the features of normal and abnormal snippets have discrepancies. The T snippets passing through the backbone become D-dim feature $\boldsymbol{B} \in \mathbb{R}^{T \times D}$ and are input to the ICA which operates temporal-wisely. As in Eq. 1, the channel size of $\boldsymbol{B}$ is expanded by the number of classes to become $\hat{\boldsymbol{B}} \in \mathbb{R}^{T \times 2D}$ where f^{ICA} is the ICA module with parameters ϕ and $\hat{\boldsymbol{B}} = \{\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2, \cdots \hat{\boldsymbol{b}}_T\}$. Then, in Eq. 2, each feature vector $\hat{\boldsymbol{b}}_i$ channel is split into $\hat{\boldsymbol{b}}1_i$ and $\hat{\boldsymbol{b}}2_i$ for the max activation.

$$\hat{\boldsymbol{B}} = f_\phi^{ICA}(\boldsymbol{B}) \qquad (1)$$

$$\hat{\boldsymbol{b}}1_i = \hat{\boldsymbol{b}}_i^{0:D}, \ \hat{\boldsymbol{b}}2_i = \hat{\boldsymbol{b}}_i^{D:2D} \qquad (2)$$

$$\boldsymbol{f}_i = \max[\hat{\boldsymbol{b}}1_i^d, \hat{\boldsymbol{b}}2_i^d]_{d=0}^D, \ (i = 1, \cdots, T) \qquad (3)$$

Through the max operation in Eq. 3, the class-representative information is implicitly aggregated from the backbone feature $\boldsymbol{B}$, and only the weight for the activated element is propagated to the gradient $\frac{\partial \boldsymbol{f}_i^d}{\partial \hat{\boldsymbol{b}}1_i^d}$ where $\hat{\boldsymbol{b}}1_i^d \geq \hat{\boldsymbol{b}}2_i^d$ and $\frac{\partial \boldsymbol{f}_i^d}{\partial \hat{\boldsymbol{b}}2_i^d}$ otherwise. This activation according to the specific characteristics of individual classes showed a powerful effect with a simple configuration of a single Conv1D layer. In addition, the ICA module brings about considerable performance improvement when added onto another VAD network in Section. 4.3, which shows that ICA is effective in distinguishing inter-classes from the backbone features. For the first time, we suggest to extract discrepancy features which are suitable for VAD in single backbone. From the ICA module, we obtain $\boldsymbol{F}^n = \{\boldsymbol{f}_1^n, \cdots, \boldsymbol{f}_T^n\}$ or $\boldsymbol{F}^a$ for normal or abnormal input snippets, respectively. Please refer to the supplementary material for details.

3.1.2 Class-Specific (CS) learning

The feature $\boldsymbol{F}$ ($\boldsymbol{F}^n$, $\boldsymbol{F}^a$) implicitly activated by the ICA module passes through the FC layers, and the anomaly score $\boldsymbol{S}$ ($\boldsymbol{S}^n, \boldsymbol{S}^a \in \mathbb{R}^T$) is predicted. During the feature extraction process, we propose a class-specific loss function L_{cs} to represent the features of each class in a similar pattern. The reconstruction-based method used in the unsupervised VAD [6, 8, 10, 17, 25, 27] reconstructs the training data composed of a single class (normal class) through the encoding–decoding process, and learns the training data pattern or distribution by minimizing the difference between input and reconstructed output data. Using this approach, shown in Fig. 2, we propose the CS modules, CSN and CSA, to reconstruct D-dim feature $\boldsymbol{F}$ with the d-dim ($d = 128$) embedded features $\boldsymbol{F}_{FC2}$ encoded by FC layers $FC1$ and $FC2$. Only the normal and abnormal features corresponding to the top-K index of the anomaly score $\boldsymbol{S}$ are input to the corresponding CS module: $\{\boldsymbol{f}_{FC2\ i}^n\}_{i=topk}$ for CSN and $\{\boldsymbol{f}_{FC2\ i}^a\}_{i=topk}$ for CSA, which each module uses for reconstruction for a single class as $\boldsymbol{F}_{topk}^{'n}$ and $\boldsymbol{F}_{topk}^{'a}$, respectively. In Eq. 4, the class-specific loss L_{cs}
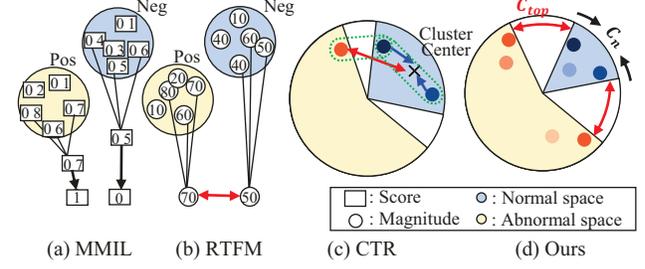
Figure 3. Illustration of loss functions. (a) MMIL [31] is an MIL-based score regression loss, (b) RTFM [33] is a feature magnitude learning loss, and (c) CTR [37] and (d) ours are a feature distance learning loss. To enlarge the class gap, RTFM and CTR adjust individual features by utilizing magnitude and center of feature, respectively. Ours, relative distance loss, intends that the distance between normal features $\boldsymbol{C}_n$ be narrower than the distance between hard negatives (normal) and positives (abnormal) $\boldsymbol{C}_{top}$.

