Multi-Scale Video Anomaly Detection by Multi-Grained Spatio-Temporal Representation Learning

Menghao Zhang, Jingyu Wang*, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, Jianxin Liao
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
zhangmenghao,wangjingyu,qiqi8266,hfsun,zhuangzirui,rpf,maruilong@bupt.edu.cn; jxlbupt@gmail.com

Abstract

Recent progress in video anomaly detection suggests that the features of appearance and motion play crucial roles in distinguishing abnormal patterns from normal ones. However, we note that the effect of spatial scales of anomalies is ignored. The fact that many abnormal events occur in limited localized regions and severe background noise interferes with the learning of anomalous changes. Meanwhile, most existing methods are limited by coarse-grained modeling approaches, which are inadequate for learning highly discriminative features to discriminate subtle differences between small-scale anomalies and normal patterns. To this end, this paper address multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. We utilize video continuity to design three proxy tasks to perform feature learning at both coarse-grained and fine-grained levels, i.e., continuity judgment, discontinuity localization, and missing frame estimation. In particular, we formulate missing frame estimation as a contrastive learning task in feature space instead of a reconstruction task in RGB space to learn highly discriminative features. Experiments show that our proposed method outperforms state-of-the-art methods on four datasets, especially in scenes with small-scale anomalies.

1. Introduction

Video Anomaly Detection (VAD) is dedicated to detecting anomalous events in videos with wide applications in public safety and intelligent surveillance. The primary challenge of VAD is the sparsity of abnormal samples, limiting the direct learning of abnormal patterns from the available data. Therefore, most previous VAD studies can be divided into the weakly supervised category [9, 29, 36, 40, 43, 50, 53] that learns with video-level annotations, or the category only learning from normal data [5, 23, 24, 39, 46]. Our work focuses on the latter, which typically models the spatio-temporal features of the normal pattern, while samples that do not conform to the model are labeled as anomalies.

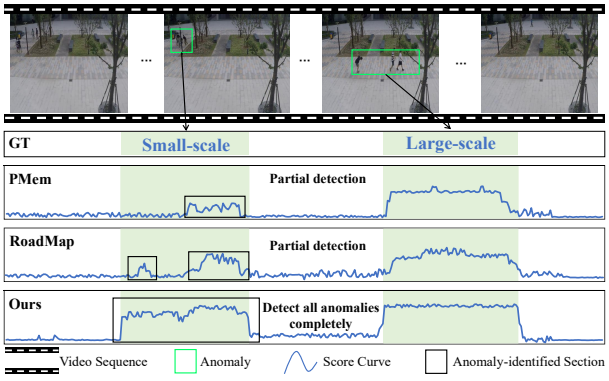

Figure 1. Examples of small-scale anomalies and comparisons of results from different methods. From top to bottom: video sequence, ground truth (green regions are abnormal), result of PMem [31] (reconstruction-based method), result of RoadMap [41] (prediction-based method) and result of ours.

In the same scene, the appearance and motion of foreground objects are the main differences between normal and abnormal patterns, and also the crucial features that the model needs to learn. We note that the effect of the spatial scale of the anomalies on feature learning is ignored. As illustrated in Fig. 1, many abnormal events occur in a limited region with only subtle differences from the normal pattern while background occupying a larger portion of the frame is the same. Therefore, severe background noise interferes with the feature learning of anomalous changes. Apart from small-scale anomalies, there are also abnormal events that span almost the entire frame, such as traffic accidents. Such changes in the scale of anomalies require the model to remain robust to anomalies of various scales.

Unfortunately, most previous methods exhibit a lower accuracy in detecting small-scale anomalies compared to larger-scale anomalies, as shown in Fig. 1. Reconstruction-

*Corresponding author

based [14, 16, 28, 30, 31, 45] and prediction-based [2, 5, 20, 23, 41] approaches are the mainstream modeling paradigms. The two approaches learn spatio-temporal features of normal patterns by reconstructing and predicting frames, respectively. However, both approaches are interfered by severe background noise since they perform reconstruction or prediction of frame in RGB space. Recent object-centric works [11, 12, 17] separate and process foregrounds and backgrounds to address the challenge of background noise. Nevertheless, modeling approaches based on prediction and binary classification struggle to learn highly discriminative features to distinguish small-scale anomalies. Some multi-scale VAD methods [41, 52] learn features at different resolutions through pooling or feature pyramiding but do not focus on appearance and motion features.

