3-8-2022

Generative Cooperative Learning for Unsupervised Video Anomaly Detection

M. Zaigham Zaheer

Arif Mahmood

Muhammad Haris Khan

Mattia Segù

Fisher Yu

See next page for additional authors

Authors
M. Zaigham Zaheer, Arif Mahmood, Muhammad Haris Khan, Mattia Segù, Fisher Yu, and Seung-Ik Lee

Generative Cooperative Learning for Unsupervised Video Anomaly Detection

M. Zaigham Zaheer[1,2,3,5], Arif Mahmood[4], M. Haris Khan[5], Mattia Segù[3], Fisher Yu[3], Seung-Ik Lee[1,2]
Electronics and Telecommunications Research Institute[1], Univ. of Science and Technology[2],
ETH Zurich[3], Information Technology Univ.[4], Mohamed bin Zayed Univ. of Artificial Intelligence[5]

Abstract

Video anomaly detection is well investigated in weakly-supervised and one-class classification (OCC) settings. However, unsupervised video anomaly detection methods are quite sparse, likely because anomalies are less frequent in occurrence and usually not well-defined, which when coupled with the absence of ground truth supervision, could adversely affect the performance of the learning algorithms. This problem is challenging yet rewarding as it can completely eradicate the costs of obtaining laborious annotations and enable such systems to be deployed without human intervention. To this end, we propose a novel unsupervised Generative Cooperative Learning (GCL) approach for video anomaly detection that exploits the low frequency of anomalies towards building a cross-supervision between a generator and a discriminator. In essence, both networks get trained in a cooperative fashion, thereby allowing unsupervised learning. We conduct extensive experiments on two large-scale video anomaly detection datasets, UCF crime and ShanghaiTech. Consistent improvement over the existing state-of-the-art unsupervised and OCC methods corroborate the effectiveness of our approach.

1. Introduction

In the real world, learning-based anomaly detection tasks are extremely challenging mainly because of the rare occurrence of such events. The challenge further exacerbates owing to the unconstrained nature of these events. Obtaining sufficient anomaly examples is thus quite cumbersome, while one may safely assume that an exhaustive set, particularly required for training fully-supervised models, will never be collected. To make learning tractable, anomalies have often been attributed as significant deviations from the normal data. Therefore, a popular approach towards anomaly detection is to train a one-class classifier which learns the dominant data representations using only normal training examples [13, 16, 24, 27, 40, 41, 44, 46, 58, 62, 68] (Fig. 1). A noticeable drawback of one-class classification (OCC) based methods is the limited availability of the normal training data, not capturing all the normalcy variations [8]. In addition, the OCC approaches are usually unsuitable for complex problems with diverse multiple classes and a wide range of dynamic situations often found in video surveillance. In such cases, an unseen normal activity may deviate significantly enough from the learned normal representations to be predicted as anomalous, resulting in more false alarms [13, 63, 64].

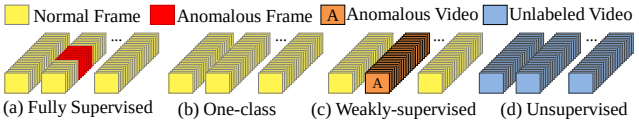


Figure 1. Different training modes for video anomaly detection: (a) Fully supervised mode requires frame-level normal/abnormal annotations in the training data. (b) One-Class Classification (OCC) requires only normal training data. (c) Weakly supervised mode requires video-level normal/abnormal annotations. (d) Unsupervised mode requires no training data annotations.

Recently, weakly supervised anomaly detection methods have gained significant popularity [23, 25, 33, 45, 54, 61] that reduce the cost of obtaining manual fine-grained annotations by employing video-level labels [49, 63–65, 70]. Specifically, a video is labeled as anomalous if *some* of its contents are anomalous and normal if *all* of its contents are normal, requiring careful manual inspection of the full video contents. Although such annotations are relatively cost-effective, yet remain impractical in many real-world applications. There is a plethora of video data, specifically raw footage, that can be leveraged for anomaly detection training if no annotation cost is incurred. Unfortunately, to the best of our knowledge, there are hardly any notable attempts in leveraging the unlabelled training data for video anomaly detection.

In the current work, we explore *unsupervised mode* for video anomaly detection that is certainly more challenging than fully, weakly or one-class supervision (Fig. 1). However, it is also more rewarding due to minimal assumptions and hence will encourage the development of novel and more practical algorithms. Note that, the term 'unsupervised' in literature often refers to OCC approaches which assume all normal training data [10,36,62]. However, it ren-

ders the overall learning problem partially supervised [18]. In approaching unsupervised anomaly detection in surveillance videos, we exploit the simple facts that videos are information-rich compared to still images and anomalies are less frequent than the normal happenings [7, 28, 50, 64], and attempt to leverage such domain knowledge in a structured manner.

To this end, we propose a *Generative Cooperative Learning (GCL)* method which takes *unlabelled* videos as input and learns to predict frame-level anomaly score predictions as output. The proposed GCL comprises two key components, a *generator* and a *discriminator*, which essentially get trained in a mutually cooperative manner towards improving the anomaly detection performance. The generator not only reconstructs the abundantly available normal representations but also distorts the possible high-confidence anomalous representations by using a novel negative learning (NL) approach. The discriminator instead estimates the probability of an instance to be anomalous. Since we approach unsupervised anomaly detection, we create pseudo-labels from generator and use these to train the discriminator and in the following step, we create pseudo-labels from the trained version of discriminator and then use these to improve the generator. The overall system is trained in an alternate training fashion where, in each iteration, both the generator and the discriminator get improved with mutual cooperation.

Contributions. We propose an anomaly detection approach capable of localizing anomalous events in complex surveillance scenarios without requiring labelled training data. To the best of our knowledge, our method is the first rigorous attempt tackling the surveillance videos anomaly detection in a fully unsupervised mode. A novel Generative Cooperative Learning (GCL) framework is proposed that comprises a generator, a discriminator, and cross-supervision. The generator network is forced not to reconstruct anomalies by using a novel negative learning approach. Extensive experiments on two large-scale complex anomalous event detection datasets, UCF-Crime and ShanghaiTech, show that our method provides visible gains over the baselines and several existing unsupervised as well as OCC methods.