uses an L1 loss that minimizes the gap between the predicted output feature and the class discriminant feature $\boldsymbol{F}_{topk}^n = \{\boldsymbol{f}^n_i\}_{i=topk}$ and $\boldsymbol{F}_{topk}^a = \{\boldsymbol{f}^a_i\}_{i=topk}$, which forces each class feature to contain predictable representative information. These CS modules with L_{cs} are auxiliary branches that are removed in the test phase.

$$L_{cs} = L1(\boldsymbol{F}_{topk}^n, \boldsymbol{F}_{topk}^{'n}) + L1(\boldsymbol{F}_{topk}^a, \boldsymbol{F}_{topk}^{'a}) \qquad (4)$$

3.2. Relative Distance Learning

The feature $\boldsymbol{F}$ extracted through CLAV is from the same backbone, but infer to be implicitly activated for each class. As normal and abnormal have ambiguous boundaries and are difficult to define, it is necessary to learn in consideration of relativity. Contrary to the previous methods where only top-k snippets were used for training, most normal snippets are ignored, and the relative difference between normal and abnormal is not considered, we propose distance learning that adjust feature distance with regard to the overall normality. In Eq. 5, where $\boldsymbol{F}^n = \{\boldsymbol{f}_1^n, \cdots, \boldsymbol{f}_T^n\}, \boldsymbol{F}^a = \{\boldsymbol{f}_1^a, \cdots, \boldsymbol{f}_T^a\}$ and cosine similarity is indicated by cos, the similarity of overall normal snippet features is $\boldsymbol{C}_n$, and between top-K $\boldsymbol{F}^n$ and $\boldsymbol{F}^a$ is C_{top}.

$$C_n = \frac{1}{T^2} \sum_{i,j}^T cos(\boldsymbol{f}_i^n, \boldsymbol{f}_j^n) \qquad (5)$$

$$C_{top} = cos(\boldsymbol{f}_{top}^n, \boldsymbol{f}_{top}^a)$$

$$\text{where } \boldsymbol{f}_{top} = \frac{1}{K} \sum^K \boldsymbol{f}_i, i = \text{top-K score index}$$

In Eq. 6, the relative similarity C_{top} between the hard negatives (normal top-K snippets) and the positives (abnormal top-K snippets) is decreased based on C_n, the similarity of all normal features, rather than a fixed value; concurrently, normal feature similarity C_n is increased as per Eq. 7. In Fig. 3, (a) simply regresses the score into fixed values 0 or 1, (b) enlarges the magnitude value of the feature, and (c) adjusts the feature distance of each class based on

the center feature. However, even if the center feature and normal features are closer than the abnormal features in (c), there are still samples with a low distance between the normal and the abnormal (in green dash circles). In contrast, (d) our relative distance learning considers normal-aware relative distances using all normal snippets.

$$C_{top} < C_n - margin \qquad (6)$$

$$L_{dist} = \max(C_{top} + margin - C_n, 0) + (1 - C_n) \quad (7)$$

3.3. Context-Motion Interrelation (CoMo) Module

While the previous method utilizes features focused on temporal dependency or motion information, we propose a novel CoMo module that extracts context information and explores relation propagation for the interrelation between context and motion. Feature $\boldsymbol{F}$ is used as an input, final relational feature $\boldsymbol{F}_R \in \mathbb{R}^{T \times st}$ and relational score $\boldsymbol{S}_R \in \mathbb{R}^T$ are the output, and CoMo performs relation modeling between context and motion by mapping the feature to the interaction space invariant to the temporal axis. Through CoMo, we ensure that not only superior performance but also relational information is robust to generalizability (Section. 4.4), which is important in real-world VAD.

