

Normality Guided Multiple Instance Learning for Weakly Supervised Video Anomaly Detection

Seongheon Park¹ Hanjae Kim¹ Minsu Kim¹ Dahye Kim¹ Kwanghoon Sohn^{1,2*}

¹Yonsei University ²Korea Institute of Science and Technology (KIST)

{sam121796, incohjk, minsukim320, dadaday, khsohn}@yonsei.ac.kr

Abstract

Weakly supervised Video Anomaly Detection (wVAD) aims to distinguish anomalies from normal events based on video-level supervision. Most existing works utilize Multiple Instance Learning (MIL) with ranking loss to tackle this task. These methods, however, rely on noisy predictions from a MIL-based classifier for target instance selection in ranking loss, degrading model performance. To overcome this problem, we propose Normality Guided Multiple Instance Learning (NG-MIL) framework, which encodes diverse normal patterns from noise-free normal videos into prototypes for constructing a similarity-based classifier. By ensembling predictions of two classifiers, our method could refine the anomaly scores, reducing training instability from weak labels. Moreover, we introduce normality clustering and normality guided triplet loss constraining inner bag instances to boost the effect of NG-MIL and increase the discriminability of classifiers. Extensive experiments on three public datasets (ShanghaiTech, UCF-Crime, XD-Violence) demonstrate that our method is comparable to or better than existing weakly supervised methods, achieving state-of-the-art results.

1. Introduction

On increasing demands of analyzing surveillance videos, Video Anomaly Detection (VAD) has become an essential algorithm for human convenience and safety, such as security [12], medical imaging [39], factory automation [8] and autonomous driving [2]. Anomalies are frequently defined as behavior or appearance patterns that depart from normal patterns [5, 7]. VAD aims to predict such anomaly score of each segment in a video sequence. A typical approach is to regard VAD as a special case of video action classification [18, 41], having two classes of normal and abnormal. However, training the classifier requires large-scale datasets with fine-grained frame-level annotations, which are expen-

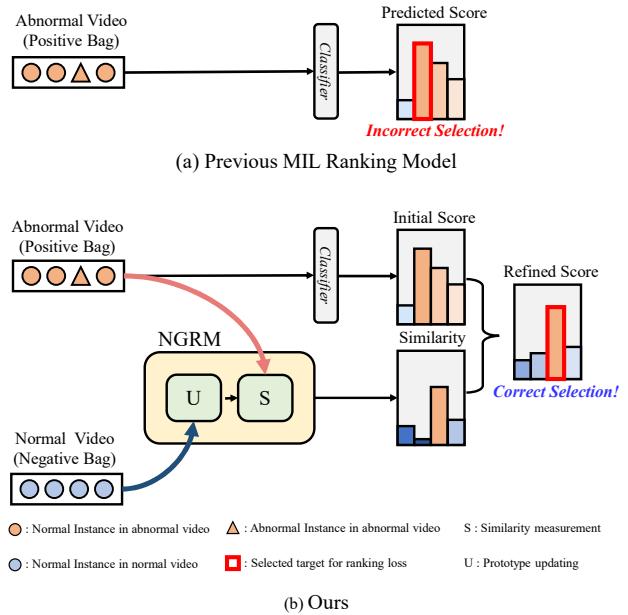


Figure 1. Illustration of previous MIL ranking model (a) and our proposed NG-MIL model (b). We refine unreliable initial prediction by Normality Guided Refinement Module (NGRM), which encodes global characteristics of normal patterns.

sive and time-consuming. To relieve this problem, some researchers have addressed weakly supervised VAD (wVAD) [33, 52, 46, 9, 38, 17, 48, 29], which needs only video-level annotations indicating whether anomalous contents exists in a video or not. Thanks to its competitive performance with lower laborious costs for annotations, the wVAD approach has attracted considerable research interest.

To detect abnormal segments of video without fine-grained labels, a common approach is to formulate wVAD as Multiple Instance Learning (MIL) problem, in which a video is represented as a bag of instances containing several consecutive frames. The bag (video) is labeled as positive if any of its instances is abnormal, and negative if it has only normal instances. After instance scores are estimated by a binary classifier, top-scoring instances are sam-

*Corresponding author

pled from positive and negative bags respectively. Then, the sampled positive and negative instances are constrained to have a large margin using MIL ranking loss [33]. Recent wVAD methods [52, 9, 38, 17, 29] are benefited from the ranking loss, as it improves discriminability of anomalies against normality. However, the selected top-k targets in the positive bag might contain some normal segments, since the MIL-based classifier is prone to produce noisy anomaly scores. The error might aggravate target instance selection in ranking loss as training continues, degrading the overall performance [27, 1]. Some works have attempted to refine the predictions with self-training [9, 17] or graph neural network [51], but the re-training is conducted based solely on the initial prediction of the classifier, which makes the model easy to be stuck with local minima [42]. Moreover, abundant normal instances inside the positive bag are ignored in MIL ranking loss, which hinders detecting hard abnormal instances surrounded by normal ones.

Meanwhile, One-Class Classification (OCC) approaches [11, 24, 50, 19, 10, 26, 21] focus on encoding frequently occurred patterns of normal data as a form of centroids [28] or latent vector [50] to train an one-class classifier. The encoded compact representation enables the model to capture the global relation between training samples and the entire normal feature distribution. The anomalies are detected based on the deviation from the learned normalities. However, due to a lack of prior knowledge of abnormalities, such OCC methods show relatively low performance compared to wVAD methods [46].