2. Related Work

Anomaly detection is a widely studied problem both in the image [6, 15, 38] and video domain [48, 49, 62, 64, 65]. We here introduce different supervision modes for video anomaly detection and traditional mutual learning strategies

Anomaly Detection as One-Class Classification (OCC). OCC based approaches have found their way in a wide range of anomaly detection problems including, medical diagnosis [56], cyber security [10], surveillance security systems [19, 28, 31, 62], and industrial inspection [4]. Some of these approaches use hand-crafted features [2, 30, 37, 53, 67], while others use deep features extracted using pre-trained models [41, 46]. With the advent of generative models, many approaches proposed variants of such networks to learn representations corresponding to normal data [11, 34, 35, 42–44, 59, 60, 62]. OCC based approaches find it challenging to avoid well-reconstruction of anomalous test inputs. This problem is attributed to the fact that since OCC approaches only use normal class data while training, an ineffective classifier boundary may be achieved which is limited in enclosing normal data while excluding anomalies [62]. In an attempt to address this limitation, some researchers recently proposed pseudo-supervised methods in which pseudo-anomaly instances are generated using normal training data [1, 62].

Weakly Supervised (WS) Anomaly Detection. Video-level binary annotations are used to train WS classifiers capable of predicting frame-level anomaly scores [39, 49, 51, 63–65, 70]. Video-level labels are provided in such a way that a normal labeled video is completely normal whereas an anomalous labeled video contains both normal and anomalous contents without any information about the temporal whereabouts (Fig. 1).

Unsupervised Anomaly Detection. Anomaly detection methods using unlabelled training data are quite sparse in literature. According to the nomenclature shown in Fig. 1, most unsupervised methods in the literature actually fall in the category of OCC. For instance, MVTecAD [4] benchmark ensures the training data to be only normal, therefore its evaluation protocol is OCC and the methods inheriting this assumption are also essentially one-class classifiers [5, 11]. In contrast to these algorithms, our proposed GCL approach is capable of learning from unlabelled training data without assuming any normalcy. The training data in the form of videos conform to several important attributes regarding anomaly detection, such as, anomalies are less frequent than normal events and events are often temporally consistent. We derive our motivation from these clues to carry out the training in a completely unsupervised fashion.

Teacher Student Networks. Our proposed GCL shares some similarities with the Teacher Student (TS) frameworks for knowledge distillation [17]. GCL is different from TS framework mainly because its aim is not knowledge distillation. Also our generator generates noisy labels while our discriminator, being relatively robust to noise, cleans these labels which is not the case in TS framework.

Mutual Learning (ML). The GCL framework also shares similarities with the ML algorithms [69]. However, the two components of the GCL learn different types of information and are trained with cross-supervision in contrast to the supervised learning used by the ML algorithms. Further in GCL, the output of each network is passed through a threshold process to produce pseudo-labels. In ML frameworks, the cohort learns to match the distributions of each member

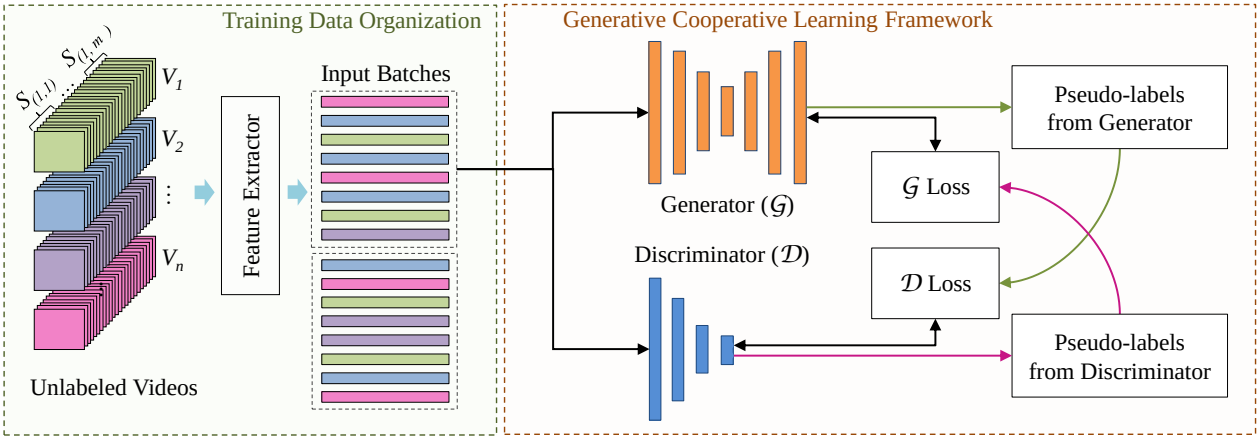


Figure 2. Proposed Generative Cooperative Learning (GCL) algorithm introduces cross-supervision for training a Generator $\mathcal{G}$ and a Discriminator $\mathcal{D}$. The pseudo-labels produced by $\mathcal{G}$ are used to compute the $\mathcal{D}$ loss and likewise, the pseudo-labels produced by the $\mathcal{D}$ are utilized to compute the $\mathcal{G}$ loss. Both $\mathcal{G}$ and $\mathcal{D}$ are trained iteratively from unlabelled training data for anomalous events detection.

while in GCL each member tries to learn from the pseudo-labels generated by the other. A mutual learning of a cohort in unsupervised mode using unlabelled training data is unexplored yet.

Dual Learning. It is also a related method in which two language translation models interactively teach each other [14]. However, the external supervision is provided using pre-trained unconditional language expert models which check the quality of translations. This way, different models have different learning tasks whereas in our proposed GCL approach the learning tasks are identical.

Another variant of **Cooperative Learning** [3] has been previously proposed to learn multiple models jointly for the same task across different domains.

For instance, object recognition is formulated by training a model on RGB images and another model on depth images which then communicate the domain invariant object attributes. Whereas, in our GCL approach both models address the same task in the same domain.

3. Method

Our proposed Generative Cooperative Learning approach for Anomaly Detection (GCL) comprises a feature extractor, a generator network, a discriminator network, and two pseudo-label generators. Fig. 2 shows the overall architecture. Each of the components are discussed next.

3.1. Training Data Organization

To minimize the computational complexity and to reduce the training time of GCL, similar to the existing SOTA [49,51,63–65,70], we also utilize a deep feature extractor to convert video data into compact features. All input videos are arranged as segments, features of which are then extracted. Furthermore, these features are randomly arranged as batches. In each iteration a randomly selected batch is used to train the GCL model (Fig. 2). Formally, given a training dataset of n videos, every video is partitioned into non-overlapping segments $S_{(i,j)}$ of p frames each, where $i \in [1, n]$ is the video index and $j \in [1, m_i]$ is the segment index. The segment size p is kept the same across all training and test videos of a dataset.