3.3.1 Context and dynamic path

To extract the context information of a scene, we first estimate the motion information through the dynamic path, and use this prediction to filter features that have relatively low motion intensity. Then, in the context path, we focus on the surroundings and instances of a static scene with low motion intensity. Conv1D(k, s, c) denotes a 1D convolution layer, where k, s, and c are the kernel, strides and channel size, respectively. The dynamic path consists of Conv1D(1, 1, 512)–Conv1D(1, 1, 1), and motion intensity score $\boldsymbol{S}^{int} \in \mathbb{R}^T$ becomes the output. The higher the motion score, the more dynamically the scene is moving and vice versa (please refer to the supplementary material for intensity score). In Eq. 8, the loss function is an $L1$ loss between the optical flow [40] intensity $\boldsymbol{I}$ and $\boldsymbol{S}^{int}$ to make the dynamic path output the score for motion information.

$$L_{int} = L1(\boldsymbol{S}^{int}, \boldsymbol{I}), \ L_{obj} = L1(\boldsymbol{S}^{obj}, \boldsymbol{O}) \qquad (8)$$

Like methods [2, 11] that learn appearance via knowledge distilling with object or mask prediction as a proxy task, context path predicts the object class score by focusing on the appearance rather than motion. For a context feature representing the appearance, we select the features of static scenes with low motion intensity. Therefore, $\{\boldsymbol{F}_i\}_{i=bottomN}$, where i is the bottom-N index of the motion intensity score $\boldsymbol{S}^{int}$, is input to the context path. The path consists of Conv1D(3, 1, D) followed by FC layers and predicts object C class scores $\boldsymbol{S}^{obj} \in \mathbb{R}^{N \times C}$ of N snippets, which is intended to explore the appearance in the scene. As in Eq. 8, the context path is trained by L_{obj}, utilizing the $\boldsymbol{O} \in \mathbb{R}^{N \times C}$ mean value of each object class score within
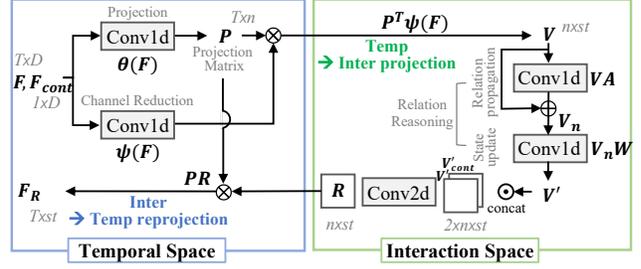
Figure 4. Architecture of CoMo graph module. Through projection matrix $\boldsymbol{P}$, $\boldsymbol{F}$ and $\boldsymbol{F}_R$ are each mapped to the node and state matrix of the interaction space. Then, through relation propagation of nodes and updation of state, each relation of context and motion is explored. Finally, $\boldsymbol{R}$ is obtained through interrelation by fusing the two pieces of information and reprojected into the final relational feature $\boldsymbol{F}_R$ in temporal space.

the snippet predicted by the MS COCO [21] pretrained object detector YOLOv5 [15] as a pseudo label. As a constraint of L_{obj}, the aggregated features from Conv1D layer become appearance-aware of the scene, and the mean of these is the context feature $\boldsymbol{F}_{cont}$ as shown in Fig. 2. The sum of both loss functions $L_{aux} = L_{int} + L_{obj}$ is used for the auxiliary task, and the layers that predict object class score are removed during testing.

