Towards Open Set Video Anomaly Detection

Yuansheng Zhu, Wentao Bao, and Qi Yu

Rochester Institute of Technology
{yz7008, wb6219 and qi.yu}@rit.edu

Abstract. Open Set Video Anomaly Detection (OpenVAD) aims to identify abnormal events from video data where both known anomalies and novel ones exist in testing. Unsupervised models learned solely from normal videos are applicable to any testing anomalies but suffer from a high false positive rate. In contrast, weakly supervised methods are effective in detecting known anomalies but could fail in an open world. We develop a novel weakly supervised method for the OpenVAD problem by integrating evidential deep learning (EDL) and normalizing flows (NFs) into a multiple instance learning (MIL) framework. Specifically, we propose to use graph neural networks and triplet loss to learn discriminative features for training the EDL classifier, where the EDL is capable of identifying the unknown anomalies by quantifying the uncertainty. Moreover, we develop an uncertainty-aware selection strategy to obtain clean anomaly instances and a NFs module to generate the pseudo anomalies. Our method is superior to existing approaches by inheriting the advantages of both the unsupervised NFs and the weakly-supervised MIL framework. Experimental results on multiple real-world video datasets show the effectiveness of our method.

Keywords: Video anomaly detection, weakly supervised learning, open set recognition, normalizing flows.

1 Introduction

Traditional video anomaly detection aims to detect abnormal events that significantly deviate from normal ones. Examples of such abnormal events include human crimes, natural disasters, and traffic accidents, to name a few. It has been successfully applied to many real-world applications [32]. However, unseen novel anomalies may occur after a well-trained supervised model has been deployed in an open world. Being aware of the unseen anomalies essentially leads to the Open-set Video Anomaly Detection (OpenVAD) problem (see Figure 1), which is under-explored in literature despite being critical to real-world applications.

Unsupervised learning is one of the typical ways to handle unseen anomalies in existing literature [28,29,47,55,16]. They aim to learn representative features of normal events only from normal videos. However, as unsupervised methods neglect the anomaly information, they are in general less effective to detect complicated anomalies that are similar to normal samples in the feature space [4]. Besides, (weakly-)supervised methods can use annotated anomaly videos so that
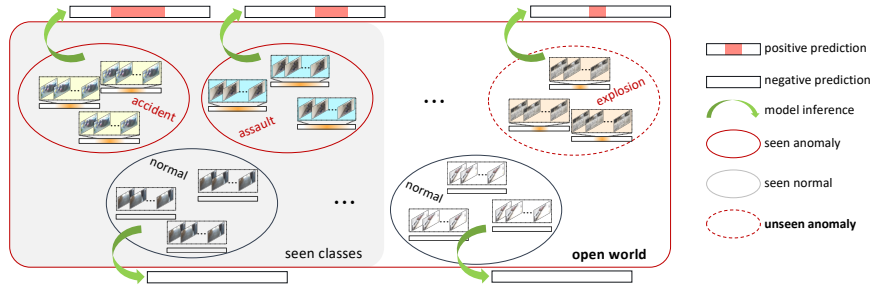
Fig. 1: **OpenVAD Task.** We propose to conduct OpenVAD through the weakly-supervised video anomaly detection in an open world. In training, only a closed-set of anomaly videos with video-level annotations are observed along with normal videos. In testing, the model is asked to identify and localize the anomaly events from videos in an open world where arbitrary unseen anomalies could exist.

their performances are generally better. However, the learned models are limited to detecting a closed set of anomaly categories and unable to handle the arbitrary unseen anomalies. Inspired by the recent advances in open set recognition [36,3,13,2], we propose to perform video anomaly detection in an open world by extending the existing weakly-supervised video anomaly detection paradigm. The goal is to achieve accurate anomaly detection in an open set by learning a model only with video-level annotations from a closed subset of anomaly types.

Most existing works [42,62,51,46] formulate weakly supervised anomaly detection as a multiple instance learning (MIL) problem [10], where a video is modeled as a bag of instances (*e.g.,* video clips). Due to the lack of fine-grained instance labels, simply assigning all instances in an abnormal (*i.e.,* positive) bag with the same anomaly label inevitably incurs severe labeling noise. Such noise further results in a high false positive rate, *i.e.,* falsely treating normal events as abnormal ones, when using a trained model for detection. Therefore, it is essential to select the clean anomaly instances in MIL. Besides, to enable the MIL model to be aware of the unseen anomaly in testing, the open space risk of MIL should be bounded while it is under-explored in the existing literature.

To tackle these challenges, we first leverage graph neural networks (GCNs) and a triplet loss to learn representative instance features. Then, we instantiate the MIL with evidential deep learning (EDL) [37,39,2] and use the predicted evidence to help select anomaly instances with high *cleanness* for robust MIL training. To bound the open space risk, an unsupervised normalizing flows (NFs) is learned from normal instances, which allows us to sample pseudo anomaly samples from the low density area of the instance distribution.

We are among the first few attempts that address video anomaly detection in the open set setting, where arbitrary unseen anomaly events could appear in testing. This is fundamentally more challenging than a using fully-supervised training process as adopted by some existing efforts in [4,34] and some of these methods are not suitable to tackle high-dimensional video data. Though there

are few works [23,12] with a similar open set assumption, our task definition is more realistic in real-world, where the model is weakly-supervised and aims to identify arbitrary unknown anomaly events in testing. By effectively leveraging any available anomaly videos through weakly supervised learning in novel ways, our approach also achieves better detection performance than few existing open set anomaly detection methods (*e.g.,* [27]) that solely rely on normal videos to initialize a detection model through unsupervised learning. In summary, the key contributions are threefold:

- We formulate a novel MIL framework for the OpenVAD problem to detect both seen and unseen video anomalies in a challenging open-world setting.
- We integrate MIL with a normalizing flow-based generative model and evidential learning for high-quality anomaly instance selection.
- We conduct extensive experiments and the results show significant superiority to existing unsupervised and weak supervised learning approaches.

