Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies

Wen Liu[1*], **Weixin Luo**[1⋆], **Zhengxin Li**[1], **Peilin Zhao**[2] and **Shenghua Gao**[1†]

[1]ShanghaiTech University

[2]Tencent AI Lab

{liuwen, luowx, lizhx}@shanghaitech.edu.cn, peilinzhao@hotmail.com, gaoshh@shanghaitech.edu.cn

Abstract

Classical semi-supervised video anomaly detection assumes that only normal data are available in the training set because of the rare and unbounded nature of anomalies. It is obviously, however, these infrequently observed abnormal events can actually help with the detection of identical or similar abnormal events, a line of thinking that motivates us to study open-set supervised anomaly detection with only a few types of abnormal observed events and many normal events available. Under the assumption that normal events can be well predicted, we propose a Margin Learning Embedded Prediction (MLEP) framework. There are three features in MLEP- based open-set supervised video anomaly detection: i) we customize a video prediction framework that favors the prediction of normal events and distorts the prediction of abnormal events; ii) The margin learning framework learns a more compact normal data distribution and enlarges the margin between normal and abnormal events. Since abnormal events are unbounded, our framework consequently helps with the detection of abnormal events, even for anomalies that have never been previously observed. Therefore, our framework is suitable for the open-set supervised anomaly detection setting; iii) our framework can readily handle both frame-level and video-level anomaly annotations. Considering that video-level anomaly detection is more easily annotated in practice and that anomaly detection with a few anomalies is a more practical setting, our work thus pushes the application of anomaly detection towards real scenarios. Extensive experiments validate the effectiveness of our framework for anomaly detection.

1 Introduction

Anomaly detection is an important task in computer vision and machine learning because of its potential applications in video surveillance, network traffic monitoring, *etc.* However, it is a challenging task because anomalies are rare. This means that abnormal events seldom happen. Further, anomalies are unbounded, which means there are many possibilities for abnormal events, even under an identical scene. Based on the availability of labeled data and abnormal events in training sets [Ienco *et al.*, 2017], anomaly detection can typically be categorized into unsupervised anomaly detection where training data labels are not given, semi-supervised anomaly detection where only normal data are provided in the training set, and supervised anomaly detection where both normal and abnormal data are provided and labeled in the training set. Further, based on where the anomalies in the testing set are included in there training set, supervised anomaly detection can be further categorized as either a closed-set setting where all types of anomalies have been included in the training set [Sultani *et al.*, 2018], or an open-set setting where some types of anomalies in the testing set are not included. In practice, once a few abnormal events occur and are recorded, these abnormal events can obviously help with the anomaly detection of identical or similar anomalies in the future. Meanwhile, because of the unbounded nature of anomalies, in this paper, we propose to study open-set supervised anomaly detection.

Now that we have both normal and abnormal data at hand, an intuitive idea is to formulate supervised video anomaly detection as an (imbalanced) binary video classification problem. Such a solution, however, only works for closed-set supervised video anomaly detection as it fails in the open-set setting because of the following two reasons: i) due to the unbounded nature of anomalies, the observed data only contains a few types of anomalies rather than all types of anomalies. Those unseen types of anomalies may be classified as normal events by a classifier, and the loss for such false negative is big; ii) because of the rare nature of anomalies, the distribution between normal and abnormal data is imbalanced, and the distributions of those unseen anomalies are unknown. As as result, it is not easy to properly train a classifier for binary video classification.

Under the assumption that normal events are predictable and abnormal events are unpredictable [Liu *et al.*, 2018; Chandola *et al.*, 2009], we propose a Margin Learning Embedded Video Prediction (MLEP) framework for open-set supervised anomaly detection. Specifically, there are three features in our MLEP-based anomaly detection framework: i)

[*]The authors contribute equally.

[†]Corresponding author.

we carefully design a ConvLSTM [Shi *et al.*, 2015] for video prediction based anomaly detection. Rather than stacking a whole video snippet as the input for future predictions [Mathieu *et al.*, 2016] [Liu *et al.*, 2018], we propose the sequential feeding of features of each frame into this ConvLSTM because this method is better able to encode both temporal and spatial information [Luo *et al.*, 2017a]. Such a network favors the prediction of normal frames and distorts the prediction of abnormal frames; ii) we embed margin learning into our network architecture for the open-set supervised setting. By enforcing the margin between the pairwise distance of a normal events pair and that of a normal and abnormal pair to be larger than a given margin, our framework tightens the distribution boundary of normal events and enlarges the gap between normal and abnormal events. Consequently, our solution helps with the detection of both observed and unobserved anomalies; iii) our framework can tackle anomaly detection with both frame-level and video-level anomaly annotations. Since video-level annotation is much easier, our solution is more practical in real scenarios.