In this paper, we propose Normality Guided Multiple Instance Learning (NG-MIL) framework for overcoming the aforementioned limitations of MIL ranking loss as shown in Fig. 1. The key idea is to leverage numerous normal instances in the negative bags, which are noise-free [46], for eliminating false positives in the anomaly prediction scores. Inspired by normality representation [10, 26] in the OCC methods, we encode normal patterns across all normal video sets into compact prototypes which are the centroids of normal instances. The generated prototypes are utilized to formulate an additional anomaly classifier, whose score is defined as inverted cosine similarity between the prototypes and unlabeled instances in the positive bag. This similarity-based classifier allows to refine the anomaly score through model ensemble [36, 35, 16, 49] with the MIL-based classifier. Also, we propose normality clustering and normality guided triplet loss to enhance the discrimination of anomalies with prototypes inside the positive bag.

The main contributions are summarized as follows:

- We propose Normality Guided Multiple Instance Learning framework to refine the anomaly prediction of the MIL-based classifier with the similarity-based classifier. It is composed of normality prototypes, taking advantage of noise-free instances in the negative

bag.

- We propose normality clustering and normality guided triplet loss to increase the discriminative ability of classifiers.
- We conduct extensive experiments to validate the effectiveness of our method and show that it outperforms the state-of-the-art methods by a large margin on three VAD benchmark datasets, namely ShanghaiTech [19], UCF-Crime [33], and XD-Violence [44].

2. Related Works

Anomaly Detection as One-Class Classification. Conventional anomaly detection frameworks formulate the task as modeling normality given numerous normal samples and declaring the anomalies based on the deviation from the normality. Early works seek to learn a discriminative decision boundary using hand-crafted features, such as OC-SVM [31], kernel OC-SVM [30], and SVDD [37]. With the advent of deep convolutional networks, many approaches adopt image reconstructive model [11, 50, 19, 23] to learn normal data representations in an unsupervised way. However, these methods even reconstruct the anomalous test samples with a small error rate [10], resulting in missed detection. Some recent papers [10, 26] solved the problem by introducing normality prototypes. Each prototype is updated by aggregating features from training samples, which is an approximated centroid of normal data. By replacing the deep features with their nearest prototype, these methods could lessen the generalization capability of the reconstruction model.

Our method also leverages the prototypical representation of normal data. Contrary to the aforementioned OCC methods, we leverage the prototypes to refine the initial noisy prediction of the MIL-based classifier.

Weakly Supervised Video Anomaly Detection. wVAD has recently received much attention because labeling video-level annotations is much faster and easier than frame-level annotations. Sultani *et al.* [33] formulated wVAD as MIL problem and proposed MIL ranking loss, which allows the training of classifier with video-level annotation [25]. Recent approaches incorporated MIL ranking loss for optimization and improvement of anomaly detection performance. For example, Wan *et al.* [48] extended MIL ranking loss to the inner positive bag to encourage the discriminability inside the bag. Zhu *et al.* [52] proposed temporal augmented MIL ranking loss considering the temporal context through an attention mechanism. Wu *et al.* [43] introduced a causal convolution for feature extraction, to capture long-range dependencies in accurate anomaly detection. Despite their plausible results, their performance is

limited by unreliable classification scores from the weak supervisory signal. To alleviate this problem, Zhong *et al.* [51] proposed to refine the noisy prediction from video-level labels using graph convolutional neural networks. Inspired by self-training [45], MIST [9] and MSL [17] iteratively generate pseudo labels based on the predicted anomaly scores and re-train the classifier to obtain further refined pseudo labels. However, these methods still rely on the prediction of the unreliable MIL-based classifier for pseudo label generation.

Concurrent to our method, Liu *et al.* [20] also learn normality in noise-free negative bags to enhance the MIL-based wVAD performance. Unlike this approach using auto-encoder, we encode normality as prototypes, *i.e.*, centroids of normal features. It allows the model to obtain compact decision boundaries for normal instances [10, 40, 26]. Also, we use the learned normality to refine the score from the MIL-based classifier in an end-to-end fashion.

3. Method

3.1. Background and Motivation

Multiple instance ranking framework [33, 52, 9, 38, 17, 29] is widely used in weakly supervised video anomaly detection, thanks to its capability to discriminate anomalous segments using only video-level label. Given a video $B = \{v_t\}_{t=1}^T$ with T non-overlapping segments, each instance $f_t \in \mathbb{R}^D$ is computed through a feature extractor $E(\cdot)$ such that $f_t = E(v_t) \in \mathbb{R}^D$. The method then defines the abnormal video as a positive bag $B^a = \{f_t^a\}_{t=1}^T$ and the normal video as a negative bag $B^n = \{f_t^n\}_{t=1}^T$. They typically aim to maximize the anomaly score between top-k highest instances in the positive and negative bag through ranking loss:

$$L_{rank} = [1 - \frac{1}{k} \sum_{i=1}^k c(f_i^a) + \frac{1}{k} \sum_{j=1}^k c(f_j^n)]_+, \quad (1)$$

where $[\cdot]_+$ is hinge function, and $c(f_i^a)$, $c(f_j^n)$ indicate i^{th} , j^{th} index of predicted anomaly scores sorted in descending order. Minimizing the ranking objective in Eq. (1) improves instance discrimination power for abnormal instances against normal instances. However, they still select the top-k instances using an anomaly classifier solely trained on the video-level label. It often results in high-confidence anomaly scores for normal instances in the abnormal video, thus accumulating errors in the subsequent learning process, as demonstrated in Sec. 4.7.