2   Related Work

Video Anomaly Detection. Existing methods to detect anomalies from videos can be categorized into two groups: 1) unsupervised learning and 2) weakly supervised learning. Unsupervised approaches are motivated by the premise that anomaly events are fundamentally challenging to completely enumerate and characterize. Therefore, these approaches conduct either dictionary learning/sparse coding [29,61] or construct a deep auto-encoder [53,8,47]. Unlike unsupervised approaches, weakly supervised approaches [27,42,17,62,51,46] use both normal videos and anomaly videos for training. They take advantage of normal videos and leverage additional anomaly videos to train a binary classifier over the video clips. Among existing weakly supervised approaches, one representative approach leverages a deep auto-encoder to locate clips in anomaly videos with low reconstruction scores and then further reduces these scores through margin learning [27]. Some other approaches [42,17,62] formulate this problem as multiple instance learning (MIL) problem and build an instance-level classifier. To train the classifier properly, these MIL models select one or a set of instances from a positive bag to feed into the classifier and either a fixed number of instances are selected [42,17] or a predefined threshold is set as the selection criterion [62]. Previous MIL based approaches achieve superior performance compared to unsupervised counterparts. The core design of these approaches lies in the measurement of *cleanness* and the instance selection strategy. In this paper, we empirically show that the top-k strategy is robust, which inspires us to develop the *cleanness* measurement and an instance selection strategy by introducing the evidence based learning method.

Open Set Recognition. Open Set Recognition (OSR) aims to identify the unknowns while keeping reasonable closed set performance in the open set testing. Since the unknowns are never seen by the model and are out-of-distribution (OOD) with respect to the in-distributional (ID) training data, they are referred

to as "unknown unknown class" in existing literature [14,54]. In contrast, the exact classes of ID data are known and are called "known known classes". The OSR approaches attempt to classify test samples from "known known classes" and detect test samples from "unknown unknown classes". The most recent work [31] encourages learning a multi-class classifier to give low activation values for external OOD data and large values for ID data. Some works do not depend on external OOD data in training by using either specialized loss functions [2,7] or unknown class generation [63]. Other works [59,3] address the OSR problem by integrating probabilistic density estimation in the latent feature space with a multi-class classifier. Our proposed OpenVAD and the traditional OSR problems can both be regarded as a generalized OOD detection problem [54], because of the semantic distribution shift between testing and training datasets. A key distinction of the proposed OpenVAD is that it does not care about if an anomaly video is from seen or unseen categories in testing. Instead, it focuses on distinguishing arbitrary anomaly activities from normal ones in an open world.

3   Methodology

Our work aims to handle the OpenVAD problem under a weakly supervised setting. Specifically, a model is trained using normal videos and a closed set of untrimmed anomaly videos with video (or bag)-level labels. Each anomaly video contains at least one anomaly event (*i.e.,* positive instance). The model is tested in an open world, where unseen anomaly events may occur, and the model is expected to robustly identify all anomaly clips (*i.e.,* segments of a video) during testing. Since we only have access to the video-level anomaly class labels, MIL provides an ideal learning framework to leverage the weakly supervised learning signals. Specifically, an untrimmed video is regarded as a bag X, which contains N temporal clips as instances in the bag: $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The learning objective of MIL is to train an instance-level binary classifier $\Phi : \mathbf{x}_n \rightarrow y_n$ with only bag-level class label $Y \in \{0, 1\}$ in the training, while the instance-level labels $y_n \in \{0, 1\}$ with $y_n \in \{y_1, \ldots, y_N\}$ are not available. In MIL, a positive bag $(X, Y = 1)$ contains at least one positive instance (*e.g.,* anomaly event), *i.e.,* $\exists n \in [1, N], y_n = 1$, while a negative bag consists of only negative instances (*e.g.,* normal event), *i.e.,* $\forall n \in [1, N], y_n = 0$.

For the OpenVAD problem, the model is expected to handle the semantically shifted testing data that contains unseen anomaly events. In testing, given an untrimmed video, the model is expected to identify the anomaly clips, which could be either seen or unseen anomaly if the video contains any anomaly events.

Overview of the Framework. Our method is developed as a MIL framework, as shown in Figure 2. The instances are structured as graph data to address the temporal consistency and feature similarity of video clips. The MIL framework consists of two graph convolutional networks (GCN) followed by an evidential deep learning (EDL) head, which are trained by weakly-supervised MIL loss and triplet loss. The classification confidence and uncertainty are used to select
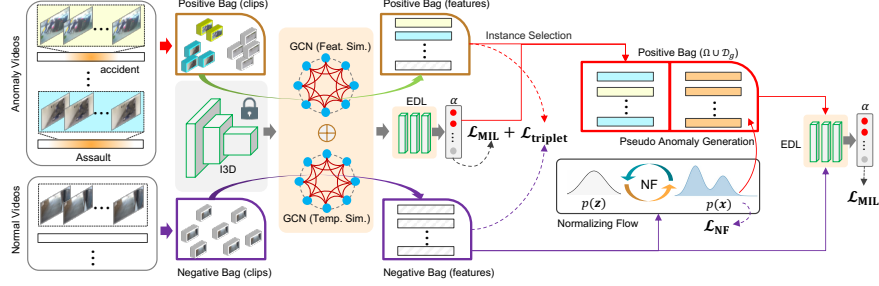
Fig. 2: Overview of the OpenVAD Framework. We first integrate two GCNs feature extractors and the evidential deep learning (EDL) into the multiple instance learning (MIL), which are learned by $\mathcal{L}_{\mathrm{MIL}} + \mathcal{L}_{\mathrm{triplet}}$. Then, the classification confidence and uncertainty from the trained EDL head are used to select instances with high *cleanness*. We further utilize the GCNs features of normal video clips to train a normalizing flow model, from which pseudo anomaly features are generated. Together with the normal videos, the selected clean anomaly and the generated pseudo anomaly are gathered to refine the EDL head by $\mathcal{L}_{\mathrm{MIL}}$.

