Scale-Aware Spatio-Temporal Relation Learning for Video Anomaly Detection

Guoqiu Li[1], Guanxiong Cai[2], Xingyu Zeng[2](✉), and Rui Zhao[2,3]

[1] Tsinghua University, Beijing, China
[2] SenseTime Research, Shanghai, China
zengxingyu@sensetime.com
[3] Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China

Abstract. Recent progress in video anomaly detection (VAD) has shown that feature discrimination is the key to effectively distinguishing anomalies from normal events. We observe that many anomalous events occur in limited local regions, and the severe background noise increases the difficulty of feature learning. In this paper, we propose a scale-aware weakly supervised learning approach to capture local and salient anomalous patterns from the background, using only coarse video-level labels as supervision. We achieve this by segmenting frames into non-overlapping patches and then capturing inconsistencies among different regions through our patch spatial relation (PSR) module, which consists of self-attention mechanisms and dilated convolutions. To address the scale variation of anomalies and enhance the robustness of our method, a multi-scale patch aggregation method is further introduced to enable local-to-global spatial perception by merging features of patches with different scales. Considering the importance of temporal cues, we extend the relation modeling from the spatial domain to the spatio-temporal domain with the help of the existing video temporal relation network to effectively encode the spatio-temporal dynamics in the video. Experimental results show that our proposed method achieves new state-of-the-art performance on UCF-Crime and ShanghaiTech benchmarks. Code are available at https://github.com/nutuniv/SSRL.

Keywords: Scale-aware · Weakly-supervised video anomaly detection · Spatio-temporal relation modeling

1 Introduction

Video Anomaly Detection (VAD) aims to automatically recognize events that deviate from normal patterns and determine the time window in which anomalies occurred [4,6,28,30]. It is invaluable in many practical applications, such as monitoring terrorist and violent events in public places, or traffic accidents on urban roads, significantly reducing the labor costs of manual surveillance.

Most previous VAD approaches [6,28,30,32] would employ a feature encoder to extract features of video frames and then identify unusual patterns as

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13664, pp. 333–350, 2022.
https://doi.org/10.1007/978-3-031-19772-7_20

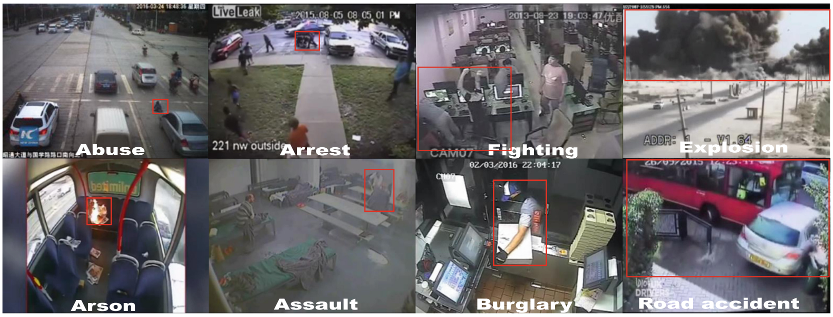

Fig. 1. Samples of video anomaly detection benchmark UCF-Crime [28]. The red boxes denote the anomalous regions in frames.

anomalies. However, we found that the effect of the spatial size of anomalies was overlooked. As illustrated in Fig. 1, in the video anomaly detection datasets, many anomalous actions such as abuse and arrest occur in small areas and are difficult to be distinguished from background normal behaviors in the frame. Therefore, the extracted features of full-resolution frames, with limited local anomaly regions inside, will be dominated by the background information, increasing the recognition difficulty of subsequent classifiers.

To address the above limitation, we propose a scale-aware video anomaly detection model to efficiently capture local anomalies from the background. Specifically, we divide the input video frames into a set of non-overlapping patches by using a sliding window. Once anomalies occur, the corresponding anomaly patches will contain more salient anomalous information due to the restricted receptive field, thus suppressing the background noise. Since the patterns of anomaly patches are likely to be distinct from normal patches, we propose a patch spatial relation (PSR) module to identify the occurrence of anomalous events by capturing inconsistencies among different spatial regions.

We also observe that anomalous events vary in size, which poses a challenge to the robustness of the patch-based methods. In Fig. 1, we can see that some anomalous events such as assault or arson occur in small regions, while others, such as road accidents or explosions, span almost the entire image. To cope with this scale variation issue, we further propose a multi-scale patch aggregation (MPA) method to effectively explore anomalous regions with different scales. Since the single-scale patches are likely to suffer from size mismatch when capturing anomalous events, we gradually adjust the size of the sliding window from small to large to generate patches with a pyramidal distribution of scales. The information of these patches will eventually be integrated to give the model a pyramid-like spatial perception of the video from local to global. This improves the scale robustness in anomaly detection.

Previous studies [30,38,49] have demonstrated the significance of long-range temporal dependencies in VAD. Inspired by them, we introduce an existing video temporal relation (VTR) module [30] to extend the relation modeling from the spatial domain to the spatio-temporal domain to effectively encode the spatio-temporal dynamics in videos. The combination of our PSR module and VTR module explicitly considers the local consistency at frame level and global coherence of temporal dynamics in video sequences.

Despite the significance of spatio-temporal feature learning, the corresponding spatio-temporal level annotations are costly. For example, Liu et al. [11] have driven the network to focus on anomalous regions by using a large number of manually labeled spatio-temporal annotations as supervision. To reduce the annotation cost, we follow Sultani et al. [28] to address the VAD task in a weakly supervised multiple instance learning setting by using training samples annotated with normal or abnormal video-level labels.

Our main contributions can be summarized as follows:

- We propose a novel scale-aware weakly supervised video anomaly detection framework, which enables local-to-global spatio-temporal perception for capturing anomalous events of various scales.
- We present a multi-scale patch aggregation method to further boost the detection performance by integrating information from various scale patches.
- We introduce a separable spatio-temporal relation network. Our PSR module learns the spatial relationships among patches and the VTR module captures the temporal dependencies in the video.
- We carry out experiments on UCF-Crime and ShanghaiTech datasets to verify the effectiveness of our method, and the experimental results show that our approach achieves significant performance boosts.