In this paper, we propose Normality Guided Multiple Instance Learning (NG-MIL), in which normality prototypes encoded with diverse normal patterns from normal videos give guidance in anomaly prediction refinement. By utilizing the similarity between normality prototypes and instances as additional classification scores, we refine the

error-prone initial noisy prediction. In the following, we elaborate Normality Guided Refinement Module (Sec. 3.2) and learning objectives (Sec. 3.3). The overall framework of NG-MIL is illustrated in Fig. 2.

3.2. Normality Guided Refinement Module

Normality Guided Refinement Module (NGRM) is designed to refine the unreliable anomaly scores using a set of normality prototypes $P = \{p_m\}_{m=1}^M$, where each normality is represented by a prototype $p_m \in \mathbb{R}^{D/4}$. It consists of two major processes, normality update and anomaly prediction refinement. The details are introduced as follows.

Normality Update. Our normality update process aims at capturing global characteristics of normality from all normal videos. It is inspired by the previous memory-based methods [32, 13, 10, 26]. To update normality prototype p_m , we first project f_t^n into \hat{f}_t^n , to align the feature dimension to p_m . We then compute the cosine similarity between each projected instance feature \hat{f}_t^n and all normality prototypes P :

$$s_{t,m}^n = \frac{\hat{f}_t^n p_m^\top}{\|\hat{f}_t^n\| \|p_m\|}, \quad m \in \{1, \dots, M\}. \quad (2)$$

It results in a 2-dimensional similarity map of size $T \times M$. Each projected instance is assigned to update the nearest normality prototype. We denote the set of projected instance indices U_m for updating the m^{th} normality prototype. Note that the projected instance features can be assigned to a single normality prototype. We then update the normality prototype using the projected instance feature as follows:

$$p_m \leftarrow (1 - \lambda)p_m + \lambda \frac{1}{|U_m|} \sum_{t \in U_m} \hat{f}_t^n, \quad (3)$$

where λ represents a momentum for exponentially weighted moving average. Note that we update the normality prototype only if the projected instance feature is assigned.

Anomaly Prediction Refinement. Unlike existing methods [33, 9, 38, 17, 29] that exploit top-scoring instances for discriminative representation of anomalous segments, we extend it by incorporating normality prototypes as guidance to refine the unreliable anomaly scores. We first compute cosine similarity between each instance and normality prototypes. We then apply softmax operation along the M normality prototypes, and use it as attention weight for similarity score as follows:

$$g(f_i, P) = \sum_{m=1}^M \frac{\exp(s_{i,m}/\tau)}{\sum_{m=1}^M \exp(s_{i,m}/\tau)} s_{i,m}, \quad (4)$$

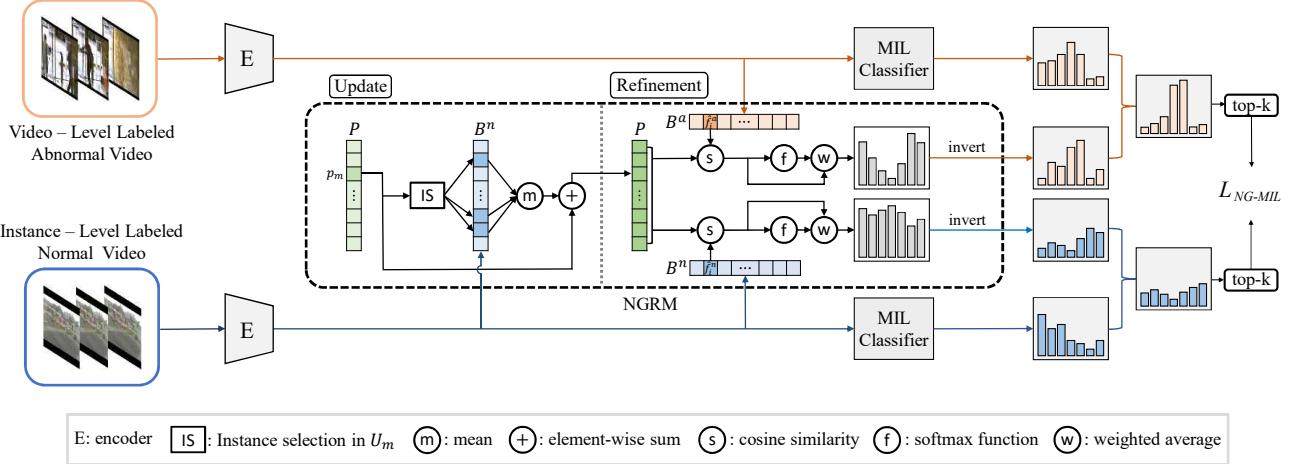


Figure 2. Overall architecture of proposed NG-MIL framework, which is composed of an encoder, NGRM, and MIL-based classifier. First We feed a pair of abnormal and normal videos into the network, generating feature embedding by a pre-trained backbone and encoder. Then, the MIL-based classifier predicts anomaly scores, which are refined by NGRM. Finally, NG-MIL ranking loss is applied with the refined scores. Note that the normal and abnormal branches share the same encoder and classifier. During the testing stage, a single unlabeled video is the input of the network.