3.3.2 Graph relation propagation

In Fig. 4, the features $\boldsymbol{F}$ and $\boldsymbol{F}_{cont}$ are embedded into the interaction space to understand the scene through relational information, which is independent from temporal consistencies. $\boldsymbol{F}$ and $\boldsymbol{F}_{cont}$ are mapped to n and st number of nodes and states, respectively, as $\boldsymbol{V}, \boldsymbol{V}_{cont} \in \mathbb{R}^{n \times st}$ through the $\boldsymbol{P}$ bi-projection matrix. Then, using Graph Convolutional Networks (GCNs) [7], we propagate the edge representing the relationship with each node with adjacent matrix $\boldsymbol{A}$ and update the state with $\boldsymbol{W}$. After that, the two relation-propagated matrices $\boldsymbol{V}'$ and $\boldsymbol{V}'_{cont}$ are concatenated and fused node-wise and state-wise to explore the interrelations. Relational information $\boldsymbol{R}$ is reprojected into temporal space, with the projection matrix $\boldsymbol{P}$, and becomes the final relational feature $\boldsymbol{F}_R$ of st-dim. Then, $\boldsymbol{F}_R$ passes through the FC layer to obtain relational anomaly score $\boldsymbol{S}_R$.

3.4. Training and Testing Phase

$$L_{mil} = \sum_{s \in \Omega_k(\boldsymbol{S}, \boldsymbol{S}_R)} -(y \log(s) + (1-y) \log(1-s)) \quad (9)$$

$$L = L_{mil} + \lambda_{cs} L_{cs} + \lambda_d L_{dist} + L_{aux} \qquad (10)$$

The overall framework described above learns from video-level weak-labeled data. In Eq. 9, L_{mil} is obtained through the cross-entropy loss of $\Omega_k(\boldsymbol{S})$ and $\Omega_k(\boldsymbol{S}_R)$, which are the top-$K$ score sets of the abnormality score $\boldsymbol{S}$ and the relational abnormality score $\boldsymbol{S}_R$. In Eq. 10, L_{dist} indicates $L_{dist}(\boldsymbol{F}) + L_{dist}(\boldsymbol{F}_R)$, and the final loss function is L. We use the same temporal smoothness ($\sum_{i=1}^{T}(s_i - s_{i-1})^2$) and

Table 1. Comparison Result with Other Methods

Methods		UCF [31]	XD [38]	ST [22]	AV [23]
Binary SVM	-	-	-	-	69.1
150FPS [23]	ICCV13'	65.5	-	72.9	62.1
SMIL [20]	CVPR15'	78.0	-	90.4	72.2
AMIL [13]	ICML18'	76.5	-	85.8	72.4
MMIL [31]	CVPR18'	75.4	-	92.2	70.4
MMIL*	CVPR18'	82.4	73.1	93.5	-
Noise-C [45]	CVPR19'	82.1	-	84.4	-
IBL [44]	ICIP19'	78.7	-	-	-
DMIL [35]	ICME19'	-	-	91.2	-
MA [47]	BMVC19'	79.0	-	-	-
MIST [9]	CVPR21'	82.3	-	94.8	-
RTFM [33]	ICCV21'	84.0	77.8	97.2	-
CTR [37]	TIP21'	84.9	75.9	97.5	-
Transformer [19]	AAAI22'	85.6	78.6	97.3	-
Openness [41]	ITS22'	85.5	-	97.5	-
WSTR [42]	SP22'	83.2	-	97.6	-
WAGCN [3]	Arxiv22'	83.1	-	95.9	-
Bayesian [30]	CVPR22'	83.4	-	96.0	80.9
S3R [36]	ECCV22'	86.0	80.3	97.5	-
SSRL [18]	ECCV22'	**87.4**	-	**98.0**	-
Ours		86.1	**81.3**	97.6	**89.8**

* denotes reproduced results with I3D backbone feature.

sparsity ($\frac{1}{T}\sum_{i=1}^{T}|s_i|$) regularization terms as [31, 33]. After the training phase, the anomaly score is $S + \lambda S_R$, and the CS branch and FC layers of the context path are removed.

4. Experimental Results

We conduct experiments and analyze our proposed method on four video anomaly detection benchmarks. UCF-Crimes [31] and XD-Violence (XD-Vio.) [38] are large-scale WVAD datasets containing normal and abnormal training data with video-level labels, while ShanghaiTech (ST) [22] and CUHK Avenue (AV) [23] are datasets for unsupervised VAD in which the training set only contains normal videos. Following the previous studies, we use the Area Under Curve of the Receiver Operating Characteristic, and for XD-Vio., we use Average Precision.