2 Related Work

Weakly Supervised Video Anomaly Detection. Relevant studies on VAD can be broadly classified into two categories: unsupervised learning methods [1,5,6,12,14–19,22,24,26,29,33,34,36,40,43,44,50] and weakly supervised learning methods [4,20,28,30,32,37–39,45,46,49,51]. Most unsupervised approaches learn the usual patterns from normal videos and then identify detection targets with large prediction errors [12,16,26,43,47,50] or reconstruction errors [3,6,17–19,24,40,44] as anomalies.

Recently, many weakly supervised approaches have been developed. Sultani et al. [28] proposed a deep multiple instance learning (MIL) ranking framework with cheap video-level annotations to detect anomalies. Zhang et al. [46] further introduced an inner bag loss constraint. Zhong et al. [49] formulated weakly supervised anomaly detection as a label noise learning problem, and used a graph convolution neural network to filter the label noise. Wu et al. [38] explored the importance of temporal cues and feature discrimination. Tian et al. [30] proposed a feature magnitude learning approach of selecting top k snippets with the largest feature magnitude as a stronger learning signal. Feng et al. [4] proposed a two-stage scheme in which a MIL-based pseudo label generator was first trained to produce snippet-level pseudo labels, which were then used to finetune a task-specific feature encoder for VAD. Liu et al. [11] re-annotated the

UCF-Crime dataset, adding fully supervised anomaly location annotations to drive the model to focus on anomalous regions. Compared with previous weakly supervised methods, our approach explores the local salience of anomalous events and identifies anomalies by capturing the inconsistencies among patches, which is relatively deficient in the VAD area. Moreover, our method addresses the problem of scale variation by aggregating features of patches with different scales.

Spatio-Temporal Relation Modeling. Recently, spatio-temporal relation learning has been successfully applied in several fields, such as object detection [7], action recognition [13,21,27] and object tracking [41]. In anomaly detection, Zhao et al. [48] proposed a spatio-temporal autoencoder to learn video representation by performing 3-dimensional convolutions. Wu et al. [37] introduced a new task to localize the spatio-temporal tube of anomalous event, they used the Faster-RCNN algorithm [25] to extract tube-level instance proposals, and then adopted the multi-head self-attention method [31] to capture the relationships between video objects. However, the pre-trained object detector cannot recognize objects of unseen categories. Our method uses a separable spatio-temporal relation network to effectively capture the inconsistencies among different regions in frames and the long-range temporal dynamics in video sequences.

3 Methodology

In this section, we first present the overall pipeline of our proposed scale-aware spatio-temporal relation learning (SSRL) method in Sect. 3.1. Then we describe the patch spatial relation module in Sect. 3.2. Section 3.3 introduces the multi-scale patch aggregation method, and the video temporal relation module is described in Sect. 3.4. Finally, we introduce the loss function in Sect. 3.5.

3.1 Overview

The overall pipeline of our SSRL is shown in Fig. 2. Given an input untrimmed video V, the corresponding weak video-level label $y \in \{0, 1\}$ indicates whether abnormal events exist in this video ($y = 1$ if there exist anomalous events in video V and $y = 0$ otherwise). Following the previous MIL-based frameworks [4,20,28,30,49,51], we divide the video V into a sequence of temporal non-overlapping video snippets $\{\mathbf{v}_t\}_{t=1}^{T}$, here we use T to denote the number of video snippets.

We can see that given a video snippet $\mathbf{v}_t \in \mathbb{R}^{H\times W\times L\times 3}$, where H and W are the height and width of the video snippet, respectively, and L denotes the temporal length of the video snippet. The video snippet $\mathbf{v}_t$ will first be split into several sets of spatial non-overlapping patch cubes with different spatial sizes. As Fig. 3 shows, we set a number of sliding window sizes $\{(h_r, w_r)\}_{r=1}^{R}$ to extract patch cubes. Thus, each set of patch cubes is represented as $\mathcal{P}_t^r = \{\mathbf{p}_{t,i}^r\}_{i=1}^{N_r}$, $r \in \{1, \ldots, R\}$, where $\mathbf{p}_{t,i}^r \in \mathbb{R}^{h_r\times w_r\times L\times 3}$ denotes the extracted patch cube with spatial patch size of $h_r \times w_r$, and $N_r = \lfloor H/h_r \rfloor \times \lfloor W/w_r \rfloor$ is the number of patch cubes. Then every patch cube is fed into a pretrained feature extractor (I3D [2]) to generate features, and then features of patch cubes with the

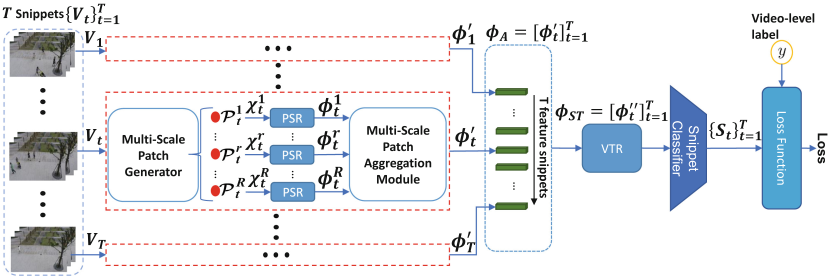
T Snippets$\{V_t\}_{t=1}^T$
V_1
V_t
V_T
Multi-Scale Patch Generator
$\mathcal{P}_t^1$ χ_t^1
$\mathcal{P}_t^r$ χ_t^r
$\mathcal{P}_t^R$ χ_t^R
PSR
PSR
PSR
ϕ_t^1
ϕ_t^r
ϕ_t^R
Multi-Scale Patch Aggregation Module
ϕ'_1
ϕ'_t
ϕ'_T
$\phi_A = [\phi'_t]_{t=1}^T$
T feature snippets
$\phi_{ST} = [\phi''_t]_{t=1}^T$
VTR
Snippet Classifier
$\{S_t\}_{t=1}^T$
Video-level label
y
Loss Function
Loss

Fig. 2. The overall pipeline of our proposed SSRL. Each video snippet is first divided into several sets of patch cubes. Then, our PSR modules learn the spatial relations among patch cubes with the same size, and the multi-scale patch aggregation method further fuses the features of multi-scale patch cubes. After that, the VTR module captures the temporal dependencies among video snippets, and the snippet classifier will predict the snippet scores. Finally, the video-level label y and the snippet scores are used to compute the loss.