clean anomaly instances for robust MIL training. To reduce the open space risk in OpenVAD tasks, we further utilize normal videos to train a normalizing flow (NF) model to generate pseudo anomalies. The selected and generated anomalies are used to fine-tune the EDL classification head. In testing, our model could correctly identify an unseen anomaly event from untrimmed video data. See Appendix Section A for the summary of notations.

3.1   Instance Feature Representation Learning

We take advantage of a pre-trained action recognition backbone (*e.g.,* I3D [5]) to extract features of raw video clips, denoted as **x**. Given the temporal and visual similarity of close clips in a video, we propose to enhance the feature **x** by constructing temporal and feature similarity graphs, respectively. Furthermore, to enlarge the feature discrepancy between normal and anomaly instances, a triplet learning constraint is introduced in the feature embedding space.

More specifically, since each bag is formed from a video, there exist natural dependencies among instances in the bag. Thus, video clips that are located closely in temporal and feature space should have similar feature representation. We leverage a graph neural network to model such dependency. Given a graph defined by the adjacency matrix A and graph node features X (*i.e.,* a bag), the graph neural network is updated according to the following rule [21]:

$$\mathcal{H}(X, A) = \sigma \left(\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} X W \right), \tag{1}$$

where $\hat{A} = A + I$ and I is the identity matrix, $\hat{D}$ is the degree matrix of $\hat{A}$, and W is the learnable weights. We employ two GCNs where the adjacent matrix

A is determined to capture feature similarity and temporal consistency, respectively [51]. The two GCN embeddings are further concatenated to form the final instance representation. To avoid a cluttered notation, we denote the refined instance representation as $\mathcal{H}(\mathbf{x})$.

The GCNs help to compactly embed the neighbouring instances in the same bag. When multiple bags are involved, we further employ a triplet loss to achieve good separability of instances from different classes and compactness of instances in the same class. A triplet consists of an anchor instance, a paired instance from the same class as the anchor, and an unpaired instance from a different class from the anchor, which are sampled from the collection Ω (defined in Eq (4)). The triplet loss is given by

$$\mathcal{L}_{triplet} = [d_{ap} - d_{an} + m]_{+} \tag{2}$$

where m is the margin that controls the desired difference between the paired distance d_{ap} and unpaired distance d_{an}.

In addition to GCNs, the consistency loss function [42,46] and other network architectures [46] (e.g., LSTM and CNN) also have been investigated to model the relationship among instances. We choose GCNs because they have been proved to be effective in close set video anomaly detection tasks [62,51]. However, the representation learning in the open world is more complex than in a close set. The triplet loss will make the normal data stay close to each other while remaining dissimilar to the seen anomaly. This facilitating a normalizing flow (NF) to learn their distribution in the following stage.

3.2   Uncertainty-aware Multiple Instance Learning

To train an instance-level classifier with only bag-level annotations, a naive solution is to assign a noisy version of label y_n for each instance $\mathbf{x}_n$ by directly using the bag-level label, *i.e., $y_n = Y$*. As a result, most of instances in the positive bag may be incorrectly labelled (due to normal video clips in anomaly videos) during model training. These mislabeled instances may degrade the model performance as the learning targets are noisy. Furthermore, a MIL model instantiated by deep neural networks (DNNs) is usually sensitive to noisy labels because DNNs are of high capacity to fit random labels [58]. Therefore, instance selection from positive bags poses a critical issue that may significantly impact the performance of a MIL model [51,62].

To address this challenge, top-k methods (*e.g.,* [24]) select the "most positive" instances from the positive bag in training. Two recent works [46,51] adapt this technique and achieve decent performance in close set video anomaly detection tasks. Most top-k methods follow a "sort-then-rank" process, where a critical step is to determine the *cleanness* of each instance to evaluate how likely the instance is a clean positive one (*i.e.,* an anomaly event). In this work, we propose to quantify the classification uncertainty by evidential deep learning. We then introduce an instance selection approach based on uncertainty and confidence.

Given an instance $\mathbf{x}$, we introduce a binary random variable $p \in \{0,1\}$ that represents the chance that it is an anomaly, which follows a Bernoulli distribution. Similar to EDL [37], we further introduce a Beta distribution $\text{Beta}(\boldsymbol{\alpha})$ as the conjugate prior of the Bernoulli likelihood. In our framework, the learning objective is to train a DNN to directly predict the Beta distribution parameter $\boldsymbol{\alpha}$ instead of p. According to Subjective Logic and evidence theory [18,38], the predicted parameters $\boldsymbol{\alpha} = (\alpha_+, \alpha_-)^\top$ can be regarded as the evidences over the positive (anomaly) and negative (normal) classes, which can be expressed as

$$\boldsymbol{\alpha} = \Phi(\mathcal{H}(\mathbf{x})) + \mathbf{a}W \tag{3}$$

where $\mathbf{a} = (a_+, a_-)^\top$ is the base rate with $a_+ = a_- = 1/2$ and $W = 2$ in our binary case; $\mathcal{H}$ and Φ denote GCNs based backbone as the feature extractor and a DNN as the EDL classification head, respectively. The GCNs encode the relationship among instances in a bag and output discriminative video clip features. In practice, the EDL classification head consists of two fully connected layers and ReLU activation to predict non-negative class-wise logits. The predictions α_+ and α_- correspond to the instance class $y = 1$ and $y = 0$ respectively, indicating the virtual quantity of evidence to support $\mathbf{x}$ to be classified as positive or negative.

