ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230417

ESSL: Enhanced Spatio-Temporal Self-Selective Learning Framework for Unsupervised Video Anomaly Detection

Qun Li[a], Xubei Pan[a], Fu Xiao[a;*] and Bir Bhanu[b]

[a]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
[b]Department of Electrical and Computer Engineering, University of California at Riverside, CA, USA
ORCiD ID: Qun Li https://orcid.org/0000-0002-8034-6030, Xubei Pan https://orcid.org/0000-0003-3517-1224,
Fu Xiao https://orcid.org/0000-0003-1815-2793, Bir Bhanu https://orcid.org/0000-0001-8971-6416

Abstract. Unsupervised Video Anomaly Detection (UVAD) utilizes completely unlabeled videos for training without any human intervention. Due to the existence of unlabeled abnormal videos in the training data, the performance of UVAD has a large gap compared with semi-supervised VAD, which only uses normal videos for training. To address the problem of insufficient ability of the existing UVAD methods to learn normality and reduce the negative impact of abnormal events, this paper proposes a novel Enhanced Spatio-temporal Self-selective Learning (ESSL) framework for UVAD. This framework is designed for capturing both the appearance and motion features through effective network structures by solving the spatial and temporal jigsaw puzzles. Specially, we develop a Self-selective Learning Module (SLM) for UVAD, which prevents the model learning abnormal features and enhances the model by selecting normal features. Experimental results on three benchmark datasets show that the proposed method not only surpasses the state-of-the-art UVAD works, but also achieves the performance comparable to the classic semi-supervised methods for video anomaly detection that needs normal videos selected manually. Code is available at: https://github.com/xusuger/ESSL.

1 Introduction

Video Anomaly Detection (VAD) is a challenging task for detecting abnormal events that deviate from normality [3], such as fires and traffic accidents. Considering most of the abnormal events endanger public safety, the research on VAD is important and significant [26]. Due to the rarity and variety of abnormal events, it is impossible to collect all anomalies for fully-supervised learning [37]. Therefore, classic semi-supervised VAD is conducted in a semi-supervised learning scheme [17] by using only the normal videos. These methods detect anomalies that do not fit the normality mode of the detection model. However, normal videos that are required for semi-supervised learning need to be manually selected. In addition, classic VAD methods have difficulty in adapting to complex scenes. For these reasons, weakly-supervised VAD methods [7, 20] are proposed, which utilize videos with only video-level labels. Although weakly-supervised VAD avoids manual selection at the fine-grained level, it needs to inspect the entire video content. Thus, it is still a time-consuming and laborious job, when facing with a large number of surveillance videos.

* Corresponding Author. Email: xiaof@njupt.edu.cn.

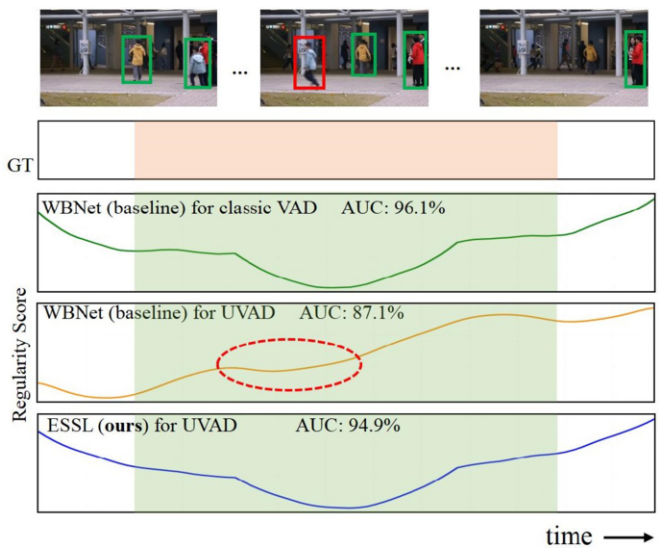

Figure 1. An example from the CUHK Avenue dataset. Here,"running" is identified as an anomaly event, and the shadow presents the ground-truth corresponding to the abnormal event where the person is running. The red dashed circle marks the failure case of the baseline for Unsupervised Video Anomaly Detection (UVAD). It can be seen that the Area Under Curve (AUC) of our method exceeds the baseline by 7.8% on UVAD tasks, and our unsupervised method without manual annotations is competitive with the classic semi-supervised method that needs manual annotations.

An ideal approach is the one that achieves VAD without any human intervention. Therefore, unsupervised VAD (UVAD), which is quite challenging, utilizes only unlabeled videos for training, and it has began to attract increasing attention from researchers. However, it leads to poor performance when the classic VAD methods are applied to UVAD tasks. As shown in Figure 1, when the baseline method for classic semi-supervised VAD is applied to UVAD tasks, the performance declines 9%. Most of the existing methods [37, 39] for UVAD focus on reconstructing the normality by using autoencoders and force the training model not to reconstruct anomalies. The problem is that, due to the strong generalization ability of autoencoders [12], some anomalies can be reconstructed as well, which re-

sults in a reduction of performance. Current research [37] has demonstrated the benefits of deep reconstruction in UVAD. Therefore, our work develops an effective framework to capture the deep normal information.

Recent UVAD works [37, 39] reduce the impact of abnormal samples by comparing the pixel error of sequential video frames to eliminate abnormal samples or assign low weights to suspicious abnormal samples, but such coarse-grained elimination does not have a good effect. Hence, another problem is how to effectively reduce the impact of abnormal samples on the learning of normal features. Improved self-paced refinement [37] based on self-paced learning [28] removes suspicious anomalies for VAD, however, requires precise parameters. Thus, it is difficult to tune a suitable set of parameters for models, which limits it to a wide range of applications. Due to the participation of abnormal samples during the training process, the accuracy of UVAD is obviously lower than classic semi-supervised VAD, as shown in Figure 1.