where τ is a temperature hyper-parameter. We further apply a ReLU function to g to make sure it is non-negative. Note that our model applies the same rule for both abnormal and normal videos, so we omit the superscripts a and n for brevity. We can simply represent anomaly score $a(f_i, P)$ as an inverted similarity score between the instance and normality prototype such that:

$$a(f_i, P) = 1 - g(f_i, P). \quad (5)$$

Finally, we refine the score by ensembling two prediction scores from each classifier, following [36, 35, 16, 49] :

$$r(f_i) = \frac{1}{2}(c(f_i) + a(f_i, P)). \quad (6)$$

3.3. Learning Objectives

We utilize three losses for optimizing our network: normality guided MIL ranking loss L_{NG-MIL} , and two auxiliary losses for NGRM, L_{clst} and L_{tri} which regularize the instance features inside the negative and positive bag respectively.

Normality Guided MIL Ranking Loss. Using NGRM introduced in Sec. 3.2, we propose NG-MIL ranking loss as follows:

$$L_{NG-MIL} = [1 - \frac{1}{k} \sum_{i=1}^k r(f_i^a) + \frac{1}{k} \sum_{j=1}^k r(f_j^n)]_+, \quad (7)$$

where i and j are the indices of score r , sorted in descending order. Compared to the base MIL ranking loss in Eq.

(1), NG-MIL ranking loss allows to select more confident target instances for ranking loss, which alleviates training instability from weak labels.

Normality Clustering Loss. Motivated by cluster loss [6], we further propose normality clustering loss to encourage clustering between each instance in the negative bag and its nearest neighbor prototype:

$$L_{clst} = \frac{1}{T} \sum_{i=1}^T \min_{p_m \in P} \|p_m - \hat{f}_i^n\|_2^2. \quad (8)$$

This clustering loss reduces the intra-class variance of normalities, which facilitates the discriminability of similarity-based anomaly classification in NGRM.

Normality Guided Triplet Loss. For accurate classification from both the MIL-based classifier and similarity-based classifier, we expect that abnormal instance features lie far apart from normal instance features in both positive and negative bags. However, NG-MIL ranking loss in Eq. (7) only considers top- k instances as optimization units, ignoring normal instances in the positive bag. As the majority of segments in abnormal video contain normal events, this hinders the classifier from detecting abnormal instances surrounded by normal ones.

From this motivation, we introduce normality guided triplet loss that penalizes the gap between normal and abnormal instance features by a large margin. We first sample pseudo abnormal set $\Omega_a = \{\hat{f}_1^a, \dots, \hat{f}_k^a\}$ and pseudo normal set $\Omega_n = \{\hat{f}_{T-k+1}^a, \dots, \hat{f}_T^a\}$, which contain top- k and

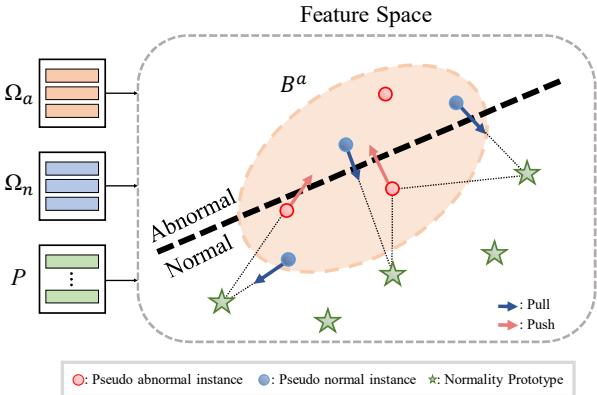


Figure 3. Illustration of normality guided triplet loss. For each sample (*i.e.*, positive sample) of the pseudo normal set, the nearest normality prototype becomes an anchor. Then the sample from the pseudo abnormal set closest to each anchor becomes the negative sample, forming a triplet set. The normality guided triplet loss minimizes the distance between the anchor and the positive sample and maximizes the distance between the anchor and the negative sample.

bottom-k instances sampled from the positive bag using refined scores from Eq. (6) sorted in descending order. Then, the loss is formulated as:

$$L_{tri} = \frac{1}{k} \sum_{i=1}^k [\|\hat{f}_{T-i+1}^a - p_j\|_2^2 - \min_{\hat{f}^a \in \Omega_a} \|\hat{f}^a - p_j\|_2^2 + \gamma]_+, \quad (9)$$

where γ is a pre-defined margin and j is an index of the nearest prototype from each sample in the pseudo normal set:

$$j = \arg \min_{m \in M} \|\hat{f}_{T-i+1}^a - p_m\|_2^2. \quad (10)$$

Our loss enhances the intra-class compactness of normalities and inter-class separability in the positive bag by penalizing triplets in (P, Ω_n, Ω_a) , which results in boosting the refinement quality of NGRM significantly.