The benefit of placing the Beta prior on the Bernoulli likelihood is that the prediction uncertainty can be efficiently and accurately quantified according to recent Dirichlet-based models [22], enabling the MIL model to *know what it does not know*. The predicted probabilities are the ratios of the evidence, *i.e.*, $\mathbb{E}[p_+] = \alpha_+/\alpha_0$ and $\mathbb{E}[p_-] = \alpha_-/\alpha_0$, where $\alpha_0 = \alpha_+ + \alpha_-$, and the evidential uncertainty (or vacuity) is quantified as $u = 2/\alpha_0$.

Based on the evidential uncertainty above, we propose to train the MIL by selecting the *clean* samples, which form a set:

$$\Omega = \left\{ \mathbf{x}_i | p_+^{(i)} \geq \tau_p, \alpha_+^{(i)} \geq \tau_u \right\}, \tag{4}$$

where $p_+^{(i)}$ and $\alpha_+^{(i)}$ are the confidence score and evidence that support the sample $\mathbf{x}_i$ being classified as an anomaly; τ_p and τ_u are thresholds to control the size of the set. The MIL model is trained by minimizing the Type II maximum likelihood loss (MLL) [37]:

$$\mathcal{L}_{\text{MIL}}(\mathbf{x}_i') = \sum_{k=1}^{2} \left[\hat{y}_k^{(i)} \left(\log(\alpha_0^{(i)}) - \log(\alpha_k^{(i)}) \right) \right] \tag{5}$$

where the training sample $\mathbf{x}_i'$ is from either the selected set Ω or the negative bag $\mathcal{N} = \{\mathbf{x}_i | y_i = 0\}$, *i.e.*, $\mathbf{x}_i' \in \Omega \cup \mathcal{N}$, $\hat{y}$ is the one-hot version of the label y. The benefit of Type II MLL lies in that it not only enforces the predicted probability to be as close to the ground-truth bag label, but also increases the evidence for observed data to support better instance selection.

MIL is intrinsically learning using the noisy labels, which is a long-standing challenge. Recent research approaches this problem generally from sample selection [30,56,57,52], sample re-weighting [50,6], and meta-learning [25,60] perspectives. We choose to explore sample selection because it is well-justified and

empirically works well. The sample selection methods often leverage the memorization effect of DNNs to identify the clean samples, which stems from the observation that DNNs tend to fit clean samples faster than noisy ones [1]. Here, τ_p in Eq. (4) is derived from the memorization effect, which encourages Ω to retain those qualified samples, *i.e.,* samples whose predicted probability p_+ is most close to the learning target, $y = 1$. However, sample selection methods could severally suffer from the bias confirmation [45] caused by incorrect selection, which is challenging to solve. The proposed uncertainty aware MIL is distinguished from exiting efforts [51,48,42] for mitigating the bias confirmation by filtering out fluctuating samples (*i.e.,* those with a small α_+ less than τ_u), which may be incorrectly selected if only p_+ is considered.

3.3   Pseudo Anomaly Generation

For the OpenVAD problem, open space risk should be bounded in training so that the model could be aware of the semantically shifted unseen data. However, unseen anomaly videos are not available during training under the open set setting. Inspired by recent generative open set modeling [13,35] and normalizing flows (NFs), we propose to leverage NFs to generate pseudo anomaly instances in the latent feature space.

NFs are the flow-based deep generative models that could explicitly learn the probability density of data [43,44]. It has been shown effective for realistic data generation, exact posterior inference, and exact density estimation. NFs consist of composition of invertible learnable functions f_l to transform between a simple distribution $p(\mathbf{z})$ and the complex data distribution $p(\mathbf{x})$, *i.e.,* $\mathbf{z} = f(\mathbf{x})$ and $\mathbf{x} = f^{-1}(\mathbf{z})$ where $f = f_1 \circ \cdots \circ f_L$. To enable exact density estimation and data generation, the invertible transformation is constructed following the change-of-variable rule:

$$p(\mathbf{x}) = p(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^\top} \right) \right| \tag{6}$$

where $\mathbf{z}$ typically follows a non-informative Gaussian distribution. Eq. (6) is associated with the determinant of Jacobian, which can be efficiently computed when a proper bijective function f_l is designed to make the Jacobian triangular. Recent NF models such as RealNVP [11], Glow [19], and IAF [20] use different coupling layers to address the functional invertibility and computational efficiency of the transformation f_l based on deep neural networks. In this paper, we adopt the IAF model because it is efficient for sampling purpose, which is critical for pseudo anomaly generation. The coupling layer f_l is designed as

$$x_i = z_i \times s(\mathbf{z}_{1:i-1}) + t(\mathbf{z}_{1:i-1}), \tag{7}$$

where x_i and z_i are the i-th entries of feature vector $\mathbf{x}$ and $\mathbf{z}$, respectively. The scaling and shifting functions s and t are instantiated by deep neural networks.

To train the flow model, the goal is to maximize the log-likelihood of the normal video data, which is equivalent to minimizing the following objective:

$$\mathcal{L}_{\mathrm{NF}} = \mathbb{E}_{\mathcal{D}(X,Y=0)} \left[-\log p(\mathcal{H}(\mathbf{x})) \right] \tag{8}$$

Note that here we learn the flow model f in the latent feature space $\mathcal{H}(\mathbf{x})$. This alleviates the difficulty of directly learning in a high-dimensional space.