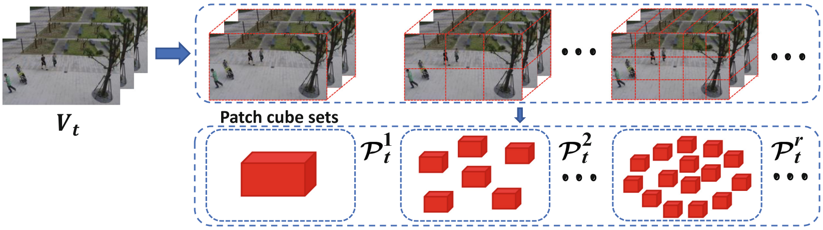

Fig. 3. Illustration of the multi-scale patch generator.

same size are stacked in horizontal dimension as $\chi_t^r \in \mathbb{R}^{N_r \times D}$, where D denotes the feature dimensions. To capture the inconsistencies between anomalous and normal patch cubes, our proposed patch spatial relation (PSR) module comes into play. It computes patch-wise correlations among patch cubes of the same size through a self-attention mechanism [35] and dilated convolutions [42] to produce spatial enhanced patch representations, denoted as $\phi_t^r \in \mathbb{R}^{N_r \times D}$. After that, we apply a multi-scale patch aggregation method to enhance the scale robustness of our model. The features of the multi-scale patch cubes will be fused to produce an aggregated snippet feature $\phi'_t \in \mathbb{R}^D$, enabling a local-to-global perception of the video snippet. Please see Sect. 3.3 for more details. After the above process, we obtain the aggregated video representation $\phi_A = [\phi'_t]_{t=1}^T$ from the T video snippets, which is then fed into a video temporal relation (VTR) module to capture the temporal dependencies among video snippets, resulting in a temporal enhanced video representation $\phi_{ST} = [\phi''_t]_{t=1}^T$, where $\phi''_t \in \mathbb{R}^D$ denotes the enhanced feature of each snippet. Finally we employ a snippet classifier [28] to generate anomaly scores $\{s_t\}_{t=1}^T$ for all video snippets.

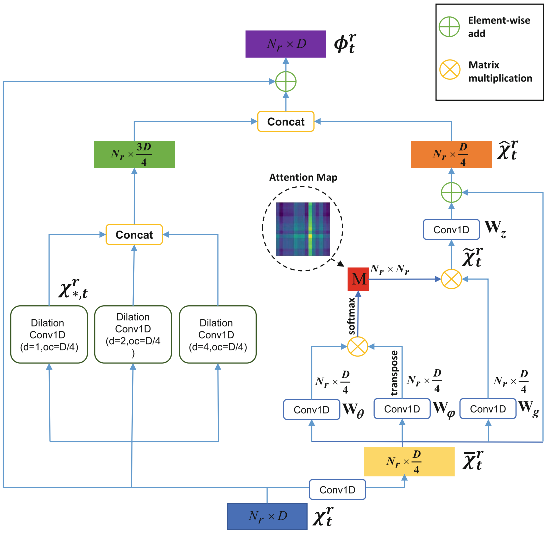

Fig. 4. Illustration of the relation network. N_r denotes the number of input patch cubes in the PSR module.

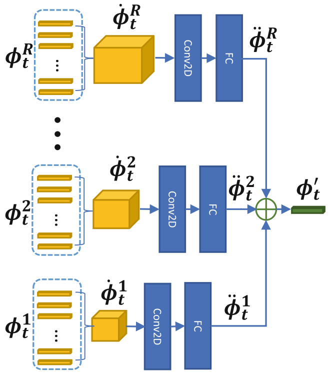

Fig. 5. Illustration of the multi-scale patch aggregation method.

3.2 Patch Spatial Relation Module

Inspired by the previous work [30], which used a combined network consisting of a pyramid of dilated convolutions [42] and a temporal self-attention module [35] to capture long- and short-range temporal dependencies between video snippets, we use the same relation network in our framework, but the difference is that we employ the relation network not only to capture the temporal dependencies among video snippets but also to learn the spatial relations among patch cubes. The detailed architecture of this relation network is illustrated in Fig. 4.

Supposing an abnormal event occurs at a limited location in a video snippet, then the local feature distribution of the corresponding anomaly patch cubes should be significantly different from other normal patch cubes. Therefore, to capture the inconsistencies among different spatial regions, we propose the patch spatial relation (PSR) module. It learns the patch-wise correlations of patch cubes across different spatial regions from the pre-extracted initial patch features $\chi_t^r \in \mathbb{R}^{N_r \times D}$, where N_r denotes the number of patch cubes extracted with a particular spatial size (h_r, w_r). The right sub-network in Fig. 4 is a non-local network that aims to model the global spatial relations among patch cubes by self-attention mechanism [35]. After we feed χ_t^r into the PSR module, the non-local network first uses a 1×1 convolution to reduce the channel dimension from $\chi_t^r \in \mathbb{R}^{N_r \times D}$ to $\bar{\chi}_t^r \in \mathbb{R}^{N_r \times D/4}$, then a non-local operation is performed to model global spatial relations among patch cubes:

$$\tilde{\chi}_t^r = \text{softmax}\left(\mathbf{W}_\theta \bar{\chi}_t^r \bar{\chi}_t^{r\top} \mathbf{W}_\varphi^\top\right) \mathbf{W}_g \bar{\chi}_t^r, \qquad (1)$$

$$\hat{\chi}_t^r = \mathbf{W}_z \tilde{\chi}_t^r + \bar{\chi}_t^r, \tag{2}$$

where $\bar{\chi}_t^r$ is first projected into the embedded space by three 1×1 convolutions with learnable weights $\mathbf{W}_\theta$, $\mathbf{W}_\varphi$ and $\mathbf{W}_g$. We then calculate the attention map $\mathbf{M} \in \mathbb{R}^{N_r \times N_r}$ by the dot product operation and softmax normalization, the attention map $\mathbf{M}$ is used as weights to compute the weighted sum $\tilde{\chi}_t^r$. Once the network captures an anomalous patch, its corresponding column in $\mathbf{M}$ is expected to be highlighted, thus passing the information of this anomalous patch to all patches. Our expectations coincide with the results of the later experiments (Fig. 6). Then we obtain the self-attention based representation $\hat{\chi}_t^r \in \mathbb{R}^{N \times D/4}$ by a 1×1 convolution with learnable weights $\mathbf{W}_z$ and a residual connection, shown in Eq. 2.

The left sub-network in Fig. 4 contains a pyramid of dilated convolutions to learn the local spatial dependencies of neighbouring patch cubes with multi-scale receptive fields. Specifically, we set up three 1-D dilated convolutions with different dilation factors $d \in \{1, 2, 4\}$. The input features χ_t^r will be simultaneously fed into three dilated convolutions to produce multi-scale dilation embedded representations $\chi_{*,t}^r \in \mathbb{R}^{N \times D/4}$, $* \in \{DC1, DC2, DC3\}$. Then, a concatenation operation and residual connection are applied to the outputs of two sub-networks to produce the spatial enhanced patch representations:

$$\phi_t^r = [\hat{\chi}_t^r, \chi_{*,t}^r] + \chi_t^r, \tag{3}$$

where [.] denotes the concatenation operation.

3.3 Multi-scale Patch Aggregation

In the video, the spatial scales of different anomalous objects vary greatly. If we directly split the input video snippet into multiple patch cubes with a single fixed spatial size, it is likely that a large anomaly object cannot be completely divided into a single patch, while a small anomaly object still occupies only a small part of the patch region. Therefore, we propose a multi-scale patch aggregation (MPA) method to deal with this size mismatch case.

As mentioned in Sect. 3.1, we will first use sliding windows of different sizes to split the input video snippet into several sets of non-overlapping patch cubes to cover anomaly objects with different sizes. These sets of patch cubes will then be passed through the I3D feature extractor and PSR modules in parallel to obtain the spatial enhanced patch representations $\{\phi_t^r\}_{r=1}^R$, $\phi_t^r \in \mathbb{R}^{N_r \times D}$. After that, the MPA method comes into play. As Fig. 5 shows, each input patch representation ϕ_t^r will first be reconstructed into a 3-D feature vector $\dot{\phi}_t^r \in \mathbb{R}^{\lfloor H/h_r \rfloor \times \lfloor W/w_r \rfloor \times D}$ according to the initial spatial location of patch cubes. Subsequent convolutional and fully connected layers transform this 3-D feature vector into a 1-D feature vector $\ddot{\phi}_t^r \in \mathbb{R}^D$. After the above steps, the local information of the same scale patches is aggregated. Finally, the multi-scale fused patch features $\{\ddot{\phi}_t^r\}_{r=1}^R$ will be aggregated together by an element-wise add operation, resulting in an aggregated snippet feature $\phi_t' = \sum_{r=1}^R \ddot{\phi}_t^r$.

3.4 Video Temporal Relation Module

In an anomalous video, it is most likely that the motion patterns of anomalous video snippets would not follow the patterns of other normal video snippets. The existing video temporal relation (VTR) module aims to learn the temporal context of video snippets by applying the relation network in Fig. 4 over the time dimension. We formulate the VTR module as follows:

$$\tilde{\phi}_A = \mathrm{softmax}\left(\mathbf{W}_\theta \bar{\phi}_A \bar{\phi}_A^\top \mathbf{W}_\varphi^\top\right) \mathbf{W}_g \bar{\phi}_A, \tag{4}$$

$$\hat{\phi}_A = \mathbf{W}_z \tilde{\phi}_A + \bar{\phi}_A, \tag{5}$$

$$\phi_{ST} = [\hat{\phi}_A, \phi_{*,A}] + \phi_A, \tag{6}$$

where $\phi_A = [\phi'_1, \phi'_2, \ldots, \phi'_T] \in \mathbb{R}^{T\times D}$ represents the input aggregated features of the T video snippets, which are then fed into the 1×1 convolution layer to produce $\bar{\phi}_A \in \mathbb{R}^{T\times D/4}$. $\phi_{*,A}$ denotes the outputs of three dilated convolution layers and $\phi_{ST} \in \mathbb{R}^{T\times D}$ denotes the output spatio-temporal enhanced video representation.

3.5 Loss Function

We chose the multiple instance learning (MIL) method for weakly supervised learning. In addition, to further improve the robustness of our SSRL in detecting anomaly events, we draw on the feature magnitude learning method presented in [30], which enables better separation between anomaly and normal videos by selecting the top k snippets with largest feature magnitudes instead of the snippet with the highest anomaly score to supervise MIL model. Specifically, for the output spatio-temporal video representation $\phi_{ST} = [\phi''_t]_{t=1}^T$, where $\phi''_t \in \mathbb{R}^D$ denotes the snippet feature, the mean feature magnitude is defined by:

$$g(\phi_{ST}) = \max_{\Omega_k(\phi_{ST}) \subseteq \{\phi''_t\}_{t=1}^T} \frac{1}{k} \sum_{\phi''_t \in \Omega_k(\phi_{ST})} \|\phi''_t\|_2 \,, \tag{7}$$

where $\Omega_k(\phi_{ST})$ contains k snippets selected from $\{\phi''_t\}_{t=1}^T$, the snippet feature magnitude is computed by ℓ_2 norm. After that, the feature magnitude based MIL ranking loss is formulated by:

$$\mathcal{L}_{FM} = \max\left(0, \epsilon - g(\phi_{ST}^+) + g(\phi_{ST}^-)\right), \tag{8}$$

where ϵ is a pre-defined margin, ϕ_{ST}^+ and ϕ_{ST}^- denote the anomaly and normal video representations, respectively.