For each $S_{(i,j)}$, a feature vector $\mathbf{f}_{(i,j)} \in \mathbb{R}^d$ is computed as $\mathbf{f}_{(i,j)}$=$\mathcal{E}(S_{(i,j)})$ using the feature extractor $\mathcal{E}(\cdot)$.

In the existing weakly supervised anomaly detection approaches, each training iteration is carried out on one or more complete videos [49, 70]. Recently, CLAWS Net [64] proposed to extract several batches of temporally consistent features, each of which was then randomly input to the network. Such configuration serves the purpose of minimizing correlation between consecutive batches. In these existing approaches, it is important to maintain temporal order at batch or video level. However, in the proposed GCL appraoch we randomize the order of input features which removes both the intra-batch and inter-batch correlations.

3.2. Generative Cooperative Learning

Our proposed Generative Cooperative Learning (GCL) approach for anomaly detection consists of a generator $\mathcal{G}$ which is an autoencoder (AE) and a discriminator $\mathcal{D}$ which is a fully connected (FC) classifier. Both these models are trained in a cooperative fashion without using any data annotations. More specifically, we neither use the normal class annotations as in one class classification (OCC) approaches [11, 36, 52], nor binary annotations used by the weakly supervised anomaly detection systems [49,64,65,70]. As discussed in Section 1, the intuition behind using an AE is that such models can somewhat capture the overall dominant data trends [11]. On the other

hand, the FC classification network used as a discriminator is known to be efficient when provided with supervised, albeit noisy, training [64]. In order to carry out the training, first pseudo annotations created using $\mathcal{G}$ are used to train $\mathcal{D}$. In the next step, pseudo annotations created by using $\mathcal{D}$ are used to improve $\mathcal{G}$. Thus, each of the two models are trained by using the annotations created by the other model, in an alternate training fashion. The training configuration aims that the pseudo-labeling is improved over training iterations which consequently results in an improved overall anomaly detection performance. Particular architecture details and several design choices are discussed next.

3.2.1 Generator Network

$\mathcal{G}$ takes features as input and produces reconstructions of those features as output. Typically, $\mathcal{G}$ is trained by minimizing the reconstruction loss $\mathcal{L}_r$ as:

$$\mathcal{L}_r = \frac{1}{b}\sum_{q=1}^{b}\mathcal{L}_G^q,\ \mathcal{L}_G^q = ||f_{i,j}^q - \widehat{f}_{i,j}^q||_2, \tag{1}$$

where $f_{i,j}^q$ is a feature vector that is input to $\mathcal{G}$ and $\widehat{f}_{i,j}^q$ is the corresponding reconstructed vector, b is the batch size.

3.2.2 Pseudo Labels from Generator

In our proposed collaborative learning, pseudo labels from $\mathcal{G}$ are created to train $\mathcal{D}$. The labels are created by keeping in view the distribution of the reconstruction loss $\mathcal{L}_G^q$ of each instance q over a batch. The main idea is to consider feature vectors resulting in higher loss values as anomalous and those generating smaller loss values as normal. In order to implement this intuition, one may consider using a threshold $\mathcal{L}_G^{th}$ as:

$$l_G^q = \begin{cases} 1, & \text{if } \mathcal{L}_G^q \geq \mathcal{L}_G^{th} \\ 0, & \text{otherwise .} \end{cases} \tag{2}$$

We have followed a simple approach for the $\mathcal{L}_G^{th}$ selection by considering a fixed percentage of the samples having maximum reconstruction error as anomalous. In the $\mathcal{L}_G^q$ histograms we empirically observed a bigger peak towards minimum error and a smaller peak towards maximum error. Due to the fact that the class boundaries often fall in low density regions, error histograms are also an effective tool for the selection of appropriate $\mathcal{L}_G^{th}$. Analysis of different alternates for $\mathcal{L}_G^{th}$ selection is given in the Supplementary.

3.2.3 Discriminator Network

The binary classification network used as the discriminator $\mathcal{D}$ is trained using the pseudo annotations from $\mathcal{G}$ by minimizing the binary cross entropy loss over a batch b as:

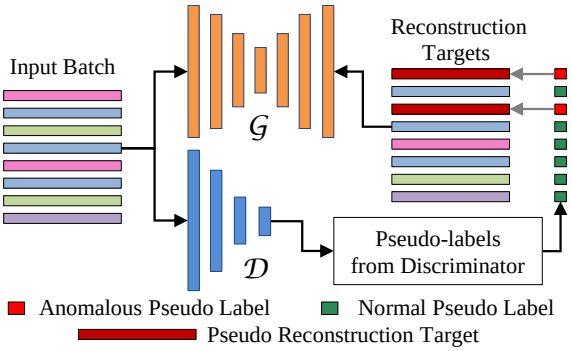


Figure 3. Negative learning in GCL: $\mathcal{G}$ is constrained to not learn the reconstruction of anomalies using Pseudo Reconstruction Targets (PRT). Based on the pseudo-labels produced by $\mathcal{D}$, PRT are generated for the anomalous inputs while normal targets are used for the normal inputs to guide the training of $\mathcal{G}$.

$$\mathcal{L}_D = \frac{-1}{b}\sum_{q=1}^{b} l_G^q \ln \widehat{l}_{i,j}^q + (1 - l_G^q)\ln(1 - \widehat{l}_{i,j}^q), \tag{3}$$

where $l_G^q \in \{0, 1\}$ is the pseudo label generated by $\mathcal{G}$ and $\widehat{l}_{i,j}^q$ is the output of $\mathcal{D}$ when a feature vector $f_{i,j}^q$ is input.

3.2.4 Pseudo Labels from Discriminator

Pseudo labels from $\mathcal{D}$ are used to improve the reconstruction discrimination capability of $\mathcal{G}$. The output $\widehat{p}_{i,j}^q$ of $\mathcal{D}$ is the probability of a feature vector $f_{i,j}^q$ to be anomalous. Therefore, the features obtaining higher probability are considered as anomalous by using a threshold mechanism on the output $\widehat{p}_{i,j}^q$ of $\mathcal{D}$. The annotations generated by $\mathcal{D}$ are then used to fine tune $\mathcal{G}$ in the next iteration.

$$l_D^q = \begin{cases} 1, & \text{if } \widehat{p}_{i,j}^q \geq \mathcal{L}_D^{th} \\ 0, & \text{otherwise ,} \end{cases} \tag{4}$$

where the threshold $\mathcal{L}_D^{th}$ is computed the same way as the threshold $\mathcal{L}_G^{th}$ is computed.