The generation is based on the assumption that pseudo anomalies are outliers with respect to the normal class. Therefore, a pseudo anomaly instance $\tilde{\mathbf{x}}$ should be sampled from the low density region of the probability distribution of the normal class $p(\mathbf{x}|y = 0)$. In this paper, we first use the tractable $p(\mathbf{z})$ to repeatedly sample the latent feature $\tilde{\mathbf{z}}_i$. Then, the pseudo samples are generated by the flow model of the normal class, $i.e.,$ $\tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}; y = 0)$ where $g(\cdot) = f^{-1}(\cdot)$. Finally, only the pseudo samples with low density ($i.e.,$ with a small $p(\tilde{\mathbf{x}}|y = 0)$) are preserved as pseudo anomaly instances. The following equations summarize this procedure:

$$\tilde{\mathbf{z}} \sim p(\mathbf{z}), \ \tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}; y = 0), \ \mathcal{D}_g = \{\tilde{\mathbf{x}}|p(\tilde{\mathbf{x}}|y = 0) \leq \epsilon\} \tag{9}$$

where $p(\tilde{\mathbf{x}}|y = 0) = p(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} \right) \right|$. The benefit of generating pseudo anomaly instances from normal class distribution is that the class boundary of the normal class is constrained by the pseudo anomalies so that the open space risk is well managed. Regarding the way of generating pseudo anomaly instances, unlike previous work [35] that generates in the pixel space, we perform generation in the feature space, similar to [13]. Spawning from the feature space has dual advantages over spawning from the pixel space: 1) the latent space has a much lower dimension, which allows fast sampling for the NFs, and 2) the latent features are more discriminative than the raw pixels [9]; thus the generated data are less likely to be negatively impacted by the background noise in the video data. In contrast, [13] uses the Gaussian mixtures to model the class-wise in-distribution data density. We argue that the Gaussian distribution is not flexible to model the complex in-distribution data. NFs can seamlessly solve this problem owing to its high capability via a deep composition of the bejiection function f.

NFs have a high capability of density estimation while free of supervision. A natural question that arises is whether we can directly ensemble the prediction results of a normalizing flow and a classifier. Ensemble is widely used yet seems to be simpler than pseudo anomaly generation. However, we remark that the ensembles can not inherit the benefit from unsupervised learning and (weakly-) supervised learning paradigms. This is because that NFs cannot detect the complicated anomaly ($e.g.,$ anomaly samples that look similar to normal ones), which is consistent with the observations in [33]. Our motivation is to only leverage the strengths ($i.e.,$ generating anomaly) and avoid weaknesses ($i.e.,$ detecting anomaly) of NFs.

3.4   Multi-Stage Model Training and Inference

Since NFs are used for generating pseudo anomaly instances for the target MIL model, we propose a multi-stage model training scheme. Specifically, we first train a feature encoder $\mathcal{H}(\cdot)$ and an EDL classifier $\Phi(\cdot)$ by minimizing the weighted sum of MIL loss and triplet loss. Then, the feature encoder $\mathcal{H}$ is frozen and the flow model $f(\cdot)$ is trained in an unsupervised way using only normal

---

Algorithm 1 Multi-stage training

Input: $\mathcal{D}_{train} = \{X, Y\}$, number of epochs for training EDL model (Φ) and NFs model (f) T_{EDL}, T_{NFs}, thresholds τ_p, τ_p, ϵ, loss weight β
Output: feature encoder $\mathcal{H}$, EDL classifier Φ.
  ▷ *Stage 1*: Warmup training
    Train $\mathcal{H}$ and Φ by minimizing $\mathcal{L}_{\text{MIL}} + \beta\mathcal{L}_{triplet}$.
  ▷ *Stage 2*: Freeze $\mathcal{H}$ and train f
  for $t \in [1, T_{\text{NFs}}]$ **do**
    Fetch m normal videos, *i.e.*, $\{\mathbf{x}_i\}_{i=1}^{m} \sim \mathcal{D}_{train}(X, Y = 0)$.
    Update NFs model f by minimizing $\mathcal{L}_{\text{NF}}$.
  end for
  ▷ *Stage 3*: Freeze $\mathcal{H}$ and f, and fine-tune Φ
  for $t \in [1, T_{\text{EDL}}]$ **do**
    Fetch m normal and anomaly videos from $\mathcal{D}_{train}$.
    Construct the clean positive set Ω using Eq. (4).
    Augment Ω by sampling m pseudo anomalies using Eq (9), i.e., $\Omega \leftarrow \Omega \cup \mathcal{D}_g$.
    Update EDL model Φ by minimizing $\mathcal{L}_{\text{EDL}}$.
  end for

---

videos. Finally, the EDL classifier is fine-tuned by using the selected positive instances, the generated pseudo anomaly instances, and the normal videos. Algorithm 1 shows the details of the training process. It is worth to note that during inference, the flow model does not involve additional overhead we use the prediction of EDL classifiers to score samples and the anomaly score is defined as the mean probability of being an anomaly: $\mathbb{E}[p_+] = \frac{\alpha_+}{\alpha_0}$.

4   Experiments

Since OpenVAD is less explored, no standard evaluation setup is available for this type of task. To adequately assess the method and inform the practical choice, we develop a benchmark and test several state-of-the-art methods. Through comprehensive comparison on various datasets, we demonstrate the effectiveness of the proposed framework. More experiments details and results can be found in the Appendix, and the link to the source code is available in https://github.com/YUZ128pitt/Towards-OpenVAD.git.