To address the above problems, our motivation is to expect the training process of UVAD to focus on modeling normality, just like classic VAD, while without human intervention and automation. To realize such an idea, we present a novel Enhanced Spatio-temporal Selection Learning (ESSL) framework for UVAD. The framework is based on constructing spatial and temporal jigsaw puzzles, and it effectively solves both jigsaw puzzles through the enhanced network structure to capture the appearance and motion features of normal events. Specially, we propose a simple yet effective and plug-and-play module named Self-selective Learning Module (SLM), which can autonomously filter out abnormal samples and guide the model to learn towards the normality.

Our main contributions are summarized as follows: (1) We propose an enhanced spatio-temporal framework that can effectively capture the normality features of normal events by solving jigsaw puzzles. (2) We propose a SLM for UVAD to force the model on learning normal features rather than abnormal features and develop the training process of UVAD approximating to classic VAD without manual annotations. (3) Experimental results on three benchmark datasets show that our method not only surpasses the State-Of-The-Art (SOTA) UVAD networks, but also achieves the competitive performance in comparison to classic methods for classic semi-supervised VAD.

2 Related Work

Unsupervised Video Anomaly Detection (UVAD). Both semi-supervised and weakly-supervised methods for VAD require manual labeling for training data, which is time-consuming and laborious. To achieve a fully automatic VAD that does not require human intervention, the UVAD has gained a growing attention [37, 39]. Ordinal regression [26] constructs a self-trained network to achieve end-to-end UVAD. Generative cooperative learning framework [39] for UVAD comprises a generator, a discriminator and a cross-supervision module, where the generator is forced not to reconstruct anomalies. The localization-based reconstruction method [37] points out that deep reconstruction is surprisingly effective for UVAD. Therefore, to reduce the interference of abnormal samples, we propose a novel ESSL framework that not only eliminates suspicious abnormal samples, but also enhances the ability of the network to learn normality.

Object-Centric Method. In surveillance videos, there are both static backgrounds and dynamic events, such as walking people and moving cars. In VAD tasks, anomalies are often dynamic events, thus we should pay more attention to those dynamic events that are more prone to anomalies than static backgrounds. Object-centric methods can largely reduce the negative effect of complex backgrounds on normality learning. Specially, the work in [13] trains object-centric convolutional autoencoders for both motion and appearance features for VAD. Localization-based reconstruction method [37] tries to reconstruct the object-centric cube for VAD. Some self-supervised works [9, 31] utilize objects in videos to build proxy tasks for VAD. In this paper, we follow existing works and apply an object detector to construct an object training set.

Normality Advantage. Recent studies [32, 37] unveil a property named "*normality advantage*". It means that anomalies are unusual events and the probability of occurrence is extremely low, while the main content of videos are normal. Exploiting the large difference in the number of normal and abnormal events just provides us with an opportunity to separate of normal events from abnormal events. Besides, there may be multiple object-level events in one frame, but only one event is abnormal. However, frame-level based methods treat the whole frame as an anomaly. Recent works further magnify normality advantage by extracting object-level events as the basic training data. In order to leverage the normality advantage, we propose the SLM that can effectively reduce the negative impact of abnormal samples in UVAD.

Self-supervised Learning. Self-supervised learning seeks supervisory signals from data for training. It commonly constructs pretext tasks, such as jigsaw puzzles and image rotations, to capture deep features. Although recent works demonstrate the high effectiveness of self-supervised learning in VAD, there are only a few works to explore self-supervised learning for VAD. A self-supervised and multi-task learning method [9] utilizes multiple proxy tasks to train a 3D Convolutional Neural Network (CNN) for VAD. Further, in order to make proxy tasks simple yet effective, the latest work in [31] tries to train the 3D CNN to solve decoupled spatio-temporal jigsaw puzzles. However, most of the existing works focus on how to build proxy tasks but only use a shallow 3D CNN to handle the proxy tasks. Considering these, we design an enhanced spatio-temporal framework that effectively captures normality features to solve jigsaw puzzles.

3 Our Method

As shown in Figure 2, our method contains three stages including the jigsaw construction, the enhanced spatial-temporal module, and SLM. First, we transform the extracted Spatio-Temporal Cubes (STCs) to construct the spatial jigsaws and temporal jigsaws. Then, we input the jigsaw STCs into the framework to predict the correct spatial and temporal order, respectively. Finally, the proposed model eliminates potential anomalous samples through SLM, and the remaining ones participate in back propagation.

3.1 Jigsaw Construction

To avoid interference from complex backgrounds and capture both appearance and motion features, we construct two different jigsaw puzzles based on objects for the network to solve.

Spatio-Temporal Cube (STC). As mentioned in Section 2, anomalies are often dynamic events, thus we are more interested