UCF-Crimes is large-scale WVAD database that contains untrimmed videos of real world safety-related anomalies acquired through various conditions such as illumination, resolution, and weather. The training set consists of 800 and 810 and the test set consists of 150 and 140 normal and abnormal videos, respectively. **XD-Violence** is the largest and most diverse dataset with 4,754 untrimmed sports, movie, and surveillance videos, and the number of training and testing sets is 3,954 and 800 videos, respectively. **ShanghaiTech** consists of 437 videos with 13 scenes. To utilize these for the weakly supervised approach, we reorganize them as 238 training videos and 199 testing videos that contain normal and abnormal videos in each training and test set. We use exactly the same splits as [19, 33, 45]. **CUHK Avenue** consists of 16 normal training sets and 21 abnormal test sets, and we reorganize 80:20 split of normal and abnormal video following [30].

Table 2. Ablation Studies of Proposed Module and Loss Function on UCF-Crimes Database [31]

(a) Module				(b) Loss Function			
ICA	CS	CoMo	AUC	$L_{mil}+$ L_{aux}	L_{dist}	L_{cs}	AUC
			82.84	✓			84.63
✓			83.75	✓	✓		83.83
✓	✓		85.22	✓		✓	85.07
✓	✓	✓	86.07	✓	✓	✓	86.07

Table 3. Results of Add-on ICA Module

Database	Methods			
	MMIL [31]	MMIL w ICA	Ours-w/o ICA	Ours
UCF-Crimes [31]	82.43	82.73	83.26	86.07
XD-Vio. [38]	73.10	76.85	75.40	81.31
ST [22]	93.46	95.04	96.19	97.59

4.1. Implementations

We extract D = 2048-dim or 1024-dim RGB features from the 'mix_5c' layer of the ResNet-50 I3D and Inception-v1 I3D [4], respectively, pretrained on the Kinetics [4] dataset. For UCF-Crimes, XD-Vio., ST, and AV datasets, the batch consists of half normal and half abnormal videos which size is used 64, 16, 64, 2 and T is 16, 16, 8, 8, respectively. For all experiments, we experimentally set the $margin = 0.3$ between [0, 1], K=3 between [1, 3, 5, 7], $\lambda_{cs} = 1$, and $\lambda_d = 10$. We set number of nodes $n = 32$, state $st = 128$, and $N = 0.8 \times T$. For training, an Adam optimizer [16] with a weight decay of 0.0005 and learning rate of 0.001 is used in an end-to-end manner. For testing, we set λ as 0.1, 1, 0.4, and 0.1. For a fair comparison, we use the same benchmark setup as [30, 33]. Please refer to the supplements for details including experimental results of hyper-parameters.

4.2. Comparison Results

Table 1 presents comparison results on four benchmarks using SOTA methods. As 150FPS [23], an unsupervised approach, has only normal videos in the training set, it is difficult to directly compare the result because the training/test set is different from WVAD approaches, but the performance gap with these methods is large. This shows the effectiveness of weak labeling, which achieves high performance with low labeling cost. The proposed method, considering the interrelation of context and motion, showed superior performance in all datasets compared to when the network focused on temporal dependency [3, 43] or motion information [47]; Congqi et al. [3] utilize GCNs to capture temporal dependencies, WSTR [42] aggregates multi-head relation through a stack of transformer encoder for temporal relations, and MA [47] focuses only on motion features. Compared with the temporal relational-focused WAGCN [3] and WSTR [42], the proposed method shows better results in the UCF-Crimes of a complex real-world scene than the ST composed of simple abnormal events, demonstrating the effectiveness of context and motion interrelations. SSRL [18] exhibits good result in UCF and ST

Table 4. Cross-database Experimental Results on Three Benchmarks

Source	UCF-Crimes	XD-Vio.	XD-Vio.	UCF-Crimes	ST	UCF-Crimes	XD-Vio.
Target	UCF-Crimes [31]		XD-Vio. [38]		ST [22]		
RTFM [33]	84.48	68.59 (↓ 18.81%)	76.62	37.30 (↓ 51.32%)	97.20	45.49 (↓ 53.2%)	**40.18** (↓ **58.66%**)
Ours	86.07	**69.89** (↓ **18.80%**)	81.31	**46.74** (↓ **42.52%**)	97.59	**52.02** (↓ **46.7%**)	38.03 (↓ 61.03%)

Table 5. Results of Relational Modeling Module

Database	Relation Modeling		
	-	RN [29]	CoMo
UCF-Crimes	85.22	84.67	**86.07**
XD-Vio.	76.99	78.33	**81.31**
ST	96.39	96.73	**97.30**

but compared to others, its input passes the backbone four times to utilize multi-scale patches and the multi-branches are trained with 8 GPUs (light version costs twice complexity but only gains 0.8 and 0.2 AUC than ours).