In this paper, we aim to explore a robust multi-scale VAD method. On the one hand, the context dependence of the anomalies necessitates the model to grasp global motion patterns and long-range features within the video. On the other hand, detecting small-scale anomalies not only demands that the model avoid the background noise interference but also requires it to discern highly discriminative short-range spatio-temporal variances between frames, owing to their subtle differences from the normal pattern. Consequently, the model needs to comprehensively learn the spatio-temporal features (spatial appearance and temporal motion features) of normal pattern in a multi-grained manner to achieve robustness to multi-scale anomalies.

To this end, we take video continuity [19] as supervision to construct three self-supervised proxy tasks, enabling VAD model to learn spatio-temporal features in both coarse-grained and fine-grained manners during training. *(i)* To determine whether a video sequence is continuous or not. Continuity judgment requires the model to learn the overall long-range temporal features and global motion patterns of the video in a coarse-grained manner. *(ii)* To locate discontinuities. Finding where the discontinuities occur drives the model to capture changes in local motion in a fine-grained manner rather than background noise. *(iii)* To estimate the missing frame in feature space instead of RGB space. We formulate the estimation task as a contrastive learning task to avoid the background noise interference and learn highly discriminative features of motion and appearances. By jointly solving three proxy tasks, the model can maintain robustess to anomalies of various scales.

We conduct experiments on four challenging datasets (Avenue [25], ShanghaiTech [27], UCF-Crime [36] and Campus [2]). The experiments show that our proposed method outperforms the State-Of-The-Art (SOTA) method, especially in scenes with small-scale anomalies.

Our contributions can be summarized as follows:

- Three straightforward yet effective multi-grained spatio-temporal representation proxy tasks are designed, enabling the video anomaly detection model to maintain robustness to multi-scale anomalies.
- A reconstruction scheme for frames based on contrastive learning is proposed, motivating the model to learn highly discriminative features from both appearance and motion, rather than simple RGB features and background noise.
- Our method achieves state-of-the-art performance on four datasets, especially outperforming previous methods by a large margin in scenes with small-scale anomalies.

2. Related Work

2.1. Video Anomaly Detection

Our work focuses on the video anomaly detection method that only learning from normal data, which is mainly grouped into the reconstruction-based methods [4, 16, 28, 33] and the prediction-based methods [20, 41, 48]. These two methods typically use autoencoders (AEs) [15, 51], memory-augmented AEs [13, 14, 23, 28, 31], or generative models [20, 32, 41] to reconstruct current frames or predict future frames so that frames with large reconstruction or prediction errors are recognized as anomalous. While some work attempts to introduce optical flow [23, 30], skeletal information [7, 10, 39], or utilize bi-directional prediction [5, 8, 18, 52] to learn motion and temporal features, these methods cannot be effective for small-scale anomalies due to disturbance from background noise. Recent object-centric methods [12, 17, 39] attempt to separate background and objects to avoid the disturbance of background noise. Although such approach can focus on the foreground objects, the way these methods model the object in terms of classification or RGB prediction lacks the exploration of fine-grained features. In addition, some efforts [1, 24] to synthesize virtual anomaly data fail to work for such anomalies owing to the lack of small-scale virtual data.

Different from these methods, we expect to enable the model to learn different scales of variation and fine-grained features of the normal pattern via self-supervised spatio-temporal representation learning.

2.2. Self-Supervised Learning in VAD

Self-supervised learning VAD methods typically performe contrastive learning or solve pretext tasks. For instance, Wang *et al.* [42] performs contrastive learning with clustered attention mechanism. However, it requires customized data enhancement strategies. For pretext tasks, Yang *et al.* [46] learn spatio-temporal representations by keyframe-based event restoration, while the work [11] designs multiple proxy tasks in an object-centric way to detect anomalies in different aspects. However, the former [46] suffers from severe background noise interference due to reconstructing events in RGB space, while the latter [11] lacks the ability to learn fine-grained features as its tasks are primarily