Total Loss Our total loss function is defined with the summation of NG-MIL ranking loss L_{NG-MIL} , normality clustering loss L_{clst} , and normality guided triplet loss L_{tri} . In addition, following Sultani *et al.* [33], we incorporate the temporal smoothness term defined as $L_{ts} = \sum_{i=1}^{T-1} (r(f_i) - r(f_{i+1}))^2$ and sparsity constraints term defined as $L_s = \sum_{i=1}^T r(f_i)$, leveraging characteristics of temporal consistency of events and rarity of abnormal events in real-world scenarios. Finally, the total loss is defined as:

$$L_{total} = L_{NG-MIL} + \lambda_t L_{clst} + \lambda_c L_{tri} + \lambda_r (L_{ts} + L_s), \quad (11)$$

where λ_t , λ_c , and λ_r assign relative importance to different loss signals.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on three video anomaly detection benchmarks, ShanghaiTech [19], UCF-Crime [33], and XD-Violence [44].

ShanghaiTech is a medium-scale dataset that contains 437 campus surveillance videos with 130 abnormal events in 13 scenes. Since the original training dataset contains only normal videos, Zhong *et al.* [51] reorganized testing videos into training data and vice versa.

UCF-Crime is a large-scale complex dataset that contains 1900 indoor and outdoors untrimmed real-world surveillance videos. The training set consists of 800 normal and 810 abnormal videos, and the test set includes 150 normal and 140 abnormal videos with 13 types of anomalous events.

XD-Violence is a large-scale diverse dataset collected from movies, in-the-wild scenes, and surveillance cameras. The dataset contains 4754 videos, which consists of 2349 normal and 2405 abnormal videos. The training set includes 3954 videos and the test set contains 800 videos.

Evaluation Metrics. Following previous works [33, 52, 46, 9, 38, 17, 48, 29], we plot the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) as evaluating the performance of our method. The ROC curve shows the performance at all classification thresholds, which is mainly used for the binary classification task. In addition, for XD-Violence dataset, we also use average precision (AP) as an evaluation metric following [44, 38, 17]. Note that a higher AUC and AP implies better anomaly detection performance.

4.2. Implementation Details

We extract 2,048D features from the ‘mixed.5c’ layer of the pre-trained I3D [4] or 4,096D features from the ‘fc6’ layer of the pre-trained C3D [14]. The encoder consists of temporal convolution layers and ReLU activation function. Following previous works [38, 33], we divided each video into 32 non-overlapping segments. The MIL-based classifier is a 3-layer MLP, where the number of nodes is 512, 128, and 1 respectively. Each layer is followed by ReLU activation function and a dropout function with a rate of 0.7. NGRM contains 32 prototypes for ShanghaiTech and 64 for UCF-Crime and XD-Violence, initialized by K-means clustering algorithm [22] across all normal videos. The momentum parameter for normality update is set to 0.1, and the temperature hyper-parameter in Eq.(4) to 0.5. Note that

Supervision	Method	Feature	AUC (%)
One-Class Classification	Conv-AE [11]	-	50.60
	Stacked-RNN [24]	-	68.00
	MNAD [26]	-	70.50
	AMMC [3]	-	73.70
	GCL [47]	-	78.93
Weakly Supervised	IBL [48]	C3D RGB	82.50
	GCN-Anomaly [51]	TSN RGB	84.44
	Sultani et al.* [33]	I3D RGB	85.33
	AR-Net [40]	I3D RGB/Flow	91.24
	CLAWS [46]	C3D RGB	89.67
	MIST [9]	C3D RGB	93.13
	MIST [9]	I3D RGB	94.83
	RTFM [38]	C3D RGB	91.51
	RTFM [38]	I3D RGB	97.21
	MSL [17]	C3D RGB	94.81
	MSL [17]	I3D RGB	97.32
	BN-SVP [29]	C3D RGB	96.00
	Ours	C3D RGB	96.02
	Ours	I3D RGB	97.43

Table 1. Comparison of frame-level AUC performance with other SOTA methods under one-class classification and weakly supervision mode on ShanghaiTech. The method with * is reported by [38]. The highest result is bolded.

we do not update normality prototypes at the testing stage. We set the margin value of normality guided triplet loss in Eq. (9) to 8. Our method is trained in an end-to-end manner using Adam optimizer [15] with a learning rate of 0.001, a weight decay of 0.0005, and a batch size of 64. Each mini-batch is composed of 32 randomly selected normal and abnormal videos. Through the cross-validation using grid-search in log-scale, we set the hyper-parameters λ_t , λ_c , and λ_r as 0.1, 0.1, and 0.5, respectively.

4.3. Performance on ShanghaiTech

The AUC results on ShanghaiTech are shown in Table 1. Our method achieves AUC score of 97.43% with I3D RGB features and 96.02% with C3D RGB features, outperforming existing state-of-the-art one-class classification (OCC) [11, 24, 26, 3, 47] and weakly supervised methods [48, 51, 33, 40, 46, 9, 38, 17, 29]. These results demonstrate the effectiveness of our proposed NG-MIL.