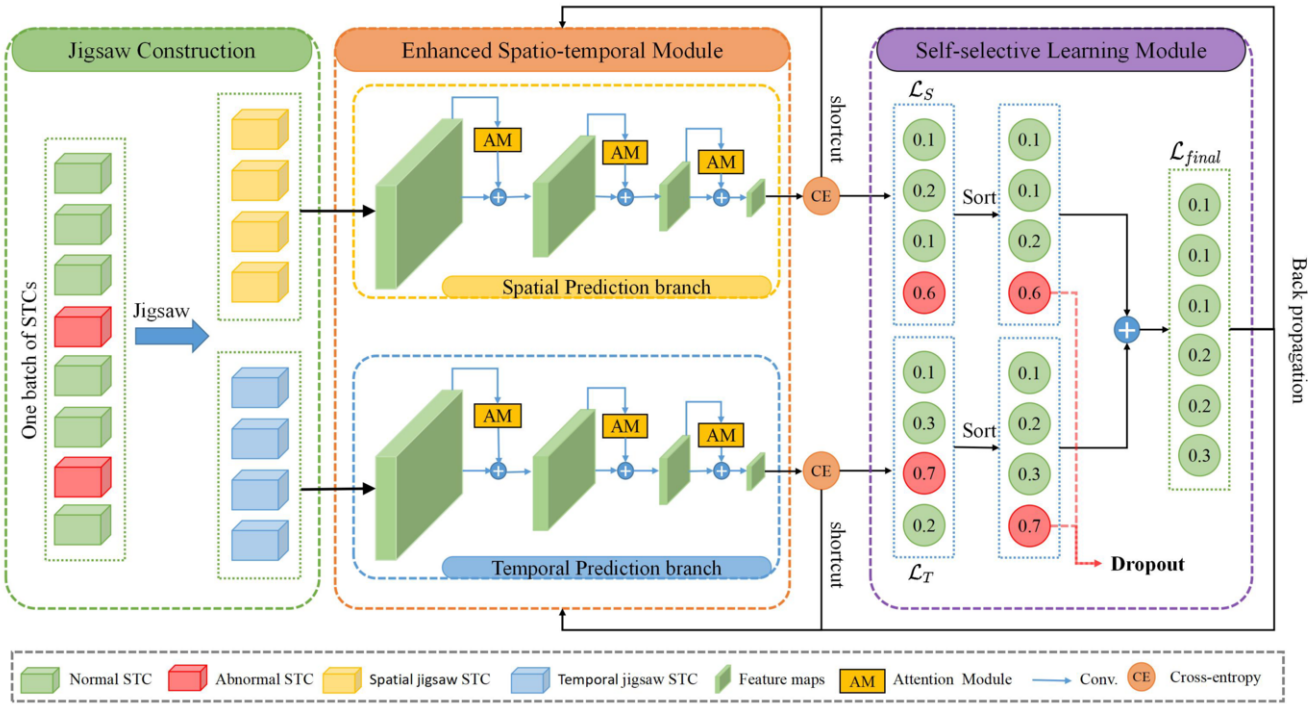

Figure 2. Overview of the proposed framework. First, the network transforms the extracted Spatio-Temporal Cubes (STCs) to construct the spatial jigsaws and temporal jigsaws. Then, we input the jigsaws into the dual-branch network to predict the correct spatial order and temporal order, respectively. Finally, the model eliminates potential anomalous through the SLM, and the remaining ones participate in back propagation. Here, $\mathcal{L}_S$ denotes the loss of the spatial branch, and $\mathcal{L}_T$ denotes the loss of the temporal branch. $\mathcal{L}_{final}$ represents the final total loss.

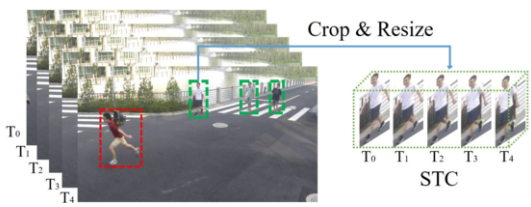

Figure 3. Extraction of the Spatio-Temporal Cube (STC). Objects are detected by YOLOv3 from temporally adjacent T frames to build the STC by cropping, resizing and stacking.

in dynamic events than static backgrounds. Considering that object-centricity can largely reduce the negative effect of complex backgrounds, similar to most of the object-centric works [9, 37], we apply a pre-trained YOLOv3 detector to detect and obtain a large number of objects. Then, for each object in a frame, we use the same bounding box to extract equal-size patches on its temporally adjacent T frames to build a stacked STC, as one STC example shown in Figure 3. We regard a STC as a basic event and apply it as a basic training sample. Noted that the number of normal events is obviously more than abnormal events even in an abnormal frame, the STC utilizes this phenomenon to further strengthen the normality advantage mentioned above, which enhances our network as conducive to learn normal features as possible.

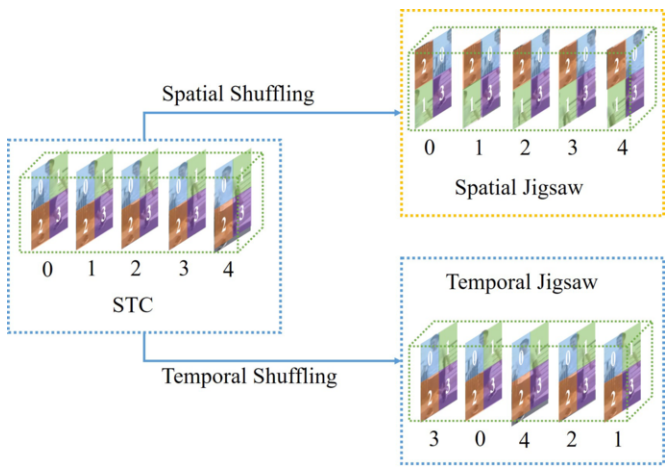

Figure 4. Construction of jigsaw puzzles. We first divide each object in the Spatio-Temporal Cube (STC) into $n \times n$ equal-size mini-patches, then randomly shuffle these mini-patches without shuffling the temporal permutation to construct the spatial jigsaw. Similarly, we shuffle the temporal permutation of T objects in the STC without shuffling the spatial permutation to construct the temporal jigsaw.

Jigsaw Puzzles. Before the training data are fed into the network, we shuffle the spatial and temporal order of the original STC to construct jigsaw puzzles. Figure 4 shows the construction process of the jigsaw puzzles. We first divide each object in the STC into $n \times n$ equal-size mini-patches, then randomly shuffle the mini-patches but keep the temporal permutation to construct the spatial jigsaw. Similarly, we randomly shuffle the T objects but keep the spatial permutation to construct the temporal jigsaw. Finally, these jigsaws are fed into the network to predict the correct permutation order. As shown in Figure 2, for a batch of STCs, we perform corresponding ran-