4.1   Datasets and Evaluation Setup

We conduct experiments on three video anomaly detection datasets of different scales. Videos in these datasets are untrimmed, and only video-level labels are provided during training. The frame-level annotations are used for evaluation purpose. Detailed information for these three datasets is as follows:

– **XD Violence** [51]: This dataset contains totally 4,754 videos collected from movies, online videos, sport streaming, surveillance cameras and CCTVs. There are five types of anomaly events, and each anomaly video may contain one or more types of anomalies, including *Abuse, Riot, Shooting, Fighting, Car accident*, and *Explosion*.
– **UCF Crime** [42]: This dataset contains totally 1,900 videos collected from surveillance cameras from variant scenes. There 13 different types of violence-related anomaly videos: *Abuse, Burglary, Robbery, Stealing, Shooting, Shoplifting, Assault, Fighting, Arson, Explosion, Arrest, Road Accident*, and *Vandalism*. There are 1,610 training videos (of which 810 are anomaly) and 290 test videos (of which 140 are anomaly).
– **ShanghaiTech Campus** [26]: This dataset consists videos from 13 scenes with different light conditions and camera angles. Anomaly events are defined as actions that contain sudden motions, such as *chasing* and *brawling*, but there is no anomaly type labels. We follow the split in [62], in which there are totally 238 training videos (of which 63 are anomaly) and 199 testing (of which 44 are anomaly) videos.

Evaluation setup. For the OpenVAD task, we consider a more realistic setting, where testing data contains unseen novel anomalies. To simulate unseen anomaly, we remove one or more types of anomaly from the training data on XD Violence and UCF Crime. We vary the size of anomaly training videos to get comprehensive results under different ratios of unseen anomaly scenarios. For example, in XD-Violence, there are six types of testing anomalies, and we train a model with 1, 2, 3, and 4 types of anomaly videos along with normal videos. We randomly remove a certain number of training anomaly classes and repeat them three times to reduce the evaluation bias of seen anomaly. These evaluation sets reflect our assumption about the most realistic video anomaly detection scenario in an open world. We further test our method with limited number of anomaly videos on the ShanghaiTech dataset. Since no anomaly labels are provided in this dataset, we randomly choose a small subset of anomaly videos for model training. We follow the previous work to use Area under Operating Characteristic curve (AUC-ROC) as the metric on UCF-Crime and ShanghaiTech, and use Area under Precision and Recall curve (AUC-PR) on the XD-Violence dataset. Both AUC-ROC and AUC-PR depict the overall performance using a sliding threshold, and a larger value indicates better performance.

4.2   Comparison with State-of-the-arts

We use the pre-trained I3D classifier [5] to extract features for every non-overlapped video clip (each clip is a collection of 16 consecutive frames) in a video. During training, we uniformly sample 200/200/32 (XD-Violence/UCF-Crime/ShanghaiTech) video clips features to create a bag with the same size.

Our model shares the core design across three datasets and different tasks. We use two layers of GCNs and construct the EDL model with two fully connected layers stacked with a ReLU activation function. For the NFs, we use the 5 consecutive IAF [20]. See Appendix Section B for more implementation details.

Table 1: AUC-PR (%) results on XD-Violence for anomaly frame detection with various number of **seen anomaly classes**. Methods with (*) in the first column are unsupervised, while the rest are weakly-supervised.

Num. seen anomaly	0	1	2	3	4
OCSVM [40] (*)	27.25	-	-	-	-
Conv-AE [16] (*)	30.77	-	-	-	-
Wu *et al.* [51](off-line)	-	40.67	50.34	60.53	67.77
Wu *et al.* [51](on-line)	-	39.13	50.20	58.87	64.29
RTFM [46]	-	43.54	50.88	58.52	63.65
Ours	-	**45.65**	**54.65**	**64.40**	**69.61**

Results on XD-Violence. Table 1 shows the AUC-PR scores in detecting anomaly frames in settings where 6 types of anomaly activities appear in testing, while only 1, 2, 3, or 4 types of them are seen during training. Our method outperforms both unsupervised and weakly supervised ones in all cases. In particular, our method achieves a significant performance gain when very few anomalies are seen, which clearly justifies its capability in the open world setting. Hasan *et al.* [16] and OCSVM [40] do not rely on anomaly videos compared to other methods. Hasan *et al.* surpasses the OCSVM by around 3% on AUC-PR. However, weakly supervised methods perform much better when there are anomaly videos available for training. Even in the extreme case when only 1 out of 6 types of anomalies are seen, the weakly supervised methods can outperform the unsupervised ones by more than 10%. Moreover, more anomaly videos for training further boost the performance of weakly supervised methods. Our method also achieves better a AUC-ROC score than other weakly supervised methods, indicating that our method makes less false positive predictions (see additional experimental results in Appendix Section D).

Results on UCF-Crime. Table 2 shows that AUC-ROC scores in detecting anomaly frames when 13 types of the anomaly appear during testing, only 1, 3, 6, and 9 types of these anomaly are seen. Results in Table 2 show that our method consistently outperforms all baselines under all training settings. To our surprise, all weakly supervised methods can achieve relatively good performance even when a small subset of the anomaly is seen. It could be due to that most anomalous behaviours are related to human activities, and there is somehow high correlation among various anomalies, *e.g., Abuse*, *Assault*, and *Vandalism*. RTFM [46] achieves better performance than Wu *et al.* [51] when fewer anomaly classes are seen, which may be owing to its feature magnitude component. Nevertheless, our method achieves the most stable performance under various training setups, implying that our proposed framework is robust by learning a compact decision boundary.

Results on ShanghaiTech. Table 3 shows the AUC-ROC under an imbalance scenario: only a few anomalies are available for training. Results indicate

Table 2: AUC-ROC(%) results on UCF-Crime for video anomaly detection with various number of **seen anomaly classes**. Methods with (*) in the first column are unsupervised, while the rest are weakly-supervised.