We feed the top k selected snippets with largest feature magnitudes into the snippet classifier to generate the corresponding snippet anomaly scores $\{s_j\}_{j=1}^k$. Then, we apply a cross-entropy-base loss function to train the snippet classifier:

$$\mathcal{L}_{CE} = \sum_{s \in \{s_j\}_{j=1}^k} -(y \log(s) + (1-y)\log(1-s)). \tag{9}$$

Following the previous work [28], we add the sparsity and temporal smoothness constraints on all predicted snippet scores$\{s_t^+\}_{t=1}^T$ of the anomaly video. To sum up, the total loss function of our model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{fm}\mathcal{L}_{FM} + \lambda_1 \sum_{t=1}^{T} |s_t^+| + \lambda_2 \sum_{t=1}^{T} (s_t^+ - s_{t-1}^+)^2, \tag{10}$$

where λ_{fm}, λ_1 and λ_2 are weighting factors used to balance the losses of each component. $\sum_{t=1}^{T} |s_t^+|$ and $\sum_{t=1}^{T} (s_t^+ - s_{t-1}^+)^2$ denote the sparsity regularization and temporal smoothness constraint, respectively.

4 Experiments

4.1 Datasets and Metrics

We validated our SSRL on two large benchmark datasets for video anomaly detection, namely UCF-Crime [28] and ShanghaiTech [18].

UCF-Crime: UCF-Crime [28] is a large-scale dataset. It has a total of 128 h with 1900 long untrimmed videos. All videos were captured from real-world surveillance, including 13 types of anomalous events that have a significant impact on public safety. The training set consists of 1610 videos, and the testing set contains 290 videos. Both training and testing sets contain all 13 anomalies at various temporal locations in the videos.

ShanghaiTech: ShanghaiTech dataset has 437 videos, including 307 normal videos and 130 anomaly videos, all collected under 13 different scenes with complex shooting angles. However, since the original dataset [18] was proposed for semi-supervised anomaly detection, only normal videos were available in the training set. Following Zhong et al. [49], we reorganize the videos into 238 training videos and 199 testing videos, making both the training and testing sets contain anomalous videos, thus adapting to the weakly supervised setting.

Evaluation Metrics. Following previous works [4,28,30], we compute the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) as the main metric to evaluate the performance of our model and comparison methods, where a larger AUC implies higher distinguishing ability.

4.2 Implementation Details

Following previous works [28,30], for a training video, we first split the video into 32 non-overlapping video snippets ($T = 32$). Then we resize each video snippet to $480 \times 840 \times 16$ pixels ($H = 480$, $W = 840$, $L = 16$). To extract multi-scale patch cubes from the video snippet, we set four different sliding window sizes: $480 \times 840, 240 \times 280, 160 \times 168$ and 120×120, where 480×840 means that we treat the entire video frame as a patch. After that, we deploy the I3D network pretrained on Kinetic-400 dataset [8] to extract initial features. For the relation network described in Fig. 4, we use the same setting as [30]. During the multi-scale patch aggregation process, for each branch, we employ a 2-D convolutional layer with 3 × 3 kernel and 2 × 2 stride and a fully connected layer with 2048 output nodes. The snippet classifier consists of three fully connected layers with output nodes of 512, 128 and 1. For hyper-parameters, we set the margin $\epsilon = 100$ in (8) and the number of selected snippets $k = 3$ in (7), and the weighting factors λ_{fm}, λ_1 and λ_2 in (10) are set to 0.0001, 0.008 and 0.0008, respectively. All hyper-parameters are the same for both UCF-Crime and ShanghaiTech.

Training. We train our network on 8 NVIDIA Tesla V100 GPUs using PyTorch [23]. We randomly sample 32 abnormal videos and 32 normal videos per batch and use the Adam optimizer [9] with the initial learning rate of 0.001 and a weight decay of 0.0005 to train our SSRL. For the MPA module, the fusion process of multi-scale patches is difficult to optimize. To reduce the difficulty of optimization, we adopt a step-by-step training strategy. We first optimize the process of single-scale patches, and then gradually introduce new-scale patches. More details about implementation are reported in Supplementary Material.

4.3 Comparisons with Related Methods

The AUC results on two benchmarks are presented in Table 1. For the UCF-Crime dataset, our method achieves the highest AUC result of 87.43%. Compared with the existing unsupervised methods [33,34], our SSRL outperforms BODS [33] by 19.17% and GODS [34] by 16.97%. Our SSRL also surpasses existing weakly supervised methods [4,20,28,30,38,39,45,46,49,51]. In particular, when using the same I3D-RGB initial features, our model exceeds Wu et al. [39] by 4.99%, MIST [4] by 5.13%, RTFM [30] by 3.13%, Wu et al. [38] by 2.54%, MSL [10] by 2.13% and WSAL [20] by 2.05%. For the ShanghaiTech dataset, as the table indicates, the detection performance of our SSRL outperforms all previous weakly supervised methods. It is worth noting that among other models that also use I3D-RGB features, the previous best method [38] achieved an AUC result of 97.48%, which is already a fairly high result considering that only video-level anomaly labels are provided, but our method still improves further on this to 97.98%, which proves the powerful anomaly detection capability of our SSRL. We further report the result when the parameters of different PSRs are shared, with an AUC of 86.85% and 97.84% on the two datasets, respectively. Although the performance is slightly dropped, the parameter amount is much reduced (Table 6), which is more favorable for practical applications.