4.4. Performance on UCF-Crime

The performances on UCF-Crime are demonstrated in Table 2. Consistent with the results on ShanghaiTech, our method outperforms all OCC [11, 34] and weakly supervised approaches [33, 48, 52, 51, 46, 9, 38, 17, 29] by large margins. For example, with I3D RGB features, our method outperforms Sultani *et al.* [33] by 7.71%, GCN-Anomaly [51] by 3.51%, MIST [9] by 3.33%, RTFM [38] by 1.60%, MSL [17] by 0.33%, and BN-SVP [29] by 2.24%. Considering C3D RGB features, our approach has also achieved competitive results. Compared to computationally costly alternative training [51] and self-training [17, 9] methods, our

Supervision	Method	Feature	AUC (%)
One-Class Classification	Conv-AE [11]	-	50.60
	ST-Graph [34]	-	72.70
Weakly Supervised	Sultani et al. [33]	C3D RGB	75.41
	Sultani et al.* [33]	I3D RGB	77.92
	IBL [48]	C3D RGB	78.66
	Motion-Aware [52]	PWC Flow	79.00
	GCN-Anomaly [51]	TSN RGB	82.12
	CLAWS [46]	C3D RGB	83.03
	MIST [9]	C3D RGB	81.40
	MIST [9]	I3D RGB	82.30
	RTFM [38]	C3D RGB	83.28
	RTFM [38]	I3D RGB	84.03
	MSL [17]	C3D RGB	82.85
	MSL [17]	I3D RGB	85.30
	BN-SVP [29]	I3D RGB	83.39
	Ours	C3D RGB	83.43
	Ours	I3D RGB	85.63

Table 2. Comparison of frame-level AUC performance with other SOTA methods under one-class classification and weakly supervision mode on UCF-Crime.

Supervision	Method	Feature	AP (%)
One-Class Classification	OC-SVM [31]	-	27.25
	Conv-AE [11]	-	30.77
Weakly Supervised	Sultani et al. [33]	C3D RGB	73.20
	Sultani et al.* [33]	I3D RGB	75.68
	Wu et al. [44]	I3D RGB	75.41
	Wu et al.† [44]	I3D RGB/Audio	78.64
	RTFM [38]	C3D RGB	75.89
	RTFM [38]	I3D RGB	77.81
Weakly Supervised	MSL [17]	C3D RGB	75.53
	MSL [17]	I3D RGB	78.28
Ours	C3D RGB	75.91	
	I3D RGB	78.51	

Table 3. Comparison of AP performance with other SOTA methods under one-class classification and weakly supervision mode on XD-Violence.

method outperforms these methods by training the model in an end-to-end fashion, proving the effectiveness of our model.

4.5. Performance on XD-Violence

The performances on XD-Violence are demonstrated in Table 3. Our model exceeds OCC methods [31, 11] by a minimum of 47.74% in AP. Moreover, comparing with other state-of-the-art weakly supervised methods, our methods performs better than Sultani *et al.* [33] by 2.83%, RTFM [38] by 0.70%, MSL [17] by 0.23% using I3D RGB features. Specifically, compared with the trained method with both RGB and Audio features by Wu *et al.* [44], it can be observed that our method even can achieve comparable performance with only RGB features.

4.6. Ablation Study

Top-k precision. To validate the effectiveness of our NG-MIL framework, we utilize top-k precision metric:

$$\text{Top-k Precision (\%)} = \frac{TPI}{TPI + FPI} \times 100, \quad (12)$$

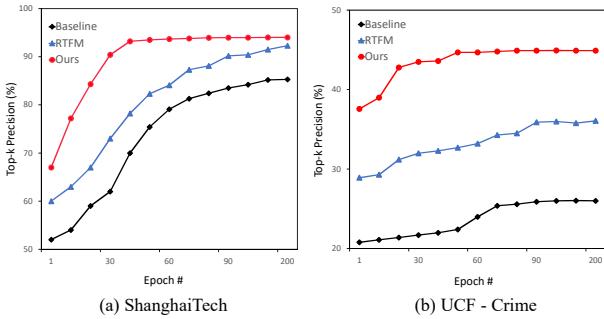


Figure 4. Comparison results on (a) ShanghaiTech and (b) UCF-Crime measured by Top-k Precision at each epoch ($k=3$).

Method	ShanghaiTech	UCF-Crime
Baseline	93.13	83.01
NGRM^{sim}	94.01	83.64
NGRM	97.43	85.63

Table 4. Comparison results of top-k selection strategy on ShanghaiTech and UCF-Crime, measured by AUC.

where TPI and FPI indicate the number of true and false positives within the top-k instances. Note that it is measured on the testing abnormal videos. A higher value of the measure indicates that the top-k instances are more accurately selected to calculate the ranking loss, which can boost the overall learning of the model. We compare the top-k precision performance with other top-k ranking models, including baseline and RTFM, using I3D RGB features on the ShanghaiTech and UCF-Crime benchmarks.

In Fig. 4, we observe that our method shows faster convergence speed and higher performance in terms of top-k precision. For ShanghaiTech, our method outperforms the baseline and RTFM by 16.03%, 8.12% at the first epoch, and 11.54%, 2.04% after convergence. A similar result is observed on UCF-Crime, showing improvement of 16.79%, 8.66% at the first epoch, and 18.81%, 8.79% for after convergence. It indicates that our NGRM contributes to sampling the top-k instances more accurately. Taking advantage of the accurate samples, our model also results in better AUC performance compared to the other methods, as presented in Tables 1 and 2.