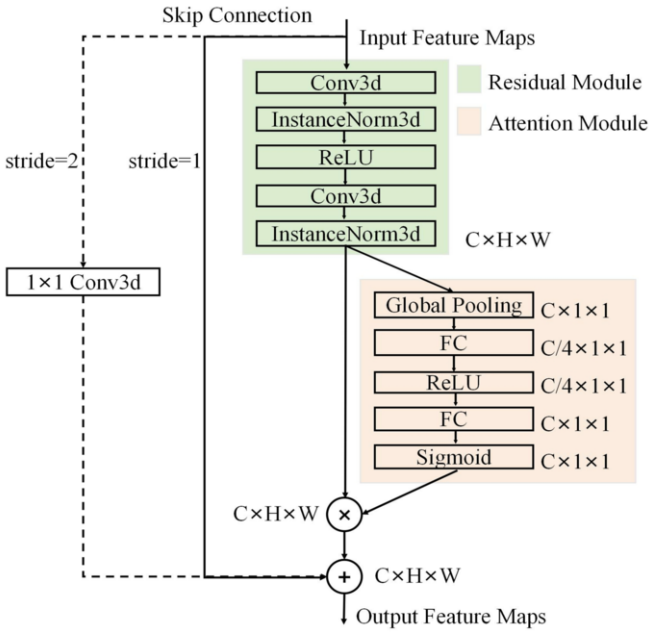

Figure 5. Structure of RAB. It mainly contains a residual module, an attention module and skip connections. The residual module captures deep features, while the attention module learns appropriate weights for the feature maps through two Fully Connected (FC) layers. Skip connections ensure that the original input can be directly mapped to the next layer. If the size of the input and output feature maps are not in equal-size, the connection of $stride = 2$ is adopted. Otherwise, the connection of $stride = 1$ is adopted.

dom shuffling operations to construct spatial and temporal jigsaws, respectively. Then, we feed the spatial and temporal jigsaws into their respective prediction branch.

3.2 Enhanced Spatio-temporal Module (ESM)

We construct pair-wise permutation branches to learn spatial and temporal normality by solving spatial and temporal jigsaw puzzles. Besides, we further embed residual attention modules in the network to enhance the ability of the network to capture deep features.

Enhanced Residual Attention Block (RAB). Figure 5 shows the detailed structure of RAB, including a residual module, an attention module and skip connections. The residual module contains two 3D convolutions (Conv3d), two 3D InstanceNorms and a $ReLU$ activation function. It adjusts mainly the number of channels and the size of the feature maps to capture deep features. The attention module first performs a global pooling operation on the feature map with the size of $C \times H \times W$ from the residual module to obtain attention weights with the size of $C \times 1 \times 1$. Then, two Full-Connected (FC) layers are applied to learn new attention weights with the size of $C \times 1 \times 1$. After passing through the $Sigmoid$ activation layer, these new weights are multiplied by the original feature maps from the residual module to obtain new feature maps with the size of $C \times H \times W$. With the help of skip connection ($stride = 1$), the input feature maps are directly added to the new feature maps to get the outputs. In addition, if the sizes of input and output feature maps are inconsistent, we choose the other skip connection ($stride = 2$) with an 1×1 Conv3d to resize the input feature maps. For a clear explanation, the complete operation is defined as follows:

$$Y = SC(X) + RM(X)AM(RM(X)), \tag{1}$$

where X and Y denote the input and output feature maps, respectively. $SC(\cdot)$ represents the skip connection, $RM(\cdot)$ represents the residual module operation, and $AM(\cdot)$ represents the attention module operation.

Spatio-temporal framework. As shown in Figure 2, in order to make the network efficient in solving different types of puzzle problems, which means that the network can better model the spatial and temporal features, the proposed framework is constructed by two prediction branches, which carry out the spatial and temporal permutation prediction, respectively. Each branch focuses on solving one jigsaw puzzle. The residual structure of the block is introduced into the dual-branch network, which allows us to build a deeper network that improves the ability of our model to learn normality features. Since these two features are equally important, the structures of these two branches are exactly the same.

Permutation Prediction. Traditional methods [11, 14, 15] regard the jigsaw puzzle as a multi-classification task and view each permutation as one class. But they suffer from an significant limitation in solving complex problems. For example, in the 3D jigsaw puzzle, a STC with 5 frames divided into 3×3 mini-patches results in $5! \times (3 \times 3)! = 43,545,600$ possible permutations. It is extremely a hard classification problem. To reduce the complexity, following the work in [31], we employ a multi-label supervision, which predicts the correct permutation order directly. Taking Figure 4 as an example, the target of the spatial jigsaw is to predict the correct spatial permutation [2, 0, 1, 3], and the target of the temporal jigsaw is to predict the correct temporal permutation [3, 0, 4, 2, 1].

Like most of the classification works, we choose to use the Cross-Entropy (CE) loss to optimize our network. For one STC, the losses of two prediction branches are defined as follows:

$$\begin{cases} \mathcal{L}_S = \frac{1}{n^2} \sum_{i=1}^{n^2} CE(s_i, \hat{s_i}) \\ \mathcal{L}_T = \frac{1}{T} \sum_{i=1}^{T} CE(t_i, \hat{t_i}) \end{cases} \tag{2}$$

where $\mathcal{L}_S$ denotes the loss of the spatial branch, and $\mathcal{L}_T$ denotes the loss of the temporal branch. n^2 represents the number of mini-patches in an object, and T represents the number of object frames in the STC. s_i and $\hat{s_i}$ are the ground-truth spatial permutation and the predicted location of one mini-patch, respectively. t_i and $\hat{t_i}$ are the ground-truth temporal permutation and the predicted order of one object frame, respectively.

3.3 Self-selective Learning Module (SLM)

Due to the above mentioned normality advantage in UVAD, the proposed network has a bias towards solving normal jigsaw puzzles, which results in larger losses for abnormal jigsaw puzzles than normal ones. It offers us an opportunity to rule out the effects of anomalies. As shown in Figure 2, the SLM selectively filters the losses $\mathcal{L}_S$ and $\mathcal{L}_T$ from two branches, and a certain amount of large losses are directly dropped without participating in back propagation, while the remaining ones are integrated to update the parameters of our model.