Compared to an MIL-based method performing abnormality score regression [13, 20, 31], DMIL [35] to which center score loss is applied, and IBL [44] that adjusts the score gap in the inner bag, better result is shown by feature learning methods [33, 37] including ours. In particular, when compared with RTFM, which proposed feature magnitude learning and multi-scale temporal module, our CoMo with distance learning performs 2.1%, 3.5%, and 0.4% higher in UCF-Crimes, XD-Vio., and ST, respectively.

4.3. Ablation Studies

We conduct ablation studies on proposed modules and loss functions. In Table 2 (a), the loss functions are the same, and when only modules are added, the highest performance improvement is shown when the ICA and CS modules of CLAV are together, which shows that CLAV helps to extract representative features of normality and abnormality within a single branch. With features from CLAV, the highest performance is achieved through relational learning. For the loss function ablation experiment with the entire framework, in Table 2 (b), when only L_{dist} is applied, the performance is lower than the baseline with L_{mil} and L_{aux}, which shows that distance learning has limitations when widening the gap of features from a single-branch model in the embedding space. However, it shows good results using features that have discrepancy through L_{cs} of CLAV.

In Table 3, we evaluate the effect of aggregating backbone features through the CLAV of the ICA module. As F obtained through ICA is used for relative distance learning and as the input of the CS and CoMo module, ICA module plays a very important role in the proposed framework, which leads to a substantial performance improvement. ICA is designed to implicitly activate features depending on whether normal and abnormal classes are being considered. When added-on to MMIL, the results are meaningful and show the biggest performance boost in the large and diverse XD-Vio. dataset. ICA with a simple structure can be applied to any other VAD networks to improve the class representation capability of features.
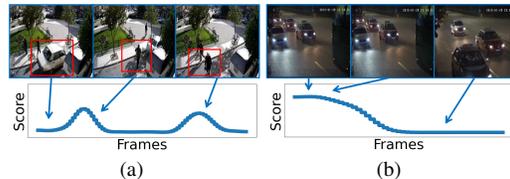
Figure 5. False alarm samples on UCF-Crime dataset [31]. A high anomaly score is shown (a) when a person passes by the road and (b) when the video is paused and cars are stopped on the road.
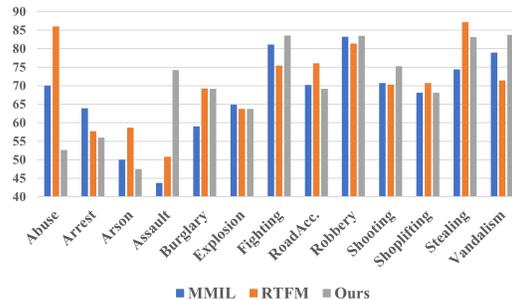
Figure 6. Comparison class-wise AUC on UCF-Crimes dataset [31] with other methods.

4.4. Discussions

Cross-database quantitative results. For a video anomaly detector to be applied to real life, the generalization ability of the trained model is very important because operating in various environments in a real-world situation is different from the training/test data. To demonstrate that the relational information between context and motion is robust to the data domain, we conducted an experiment for the verification of generalization ability in Table 4. We show the adaptability to differences in the domain gap through the performance of the model trained on the source dataset and evaluate it on the target dataset. As some abnormal events in ST such as running or jumping on a sidewalk are included as normal situations on UCF-Crimes and XD-Vio., it is difficult to evaluate the generalization ability. In particular, as most normal scenes (e.g., a movie scene running or riding bike on a sidewalk) in XD-Vio. are anomalies in ST, ours which narrowed the gap of normal features by distance learning, showed more severe degradation than RTFM.