Top-k selection strategy. In Table 4, we investigate the contribution of our refining strategy on ShanghaiTech and UCF-Crime using I3D RGB features. We consider three types of top-k selection strategy for MIL ranking loss: (1) using only MIL-based classifier (Baseline) (2) NGRM using only similarity-based classifier (NGRM^{sim}), and (3) NGRM using both MIL-based classifier and similarity-based classifier (NGRM). Selecting top-k instances by refined score largely outperforms score-only, and similarity-

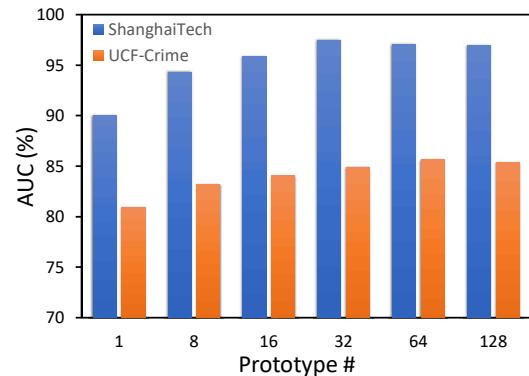


Figure 5. AUC with respect to the different number of prototypes on ShanghaiTech and UCF-Crime.

$L_{\text{NG-MIL}}$	L_{tri}	L_{clst}	ShanghaiTech	UCF-Crime
✓			93.13	83.03
✓	✓		95.51	84.06
✓		✓	96.59	84.87
✓	✓	✓	97.43	85.63

Table 5. AUC results of loss function analysis on ShanghaiTech and UCF-Crime.

only methods by 4.30%, 3.42% for ShanghaiTech and 2.62%, 1.99% on UCF-Crime. This indicates that NG-MIL framework contributes to the overall performance, which learns complementary information between the similarity-based branch and MIL-based branch, thus avoiding being trapped in local minima.

Number of prototypes. We use ShanghaiTech and UCF-Crime to study the effect of the number of prototypes M . We conduct the experiments by using a different number of prototypes with I3D RGB features and show the results in Fig. 5. It can be observed that the predicted anomaly scores have the highest AUC of 97.43% with $M = 32$ on ShanghaiTech and 85.63% with $M = 64$ on UCF-Crime. This indicates that UCF-Crime, captured from real-world surveillance, has more diversity of normal patterns than ShanghaiTech captured only on campus. Also, an insufficient number of prototypes ($M < 10$) degrades the performance significantly, which validates the importance of modeling diverse normal patterns.

Effects of loss components. We conduct component analysis of each proposed loss function on ShanghaiTech and UCF-Crime using I3D RGB features in Table 5. The baseline with NG-MIL ranking loss achieves 93.13%, and 83.03% AUC for each dataset. The proposed normality guided triplet loss L_{tri} improves 2.38%, 1.03% from NG-MIL ranking loss on ShanghaiTech and UCF-Crime, while

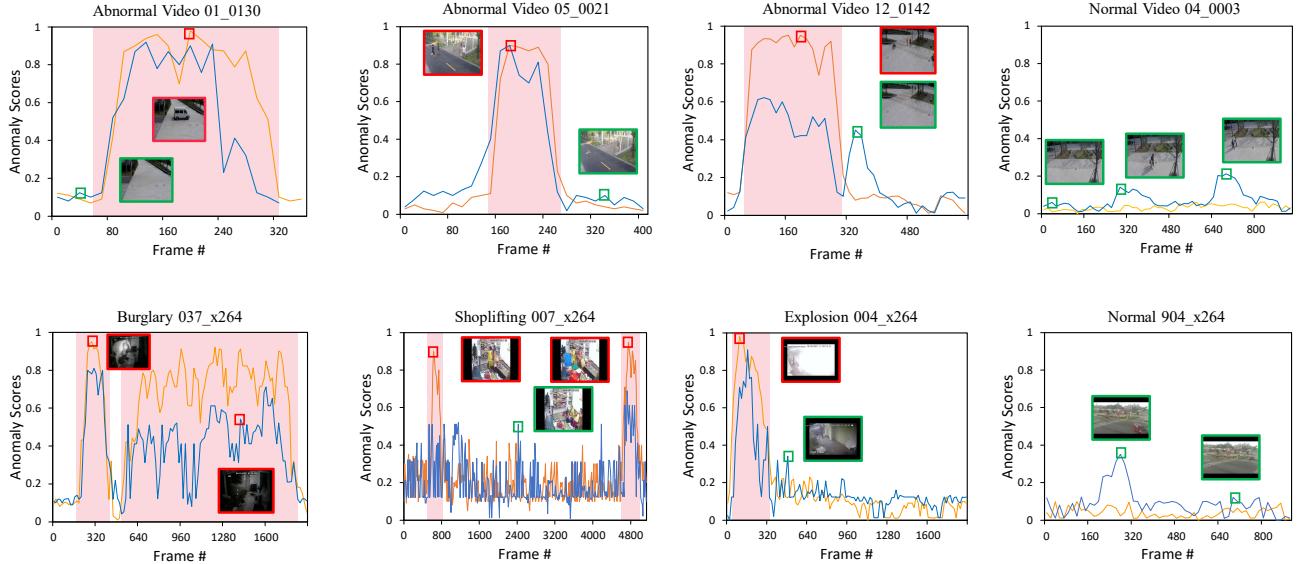


Figure 6. Visualization of anomaly scores on ShanghaiTech and UCF-Crime test videos. Orange curves show anomaly scores of our method and blue curves show anomaly scores of the baseline model without NGRM. Pink areas indicate the ground-truth abnormal frames. Each red and green box shows the abnormal and normal event. Best viewed in color.

the normality clustering loss L_{clst} achieves improvement of 3.46% and 1.84%. The model with all of the loss components performs best with 97.43% and 85.63%. It demonstrates that both losses are effective for boosting performance along with NG-MIL ranking loss, by increasing the discriminability of anomalies in a video.