According to the analysis above, "*the losses of abnormal samples are larger than normal ones*", we introduce a selection factor λ for

dropout large losses to prevent abnormal samples from participating in training. In a batch, let $\mathcal{D}_S = \{\mathcal{D}_S^i\}_{i=1}^{N_S}$ be the set of $\mathcal{L}_S$ and $\mathcal{D}_T = \{\mathcal{D}_T^i\}_{i=1}^{N_T}$ be the set of $\mathcal{L}_T$, where $\mathcal{D}_S^i$ denotes the i^{th} element in $\mathcal{L}_S$, and $\mathcal{D}_T^i$ denotes the i^{th} element in $\mathcal{L}_T$, respectively. N_S is the number of $\mathcal{L}_S$, and N_T is the number of $\mathcal{L}_T$. First, we sort $\mathcal{D}_S$ and $\mathcal{D}_T$ as follows:

$$\begin{cases} \mathcal{Q}_S = sort(\mathcal{D}_S) \\ \mathcal{Q}_T = sort(\mathcal{D}_T) \end{cases} \tag{3}$$

where $sort(\cdot)$ is an ascending sort operator. $\mathcal{Q}_S$ and $\mathcal{Q}_T$ denote the sorted $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. With the help of the selection factor λ, the numbers of retained losses are calculated as:

$$\begin{cases} R_S = \lceil \lambda_S \cdot N_S \rceil \\ R_T = \lceil \lambda_T \cdot N_T \rceil \end{cases} \tag{4}$$

where R_S and R_T represent the number of retained $\mathcal{L}_S$ and the number of retained $\mathcal{L}_T$, respectively. $\lceil \cdot \rceil$ denotes a rounding up operator. λ_S and λ_T are the selection factors for the spatial branch and the temporal branch, respectively. We drop the large losses and retain the small ones in a batch. Before applying the SLM, the model updates its parameters by minimizing the original loss, which is defined as:

$$\mathcal{L}_{org} = min(\frac{1}{N_S}\sum_{i=1}^{N_S}\mathcal{D}_S^i + \frac{1}{N_T}\sum_{i=1}^{N_T}\mathcal{D}_T^i) \tag{5}$$

where $min(\cdot)$ is the minimization operator. After applying SLM to exclude suspicious anomalies, the learning objective of the model is adapted as follows:

$$\mathcal{L}_{final} = min(\frac{1}{R_S}\sum_{i=1}^{R_S}\mathcal{Q}_S^i + \frac{1}{R_T}\sum_{i=1}^{R_T}\mathcal{Q}_T^i) \tag{6}$$

where $\mathcal{Q}_S^i$ represents the i^{th} element of the sorted $\mathcal{L}_S$, and $\mathcal{Q}_T^i$ represents the i^{th} element of the sorted $\mathcal{L}_T$.

Warm-up. The working principle of SLM is based on the extreme imbalance in the number of normal and abnormal data in VAD training data, in which normal data accounts for the majority. Therefore, the model tends to be better at solving normal jigsaws. However, in the early stage of training, the ability of the model to solve puzzles is still weak, and this bias has not been established yet. Therefore, due to the model not having a tendency towards normality at the initial stage, we apply the proposed SLM after a few warm-up epochs, which ensures that the model has a preliminary bias towards normality. During the warm-up epochs, the losses will bypass SLM and directly update the network parameters through the shortcuts, as shown in Figure 2. We will discuss how to set the number of warm-up epochs in section of ablation study.

3.4 Inference

During inference, we first construct a series of STCs. Different from training, we directly feed the original STCs into two branches to predict the orders without spatial and temporal shuffling. Then, for each tested STC, we obtain the spatial confidence matrix and temporal confidence matrix. Since the original order of the STC is not shuffled, we take the minimum values of the diagonal term of two matrices as predicted spatial and temporal scores of the STC, respectively. Due to the irregularity of anomalies and their absences for training, their predicted scores are low. Finally, we take the minimum of the STC scores in a frame as the frame-level score. After normalization, the final score of a frame is defined as:

$$Score = Score_S + \omega \cdot Score_T. \tag{7}$$

where $Score_S$ and $Score_T$ are the frame-level appearance score and motion score, respectively. ω is the balance parameter.

4 Experiments and Analysis

4.1 Datasets and Evaluation Metric

We conduct experiments on three benchmark datasets including the UCSD Ped2 dataset [24], the CUHK Avenue dataset [21], and the ShanghaiTech dataset [17]. These datasets are originally proposed for semi-supervised VAD. For performing UVAD and a fair comparison, following the latest UVAD works [37, 39], we reorganize the datasets by mixing normal and anomalous videos in both training set and testing set.

Datasets. The UCSD Ped2 dataset contains 16 normal and 12 anomalous videos with 12 irregular events, including riding a bike and driving a vehicle. After the reorganization, its training set contains 9 normal and 7 anomalous videos, while its testing set consists of 7 normal and 5 anomalous videos. The CUHK Avenue dataset consists of 16 normal and 21 anomalous videos with 47 abnormal events, such as running and throwing stuff. After the reorganization, its training set consists of 8 normal and 12 anomalous videos, while the remaining ones as the testing set. Compared to two datasets above, the ShanghaiTech dataset is multi-view and more challenging. It contains 330 normal and 107 anomalous videos with 130 anomalies on 13 scenes. Its training set contains 175 normal and 63 anomalous videos, while the remaining ones as its testing set.

Evaluation Metric. Following most of the existing works [19, 27, 38], we calculate the frame-level Area Under Curve (AUC) of receiver operation characteristic for evaluating our methods. The higher the AUC is, the better the performance is.

4.2 Implementation Details

We adopt the same implementation of YOLOv3 in [9] to filter out the detected objects with low confidences. We set the size of STC to $T \times 64 \times 64 \times 3$, where T is the length of the STC. We set $T = 9$ on the ShanghaiTech dataset, while $T = 7$ on the UCSD Ped2 and CUHK Avenue datasets. For the mini-patches in the STC, we set $n = 3$ for all three datasets. We empirically set ω to be 0.1 for the UCSD Ped2 dataset, 10 for the CUHK Avenue dataset, and 1 for the ShanghaiTech dataset. The depth of each prediction branch in this enhanced framework is 8. The selection factors (λ_S, λ_T) are set to be $(0.9, 0.9)$ for the UCSD Ped2 and CUHK Avenue datasets, while $(0.9, 0.7)$ for the ShanghaiTech dataset. The whole network is optimized by the Adam optimizer with the initial learning rate of $1e^{-4}$, and it is decayed by using a cosine annealing method. On these three datasets, the number of training epochs is set to 100, while 5 epochs are used for warm-up. The batch size in training is 64.