NUM. SEEN ANOMALY	0	1	3	6	9
SOHRAB*et al.* [41](*)	58.50	-	-	-	-
LU *et al.* [29](*)	65.51	-	-	-	-
BODS [49](*)	68.26	-	-	-	-
GODS [49](*)	70.46	-	-	-	-
CONV-AE [16] (*)	50.60	-	-	-	-
WU *et al.* [51](OFF-LINE)	-	73.22	75.15	78.46	79.96
WU *et al.* [51](ON-LINE)	-	73.78	74.64	77.84	79.11
RTFM [46]	-	75.91	76.98	77.68	79.55
OURS	-	**76.73**	**77.78**	**78.82**	**80.14**

Table 3: AUC-ROC(%) results on ShanghaiTech for video anomaly detection with various number of **seen anomaly videos**. Methods with (*) in the first column are unsupervised, while the rest are weakly-supervised.

NUM.SEEN ANOMALY	0	5	10	15	25
FRAME-PRED [28](*)	73.40	-	-	-	-
MEM-AE [15] (*)	71.20	-	-	-	-
MNAD [15] (*)	70.50	-	-	-	-
VEC [55](*)	74.80	-	-	-	-
CONV-AE [16](*)	60.85	-	-	-	-
WU *et al.* [51](OFF-LINE)	-	65.83	81.54	83.47	88.81
WU *et al.* [51](ON-LINE)	-	66.27	81.03	82.42	88.61
RTFM [46]	-	70.59	83.42	81.50	86.33
OURS	-	**80.40**	**88.24**	**85.58**	**93.99**

that our method performs better in this challenging setting when only 25 or fewer anomaly videos are available for training. In that case, our method outperforms both RTFM [46] and Wu *et al.* [51] by a large margin. In most cases, the performance advantage is more significant when fewer videos are available. This clearly indicates the effectiveness of the proposed method.

4.3   Ablation Study

We conduct ablation study to validate the proposed three components in our framework, *i.e.,* (i) triplet loss defined in Eq. (2), (ii) evidence criteria defined in Eq. (4), and (iii) pseudo anomaly generation. Specifically, we incrementally remove/replace them and compare the performance on XD-Violence, which is the largest dataset with most videos. When removing the evidence criteria, we

Table 4: Ablation study results (%) on XD-Violence

Triplet Loss	Evidence Criteria	Pseudo Anomaly	1 seen anomaly		4 seen anomaly	
			AUC-PR	AUC-ROC	AUC-PR	AUC-ROC
√	√	√	**45.65**	**72.50**	**69.61**	**88.25**
	√	√	45.26	61.30	69.14	83.57
		√	41.79	67.07	68.82	84.78
			40.94	66.83	66.61	84.09

set the Ω to include all samples. To evaluate the quality of pseudo anomalies, we replace it with a Gaussian noise.

Table 4 validates the contribution of each proposed technique. It shows that the combination of the three components achieves the best AUC-PR and AUC-ROC scores, indicating that all these three components positively contribute to the performance of our framework. Notably, the triplet loss contributes the most to the performance gain. This demonstrates the important role of instance representation learning for video anomaly detection. Besides, we note that the evidence criteria produce better AUC-PR scores while worse AUC-ROC scores. This can be explained that the evidence criteria drive the decision boundary toward the positive side in the feature space as it filters out the noisy labels. As a result, the false positive predictions are less while the false negative predictions might increase. Lastly, once the pseudo anomaly generation is applied, the performances of both AUC-PR and AUC-ROC are improved.

5   Conclusion

In this paper, we present a new approach for the OpenVAD problem under weak supervision, which is a highly challenging task by previous unsupervised and weakly supervised methods when being used alone. The OpenVAD scenario is the most realistic scenario in real-world applications, as the anomaly events are fundamentally difficult to be fully enumerated, characterized, and modeled. To address the unique challenges, the proposed framework integrates unsupervised learning (*i.e.,* NFs) and weakly supervised learning (*i.e.,* MIL) in novel ways to benefit from the advantages of both while not suffering from their individual limitations. Different from existing unsupervised methods, the proposed approach makes use of any available anomaly videos without expensive labelling costs. Unlike other weakly supervised methods, our method can detect any type of seen and unseen anomalies in an open world.

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
2. Bao, W., Yu, Q., Kong, Y.: Evidential deep learning for open set action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13349–13358 (October 2021)
3. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1563–1572 (2016)
4. Borisyak, M., Ryzhikov, A., Ustyuzhanin, A., Derkach, D., Ratnikov, F., Mineeva, O.: (1 + epsilon)- class classification: An anomaly detection method for highly imbalanced or incomplete data sets. arXiv preprint arXiv:1906.06096 (2019)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Chang, H.S., Learned-Miller, E., McCallum, A.: Active bias: Training more accurate neural networks by emphasizing high variance samples. arXiv preprint arXiv:1704.07433 (2017)
7. Charpentier, B., Zügner, D., Günnemann, S.: Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. arXiv preprint arXiv:2006.09239 (2020)
8. Chen, Z., Tian, Y., Zeng, W., Huang, T.: Detecting abnormal behaviors in surveillance videos based on fuzzy clustering and multiple auto-encoders. In: 2015 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2015)
9. Cho, M., Kim, T., Kim, W.J., Cho, S., Lee, S.: Unsupervised video anomaly detection via normalizing flows with implicit latent features. Pattern Recognition **129**, 108703 (2022)
10. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)
11. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. In: ICLR (2017)
12. Doshi, K., Yilmaz, Y.: Rethinking video anomaly detection-a continual learning approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3961–3970 (2022)
13. Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197 (2022)
14. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence (2020)
15. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proc. of ICCV. pp. 1705–1714 (2019)
16. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 733–742 (2016)
17. He, C., Shao, J., Sun, J.: An anomaly-introduced learning method for abnormal event detection. Multimedia Tools and Applications **77**(22), 29573–29588 (2018)
18. Jøsang, A.: Subjective logic. Springer (2016)

19. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Advances in Neural Information Processing Systems (NIPS) (2018)
20. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. Advances in neural information processing systems **29**, 4743–4751 (2016)
21. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
22. Kopetzki, A.K., Charpentier, B., Zügner, D., Giri, S., Günnemann, S.: Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In: International Conference on Machine Learning. pp. 5707–5718. PMLR (2021)
23. Kopuklu, O., Zheng, J., Xu, H., Rigoll, G.: Driver anomaly detection: A dataset and contrastive learning approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 91–100 (2021)
24. Li, W., Vasconcelos, N.: Multiple instance learning for soft bags via top instances. In: CVPR. pp. 4277–4285 (2015). https://doi.org/10.1109/CVPR.2015.7299056
25. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
26. Liu, W., W. Luo, D.L., Gao, S.: Future frame prediction for anomaly detection – a new baseline. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
27. Liu, W., Luo, W., Li, Z., Zhao, P., Gao, S., et al.: Margin learning embedded prediction for video anomaly detection with a few anomalies. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 3023–3030. AAAI Press (2019)
28. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection– a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6536–6545 (2018)
29. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proc. of ICCV. pp. 2720–2727 (2013)
30. Malach, E., Shalev-Shwartz, S.: Decoupling" when to update" from" how to update". arXiv preprint arXiv:1706.02613 (2017)
31. Perera, P., Patel, V.M.: Deep transfer learning for multiple class novelty detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11544–11552 (2019)
32. Ramachandra, B., Jones, M., Vatsavai, R.R.: A survey of single-scene video anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
33. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. Advances in Neural Information Processing Systems **32** (2019)
34. Ryzhikov, A., Borisyak, M., Ustyuzhanin, A., Derkach, D.: Normalizing flows for deep anomaly detection. arXiv preprint arXiv:1912.09323 (2019)
35. Ryzhikov, A., Borisyak, M., Ustyuzhanin, A., Derkach, D.: Nfad: fixing anomaly detection using normalizing flows. PeerJ Computer Science **7**, e757 (11 2021). https://doi.org/10.7717/peerj-cs.757
36. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE transactions on pattern analysis and machine intelligence **35**(7), 1757–1772 (2012)
37. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. arXiv preprint arXiv:1806.01768 (2018)

38. Sentz, K., Ferson, S., et al.: Combination of evidence in Dempster-Shafer theory, vol. 4015. Sandia National Laboratories Albuquerque (2002)

39. Shi, W., Zhao, X., Chen, F., Yu, Q.: Multifaceted uncertainty estimation for label-efficient deep learning. Advances in Neural Information Processing Systems **33** (2020)

40. Smeureanu, S., Ionescu, R.T., Popescu, M., Alexe, B.: Deep appearance features for abnormal behavior detection in video. In: International Conference on Image Analysis and Processing. pp. 779–789. Springer (2017)

41. Sohrab, F., Raitoharju, J., Gabbouj, M., Iosifidis, A.: Subspace support vector data description. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 722–727. IEEE (2018)

42. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6479–6488 (2018)

43. Tabak, E.G., Turner, C.V.: A family of nonparametric density estimation algorithms. Communications on Pure and Applied Mathematics **66**(2), 145–164 (2013)

44. Tabak, E.G., Vanden-Eijnden, E.: Density estimation by dual ascent of the log-likelihood. Communications in Mathematical Sciences **8**(1), 217–233 (2010)

45. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)

46. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision (2021)

47. Tran, H.T., Hogg, D.: Anomaly detection using a convolutional winner-take-all autoencoder. In: Proceedings of the British Machine Vision Conference 2017. British Machine Vision Association (2017)

48. Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: ICME. pp. 1–6. IEEE (2020)

49. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: Proc. of ICCV. pp. 8201–8211 (2019)

50. Wang, R., Liu, T., Tao, D.: Multiclass learning with partially corrupted labels. IEEE transactions on neural networks and learning systems **29**(6), 2568–2580 (2017)

51. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European Conference on Computer Vision. pp. 322–339. Springer (2020)

52. Xia, X., Liu, T., Han, B., Gong, M., Yu, J., Niu, G., Sugiyama, M.: Sample selection with uncertainty of losses for learning with noisy labels. arXiv preprint arXiv:2106.00445 (2021)

53. Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., Guo, B.: Unsupervised extraction of video highlights via robust recurrent auto-encoders. In: Proc. of ICCV. pp. 4633–4641 (2015)

54. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334 (2021)

55. Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., Kloft, M.: Cloze test helps: Effective video anomaly detection via learning to complete video events. arXiv preprint arXiv:2008.11988 (2020)

56. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International Conference on Machine Learning. pp. 7164–7173. PMLR (2019)
57. Zaheer, M.Z., Mahmood, A., Shin, H., Lee, S.I.: A self-reasoning framework for anomaly detection using video-level labels. IEEE Signal Processing Letters **27**, 1705–1709 (2020)
58. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
59. Zhang, H., Li, A., Guo, J., Guo, Y.: Hybrid models for open set recognition. In: European Conference on Computer Vision. pp. 102–117. Springer (2020)
60. Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T.: Distilling effective supervision from severe label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9294–9303 (2020)
61. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: CVPR 2011. pp. 3313–3320. IEEE (2011)
62. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: 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. pp. 1237–1246 (2019)
63. Zhou, D.W., Ye, H.J., Zhan, D.C.: Learning placeholders for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2021)