The UCF-Crimes and XD-Vio. datasets have similar definitions for abnormal events, but compared to UCF-Crimes, which contains surveillance videos, XD-Vio. is a larger-scale dataset with a variety of videos such as sports, surveillance, and movies. Therefore, when XD-Vio. is the source and UCF-Crimes is the target, the performance degradation is small. The opposite case is most suitable for validating the generalizability to the real-world, where the proposed

(a) I3D [4] (b) RTFM [33] (c) Ours
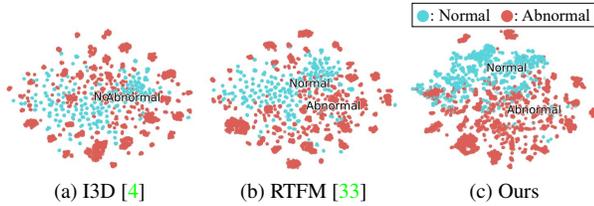
Figure 7. Visualization of final features on UCF-Crimes [31].

(a) Explosion video (b) Burglary video (c) Normal video

Figure 8. Anomaly score within the UCF-Crimes [31] video. The x-axis is the frame range, the red highlighted ranges are ground-truth abnormal frames. The green, blue, and red lines indicate the score of MMIL [31], RTFM [33], and ours, respectively.

method shows lower performance degradation than RTFM. This demonstrates that relational modeling between context and motion is more suitable to adapt to domain discrepancy than considering only temporal dependency as RTFM.

Interrelation modeling. Table 5 presents a comparison between the CoMo and RN [29] used in relational reasoning. RN is a method of concatenating each feature pair-wisely and embedding a relations through a shared FC layer to model the relationship between them. With RN, we extract each relation vector by creating a pair set between motion features and context features. As a method to explore the relationship, there is no significant difference in the results of (a) baseline (overall framework without CoMo) and using (b) RN, but in the case of (c) CoMo, there are 0.44%, 0.97%, and 0.73%, performance improvements in UCF-Crimes, ST, and XD-Vio., respectively. This shows that it is more effective to explore interrelation between separately relation-propagated context and motion features, rather than simply performing pair-wise concatenate reasoning.

Failure scenarios. Fig. 5 shows examples of cases where false alarms occurred in the normal videos. In (a), when a person passed by the roadway, the walking behavior is normal, but shows a high anomaly score through the relation of the surroundings. This scene is an ambiguous case that can be viewed as an abnormal event (jaywalking) depending on the definition; this problem of ambiguity boundaries is a limitation to be solved from the point of view of applicability to the real world. In (b), there is an error that the frames of the first certain interval of the video stopped, which results in a high abnormal score owing to the relation with the stopped motion of cars although the appearance of the car on the road is normal. In contrast, in Fig. 8 (c), the normal video in which a still scene where objects are placed in a room and a person appears, has a low score through the relationship between the object and motion. By focusing on relational information, there is a limitation of a false alarm for a paused video, which may occur in the real-world, but this issue can be solved by simple preprocessing.

Class-wise AUC. In Fig. 6, compared with other methods, the classes (e.g., abuse or road accident) that the our method shows lower AUC on are mainly anomalies with strong motion, which are relatively simple that can be detected only with additional information such as optical flow. In contrast, in the case of assault, burglary, and fighting, which require an understanding of the relations between motion and surroundings, the result is superior by 30.5%, 10.2%, 2.4% to
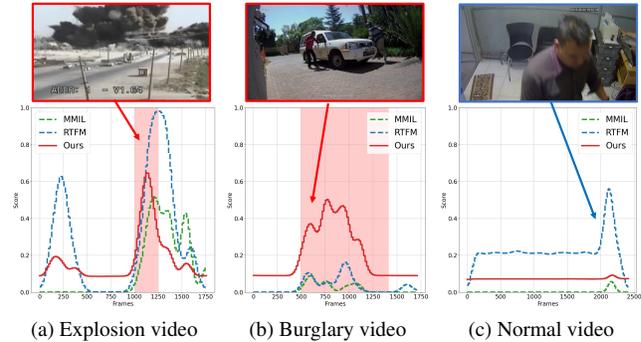
MMIL, and 23.4%, 0.1%, and 8.1% to RTFM, respectively.

4.5. Qualitative Analysis

Embedding features. In Fig. 7, the final embeddings before FC layers of I3D [4], RTFM [33], and ours is visualized using the t-SNE [34] algorithm. By utilizing (a) the backbone feature, (c) the proposed method that adjusted the relative distance between normal and abnormal features shows a distinguishable result compared to the (b) RTFM that enlarges the magnitude of normal and abnormal features.