We summarize our contributions as follows: i) we design an MLEP framework for open-set supervised video anomaly detection. By integrating a video prediction module with margin learning, our framework learns a more compact normal data distribution and enlarges the margin between normal and abnormal distributions. Consequently, our framework is robust to unbounded anomalies. As far as we know, this is the first work on open-set supervised video anomaly detection; ii) we delicately design a future prediction framework which favors the prediction of normal events and distorts the prediction of abnormal events, thereby facilitating the detection of abnormal events; iii) our framework can tackle both frame-level and video-level annotations, thus our solution is more practical in real scenarios; iv) extensive experiments validate the effectiveness of our proposed framework for anomaly detection.

2 Related Work

Deep Learning Based Semi-Supervised Anomaly Detection. Almost all of the previous work in computer vision focuses on semi-supervised anomaly detection in which the training data only contain normal data. Hasan *et al.* [Hasan *et al.*, 2016] propose a Convolutional Auto-Encoder (Conv-AE) by stacking multiple frames. Further, inspired by the ability of the spatial representation of Convolutional Neural Networks (CNN) and the temporal encoding of Recurrent Neural Networks (RNN), some works [Luo *et al.*, 2017a] combine these two networks to model normal spatial and temporal patterns. In [Luo *et al.*, 2017b], a stacked RNN based Auto-Encoder is proposed for anomaly detection. Liu *et al.* [Liu *et al.*, 2018] propose a future frame prediction framework for anomaly detection, which avoids the identity mapping and significantly improves their performance on anomaly detection. In addition, some more popular methods are based on generative models. For instance, Sabokroul *et al.* [Sabokrou1 *et al.*, 2018] use Generative Adversarial Networks (GANs) with a generator and a discriminator to learn the normal distribution.

Supervised Anomaly Detection. Even though supervised anomaly detection has been well studied in the network intrusion detection [Li and Guo, 2007][Ahmed *et al.*, 2016], it is seldom studied within computer vision. As far as we know, the only research along this direction is from [Sultani *et al.*, 2018], where multiple instance learning is used to tackle data with video-level ground-truths, but for which all types of anomalies in the testing set are included in the training set, thus resulting in a closed-set setting. Because of the rare and unbounded nature of anomalies, open-set anomaly detection is a more practical setting. Further, the amounts of normal and abnormal data are roughly balanced, whereas abnormal data are more difficult to collect and are thus limited in the training set. Compared with [Sultani *et al.*, 2018], our open-set video anomaly detection is more practical, thus it pushes the application of video anomaly detection towards real applications.

Imbalanced Classification. Most of the previous research on imbalanced classification was usually based on a data resampling strategy [Li *et al.*, 2011; Tian *et al.*, 2011; Dong *et al.*, 2017; Yan *et al.*, 2015], cost-sensitive learning[Krawczyk *et al.*, 2014; Ren *et al.*, 2018], as well as the combination of the two [Tang *et al.*, 2009; López *et al.*, 2012; Huang *et al.*, 2018] . For example, Krawczyk *et al.* [Krawczyk *et al.*, 2014] introduce an effective ensemble of cost-sensitive decision trees for imbalanced classification. Recently, many strategies based on CNNs to ease the issue of imbalance have been studied. Dong *et al.* [Dong *et al.*, 2017] develop an end-to-end deep learning framework that is able to avoid the dominant effect of majority classes by using batch incremental hard sample mining of minority classes. Ren *et al.* [Ren *et al.*, 2018] propose a novel meta-learning algorithm that learns to assign weights to the training examples based on their gradient directions.

3 Our Approach

Our MLEP leverages a prediction-based framework for video anomaly detection, as well as a large margin constraint for the open-set supervised anomaly detection setting. The whole framework is illustrated in Figure 1 which includes a future prediction and a margin learning module.