4.7. Qualitative Analysis

In Fig. 6, we visualize some representative results on several challenging cases in ShanghaiTech and UCF-Crime. We compare our model with the baseline following Sec. 4.6. The baseline model fails to distinguish abnormal from normal events where confusing abnormal events (12_0142, Burglary 037, Shoplifting 007) which are similar to any other context information, and also miss detects false positive normal events (04_0003, Explosion 004, Normal 904). In contrast to the baseline model, our method successfully predicts long-term abnormal event (01_0130, 05_0021, 12_0142, Burglary 037), single short-term abnormal event (Explosion 004), multiple abnormal events (Burglary 037, Shoplifting 007), and only normal events (04_0003, Normal 904), with large score margins between the normal and abnormal events. Furthermore, our model detects some challenging abnormal events that are similar to normal events (Burglary 037, Shoplifting 007), showing the effectiveness of our normality guided triplet loss.

5. Conclusion

In this work, we identified the inherent limitations of existing weakly supervised video anomaly detection methods

based on multiple instance learning ranking model. We observed that the majority of the methods solely relied on unreliable anomaly scores for high-confidence anomalous instance selection, which might lead to erroneous anomaly prediction. To address this problem, we proposed to refine anomaly scores from the MIL-based classifier by normality prototypes which describe global characteristics of normal information. Furthermore, we introduced normality clustering and normality guided triplet loss to boost the quality of the refinement process. Experimental results on three popular VAD datasets show the effectiveness of our method, demonstrating improved performance over the state-of-the-art methods.

Broader Impacts. Our method can be used in the real-time intelligent video surveillance system, which significantly increases monitoring efficiency. The video anomaly detection system is designed to enhance social safety, however, it can also have some potential negative societal impacts. The surveillance data and VAD datasets may cause privacy issues on irrelevant individuals. Therefore, the collection process of these data should inform the persons who are in the collection, and it must be well institutionalized for using VAD algorithms.

Acknowledgements. This work was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002) and the KIST Institutional Program (Project No.2E31051-21-203).

References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O’Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *IJCNN*, 2020.
- [2] Daniel Bogdolla, Maximilian Nitsche, and J Marius Zöllner. Anomaly detection in autonomous driving: A survey. *CVPR*, 2022.
- [3] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. *AAAI*, 2021.
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, 2017.
- [5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *CSUR*, 2009.
- [6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *NeurIPS*, 2019.
- [7] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. *CVPR*, 2015.
- [8] Laura Erhan, M Ndubuaku, Mario Di Mauro, Wei Song, Min Chen, Giancarlo Fortino, Ovidiu Bagdasar, and Antonio Littona. Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion*, 2021.
- [9] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. *CVPR*, 2021.
- [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. *ICCV*, 2019.
- [11] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. *CVPR*, 2016.
- [12] Helen M Hodgetts, François Vachon, Cindy Chamberland, and Sébastien Tremblay. See no evil: Cognitive challenges of security surveillance and monitoring. *JARMAC*, 2017.
- [13] Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. Relational prototypical network for weakly supervised temporal action localization. *AAAI*, 2020.
- [14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. *CVPR*, 2014.
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [16] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. *CVPR*, 2021.
- [17] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *AAAI*, 2022.
- [18] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. *ACM MM*, 2019.
- [19] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. *CVPR*, 2018.
- [20] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. Collaborative normality learning framework for weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2022.
- [21] Zhiyan Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. *ICCV*, 2021.
- [22] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 1982.
- [23] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. *ICME*, 2017.
- [24] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *ICCV*, 2017.
- [25] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *NeurIPS*, 1997.
- [26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. *CVPR*, 2020.
- [27] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.
- [28] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoail Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. *ICML*, 2018.
- [29] Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. *CVPR*, 2022.
- [30] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 2001.
- [31] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *NeurIPS*, 1999.
- [32] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *NeurIPS*, 2015.
- [33] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *CVPR*, 2018.
- [34] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. *ACM MM*, 2020.
- [35] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE TPAMI*, 2018.
- [36] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. *CVPR*, 2017.

- [37] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 2004.
- [38] 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. *ICCV*, 2021.
- [39] Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan W Verjans, and Gustavo Carneiro. Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. *arXiv preprint arXiv:2203.12121*, 2022.
- [40] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. *ICME*, 2020.
- [41] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 2021.
- [42] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. *CVPR*, 2019.
- [43] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE TIP*, 2021.
- [44] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhao Yang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. *ECCV*, 2020.
- [45] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. *CVPR*, 2020.
- [46] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. *ECCV*, 2020.
- [47] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. *CVPR*, 2022.
- [48] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. *ICIP*, 2019.
- [49] Yuhang Zhang, Xiaopeng Zhang, Jie Li, Robert Qiu, Haohang Xu, and Qi Tian. Semi-supervised contrastive learning with similarity co-calibration. *IEEE Transactions on Multimedia*, 2022.
- [50] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. *ACM MM*, 2017.
- [51] 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. *CVPR*, 2019.
- [52] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.