3.2.5 Negative Learning of Generator Network

Training of $\mathcal{G}$ is carried out by using pseudo labels from $\mathcal{D}$ by employing negative learning (NL). In order to increase the discrimination among reconstructions of normal and anomalous inputs, $\mathcal{G}$ is encouraged to poorly reconstruct the samples which have anomalous pseudo labels whereas, the samples having normal pseudo labels are aimed to be reconstructed as usual with minimum error.

Some variants of NL have already been explored in the literature. For instance, Munawar *et al.* [32] and Astrid *et al.* [1] make the loss negative for a full batch of known anomalous inputs. However, this configuration requires a

prior knowledge of the whole dataset and its labels. In the proposed GCL approach, pseudo labels are generated iteratively as the training proceeds, therefore it may encounter both normal and anomalous samples in the same batch. In addition, instead of making the loss negative, we enforce the abnormal samples to be poorly reconstructed by using a pseudo reconstruction target. Therefore, as illustrated in Fig. 3, for each feature vector which is pseudo-labeled as anomalous by $\mathcal{D}$, its reconstruction target is replaced by a different feature vector. In order to extensively explore this concept, we propose the following different types of pseudo targets: **1) All Ones Target**: The original reconstruction target is replaced by a similar dimensional vector of all 1's. **2) Random Normal Target**: The original reconstruction target is replaced by a normal labeled feature vector selected arbitrarily. **3) Random Gaussian Noise Target**: The original reconstruction target is perturbed by adding Gaussian noise. **4) No Negative Learning**: No negative learning is applied to $\mathcal{G}$. Instead only feature vectors pseudo-labeled as normal are used for the training of $\mathcal{G}$. Extensive analysis of different pseudo targets is shown in Fig. 5. We empirically observe that 'ones' as pseudo target yields more discriminative reconstruction capability, thus better differentiating normal and anomalous inputs. The loss function given by Eq. (1) is modified to include negative learning:

$$\mathcal{L}_G = \frac{1}{b}\sum_{q=1}^{b} ||t_{i,j}^q - \widehat{f}_{i,j}^q||_2, \tag{5}$$

where the pseudo target t_q is defined as:

$$t_{i,j}^q = \begin{cases} f_{i,j}^q, & \text{if } l_D^q = 0 \\ \mathbf{1} \in \mathbb{R}^d, & \text{if } l_D^q = 1, \end{cases} \tag{6}$$

3.3. Self-Supervised Pre-training

The proposed GCL approach is trained using unlabelled videos by utilizing the cooperation of $\mathcal{G}$ and $\mathcal{D}$. Since anomaly detection is an ill-defined problem, the lack of constraints may adversely affect the convergence and the system may get stuck in local minima. In order to improve the convergence, we explore to *jump-start* the training process by pre-training both $\mathcal{G}$ and $\mathcal{D}$. We empirically observe that using a pre-trained $\mathcal{G}$ (based on Eq. (1)) is beneficial for the overall stability of the learning system and it also improves the convergence as well as the performance of the system (see Section 4).

Autoencoders are known to capture dominant representations of the training data [11, 62]. Despite the fact that anomalies are sparse and normal features are abundant in the training data, we experimentally observe that simply utilizing all training data to pre-train $\mathcal{G}$ may not provide an effective jump-start. Using the fact that events in videos happen in temporal sequence and that anomalous frames are usually more eventful than the normal ones, we utilize temporal difference between the consecutive feature vectors as an estimator to initially clean the training dataset for the pre-training of $\mathcal{G}$. That is, a feature vector $f_{i,j}^{t+1}$ will only be used for pre-training if $||f_{i,j}^{t+1} - f_{i,j}^{t}||_2 \leq D_{th}$, where the superscripts t and $t+1$ show the temporal order of features in a video and D_{th} is the threshold. This approach does not guarantee complete removal of anomalous events however, it cleans the data for an effective initialization of the $\mathcal{G}$ to give a jump-start to the training. Once $\mathcal{G}$ is pre-trained, it is used to generate psuedo labels which are then used to pre-train the discriminator. In this step, the role of $\mathcal{G}$ is similar to a lousy teacher because the generated pseudo-labels are quite noisy and the role of $\mathcal{D}$ is like an efficient student because it learns to discriminate normal and anomalous features better even with noisy labels. In the following steps, both pre-trained $\mathcal{G}$ and $\mathcal{D}$ are plugged into our collaborative learning loop.

3.4. Anomaly Scoring

In order to compute final anomaly score at test time, several configurations are possible, i.e., using reconstruction error of $\mathcal{G}$ or prediction scores of $\mathcal{D}$. We experimentally observed that $\mathcal{G}$ remains relatively lousy while $\mathcal{D}$ remains efficient across consecutive training iterations. Therefore for simplicity, unless stated otherwise, all results reported in this work are computed using the predictions of $\mathcal{D}$.

4. Experiments

In this section, we first provide important experimental details, then draw comparisons with the existing state-of-the-art methods, and finally study different components of our GCL framework.

Datasets. UCF-Crime (UCFC) dataset contains 13 different categories of real-world anomalous events which were captured by CCTV surveillance cameras spanning 128 hours [49]. This dataset is complex because of the unconstrained backgrounds. The training split contains 810 abnormal and 800 normal videos, while the testing split has 140 anomalous and 150 normal videos. In training split, video-level labels are provided while in test split frame-level binary labels are provided. In our unsupervised setting, we discard the training-split labels and train the proposed GCL using unlabelled training videos.

ShanghaiTech contains staged anomalous events captured in a university campus at 13 different locations spanning 437 videos. This dataset was originally proposed for OCC with only normal videos provided for training. Later, Zhong *et al*. [70] reorganized this dataset to facilitate training of weakly-supervised algorithms. Normal and anomalous videos were mixed in both the training and the testing splits. The new training split contains 63 anomalous and

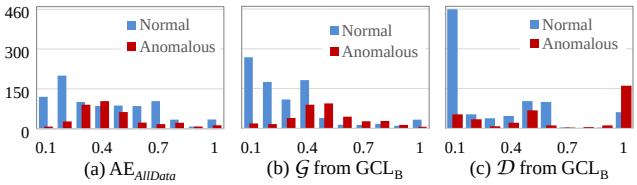


Figure 4. Distribution of scores predicted on the test split of UCF-crime dataset by (a) $\mathcal{AE}$ trained on all training data, (b) $\mathcal{G}$ trained in GCL$_B$, and (c) $\mathcal{D}$ trained in GCL$_B$. Although $\mathcal{G}$ and $\mathcal{D}$ are trained cooperatively, $\mathcal{D}$ being more robust to noise, demonstrates superior discrimination between normal and anomalous examples.