We also compare our SSRL with other spatio-temporal relation modeling methods on UCF-Crime. As shown in the Table 2, our approach significantly exceeds STC-Graph [29] by 14.73% and STAD [37] by 4.7%. Video swin transformer [13] is a transformer-based backbone architecture that has recently achieved strong performance on a broad range of video-based recognition tasks. We implemented it in VAD field, specifically, we used video swin transformer (tiny version, due to memory limitations) as backbone, and performed multi-instance weakly supervised learning using the classifier and loss function in this

Table 1. Quantitative comparisons with other state-of-the-art methods on UCF-Crime and ShanghaiTech. Share parameters denotes different PSRs' parameters are shared.

Method	Supervised	Feature	AUC (%)	
			UCF-Crime	ShanghaiTech
BODS [33]	Un	I3D-RGB	68.26	–
GODS [34]	Un	I3D-RGB	70.46	–
Sultani et al. [28]	Weak	C3D-RGB	75.41	86.30
Zhang et al. [46]	Weak	C3D-RGB	78.66	82.50
Motion-Aware [51]	Weak	PWC-Flow	79.00	–
Zhong et al. [49]	Weak	TSN-RGB	82.12	84.44
Wu et al. [39]	Weak	I3D-RGB	82.44	–
MIST [4]	Weak	I3D-RGB	82.30	94.83
CLAWS [45]	Weak	C3D-RGB	83.03	89.67
RTFM [30]	Weak	I3D-RGB	84.30	97.21
Wu et al. [38]	Weak	I3D-RGB	84.89	97.48
WSAL [20]	Weak	I3D-RGB	85.38	–
MSL [10]	Weak	I3D-RGB	85.30	96.08
MSL [10]	Weak	VideoSwin-RGB	85.62	97.32
Our SSRL (share parameters)	Weak	I3D-RGB	86.85	97.84
Our SSRL	Weak	I3D-RGB	**87.43**	**97.98**

Table 2. Quantitative comparisons with other spatio-temporal relation modeling methods on UCF-Crime. * indicates the result implemented by us.

Method	Supervised	AUC (%) - UCF
STC-Graph [29]	Un	72.70
STAD [37]	Weak	82.73
Video swin transformer [13]	Weak	81.62*
Our SSRL	Weak	**87.43**

paper, with the rest of the setup as in [13]. The test results on UCF-Crime show that our method outperforms video swin transformer by 5.81%, which may be due to the lack of long-range temporal dependencies in video swin transformer.

4.4 Ablation Study

In this section, we conduct ablation studies to study the impact of important designed elements in our SSRL.

Analysis of Multi-scale Patch Aggregation. To investigate the influence of our proposed multi-scale patch aggregation method, we conduct ablation studies on the UCF-Crime and ShanghaiTech datasets. The detailed comparison results are shown in Table 3. Specifically, we employ RTFM [30] as our baseline, which

Table 3. Ablation studies on the multi-scale patch aggregation method (see Sect. 3.3) on two benchmarks. * indicates we use the method in [30] as our baseline.

Patch size				AUC (%)	
480 × 840	240 × 280	160 × 168	120 × 120	UCF-Crime	ShanghaiTech
✓				84.30^{*}	97.21^{*}
✓	✓			86.38	97.60
✓		✓		85.69	97.69
✓			✓	85.29	97.55
✓	✓	✓		86.70	97.85
✓		✓	✓	86.32	97.77
✓	✓		✓	86.97	97.71
✓	✓	✓	✓	**87.43**	**97.98**

Table 4. Ablation studies on two benchmarks for investigating the effect of spatio-temporal relation learning. Baseline is [30] trained with video temporal relation network. SSRL is our whole model. SSRL$^{\text{w/o PSR}}$ is SSRL trained without patch spatial relation module but with multi-scale patch aggregation module. For the MPA module, we use all four patch sizes in Table 3.

Methods	VTR	MPA	PSR	AUC (%)	
				UCF-Crime	ShanghaiTech
Baseline	✓			84.30	97.21
SSRL$^{\text{w/o PSR}}$	✓	✓		85.98	97.45
SSRL	✓	✓	✓	**87.43**	**97.98**

performs anomaly detection directly on the input video snippet and achieves 84.30% and 97.21% AUC results on UCF-Crime and ShanghaiTech, respectively. Then, we set up three different patch sizes: 240 × 280, 160 × 168 and 120 × 120 for capturing anomalous events on the corresponding spatial scales. When we extract patch cubes using only one of the sizes, the corresponding experimental results are shown in the second to fourth rows of Table 3. We observe 0.99% to 2.08% and 0.34% to 0.48% improvement on the two datasets respectively. This verifies the effectiveness of patch-based feature learning. After that, we start integrating multiple sizes of patch cubes to capture anomaly events of different sizes. We report the corresponding experimental results in the fifth to the last row of the table. We observe the AUC results increase gradually as more sizes of patch cubes are introduced. The AUC increases from 86.32% to 87.43% on the UCF-Crime dataset and improves from 97.71% to 97.98% on the ShanghaiTech dataset. Our MPA module fuses patch features from multiple branches by an element-wise add operation, so that anomalies will be identified if they are captured by any of the branches. The above observations reveal the complementary effect among the patch information at multiple scales.

Analysis of Spatio-Temporal Relation Learning. The results in Table 4 verify the effect of spatio-temporal relation learning. Compared with the baseline that only considers the temporal context in the video and treats the video frame as a whole, SSRL$^{\text{w/o PSR}}$ achieves a significant improvement when the MPA-enabled spatial local patterns are utilized. In particular, we observe 1.68% and 0.24% improvement in AUC on two benchmarks, respectively, which shows that spatio-temporal features are more discriminatory than simple temporal features. Moreover, the PSR module also plays an important role in spatio-temporal feature learning by capturing the inconsistencies among different patches. Compared with SSRL$^{\text{w/o PSR}}$, with the help of the PSR module, SSRL further increases 1.45% AUC on the UCF-Crime and 0.53% AUC on the ShanghaiTech.