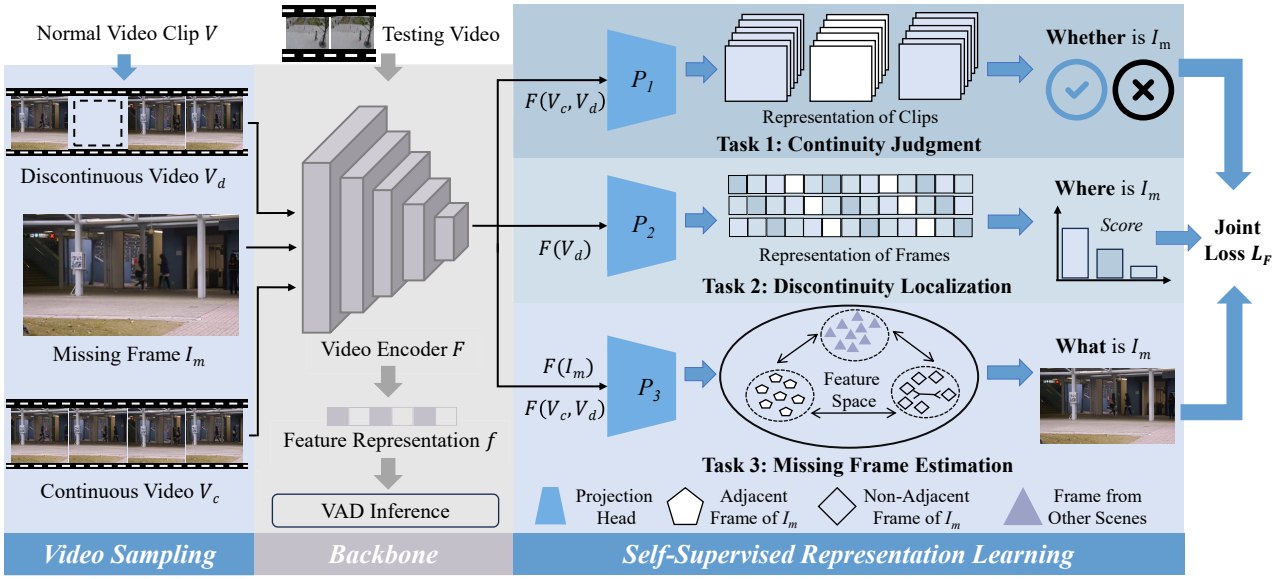

Figure 2. Method overview. The overall framework consists of a shared backbone network and three branches. The three branches perform continuity judgment, discontinuity localization, and missing frame estimation tasks to learn the spatio-temporal features of normal patterns at coarse- and fine-grained levels, respectively. The feature representation f generated after self-supervised training can effectively detect anomalies of various scales. Blue arrows and grey arrows represent training and inference processes, respectively. Best viewed in color.

binary classifications.

In this work, we design proxy tasks based on video continuity [19] to learn spatio-temporal features of normal events in a multi-grained manner to perceive anomalies of various scales. In addition, we formulate the reconstruction of frames as a contrastive learning task to learning highly discriminative features instead of background noise.

2.3. Multi-Scale Learning

Multi-scale learning typically considers information from different scales or granularities. Similar to our work, Xiao *et al.* [44] hierarchically learns the spatio-temporal features through event mask prediction task. However, event-level features are not sufficient for detecting subtle variances between frames. There is also some multi-scale works [41, 52] in VAD. For instance, Wang *et al.* [41] obtain image features of different resolutions to perform prediction through different pooling operations, while Zhong *et al.* [52] obtain features at different scales through feature pyramid. Although pooling and feature pyramids can provide feature representations of different resolutions, they cannot specifically capture the motion and appearance features which is crucial for anomaly detection. In contrast, we design the self-supervised proxy tasks to direct the model's attention towards changes in appearance and motion between frames, enabling it to learn features in a multi-grained manner.

3. Method

3.1. Overview

Our proposed multi-scale video anomaly detection method attempts to learn spatio-temporal features of normal patterns at both coarse-grained and fine-grained levels by jointly solving multiple proxy tasks of spatio-temporal representation learning. For a normal video clip V, our approach is devoted to learning a video encoder F that transforms the video clip V into the scale-aware high-quality spatio-temporal feature f. We define f as the high-quality feature in terms of anomalous information of different scales if it is able to easily address the following questions: (1) Whether there are missing frames in video clips V, i.e., whether feature representation f remains sensitive to large-scale anomalies. (2) If the video clip V is not continuous, where are the missing frames, i.e., whether feature representation f remains sensitive to small-scale anomalies. (3) If video clip V is discontinuous, what are missing frames, i.e., whether the feature representation f contains high-level motion feature and contextual feature.