3.1 Future Prediction Module for Videos

Our assumption is that normal events are predictable while abnormal ones are unpredictable [Liu *et al.*, 2018; Chandola *et al.*, 2009]. Most of the existing networks used for video prediction can be roughly divided into three categories: i) UNet [Liu *et al.*, 2018]; ii) Auto-Encoder without skip connection [Mathieu *et al.*, 2016]; iii) ConvLSTM [Villegas *et al.*, 2017]. However, all of these networks are not suitable for our anomaly detection with only a few anomalies. Specifically, i) UNet with skip connection favors the prediction of abnormal events; ii) the conventional encoder [Mathieu *et al.*, 2016] with several convolution layers does not have enough capacity to encode motion information for prediction even for normal frames; iii) [Villegas *et al.*, 2017] uses historical motion information of the observed frames for future video prediction on the test data, so it may perhaps be able to predict

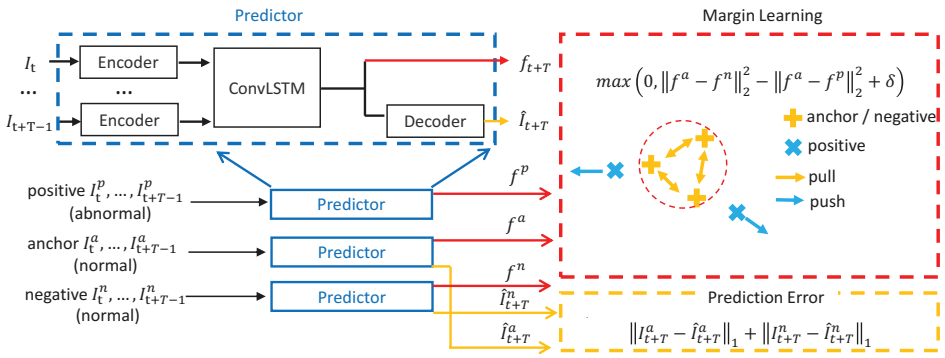


Figure 1: The network architecture of MLEP contains an encoder, a ConvLSTM and a decoder. Here, features f correspond to the hidden state of the last input in the ConvLSTM. f^a, f^p and f^n respectively correspond to features of anchor, positive and negative videos. It should be noted that we do not enforce a larger reconstruction error for abnormal frames.

anomalies.

Some existing research [Donahue *et al.*, 2015] [Luo *et al.*, 2017a] has shown that combining a 2D convolution and ConvLSTM can encode both spatial and temporal information for action recognition and anomaly detection. Therefore, we propose the use of such a scheme to favor the prediction of normal data while distorting the prediction of abnormal data. Specifically, given a video with normal events of length $T+1$, we propose to encode each frame of the first T frames with an encoder, then we will sequentially feed these features into a ConvLSTM, which has previously demonstrated its strengths in encoding motion features and spatial information from videos. Finally, we will feed the output of the ConvLSTM into a decoder to predict the last frame of a given video. For the encoder, we leverage the architecture used in [Zhu *et al.*, 2017] consisting of 3 convolution layers and 6 residual blocks. For a decoder, we use three deconvolution layers to gradually upscale the resolution and predict the future frame. Consequently, as is experimentally shown, our prediction framework will have a smaller prediction error for normal frames while having a larger one for abnormal frames. Thus, it is more suitable for anomaly detection when there are rare anomalies.

3.2 Margin Learning Module for the Open-set Supervised Setting

Regularizing normal data under a prediction framework is not enough to properly discriminate abnormal events. Taking into consideration that we only have a few abnormal events at hand and that many types of anomalies are unseen, we, in addition to the minimization of prediction errors for normal events, further utilize margin learning to enlarge the margins between normal-normal and abnormal-normal distance in the feature space. Last but not least, we propose to embed margin learning into our future frame prediction framework and to arrive at a Margin Learning Embedded Prediction (MLEP) framework.

Abnormal events come with two types of annotated ground-truths: video-level annotation and frame-level annotation. In video-level annotation, a video comes with a label to indicate whether it contains abnormal events, but it does not indicate which frames correspond to the abnormal events. Frame-level annotation, on the other hand, further indicates which frames are normal and abnormal. Our MLEP can readily handle annotations of both types as well as mixtures of the two.