Abnormality score plot. In Fig. 8, as the score distribution is different for each method, to easily compare the abnormality score, it is normalized to the range of [0, 1]. Through the plot, the proposed method shows more accurate temporal localization compared to other methods [31, 33]. In particular, ours results in a high abnormal score even for a (b) burglary scene where relational information is important. Furthermore, it shows a stable score for (c) a normal scene in which a person appears in an empty room. Please refer to the supplementary material for more examples.

5. Conclusion

In real-world scenarios, to detect anomalous events in normal scenes, for the first time, we address the importance of understanding the relationship between context and motion rather than focusing only on temporal dependencies, specific motion, or appearance information for undefined and unseen events. Therefore, we proposed CoMo which focuses not only on motion but also the relationship with static surroundings and contexts depending on the normal or anomalous situation. Furthermore, using a single pre-trained backbone branch and minimal supervision of weak labels of WVAD, we proposed a CLAV to implicitly activate each class for representative features and relative distance learning to enlarge their gap. It outperformed in four benchmarks and showed high performance for complex situations such as robbery and burglar, and stable scores for confusing normal situations. There is a limitation of false positives because of ambiguous and complex situations in our daily life, but we expect high-level relational information can help immensely with generalization ability.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2019. 2

[2] Baradaran and Bergevin. Multi-task learning based video anomaly detection with attention. *arXiv:2210.07697*, 2022. 5

[3] Congqi Cao, Xin Zhang, Shizhou Zhang, Peng Wang, and Yanning Zhang. Adaptive graph convolutional networks for weakly supervised anomaly detection in videos. *arXiv preprint arXiv:2202.06503*, 2022. 3, 6

[4] 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. 6, 8

[5] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. 1

[6] Dongyue Chen, Pengtao Wang, Lingyi Yue, Yuxin Zhang, and Tong Jia. Anomaly detection in surveillance video based on bidirectional prediction. *Image and Vision Computing*, page 103915, 2020. 4

[7] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019. 5

[8] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017. 4

[9] 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. 2, 3, 6

[10] Thittaporn Ganokratanaa, Supavadee Aramvith, and Nicu Sebe. Unsupervised anomaly detection and localization based on deep spatiotemporal translation network. *IEEE Access*, 8:50312–50329, 2020. 4

[11] Georgescu, Ionescu, Khan, Popescu, and Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE TPAMI*, 2021. 5

[12] 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 International Conference on Computer Vision*, pages 1705–1714, 2019. 2

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 6, 7

[14] 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 Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019. 2

[15] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, (Zeng Yifu), Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and xylieong. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, Aug. 2022. 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 6

[17] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatio-temporal adversarial networks for abnormal event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1323–1327. IEEE, 2018. 4

[18] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 333–350. Springer, 2022. 6

[19] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *Proceedings of the AAAI, Virtual*, 24, 2022. 2, 3, 6

[20] Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4277–4285, 2015. 1, 6, 7

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[22] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 2, 6, 7

[23] 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. 6

[24] Zhen Ma, José JM Machado, and João Manuel RS Tavares. Weakly supervised video anomaly detection based on 3d convolution and lstm. *Sensors*, 21(22):7508, 2021. 2, 3

[25] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283, 2019. 4

[26] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 1

[27] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2249–2259, 2022. 4

[28] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 2

[29] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017. 7, 8

[30] Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3212–3221, 2022. 2, 3, 6

[31] 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. 1, 2, 3, 4, 6, 7, 8

[32] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020. 2

[33] 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*, pages 4975–4986, 2021. 2, 3, 4, 6, 7, 8

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[35] 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. 1, 2, 6, 7

[36] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 729–745. Springer, 2022. 6

[37] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. 2, 3, 4, 6, 7

[38] 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. 2, 6, 7

[39] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 3

[40] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 5

[41] Chen Zhang, Guorong Li, Qianqian Xu, Xinfeng Zhang, Li Su, and Qingming Huang. Weakly supervised anomaly detection in videos considering the openness of events. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 2, 6

[42] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. *IEEE Signal Processing Letters*, 2022. 2, 3, 6

[43] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. *IEEE Signal Processing Letters*, 2022. 2, 3, 6

[44] 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. 2, 6, 7

[45] 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 Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 2, 6

[46] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550, 2019. 2

[47] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *BMVC*, 2019. 2, 3, 6