175 normal videos whereas, the new testing split contains 44 anomalous and 155 normal videos. In order to train our proposed GCL, we follow the latter split both for training and testing, without using training split video labels.

Evaluation Measures. Following the existing methods [13, 26, 49, 70], we use area under ROC curve (AUC) for evaluation and comparisons. AUC is computed based on frame-level annotations of the test videos in both datasets.

Implementation Details. To demonstrate the concept of cooperative learning in its true essence, we select fairly simple architectures, without any bells and whistles, as our $\mathcal{G}$ and $\mathcal{D}$ networks. Architectures of $\mathcal{G}$ and $\mathcal{D}$ are set as FC[2048, 1024, 512, 256, 512, 1024, 2048] and FC[2048, 512, 32, 1]. We train both networks using RMSprop optimizer with a learning rate of 0.00002, momentum 0.60, for 15 epochs on training data with batch size 8192. Thresholds for pseudo-label generation are data driven. For $\mathcal{G}$ pseudo-labels $\mathcal{L}_{\mathcal{G}}^{th} = \mu_R + \sigma_R$ where μ_R and σ_R are the mean and the standard deviation of reconstruction error as given by Eq. (1) for each batch. For $\mathcal{D}$, $\mathcal{L}_{D}^{th} = \mu_P + 0.1\sigma_P$, where μ_P and σ_P are the mean and standard deviations of the probabilities $\widehat{p}_{i,j}^{q}$ generated by D for each batch. The value of D_{th}=0.70 is used in unsupervised pre-training. As feature extractor, we use a popular framework ResNext3d proposed by Hara *et al*. [12] in default mode. Segment size p for feature extraction is set to 16 non-overlapping frames. All experiments are performed on NVIDIA RTX 2070 with Intel Core i7, 8th gen and 16GB RAM. Code will be released upon acceptance.

4.1. Comparisons with State-Of-The-Art (SOTA)

The proposed GCL approch is trained in an unsupervised fashion without using any video-level or frame-level annotations. GCL with no pre-training, GCL$_B$, is considered as the baseline. In addition, GCL with pre-training, GCL$_{PT}$, GCL combined with OCC based pre-trained autoencoder, GCL$_{OCC}$, and GCL weakly-supervised, GCL$_{WS}$ are also trained and evaluated on UCFC and ShanghaiTech datasets.

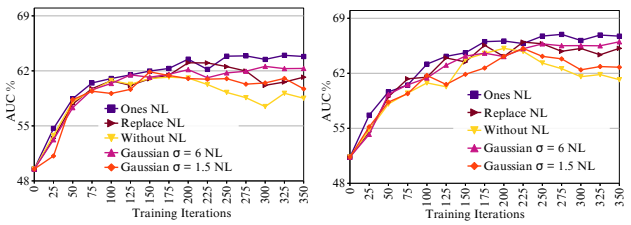
As seen in Table 1, on **UCFC dataset**, the proposed GCL$_B$ obtained an overall AUC of 68.17 % which is 11.85 % higher than the Autoecnoder (AE$_{AllData}$) trained on complete training data including both normal and anomalous training samples in an unsupervised fashion. Histogram plotted over reconstructions in Fig. 4(a) also provides insights that AE$_{AllData}$ is not able to learn discriminative reconstruction. Also in the GCL, the discrimination ability of $\mathcal{D}$ (Fig. 4(c)) is much enhanced than $\mathcal{G}$ (Fig. 4(b)). Experiments on kim *et al*. [20] are conducted on a re-implementation of the method for unlabelled training data.

GCL$_{PT}$ is the version of proposed GCL with an autoencoder pre-trained in an unsupervised fashion. In this experiment, an AUC performance of 71.04% is obtained which is 2.87% better than the baseline GCL$_B$. The two methods are also compared in Fig. 10 using multiple random seed initialization and GCL$_{PT}$ demonstrates consistent performance gains. Table 1 also shows that the pro-

Table 1. Performance comparison with existing state-of-the-art methods on UCF-Crime (UCFC) and ShanghaiTech (STech) datasets. We divide the methods into three categories based on the supervision used in training. Best results are in bold.

Supervision Type	Method	Features	UCFC AUC%	STech AUC%
One Class Classification	SVM [49]	-	50.00	-
	Hasan *et al*. [13]	-	50.60	60.85
	Sohrab *et al*. [47]	-	58.50	-
	Lu *et al*. [26]	-	65.51	68.00
	BODS [52]	I3D	68.26	-
	OGNet** [62]	ResNext	69.47	69.90
	GODS [52]	I3D	70.46	-
	TSC [27]	-	-	67.94
	Frame Prediction [24]	-	-	73.40
	MemAE [10]	-	-	71.20
	MNAD [36]	-	-	70.50
	Cho *et al*. [9]	-	-	74.70
	LNTR [1]	-	-	75.97
	RUVAD [55]	-	-	76.67
	BMAN [21]	-	-	76.20
	Proposed GCL$_{OCC}$	ResNext	**74.20**	**79.62***
Weak Supervision	Sultani *et al*. [49]	C3D	75.41	-
	Zhang *et al*. [66]	C3D	78.66	82.50
	Zhu *et al*. [71]	C3D	79.00	-
	Noise Cleaner [63]	C3D	78.27	84.16
	SRF [65]	C3D	79.54	84.16
	DUAD*** [22]	C3D	72.90	-
	GCN [70]	C3D	81.08	76.44
	GCN [70]	TSN^{RGB}	82.12	84.44
	Wu *et al*. [57]	I3D	82.44	-
	DAM [29]	I3D	82.67	88.22
	CLAWS [64]	C3D	83.03	89.67
	CLAWS** [64]	ResNext	82.61	-
	Yu *et al*. [51]	C3D	83.28	91.51
	Yu *et al*. [51]	I3D	84.30	**97.27**
	Purwantu *et al*. [39]	TRN	**85.00**	96.85
	Proposed GCL$_{WS}$	ResNext	79.84	86.21
Unsupervised	kim *et al*. ** [20]	ResNext	52.00	56.47
	AE$_{AllData}$	ResNext	56.32	62.73
	Proposed GCL$_B$	ResNext	68.17	72.41
	Proposed GCL$_{PT}$	C3D	70.74	-
	Proposed GCL$_{PT}$	ResNext	**71.04**	**78.93**

* We follow the evaluation protocol of Zhong *et al*. [70].
**We implemented the models and computed these scores.
*** [22] computes scores by taking average over videos.