For convenience, we define the notation used in this paper as follows: $\{I_t, \ldots, I_{t+T}\}$ is a video snippet with $T+1$ frames sampled consecutively from the training set, and $\{S_t, \ldots, S_{t+T} | S_t \in [0,1]\}$ correspond to their normal confidence levels, with higher values indicating that the frames are more likely to be normal. For frame-level annotation, we use the ground truth label to define S_t where $S_t = 1$ denotes a normal case and $S_t = 0$ denotes an abnormal one. For video-level annotation, since the whole video is annotated with 1 or 0, we also use S_t to denote a normalized score for the t^{th} frame. In our implementation, we set S_t based on the output of the prediction network trained with only normal data. $(\{I^a_t, \ldots, I^a_{t+T}\}, \{I^p_t, \ldots, I^p_{t+T}\}, \{I^n_t, \ldots, I^n_{t+T}\})$ denotes a triplet in margin learning, where a refers to an anchor (normal) frame, p refers to a positive (abnormal) one and n refers to a negative (normal) one. We further use $\hat{\cdot}$ to represent the output of the prediction network. For example, $\hat{I}^a_{t+T}$ and $\hat{I}^n_{t+T}$ represent the predicted $(t+T)^{th}$ frame for anchor and negative snippet, respectively. We denote normal scores as S^n_{t+T} and S^p_{t+T} for negative and positive snippets, respectively. Finally, we denote the output features of these video snippets as f^a, f^p and f^n, respectively.

The goal of the prediction network is to ensure that the predictions ($\hat{I}^a_{t+T}$ and $\hat{I}^n_{t+T}$) are close to the ground-truth (I^a_{t+T} and I^n_{t+T}) for normal events. We follow [Liu *et al.*, 2018; Mathieu *et al.*, 2016] and use a l_1 loss to measure the difference between predictions and ground-truths, and we then arrive at the following prediction loss:

$$L_{Pred}(\hat{I}^a_{t+T}, \hat{I}^n_{t+T}) = \|\hat{I}^a_{t+T} - I^a_{t+T}\|_1 + S^n_{t+T}\|\hat{I}^n_{t+T} - I^n_{t+T}\|_1 \quad (1)$$

It should be noted that we do not enforce a $-\|\hat{I}^p_{t+T} - I^p_{t+T}\|_1$ constraint in the prediction loss, because in any given scene normal and abnormal frames share the same background, enforcing abnormal data with a large prediction error to fit this

constraint will distort the prediction of normal data. Thus, we put the constraint that enlarges the gap between normal and abnormal distributions in the following margin module.

In the margin learning module, the goal is to decrease the distance between normal features while increasing the gap between normal and abnormal features, by leveraging a few anomalies and a multitude of normal data. Since the assumption is that a large volume of normal data is given, enforcing a small distance between normal data would tighten the distribution of normal data in the feature space, thus making the separation of normal and abnormal data easy. This consequently facilitates the detection of unknown anomalies in the open-set setting. Inspired by the successes of triplet loss [Weinberger and Saul, 2009] in face verification [Schroff *et al.*, 2015] and person re-id [Cheng *et al.*, 2016], our proposed margin learning method includes a weighted triplet loss:

$$\begin{aligned} L_{Triplet}(f^a, f^p, f^n) = (S^n_{t+T} - S^p_{t+T}) \\ * \max(0, \|f^a - f^n\|_2^2 - \|f^a - f^p\|_2^2 + \delta) \end{aligned} \tag{2}$$

Normality Confidence Calculation. In our implementation, for frame-level annotation, $S^n_{t+T} = 1, S^p_{t+T} = 0$. For video-level annotation, we use a prediction network trained with only normal data to predict a normal score for each frame. If some frames from abnormal videos have normal scores larger than 0.5, we add them to the candidates of negative snippets. If some frames from abnormal videos have normal scores less than 0.5, we add them to the candidates of positive snippets. Again, for video-level annotation, the score of a negative (positive) frame S^n_{t+T} (S^p_{t+T}) ranges from 0 to 1, and a smaller S^n_{t+T} (S^p_{t+T}) indicates that the frame is more likely to be an abnormal frame, and thus the triplet loss corresponding to this triplet should have a larger weight ($(S^n_{t+T} - S^p_{t+T})$ is larger).

3.3 Total Loss

By combining the losses corresponding to the future prediction module and the margin learning module, we arrive at the following objective function for our MLEP:

$$L = L_{Pred} + \lambda L_{Triplet} \tag{3}$$

where λ balances the weight of triplet loss. Here, $\lambda = 1$ in our experiment.

3.4 Training Phase

Since our objective contains a weighted triplet loss, how triplets are sampled is important to the performance of our method. Many previous works have discussed how to sample triplets in different tasks [Cheng *et al.*, 2016]. In our framework, for frame-level annotation and for samples in each mini-batch, we simply randomly choose an anchor, and a positive and a negative snippet to train the network. For video-level annotation, we first train a prediction-based anomaly detection network with our method using normal data only, and in this case, we set $\lambda = 0$. Then, we use the trained model to predict normal scores for both normal and abnormal data. Finally, we use the sampled triplets to retrain the whole framework.