Analysis on Two Sub-networks in the PSR Module. The results in Table 5 verify the effect of two sub-networks in PSR. We employ the pyramid of dilated convolutions to learn the local spatial dependencies of neighboring patches with multi-scale receptive field, and it brings 0.24% and 0.18% AUC gains on UCF-Crime and ShanghaiTech, respectively. We also use the no-local network to capture the global spatial relations of different patches, and it improves the AUC by 0.69% and 0.37% on UCF-Crime and ShanghaiTech, respectively. When both sub-networks are added, we observe an increase in AUC of 1.45% and 0.53% on UCF-Crime and ShanghaiTech, respectively. This indicates that two

Table 5. Ablation studies on two benchmarks for verifying the effectiveness of sub-networks in the PSR module.

Methods	AUC (%)	
	UCF-Crime	ShanghaiTech
Full	87.43	97.98
w/o dilated convolutions	86.57	97.82
w/o no-local network	86.22	97.63
w/o PSR module	85.98	97.45

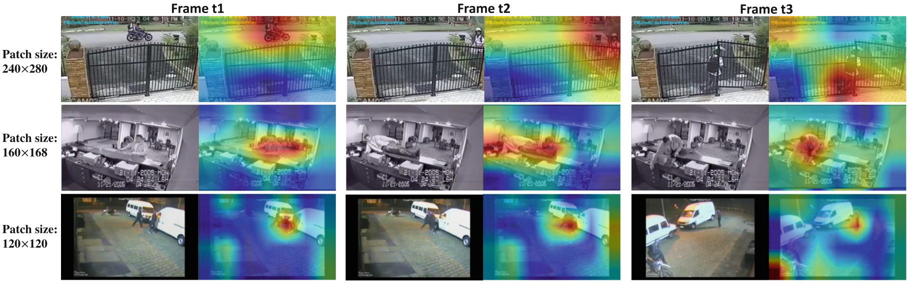

Fig. 6. Visualization results of attention heatmaps with three patch sizes on UCF-Crime (*Stealing079*, *Burglary017*, *Arson010*, from top to bottom) test videos.

sub-networks complement each other in capturing both local and global range spatial dependencies, and contribute to the overall performance.

4.5 Visual Results

To further evaluate the performance of our method, we visualize the attention map $\mathbf{M} = [w_{m,n}]_{m=1,n=1}^{N_r,N_r}$ in the PSR module. As we mentioned in Sect. 3.2, the columns corresponding to anomalous patches are expected to be highlighted in $\mathbf{M}$. We accumulate the weights $[w_{m,n}]_{m=1,n=1}^{N_r,N_r}$ row by row to generate a 1-D attention vector $[w'_n]_{n=1}^{N_r}$, $w'_n = \sum_{m=1}^{N_r} w_{m,n}$, which is normalized and then rearranged into a 2-D attention mask $\mathbf{M}' \in \mathbb{R}^{\lfloor H/h_r \rfloor \times \lfloor W/w_r \rfloor}$ according to the initial spatial location of the patches. We then use $\mathbf{M}'$ to generate the attention heatmap for spatial explanation. As Fig. 6 shows, our PSR module is able to sensitively focus on salient anomalous regions and suppress the background, even if the anomalous objects keep moving over time. In addition, the three rows in Fig. 6 correspond to three different patch sizes, and from top to bottom, we can see that as the patch size decreases, the focus of attention becomes more concentrated. This scale-aware pyramidal attentional vision can effectively improve the scale robustness of detecting anomalies. We also compare the predicted anomaly scores of our SSRL and the baseline [30] in Fig. 7. Two anomalous videos (*Shooting022*, *01_0141*) and one normal video (*Normal210*) are used. As we can see, compared with the baseline, our SSRL can effectively detect small anomalous events (e.g., *Shooting022* and *01_0141*). Moreover, our SSRL produces much less

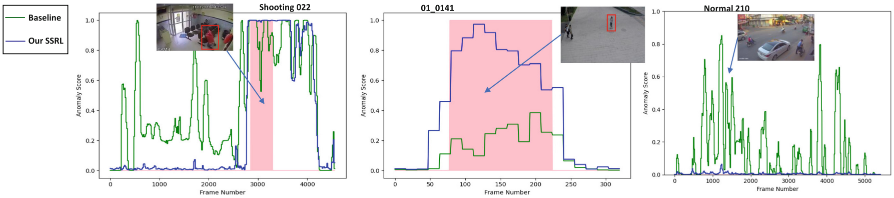

Fig. 7. Visualization of the anomaly scores of our SSRL and the baseline [30] on UCF-Crime (*Shooting022*, *Normal210*), and ShanghaiTech (*01_0141*) test videos. Pink areas are temporal ground truths of anomalies. The red boxes denote anomalous regions. (Color figure online)

Table 6. Computational complexity comparisons with other methods.

Method	Feature encoder	Param	FLOPs
RTFM [30]	I3D	28M	186.9G
Zhong et al. [49]	C3D	78M	386.2G
Our SSRL	I3D	191M	214.6G
Our SSRL (share parameters)	I3D	136M	214.6G

false positives on normal videos (e.g., *Normal210*) and anomalous videos (e.g., *Shooting022*).

4.6 Computational Complexity

We provide detailed information of parameter amount and computational cost in Table 6, and we acknowledge that large parameter amount is a potential limitation of our approach. Since our SSRL uses the same VTR module as baseline, the extra computational cost and parameters come from the PSR and MPA modules, and we can address this limitation by sharing parameters between different PSR modules. As the Table 6 shows, with a slight decrease in performance (Table 1), the parameter amount drops by 55 megabytes, which facilitates the real-world application of our method.