(a) Training of $\mathcal{G}$ in GCL$_B$ (b) Training of $\mathcal{D}$ in GCL$_B$

Figure 5. Convergence of $\mathcal{G}$ and $\mathcal{D}$ in GCL with/without Negative Learning (NL). We test different pseudo reconstruction targets in NL. Best performance is observed for 'ones' NL target.

posed GCL$_{PT}$ outperforms all existing one-class classification based anomaly detection methods. It is despite the fact that while training GCL$_{PT}$, no labeled supervision is used. In contrast, OCC methods use clean normal class for training which provides extra information compared to our unsupervised training based GCL.

In another experiment, the autoencoder is pre-trained on only the normal class of the training data, which makes the setting comparable with the one-class classifiers. This scheme of extra information provided in the form of normal class labels, referred as GCL$_{OCC}$ in Table 1, obtains an improved performance of 74.20% on UCFC which is significantly better than all existing state-of-the-art OCC methods. It is interesting to note that GCL$_{OCC}$ yields comparable performance to the approach proposed by Sultani *et al*. [49] which utilizes video-level labels for training.

Although GCL aims at unsupervised cooperative learning, we also extended it to incorporate weak-supervision. The results for this version are reported as GCL$_{WS}$ in Table 1. Despite using fairly simple networks of $\mathcal{G}$ and $\mathcal{D}$ without any bells and whistles, GCL$_{WS}$ obtains comparable results to several existing weakly-supervised learning methods.

We also evaluated our approach on **ShanghaiTech dataset** [28] and the results are compared with the existing SOTA methods in Table 1. On this dataset, our proposed GCL$_B$ obtained 72.41% AUC which is more than 10% better than AE$_{AllData}$ showing the effectiveness of the baseline approach. GCL$_{PT}$ obtained 78.93% AUC which is 6.5% better than GCL$_B$ demonstrating the importance of unsupervised pre-training for jump-start. Also, although unsupervised, GCL$_{PT}$ outperformed all existing OCC methods. The experiments on ShanghaiTech dataset also demonstrate the effectiveness of the proposed GCL$_B$ and GCL$_{PT}$ algorithms for anomalous events detection using unlabelled training data.

4.2. Ablation Study and Analysis

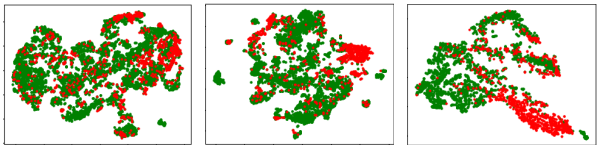
Analysis of different components, design choices, qualitative results and inclusion of supervision are discussed next.

Component-wise ablation study. A detailed ablation analysis of GCL framework with various design choices is reported in Table 2 on the UCFC dataset. It can be seen that an autoencoder trained using all training dataset without any supervision AE$_{AllData}$ yields a significantly low performance of 56.32% compared to the one trained on clean normal data in OCC setting AE$_{OCC}$ yielding AUC of 65.76%. Training an autoencoder AE$_{TD}$ with our proposed frame temporal difference based unsupervised pre-processing brings the performance closer to AE$_{OCC}$, which demonstrates the superiority of our proposed pre-processing approach. Using negative learning enhances the overall performance of GCL$_B$ over the counterpart training without negative learning GCL$_{w/oNL}$ by 3.94%. Our complete unsupervised system GCL$_{PT}$ which utilizes negative learning and unsupervised pre-training enhances the overall performance to 71.04%. In addition, in GCL$_{OCC}$ adding one-class supervision improves this performance even further by demonstrating an AUC of 74.20%. This also re-validates our claim of the overall benefit that OCC may have over a completely unsupervised setting, making them different from the unsupervised approaches.

Table 2. Ablation analysis of GCL Algorithm: performance of different components with varying supervision levels.

	Supervision		Negative learning	Unsup. pre-training	AUC%
	OCC	Unsup.			
AE$_{AllData}$	-	✓	-	-	56.32
AE$_{OCC}$	✓	-	-	-	65.76
AE$_{TD}$	-	-	-	✓	63.84
GCL$_{w/oNL}$	-	✓	-	-	64.23
GCL$_B$	-	✓	✓	-	68.17
GCL$_{PT}$	-	✓	✓	✓	71.04
GCL$_{OCC}$	✓	-	✓	✓	74.20

Evaluating negative learning approaches. Experiments are performed with and without Negative Learning (NL) with GCL framework on UCFC dataset. For the case of NL, GCL$_B$, the performances of different pseudo targets (discussed in section 3.2.5) are also compared in Fig. 5. Three different types of pseudo targets are compared: 'ones' for



(a) AE$_{AllData}$ (b) AE in GCL$_{w/oNL}$ (c) AE in GCL$_B$

Figure 6. Feature reconstructions, using tSNE, with a) AE$_{AllData}$, b) AE in GCL$_{w/oNL}$ without NL, and c) AE in GCL$_B$ with NL using 'ones' pseudo targets. Red and green points represent ground truth anomalous and normal samples, respectively. Using negative learning (NL), most of the anomalous samples get clustered separately from the normal samples, which is the underlying desideratum of providing pseudo reconstruction targets.

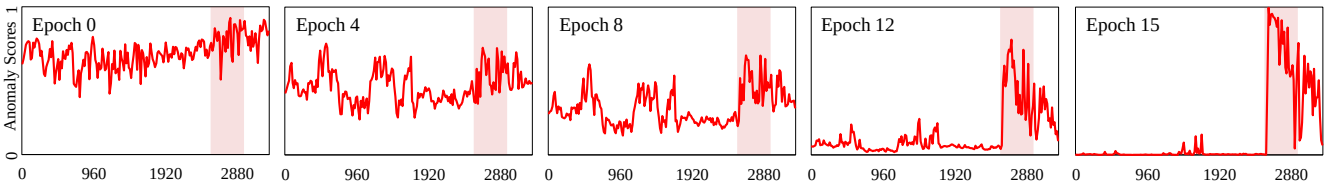


Figure 7. Evolution of the frame-level anomaly scores in GCL$_B$ framework during training. Note that our unsupervised approach successfully produces significantly higher scores in the anomalous portions whereas lower scores in the normal portions. Anomaly ground truth is shown as red boxes, and the video is *Explosion013* from UCF-Crime. Interestingly, the anomaly score stays higher after the anomalous ground truth is over which is essentially due to aftermath of the explosion that network figures to be anomalous.