3.5 Inference Phase

According to [Liu *et al.*, 2018], the quality of predicted future frames can measure the degree of an anomaly. The Peak Signal to Noise Ratio (PSNR), as shown in Equation 4, is a widely-used measurement for image quality assessment [Mathieu *et al.*, 2016] in which, a higher PSNR for the t-th frame indicates that it is more likely to be normal.

$$\begin{aligned} MSE(I_t, \hat{I}_t) &= \frac{1}{N^2} \sum_{i=0}^{N} \sum_{j=0}^{N} \|I_t(i,j) - \hat{I}_t(i,j)\|_2^2 \\ P_t(I_t, \hat{I}_t) &= 10 \log_{10} \frac{1}{MSE(I_t, \hat{I}_t)} \end{aligned} \tag{4}$$

Here, N denotes the number of rows (columns) in a frame and i, j denotes the i^{th} row and j^{th} column in this frame. Since different surveillance scenes may cause different magnitudes of PSNR, we normalize the PSNR of all frames under a single scene into scores, as illustrated in Equation 5.

$$S_t = \frac{P_t(I_t, \hat{I}_t) - \min\limits_{P_t \in B} P_t(I_t, \hat{I}_t)}{\max\limits_{P_t \in B} P_t(I_t, \hat{I}_t) - \min\limits_{P_t \in B} P_t(I_t, \hat{I}_t)} \tag{5}$$

where B is a set of PSNR of total frames under a single surveillance camera of view. The backgrounds vary highly due to different camera views and angles.

4 Experiments

All frames are resized to 224×224 pixels, and intensity values are normalized to $[-1, 1]$. The length of each video snippet T is 4. An Adam [Kingma and Ba, 2015] based stochastic gradient decent is used for parameter optimization, and the learning rate α is set to 0.0002. The margin δ is 1.0 and the coefficient λ is 1.0. An over-sampling strategy is utilized to deal with the imbalance of data of both our method and all baselines in this paper. In each iteration, two video snippets are randomly sampled from a normal data pool as an anchor and as a negative sample, and one video snippet is randomly sampled from the abnormal data pool as a positive sample. All source code will be released as a late date.

4.1 Datasets and Evaluation Metric

In this paper, we use two recently proposed anomaly datasets CUHK Avenue [Lu *et al.*, 2013] and ShanghaiTech Campus [Luo *et al.*, 2017b] for evaluation. In addition, these two datasets are the most challenging ones in terms of the types of anomalies and the sizes of the datasets.[1]

Training/Testing Split. Since only normal data are provided in the training set in the standard training/testing split for both datasets, in order to validate the performance of open-set supervised anomaly detection, we evenly split the abnormal data in the original testing set into K folds: $\{A_1, \ldots, A_k, \ldots, A_K\}$. Our criteria is that the k^{th} fold A_k

[1] In the UCF Crime dataset [Sultani *et al.*, 2018], the proportions of normal and abnormal data are equal. In addition, many camera angle changes are present in the videos, which is not ideal for prediction purposes. Thus, we do not evaluate this dataset in this paper.

only contains a few types of abnormal events rather than all types of anomalies in the training set. Then we conduct $K(=10)$-folders cross-validation, *i.e.*, and for each instance, we choose one fold and add it into the training set, and we use the remainder as the testing set. In other words, we divide the training and testing data K ways. Consequently, the ratio of normal to abnormal frames in the Avenue training set is around 50:1 and that of the ShanghaiTech training set is around 85:1. Therefore the training/testing split strategy guarantees that 1) the testing set must contain the certain types of abnormal events that are not included in the training set; and 2) the testing set may contain certain types of abnormal events observed in the training set.

Following [Hasan *et al.*, 2016; Liu *et al.*, 2018], we also leverage a frame-level AUC for a performance evaluation.