Table 1. Comparisons of results on three benchmark datasets. **bold** and underline indicate the best and second-best results for UVAD.

Methods	Datasets		
	Ped2	Avenue	SHTech
Classic semi-supervised VAD			
AnoPCN [36]	96.8	86.2	73.6
Attention [41]	96.0	86.0	–
PDE-AE [1]	95.4	–	72.5
AM-Corr. [25]	96.2	86.9	–
AnomalyNet [40]	94.9	86.1	–
Object-Centric [13]	97.8	90.4	84.9
BMAN [16]	96.6	90.0	76.2
Clustering-AE [4]	96.5	86.0	73.3
r-GAN [22]	96.2	85.8	77.9
DeepOC [35]	96.9	86.6	–
Multipath-Pred. [33]	96.3	88.3	76.6
Mem-Guided [27]	97.0	88.5	70.5
CAC [34]	–	87.0	79.3
Scene-Aware [29]	–	89.6	74.7
VEC [38]	97.3	90.2	74.8
BAF [10]	98.7	92.3	82.7
AMMCN [2]	96.6	86.6	73.7
SSMTL [9]	97.5	91.5	82.4
MPN [23]	96.9	89.5	73.8
HF^2 [19]	99.3	91.1	76.2
CT-D2GAN [8]	97.2	85.9	77.7
STJP [31]	99.0	92.2	84.3
Bi-Pred [5]	98.3	90.3	78.1
UVAD			
DF [6]	63.0	78.3	–
UM [30]	82.2	80.6	–
CTS [18]	87.5	84.4	–
OR [26]	83.2	–	–
GCL [39]	–	–	78.9
LBR-SPR [37]	97.2	90.7	72.6
ESM (Ours)	97.8	94.4	78.3
ESSL (Ours)	**98.6**	**94.9**	**80.9**

Table 2. Ablation analysis of RAB and SLM on the Ped2 dataset. The "single" and "dual" represent the networks are single-branch or dual-branch, respectively.

Model	RAB	SLM	Dual-branch	AUC (%)
WBNet (baseline)	✗	✗	✗	97.4
ESM (single)	✓	✗	✗	97.7
ESM (dual)	✓	✗	✓	97.8
ESSL (single)	✓	✓	✗	98.3
ESSL (dual)	✓	✓	✓	**98.6**

Table 3. Ablation studies on selection factors on the ShanghaiTech dataset.

Selection factor						
λ_S	1.0	0.9	0.8	0.9	0.7
λ_T	1.0	0.9	0.8	0.7	0.7
AUC (%)	78.3	78.7	79.4	**80.9**	78.7

Table 4. Ablation studies on warm-up epochs on the Ped2 dataset.

warm-up epochs	0	5	10	20	30
AUC (%)	97.4	98.6	**98.9**	97.5	97.8

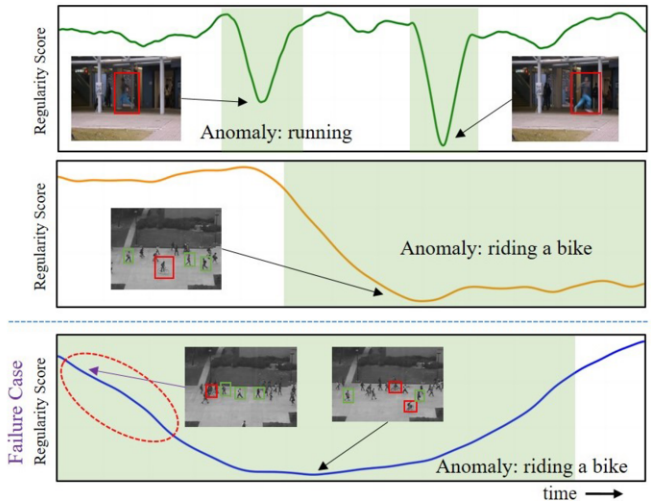

Figure 6. Example results of ESSL. The green shadow presents the ground-truth corresponding to anomalies, and the last one is a failure case of our method as marked in a red dashed circle.

4.3 Comparison with SOTA Methods

Comparison with UVAD Methods. As seen in Table 1, from the comparison with SOTA methods for UVAD, we can observe that the proposed ESSL achieves the best results on all three datasets. Especially, ESSL is the first one that exceeds 80% and reaches 80.9% in AUC on the ShanghaiTech dataset. Compared with LBR-SPR, which is also an object-centric method, it achieves the significant improvement on the performance by 1.4% on the UCSD Ped2 dataset, 4.2% on the CUHK Avenue dataset and 8.3% on the ShanghaiTech dataset. Even the ESM without the SLM, it surpasses LBR-SPR [37] by 0.6% on the UCSD Ped2 dataset and 3.7% on the CUHK Avenue dataset, due to the strong ability of ESM for capturing normal features. On the other hand, we notice an obtained AUC of 78.3%, which is 0.6% lower than the GCL method [39] on the ShanghaiTech dataset. Although the ShanghaiTech dataset is large and challenging, some abnormal features are learned due to the powerful learning ability of ESM, which leads to a lower performance. It just proves that the ESM can effectively capture deep features. Therefore, in order to reduce the negative influence of anomalies, we introduce the SLM. After applying the SLM, the performance of our model are improved with varying degrees on all three datasets as shown in Table 1. We gives three example results of ESSL in Figure 6, where the green shadow denotes the ground-truth corresponding to anomalies. The last one is a failure case. As the red dashed circle marked, the bicycle is partially blocked by pedestrians, thus it is not detected by the object detector.

Comparison with classic VAD Methods. Although our method belongs to UVAD, we also list the performance of classic semi-supervised VAD methods in Table 1. Comparing to the SOTA semi-supervised VAD works, ESSL boosts the best one named BAF [10] by 1.9% on the Avenue dataset, while only 0.7% lower on the Ped2 dataset. Take into account the different partition of the datasets, although such a simple comparison cannot prove that our UVAD method is better than the classic semi-supervised VAD methods, UVAD is more challenging and advantaged. Thus, it can say that

our method is competitive with classic semi-supervised VAD methods. Figure 1 demonstrates the superior performance of our ESSL in UVAD, even approaching classic semi-supervised VAD method.