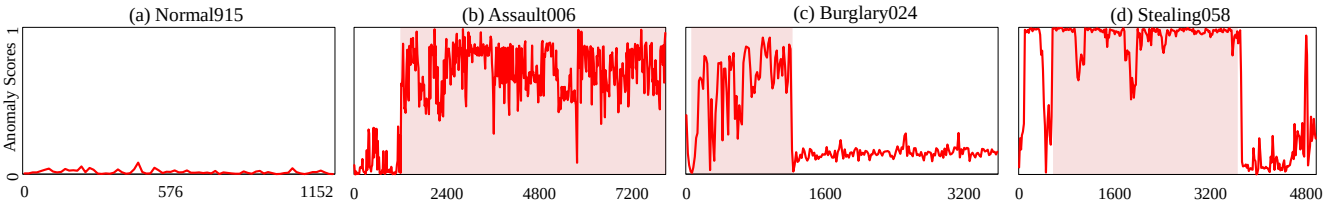


Figure 8. Anomaly scores by GCL$_{PT}$ are low in normal regions and high in abnormal regions on four different UCFC videos.

all ones, 'replace' with random normal, and 'gaussian' with $\mu = 0$ and $\sigma = \{1.5, 6.0\}$. Fig. 5 shows that the 'ones' pseudo target works better than its counterpart approaches. Gaussian perturbations with $\sigma = 1.5$ demonstrate almost identical trend to the model without any negative learning GCL$_{w/oNL}$, and with $\sigma = 6$ the performance improves but still lower than the 'ones' performance, GCL$_B$. This could be that the fixed pseudo-target helps consistent learning of GCL framework resulting in better discrimination.

To further explore the significance of NL, we provide tSNE visualizations of the reconstructions produced by AE$_{AllData}$, GCL$_{w/oNL}$ AE without NL, and GCL$_B$ AE with NL (trained using 'Ones' pseudo label) in Fig. 6. AE$_{AllData}$ is trained using all training data without any labels. Both GCL$_B$ AEs with and without NL demonstrate a superior discrimination over AE$_{AllData}$. Moreover, in AE with NL (Fig. 6(c)), the anomalous features are forming a distinct cluster which shows that the use of NL with pseudo reconstruction target is effective than using no NL option.

Qualitative analysis. A step by step evolution of our GCL approach is visualized in Fig. 7. As the training proceeds, GCL$_B$ learns to predict true anomalous portions within the video in a completely unsupervised fashion. Fig. 8 shows final anomaly scores predicted by our GCL$_{PT}$ on four different videos taken from UCFC dataset. In Fig. 8(d), some normal portions are also predicted as anomalous. A visual inspection of this video reveals that the beginning and ending frames contain floating text, which is unusual in the training data.

On convergence. We (empirically) validate the convergence of both GCL$_B$ and GCL$_{PT}$ using multiple (10) random seed initialization in Fig. 10. GCL_B and GCL_{PT} obtain an average AUC of 67.09 ± 0.65 and 70.13 ± 0.52, respectively. GCL_{PT} not only improves the overall performance but also reduces the variation over different seeds, thereby demonstrating better convergence.

On adding weak-supervision. In a series of experiments using UCFC, weak video-level labels are infused to the GCL ranging from 33% to 100%. Fig. 9 demonstrates that both $\mathcal{G}$ and $\mathcal{D}$ benefit from the added supervision. Noticeably, there is a significant jump in AUC% upon only providing 33% videos with weak labels which demonstrates the fact that even minimal supervision is quite beneficial for the proposed GCL.

On training $\mathcal{G}$ using its own pseudo-labels. To further understand proposed collaborative training, we also explore a

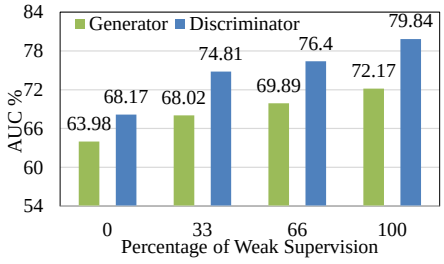


Figure 9. Performance evaluation of $\mathcal{G}$ and $\mathcal{D}$ in weakly supervised GCL$_{WS}$ by increasing supervision level from 0 to 100%.

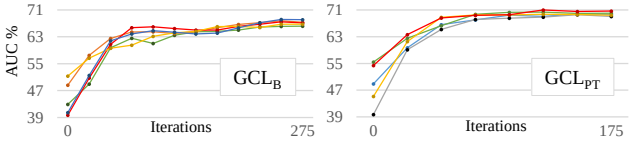


Figure 10. Convergence of both GCL_B and GCL_{PT} by initiating training using several random seeds. GCL_B and GCL_{PT} obtain average AUC of 67.09 ± 0.65 and 70.13 ± 0.52 respectively. GCL_{PT} not only improves the overall performance but also reduces the variation over different seeds, thereby demonstrating better convergence.

possibility of training $\mathcal{G}$ using its own pseudo-labels. We employ negative learning to generate labels for training of $\mathcal{G}$ using the reconstruction error of $\mathcal{G}$ itself. Under this configuration, we observed a performance of 62.28% on UCF crime dataset using ResNext3d features. It is better than 56.32%, the performance of $AE_{AllData}$ (Table 1), however noticeably lower than 71.04%, the performance of our proposed GCL_{PT}. This demonstrates that the usage of $\mathcal{D}$ for pseudo-labeling is critical due to its robust learning under noisy labels [64, 65]. Since $\mathcal{G}$ creates noisy pseudo-labels, $\mathcal{D}$ being robust to noise effectively cleans these labels ensuring the success of the overall collaborative learning.

On using soft labels. In our current configuration, while using pseudo-labels of $\mathcal{G}$ to train $\mathcal{D}$, a threshold is applied to create binary labels from the reconstruction error (eq. (2)). However, it is also possible that we use soft labels instead of thresholding. Carrying out this experiment on UCF crime dataset using ResNext3d features resulted in a AUC of 63.58%. Interestingly, the performance is almost identical to that of AE_{TD} in Table 2. Intuitively, it is because without threshold, $\mathcal{D}$ simply starts replicating the output of $\mathcal{G}$, thereby demonstrating identical performance.

Limitations. The proposed unsupervised setting enables an anomaly detection system to start detecting abnormalities just based on the observed data without any human intervention. In case there is no abnormal event so far, the system may consider the rare normal events as abnormal. However, if a system remains operational for a significant time, the probability of having no abnormal event will be rather very small.