4.2 Comparison with the State-of-the-art

We conduct experiments under two different ground-truth training set annotations, *i.e.,* frame-level annotation and video-level annotation. The results of their mixture can be seen in the supplementary materials. For the test set, all frame-level annotations are provided for performance evaluation. We compare our method with the following baselines:[2]

Semi-supervised anomaly detection methods. State-of-the-art semi-supervised anomaly detection methods, including Conv-AE [Hasan *et al.*, 2016], stacked RNNs [Luo *et al.*, 2017b], Unmasking [Ionescu *et al.*, 2017], and Future Prediction [Liu *et al.*, 2018] (for a fair comparison, we replace the backbone of [Liu *et al.*, 2018] with ours and denote this baseline as Future Prediction*). All of these methods only leverage normal data for anomaly detection in the training set. For a fair comparison, the testing set is the same over all baselines.

Imbalanced video classification methods. Closed-set supervised anomaly detection is essentially imbalanced classification, which typically uses a re-sampling strategy or cost sensitive learning. [Yan *et al.*, 2015] uses a bootstrapping method for CNN-based imbalanced image classification. Thus, we extend it to the video setting and formulate our task as imbalanced video classification with data re-sampling. We denote this baseline as Imbalanced Video Classification with Over-Sampling (IVC with OS). We also append a Focal Loss [Lin *et al.*, 2018], which is a form of cost-sensitive learning for imbalanced object classification, and we denote this baseline as IVC with OS&Focal. Further, we leverage an prevalent two-stream based action recognition [Simonyan and Zisserman, 2014] network to extract features and only train a binary classifier, which we call IVC with OS&Focal&Two-stream.

Triplet Loss + One Class SVM. To further discuss the interlinked effects of the future prediction module and the margin learning module, we also design a baseline using only a margin learning module, namely, Triplet Loss. Since margin learning by itself cannot classify anomalies, we append an One Class SVM and denote this baseline as Triplet Loss + OCSVM. This baseline also uses both normal and abnormal data for anomaly detection.

[2][Sultani *et al.*, 2018] is a closed-set supervised anomaly detection with even normal/abnormal data in its training set, so we did not make comparisons with it.

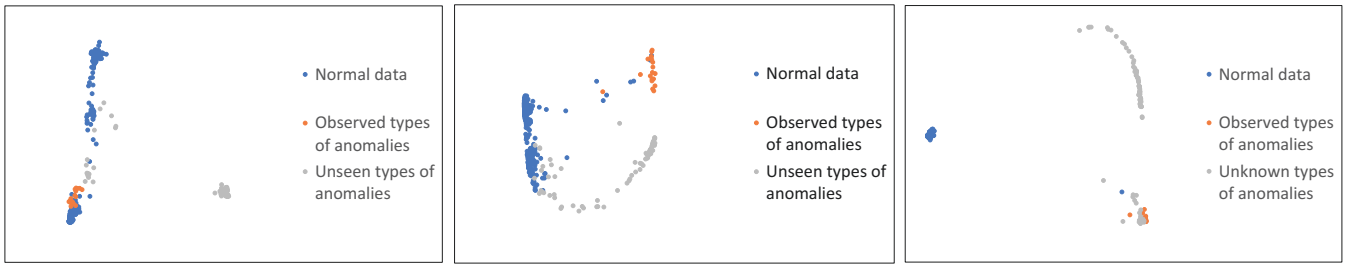
	Avenue	Shanghai Tech
Conv-AE [Hasan *et al.*, 2016]	80.0%	60.9%
Unmasking [Ionescu *et al.*, 2017]	80.6%	N/A
Stacked RNN [Luo *et al.*, 2017b]	81.7%	68.0%
Future Prediction [Liu *et al.*, 2018]	84.9%	72.8%
Future Prediction* [Liu *et al.*, 2018]	89.2%	73.4%
IVC with OS	82.8%	55.6%
IVC with OS & FL	83.1%	50.0%
IVC with OS & FL & Two-stream	80.6%	49.5%
Triplet loss + OCSVM	80.4%	50.4%
MLEP (Video-level annotations)	**91.3%**	**75.6%**
MLEP (Frame-level annotations)	**92.8%**	**76.8%**

Table 1: Performance comparison on Avenue and ShanghaiTech Campus.

The performances of these methods under different settings are reported in Table 1. We can make the following observations from this table: i) Introducing a few points of abnormal data in the training set significantly boosts the performance of anomaly detection. Specifically, our MLEP significantly outperforms the state-of-the-art semi-supervised methods (achieving around an 8% and a 4% higher AUC on the Avenue and ShanghaiTech datasets). Therefore, it demonstrates the effectiveness of supervised anomaly detection; ii) The improvement of our method over other imbalanced video classification (closed-set supervised anomaly detection) methods demonstrates the effectiveness of our solution for the open-set setting; iii) By comparing the performance of MLEP with different annotations, we can see that video-level annotation achieves a comparable performance with frame-level annotation, demonstrating how our method can handle data that has been more easily annotated; iii) By replacing the backbone of [Liu *et al.*, 2018] with ours, the performance on the Avenue and ShanghaiTech dataset increases by around 4.3% and 0.6%, respectively. Therefore, our proposed network is more suitable for anomaly detection.