5 Conclusion

In this work, we propose a scale-aware weakly supervised video anomaly detection framework that uses only video-level labeled training videos to learn to focus on locally salient anomalous regions. We adopt a separable spatio-temporal relation network which explores the spatio-temporal context in the video to generate discriminative spatio-temporal features. We also introduce a multi-scale patch aggregation method to enable the local-to-global perception in frames and to enhance the scale robustness of our model. Remarkably, our proposed method achieves significant improvements on two public benchmarks.

References

1. Cai, R., Zhang, H., Liu, W., Gao, S., Hao, Z.: Appearance-motion memory consistency network for video anomaly detection. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, pp. 938–946 (2021)
2. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4724–4733 (2017)
3. Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F.: Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Comput. Vis. Image Underst. **195**, 102920 (2020)
4. Feng, J., Hong, F., Zheng, W.: MIST: multiple instance self-training framework for video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14009–14018 (2021)
5. Georgescu, M., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12742–12752 (2021)
6. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 733–742 (2016)

7. He, L., et al.: End-to-end video object detection with spatial-temporal transformers. CoRR abs/2105.10920 (2021). https://arxiv.org/abs/2105.10920
8. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations (2015)
10. Li, S., Liu, F., Jiao, L.: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In: Thirty-Sixth AAAI Conference on Artificial Intelligence (2022)
11. Liu, K., Ma, H.: Exploring background-bias for anomaly detection in surveillance videos. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1490–1499 (2019)
12. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection - a new baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6536–6545 (2018)
13. Liu, Z., et al.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
14. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. CoRR abs/2108.06852 (2021). https://arxiv.org/abs/2108.06852
15. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in MATLAB. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2720–2727 (2013)
16. Lu, Y., Kumar, K.M., Nabavi, S.S., Wang, Y.: Future frame prediction using convolutional VRNN for anomaly detection. In: 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 1–8 (2019)
17. Luo, W., Liu, W., Gao, S.: Remembering history with convolutional LSTM for anomaly detection. In: 2017 IEEE International Conference on Multimedia and Expo, pp. 439–444 (2017)
18. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 341–349 (2017)
19. Luo, W., et al.: Video anomaly detection with sparse coding inspired deep neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **43**(3), 1070–1084 (2021)
20. Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J.: Localizing anomalies from weakly-labeled videos. IEEE Trans. Image Process. **30**, 4505–4515 (2021)
21. Pan, J., Chen, S., Shou, M.Z., Liu, Y., Shao, J., Li, H.: Actor-context-actor relation network for spatio-temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 464–474 (2021)
22. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14360–14369 (2020)
23. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32**, 8026–8037 (2019)
24. Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C.S., Sebe, N.: Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE International Conference on Image Processing, pp. 1577–1581 (2017)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28**, 91–99 (2015)

26. Rodrigues, R., Bhargava, N., Velmurugan, R., Chaudhuri, S.: Multi-timescale trajectory prediction for abnormal human activity detection. In: IEEE Winter Conference on Applications of Computer Vision, pp. 2615–2623 (2020)
27. Song, L., Zhang, S., Yu, G., Sun, H.: TACNet: transition-aware context network for spatio-temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11987–11995 (2019)
28. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6479–6488 (2018)
29. Sun, C., Jia, Y., Hu, Y., Wu, Y.: Scene-aware context reasoning for unsupervised abnormal event detection in videos. In: MM 2020: The 28th ACM International Conference on Multimedia, pp. 184–192 (2020)
30. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
31. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
32. Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2020)
33. Wang, J., Cherian, A.: GODS: generalized one-class discriminative subspaces for anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8201–8211 (2019)
34. Wang, J., Cherian, A.: GODS: generalized one-class discriminative subspaces for anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8200–8210 (2019)
35. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
36. Wang, Z., Zou, Y., Zhang, Z.: Cluster attention contrast for video anomaly detection. In: MM 2020: The 28th ACM International Conference on Multimedia, pp. 2463–2471 (2020)
37. Wu, J., et al.: Weakly-supervised spatio-temporal anomaly detection in surveillance video. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 1172–1178 (2021)
38. Wu, P., Liu, J.: Learning causal temporal relation and feature discrimination for anomaly detection. IEEE Trans. Image Process. **30**, 3513–3527 (2021)
39. Wu, P., et al.: Not only look, but also listen: learning multimodal violence detection under weak supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 322–339. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_20
40. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. In: Proceedings of the British Machine Vision Conference 2015, pp. 8.1–8.12 (2015)
41. Xu, J., Cao, Y., Zhang, Z., Hu, H.: Spatial-temporal relation networks for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3987–3997 (2019)
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: 4th International Conference on Learning Representations (2016)

43. Yu, G., et al.: Cloze test helps: effective video anomaly detection via learning to complete video events. In: MM 2020: The 28th ACM International Conference on Multimedia, pp. 583–591 (2020)
44. Zaheer, M.Z., Lee, J., Astrid, M., Lee, S.: Old is gold: redefining the adversarially learned one-class classifier training paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14171–14181 (2020)
45. Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.-I.: CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12367, pp. 358–376. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58542-6_22
46. Zhang, J., Qing, L., Miao, J.: Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In: 2019 IEEE International Conference on Image Processing, pp. 4030–4034 (2019)
47. Zhang, Y., Nie, X., He, R., Chen, M., Yin, Y.: Normality learning in multispace for video anomaly detection. IEEE Trans. Circuits Syst. Video Technol. **31**(9), 3694–3706 (2021)
48. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 1933–1941 (2017)
49. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1237–1246 (2019)
50. Zhou, J.T., Zhang, L., Fang, Z., Du, J., Peng, X., Xiao, Y.: Attention-driven loss for anomaly detection in video surveillance. IEEE Trans. Circuits Syst. Video Technol. **30**(12), 4639–4647 (2020)
51. Zhu, Y., Newsam, S.D.: Motion-aware feature for improved video anomaly detection. In: 30th British Machine Vision Conference 2019, p. 270 (2019)