4.4 Ablation Study and Analysis

Effectiveness of RAB and SLM. To verify the effectiveness of our constructed ESM, we perform ablation studies on the Ped2 dataset. we use the common Wide-Branch Network (WBNet) [31] as the baseline. We replace the convolutional block in the WBNet with the designed RAB to construct a single-branch ESM, in which the appearance features and motion features are extracted by a shared backbone network. It can be seen from Table 2 that the performance of single-branch ESM improves by 0.3% compared to the baseline, which proves the effectiveness of the designed RAB. After applying the SLM, single-branch ESSL obtains the AUC of 98.3%, which is 0.6% higher than the single-branch ESM.

Effectiveness of Dual-branch. Considering that the backbone network struggles to capture two different features, we build the dual-branch framework to solve the spatial jigsaw puzzles and temporal jigsaw puzzles, separately. To further verify the effectiveness of the dual-branch framework, we perform ablation studies on single-branch and dual-branch frameworks. It can be seen from Table 2 that ESSL and ESM with dual-branch architectures are better than those with single-branch architectures. Combining the above three improvements, the final dual-branch ESSL achieves the best AUC of 98.6% that is 1.2% higher than the baseline.

Selection Factors in SLM. To address the problem that appearance anomalies and motion anomalies are in different numbers, we choose a separate selection factor for each branch. To further explore the best pairwise of selection factors (λ_S, λ_T), we conduct multiple ablation studies on the ShanghaiTech dataset. The results of ESSL with different selection factors are shown in Table 3. We both set λ_S and λ_T to start from 1.0 to conduct our experiments, which means that no training samples are dropped. Causally, the setting of $(1.0, 1.0)$ shows a lowest performance 78.3%. The decrease of the selection factors means that fewer samples are selected for training. As (λ_S, λ_T) gradually decreases from $(1.0, 1.0)$ to $(0.8, 0.8)$, the AUC increases steadily to 79.4%. It reflects the effectiveness of the proposed SLM, which continuously eliminates abnormal samples in the training process and enforces the model focuses on learning normal features. However, as (λ_S, λ_T) decreases to $(0.7, 0.7)$, the AUC drops to 78.7%. It is easy to explain that too small selection factor discards too much training data even including normal samples. Considering the rarity of abnormal events, we do not further decrease the selection factors for the experiments. While the combination of $(0.9, 0.7)$ achieves the best performance 80.9%, we choose this pairwise for the ShanghaiTech dataset.

Take advantage of the proposed SLM, we achieve a jump in performance from 78.3% to 80.9%, while the core of SLM is only two selection factors. Hence, the SLM is a simple yet effective and plug-and-play method. It is not only effective in UVAD tasks, but also can be simple drop-in various unsupervised learning tasks in other fields to improve the performance under limited computational cost.

When to adopt SLM. As analyzed earlier, the application of SLM requires the establishment of a preliminary tendency towards model normality. Therefore, we conduct ablation studies on the Ped2 dataset to explore the appropriate application period of SLM. As shown in Table 4, when we apply SLM at the beginning of training, the model has not yet established the normality advantage, which results in poor performance 97.4%. When applying SLM after 5 or 10 enpochs, the AUC increases to 98.6% and 98.9%. However, when applying SLM after 20 or 30 epochs, the effectiveness of the model decreases because the model is already approaching convergence that makes it difficult to correct the training direction of the model by using SLM. Therefore, the intervention of SLM should be conducted as early as possible when the model has preliminary detection abilities but has not yet been fitted. And we recommend that SLM should be started during the first 5% to 10% epochs of the total training process. In this paper, for a fair comparison with the previous method [37], the number of warm-up epochs in all other experiments is set to 5.

Table 5. Sensitivity analysis of the hyperparameter ω on the Ped2 dataset.

ω	0.1	0.3	0.5	0.7	0.9
AUC (%)	**98.6**	98.3	98.5	98.5	98.3

Sensitivity Analysis. We also conduct ablation studies on the Ped2 dataset to explore the effect of hyperparameter ω on the performance. We adjust the variation of ω from 0.1 to 0.9. From Table 5, it can be seen that AUC remains stable above 98% and the change is gentle with a maximum amplitude not exceeding 0.3%. Therefore, it shows that the performance of the model is stable and not sensitive to ω.

Computational cost analysis. We conduct all experiments on an NVIDIA RTX 3070 Ti GPU and an Intel Core (TM) i7-12700K @ 3.60GHz. Compared to the previous works [19], our method is more lightweight and efficient with 9.53M #Params and 3.66G FLOPs. When there are an average of 5 objects in per frame, the model runs at a speed of 20 FPS, where the time of object detection and inference for per frame are approximately 20ms and 30ms, respectively.

Limitations. Our method is based on object-centricity, thus it is limited by the performance of object detector. In future work, we shall consider how to reduce the impact of inefficient object detection.

5 Conclusion

In this paper, in order to address the problems that previous models for UVAD were coarse-grained for the removal of suspicious anomalies and lacked effective normal information, we presented an ESSL framework to capture deep normal features. Our framework achieved impressive efficiency on three benchmark datasets, due to the effectiveness of ESM and SLM. The dual-branch framework performed jigsaw puzzles in both spatial–temporal domains to capture effective normal information by embedding multiple attention module, while the SLM aimed to achieve the fine-grained removal of suspicious anomalies by strengthening the network to focus on capturing normal features and neglecting anomalies automatically. The proposed SLM are plug-and-play and can be easily embedded into other networks or frameworks for UVAD, even the SLM can further expand to other unsupervised learning tasks.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grant No. 62276143, Natural Science Foundation of Jiangsu Province under Grant No. BK20230356, the National Science Fund for Distinguished Young Scholars of China under Grant No. 62125203, the Key Program of the National Natural Science Foundation of China under Grant No. 61932013.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara, 'Latent space autoregression for novelty detection', in *CVPR*, pp. 481–490, (2019).