4.3 Design of Network Architecture

We compare our proposed network with some state-of-the-art video prediction networks, including [Villegas *et al.*, 2017] [Ronneberger *et al.*, 2015] [Isola *et al.*, 2017]. For fair comparison, we only replace our proposed predictor in the MLEP with other networks and keep other modules unchanged. Table 3 shows that our method corresponds to be the best performance. The reason for the improvement of our network over UNet is that we do not adopt skip connection that may ignore the margin learning and favor the prediction of abnormal frames. In addition, our network has a larger gap between normal and abnormal scores than other advanced prediction networks [Villegas *et al.*, 2017] [Isola *et al.*, 2017], where a larger gap helps the classification between normal and abnormal events. Thus, our predictor is more suitable for open-set setting than other ones.



(a) MLEP ($\lambda = 0$) (b) Margin Learning Only (c) MLEP

Figure 2: The visualization of learned features by different models on the Avenue dataset. The left one is the features learned with only prediction loss, the middle one is that with only triplet loss, and the right one is that with MLEP. Blue ones presents the features of normal data, orange ones are the features of observed types of anomalies, and gray ones correspond to that of unseen types of anomalies. We can see that normal events are grouped together while there is a large gap between normal and abnormal data, even for unobserved types of ones. Best view in color.

	MCNet	UNet	Cycle GAN Generator	**Ours**
Mean AUC	88.1%	87.2%	88.0%	**92.8%**
Score on normal frames	0.726	0.771	0.810	0.799
Score on abnormal frames	0.402	0.447	0.485	0.250
Gap between normal and abnormal scores	0.324	0.324	0.325	**0.548**

Table 2: Evaluation of different network architectures on Avenue. A larger gap means it is easier to tell anomalies from normal events.

	N+S	N+U	N+S+U
MLEP	95.6%	87.8%	92.8%
IVC with OS	96.4%	71.1%	82.8%
IVC with OS & FL	84.3%	82.2%	83.1%

Table 3: Evaluation on different testing subsets of the Avenue dataset. N denotes Normal while S and U denotes Seen and Unseen types of anomalies, respectively.

	MLEP	IVC with OS	IVC with OS & FL
10-fold	92.8%	82.8%	83.1%
5-fold	93.7%	93.9%	91.5%

Table 4: Comparison with Different Portions of Anomalies.

4.4 Robustness to the Open-set Setting

According to the partition criteria of testing data in our experiments, anomalies in the testing set contain both the observed types of anomalies as well as other unseen types of anomalies. Once our MLEP is trained, we evaluate its performance on different subsets of the test data: subset 1 (N + S) contains both normal data and observed types of anomalies in the testing set; subset 2 (N + U) contains unseen types of anomalies and normal data; and subset 3 (N + S + U) is the set containing both of types of anomalies. We show the results in Table 3 where our method detects both observed as well as unseen types of anomalies very well. The reason for this is that we decrease the distances among normal data while pushing anomalies far from the normal data in our MLEP.

4.5 The Effect of The Portion of Uneen Anomalies

Next, we further discuss how the proportion of normal and abnormal data in the training set impacts binary classification. For 5-fold cross-validation, the training set contains 54.3% of the anomalies in the testing set in terms of the types of anomalies present. As for the 10-fold case, the number is 34.4%, which means the most of the anomalies in the testing set have never been seen before. We show the performance under both settings in Table 4 where introducing more anomalies into the training set will lead to a higher AUC for all methods. Further, by comparing our method with binary classifiers, we can see that our method is more robust to the case where only a few types of anomalies are present. In the training phase, when all types of anomalies are observed, the performance of our MLEP is similar to that of binary classifiers. In practice, due to the rare and unbounded nature of anomalies, it is not feasible to include all types of anomalies in the training set. Thus, our MLEP is a better solution than binary classifiers in the open-set setting.