5. Conclusion

We proposed an unsupervised anomaly detection approach (GCL) using unlabeled training videos, which can be deployed without providing any manual annotations. GCL shows excellent performance on two public benchmark datasets with varying supervision levels, including no-supervision, one class and weak-supervision. Finally, we discussed the limitations of unsupervised settings, i.e., the assumption of having anomalies in the training dataset. However, this is more realistic than OCC methods as it is natural to have anomalies in the real-world scenarios.

6. Acknowledgements

This work was supported by the seed-type challenge research project grant funded by Electronics and Telecommunications Research Institute (ETRI) (No. 21YS2700, Development of learning model and data generation/augmentation techniques for data efficient deep learning, 50%) and also supported by ETRI with a grant funded by Ulsan Metropolitan City (22AS1600, the development of intelligentization technology for the main industry for manufacturing innovation and Human-mobile-space autonomous collaboration intelligence technology development in industrial sites, 50%)

References

[1] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. *arXiv preprint arXiv:2110.09742*, 2021. 2, 4, 6

[2] Arslan Basharat, Alexei Gritai, and Mubarak Shah. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[3] Tanmay Batra and Devi Parikh. Cooperative learning with visual attributes. *arXiv preprint arXiv:1705.05512*, 2017. 3

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019. 2

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[6] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. 2

[7] Antoni Chan and Nuno Vasconcelos. Ucsd pedestrian dataset. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(5):909–926, 2008. 2

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 1

[9] MyeongAh Cho, Taeoh Kim, Ig-Jae Kim, and Sangyoun Lee. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *arXiv preprint arXiv:2010.07524*, 2020. 6

[10] 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. 1, 2, 6

[11] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 5

[12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *arXiv preprint*, arXiv:1711.09577, 2017. 6

[13] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 1, 6

[14] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in neural information processing systems*, 29:820–828, 2016. 3

[15] Matthäus Heer, Janis Postels, Xiaoran Chen, Ender Konukoglu, and Shadi Albarqouni. The ood blind spot of unsupervised anomaly detection. In *Medical Imaging with Deep Learning*, 2021. 2

[16] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017. 1

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[18] John Taylor Jewell, Vahid Reza Khazaie, and Yalda Mohsenzadeh. Oled: One-class learned encoder-decoder network with adversarial context masking for novelty detection. *arXiv preprint arXiv:2103.14953*, 2021. 2

[19] Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, and Masao Sakauchi. Traffic monitoring and accident detection at intersections. *IEEE transactions on Intelligent transportation systems*, 1(2):108–118, 2000. 2

[20] Jin-Hwa Kim, Do-Hyeong Kim, Saehoon Yi, and Taehoon Lee. Semi-orthogonal embedding for efficient unsupervised anomaly segmentation. *arXiv preprint arXiv:2105.14737*, 2021. 6

[21] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2019. 6

[22] Tangqing Li, Zheng Wang, Siying Liu, and Wen-Yan Lin. Deep unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3636–3645, 2021. 6

[23] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2019. 1

[24] 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. 1, 6

[25] Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through contrast based evaluation networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3899–3908, 2019. 1

[26] 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

[27] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 1, 6

[28] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 7

[29] Snehashis Majhi, Srijan Das, and François Brémond. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection. 6

[30] Gérard Medioni, Isaac Cohen, François Brémond, Somboon Hongeng, and Ramakant Nevatia. Event detection and analysis from video streams. *IEEE Transactions on pattern analysis and machine intelligence*, 23(8):873–889, 2001. 2

[31] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. Angry crowds: Detecting violent events in videos. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016. 2

[32] Asim Munawar, Phongtharin Vinayavekhin, and Giovanni De Magistris. Limiting the reconstruction capability of generative neural network using negative learning. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017. 4

[33] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8679–8687, 2019. 1

[34] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[35] Trong Nguyen Nguyen and Jean Meunier. Hybrid deep network for anomaly detection. *arXiv preprint arXiv:1908.06347*, 2019. 2

[36] 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. 1, 3, 6

[37] Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, 18(11):1544–1554, 2008. 2

[38] Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*, 2020. 2

[39] Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 173–183, October 2021. 2, 6

[40] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018. 1

[41] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets.

In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017. 1, 2

[42] Huamin Ren, Weifeng Liu, Søren Ingvor Olsen, Sergio Escalera, and Thomas B Moeslund. Unsupervised behavior-specific dictionary learning for abnormal event detection. In *BMVC*, pages 28–1, 2015. 2

[43] Mohammad Sabokrou, Mahmood Fathy, Guoying Zhao, and Ehsan Adeli. Deep end-to-end one-class classifier. *IEEE transactions on neural networks and learning systems*, 2020. 2

[44] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017. 1, 2

[45] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018. 1

[46] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. Deep appearance features for abnormal behavior detection in video. In *International Conference on Image Analysis and Processing*, pages 779–789. Springer, 2017. 1, 2

[47] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 722–727. IEEE, 2018. 6

[48] Jessie James P Suarez and Prospero C Naval Jr. A survey on deep learning techniques for video anomaly detection. *arXiv preprint arXiv:2009.14146*, 2020. 2

[49] 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, 5, 6, 7

[50] Waqas Sultani and Jin Young Choi. Abnormal traffic detection using intelligent driver model. In *2010 20th International Conference on Pattern Recognition*, pages 324–327. IEEE, 2010. 2

[51] 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. *arXiv preprint arXiv:2101.10030*, 2021. 2, 3, 6

[52] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019. 3, 6

[53] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014. 2

[54] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 1

[55] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 6

[56] Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751M. International Society for Optics and Photonics, 2018. 2

[57] 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. 6

[58] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1519, 2015. 1

[59] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 2

[60] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 2

[61] Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, and Junsong Yuan. Temporal structure mining for weakly supervised action detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5522–5531, 2019. 1

[62] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020. 1, 2, 5, 6

[63] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, Arif Mahmood, and Seung-Ik Lee. Cleaning label noise with clusters for minimally supervised anomaly detection. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 1, 2, 3, 6

[64] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*, pages 358–376. Springer, 2020. 1, 2, 3, 4, 6, 9

[65] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020. 1, 2, 3, 6, 9

[66] 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. 6

[67] Tianzhu Zhang, Hanqing Lu, and Stan Z Li. Learning semantic scene models by object classification and trajectory clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1947. IEEE, 2009. 2

[68] Ying Zhang, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Shun Sakai. Video anomaly detection based on locality sensitive hashing filters. *Pattern Recognition*, 59:302–311, 2016. 1

[69] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 2

[70] 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. 1, 2, 3, 5, 6

[71] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019. 6