[2] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao, 'Appearance-motion memory consistency network for video anomaly detection', in *AAAI*, pp. 938–946, (2021).

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar, 'Anomaly detection: A survey', *ACM Comput. Surv.*, **41**(3), 1–58, (2009).

[4] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan, 'Clustering driven deep autoencoder for video anomaly detection', in *ECCV*, pp. 329–345, (2020).

[5] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma, 'Comprehensive regularization in a bi-directional predictive network for video anomaly detection', in *AAAI*, pp. 230–238, (2022).

[6] Allison Del Giorno, J. Andrew Bagnell, and Martial Hebert, 'A discriminative framework for anomaly detection in large videos', in *ECCV*, pp. 334–349, (2016).

[7] Jiachang Feng, Fating Hong, and Wei-Shi Zheng, 'Mist: Multiple instance self-training framework for video anomaly detection', in *CVPR*, pp. 14009–14018, (2021).

[8] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen, 'Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection', in *ACM MM*, pp. 5546–5554, (2021).

[9] Mariana Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah, 'Anomaly detection in video via self-supervised and multi-task learning', in *CVPR*, pp. 12742–12752, (2021).

[10] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah, 'A background-agnostic framework with adversarial training for abnormal event detection in video', *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**(9), 4505–4523, (2021).

[11] Izhak Golan and Ran El-Yaniv, 'Deep anomaly detection using geometric transformations', in *NeurIPS*, pp. 1–12, (2018).

[12] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, 'Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection', in *CVPR*, pp. 1705–1714, (2019).

[13] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao, 'Object-centric auto-encoders and dummy anomalies for abnormal event detection in video', in *CVPR*, pp. 7842–7851, (2019).

[14] Dahun Kim, Donghyeon Cho, and In So Kweon, 'Self-supervised video representation learning with space-time cubic puzzles', in *AAAI*, pp. 8545–8552, (2019).

[15] Hsin Ying Lee, Jia Bin Huang, Maneesh Singh, and Ming Hsuan Yang, 'Unsupervised representation learning by sorting sequences', in *CVPR*, pp. 667–676, (2017).

[16] Sangmin Lee, Hak Gu Kim, and Yong Man Ro, 'Bman: bidirectional multi-scale aggregation networks for abnormal event detection', *IEEE Trans. Image Process.*, **29**(1), 2395–2408, (2019).

[17] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, 'Future frame prediction for anomaly detection–a new baseline', in *CVPR*, pp. 6536–6545, (2018).

[18] Yusha Liu, Chun-Liang Li, and Barnabás Póczos, 'Classifier two sample test for video anomaly detections.', in *BMVC*, p. 71, (2018).

[19] Zhian 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', in *ICCV*, pp. 13588–13597, (2021).

[20] Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua, 'Weakly supervised temporal action localization through contrast based evaluation networks', in *ICCV*, pp. 3899–3908, (2019).

[21] Cewu Lu, Jianping Shi, and Jiaya Jia, 'Abnormal event detection at 150 fps in matlab', in *ICCV*, pp. 2720–2727, (2013).

[22] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang, 'Few-shot scene-adaptive anomaly detection', in *ECCV*, pp. 125–141, (2020).

[23] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang, 'Learning normal dynamics in videos with meta prototype network', in *CVPR*, pp. 15425–15434, (2021).

[24] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, 'Anomaly detection in crowded scenes', in *CVPR*, pp. 1975–1981, (2010).

[25] Trong-Nguyen Nguyen and Jean Meunier, 'Anomaly detection in video sequence with appearance-motion correspondence', in *ICCV*, pp. 1273–1283, (2019).

[26] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai, 'Self-trained deep ordinal regression for end-to-end video anomaly detection', in *CVPR*, pp. 12173–12182, (2020).

[27] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, 'Learning memory-guided normality for anomaly detection', in *CVPR*, pp. 14372–14381, (2020).

[28] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe, 'Curriculum learning: A survey', *Int. J. Comput. Vis.*, **130**(6), 1526–1565, (2022).

[29] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu, 'Scene-aware context reasoning for unsupervised abnormal event detection in videos', in *ACM MM*, pp. 184–192, (2020).

[30] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu, 'Unmasking the abnormal events in video', in *ICCV*, pp. 2895–2903, (2017).

[31] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang, 'Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles', in *ECCV*, pp. 494–511, (2022).

[32] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft, 'Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network', in *NeurIPS*, pp. 1–14, (2019).

[33] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi, 'Robust unsupervised video anomaly detection by multipath frame prediction', *IEEE Trans. Neural Netw. Learn. Syst.*, **33**(6), 2301–2312, (2021).

[34] Ziming Wang, Yuexian Zou, and Zeming Zhang, 'Cluster attention contrast for video anomaly detection', in *ACM MM*, pp. 2463–2471, (2020).

[35] Peng Wu, Jing Liu, and Fang Shen, 'A deep one-class neural network for anomalous event detection in complex scenes', *IEEE Trans. Neural Netw. Learn. Syst.*, **31**(7), 2609–2622, (2019).

[36] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao, 'Anopcn: Video anomaly detection via deep predictive coding network', in *ACM MM*, pp. 1805–1813, (2019).

[37] Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, and Chengkun Wu, 'Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement', in *CVPR*, pp. 13987–13998, (2022).

[38] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft, 'Cloze test helps: Effective video anomaly detection via learning to complete video events', in *ACM MM*, pp. 583–591, (2020).

[39] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee, 'Generative cooperative learning for unsupervised video anomaly detection', in *CVPR*, pp. 14744–14754, (2022).

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

[41] Joey Tianyi Zhou, Le Zhang, Zhiwen Fang, Jiawei Du, Xi Peng, and Yang Xiao, 'Attention-driven loss for anomaly detection in video surveillance', *IEEE Trans. Circuits Syst. Video Technol.*, **30**(12), 4639–4647, (2019).