4.6 Visualization of Features

We project features of the test data from the Avenue dataset onto a 2D space with PCA and show the results in Figure 2. We can see that our model does a good job of separating normal and abnormal events, even for unobserved types of abnormal events. The reason for this phenomenon is that we explicitly reduce the distance between normal data. Thus, the distribution boundary for normal events is tightened. Meanwhile, we enlarge the gap between normal and abnormal data so that the distance between normal and the observed types of anomalies is further enlarged, which also helps in pushing the unobserved types of anomalies far from the normal data. In

sum, it demonstrates the effectiveness of our proposed MLEP for anomaly detection with a few anomalies.

5 Conclusion

This paper presents a Margin Learning Embedded Prediction (MLEP) framework for open-set supervised video anomaly detection. Specifically, we propose the combination of a 2D convolution encoder with a ConvLSTM for future frame prediction for video anomaly detection, with triplet loss imposed to guarantee a large margin between normal and abnormal events as well as tighten the boundary of the normal data distribution, thus helping with the open-set supervised anomaly detection. Because of the rare and unbounded nature of anomalies, our problem setting is more closely aligned with real applications. Extensive experiments validate the effectiveness of our method.

Acknowledgments

This work was supported by NSFC #61502304.

References

[Ahmed *et al.*, 2016] M. Ahmed, A. N. Mahmood, and J. Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 2016.

[Chandola *et al.*, 2009] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[Cheng *et al.*, 2016] D. Cheng, Y. Gong, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[Donahue *et al.*, 2015] J. Donahue, L. A. Hendricks, et al. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[Dong *et al.*, 2017] Q. Dong, S. Gong, and X Zhu. Class rectification hard mining for imbalanced deep learning. *ICCV*, 2017.

[Hasan *et al.*, 2016] M. Hasan, J. Choi, et al. Learning temporal regularity in video sequences. In *CVPR*, 2016.

[Huang *et al.*, 2018] C. Huang, Y. Li, C. Loy, and X. Tang. Deep imbalanced learning for face recognition and attribute prediction. *ECCV*, 2018.

[Ienco *et al.*, 2017] D. Ienco, R. G Pensa, and R. Meo. A semisupervised approach to the detection and characterization of outliers in categorical data. *IEEE TNNLS*, 2017.

[Ionescu *et al.*, 2017] R. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu. Unmasking the abnormal events in video. In *ICCV*, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[Kingma and Ba, 2015] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Krawczyk *et al.*, 2014] B. Krawczyk, M. Woźniak, et al. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 2014.

[Li and Guo, 2007] Y. Li and L. Guo. An active learning based tcm-knn algorithm for supervised network intrusion detection. *Computers & security*, 2007.

[Li *et al.*, 2011] S. Li, Z. Wang, G. Zhou, and S. Lee. Semi-supervised learning for imbalanced sentiment classification. In *IJCAI*, 2011.

[Lin *et al.*, 2018] T. Lin, P. Goyal, et al. Focal loss for dense object detection. *TPAMI*, 2018.

[Liu *et al.*, 2018] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, 2018.

[López *et al.*, 2012] V. López, A. Fernández, Jose G Moreno-Torres, and F. Herrera. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608, 2012.

[Lu *et al.*, 2013] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013.

[Luo *et al.*, 2017a] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *ICME*, 2017.

[Luo *et al.*, 2017b] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, 2017.

[Mathieu *et al.*, 2016] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016.

[Ren *et al.*, 2018] M. Ren, W. Zeng, et al. Learning to reweight examples for robust deep learning. *ICML*, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[Sabokrou1 *et al.*, 2018] M. Sabokrou1, M. Khalooei2, M. Fathy1, and E. Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, 2018.

[Schroff *et al.*, 2015] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[Shi *et al.*, 2015] X. Shi, Z. Chen, et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[Sultani *et al.*, 2018] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018.

[Tang *et al.*, 2009] Y. Tang, Y. Zhang, N. V Chawla, and S. Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009.

[Tian *et al.*, 2011] J. Tian, H. Gu, and W. Liu. Imbalanced classification using support vector machine ensemble. *Neural Computing and Applications*, 2011.

[Villegas *et al.*, 2017] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017.

[Weinberger and Saul, 2009] K. Q Weinberger and L. K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009.

[Yan *et al.*, 2015] Y. Yan, M. Chen, et al. Deep learning for imbalanced multimedia data classification. In *Multimedia (ISM)*. IEEE, 2015.

[Zhu *et al.*, 2017] J. Zhu, T. Park, P. Isola, and A. A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.