The overall architecture of our proposed method is illustrated in Fig. 2. The network consists of three branches and a video encoder F, all of the branches share the same backbone encoder F. Given two nonoverlapping video clips: continuous clip V_c, discontinuous clip V_d, and the corresponding missing frame I_m, each of the three branches solves a proxy task and operates together to optimize the

backbone. The first branch is used to discriminate whether a video clip is continuous or not. The second branch is used to locate the position of the missing frame in the discontinuous clip V_d. The last branch is designed to learn high-level features through contrastive learning mechanism, so that the feature representation of the discontinuous clip V_d can effectively approach the feature representation of the missing frame I_m. After joint optimization of self-supervised tasks, the backbone F can generate scale-aware feature representations f for anomaly inference.

3.2. Data Acquisition

To perform the proposed scale-aware proxy tasks, we select continuous video clips and discontinuous clips containing only one missing frame I_m, respectively. The continuous clip contains T continuous frames, while the discontinuous clip has a length of $T-1$. Given an initial continuous clip of length T, we uniformly select a frame from the interval $[1, T-2]$ as the missing frame I_m to form the breakpoint (counting from 0). And the leading and alternate of the missing frames are concatenated together to form discontinuous video clip V_d. The restriction of missing frame inside the clip sequence ensures that the construction of missing frames and discontinuous clips remains consistent.

As shown in Fig. 2, we formulate the continuity discrimination task as a binary classification task while the missing frame localization task is formulated as a $T-2$ class classification task labeled by I_m. The estimation task is formulated a contrastive learning task. Both initial and continuous clips are formed by randomly sampling non-overlapping clips of length T from the same video.

3.3. Network Architecture

We employ a video encoder (defaults a 3D-ConvNet) as the backbone network F and use it as a shared base for three proxy tasks. Each task introduces into its top a projection network, denoted P_1, P_2, and P_3, which are used to process the video clip embeddings f extracted from the backbone network. All projection heads contain spatio-temporal averaging pooling layers to ensure that the dimensionality of the deep feature embeddings remains consistent, even though the temporal lengths may be different.

In continuity judgment and missing frame localization, we use fully connected layers as classifiers at the end of P_1 and P_2 for performing classification operations. As for the missing frame estimation task, we use P_3 to embed the video features f extracted by the backbone network into the low-dimensional features to get the representation, i.e., the feature representation of the missing frame.

3.4. Proxy Tasks and Joint Learning

Task 1: Continuity Judgment. The continuity judgment is intended to help the model learn spatio-temporal features and motion patterns in the video in a coarse-grained manner to perceive large-scale anomalies. In this task, in order for the model to learn by video features rather than other cues such as context, the positive and negative samples in each batch in our training come from the same video scene. With N samples in a training batch, cross-entropy loss $\mathcal{L}_{CE}$ is used to optimize the model, and the loss for continuity judgment is $\mathcal{L}_1$:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{L}_{CE}\left(P_1\left(F\left(v_d^i\right)\right)_1 + P_1\left(F\left(v_c^i\right)\right)_0\right)\right), \tag{1}$$

where $(\cdot)$ represents the process of feature processing by the network, F and P_1 represent the shared backbone network and the continuity judgment projection head, respectively, and v_c^i and v_d^i represent the i_{th} continuous and discontinuous video clip, respectively.

Task 2: Discontinuity Localization. While the continuity judgment task only learns spatio-temporal features in a coarse-grained manner, the discontinuous localization task expects to learn in a fine-grained manner in order to perceive small-scale anomalies. Compared to binary labels, labeling with missing frames I_m in this task can drive the model to learn changes in motion between adjacent frames in a fine-grained manner. Similarly, the discontinuous localization loss $\mathcal{L}_2$ can be expressed as an average of the cross-entropy loss of N samples in each batch:

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{CE}\left(P_2\left(F\left(v_d^i\right)\right)_{I_i}\right), \tag{2}$$

where P_2 represents the discontinuity localization projection head and I_i represents the missing frame position corresponding to discontinuous clip v_d^i.

Task 3: Missing Frame Estimation. In order to further learn the highly discriminative spatio-temporal features of the video, we estimate missing frames in terms of both appearance and motion. Meanwhile, we choose to estimate frames in feature space rather than RGB space to avoid background noise interference. We design a contrastive learning mechanism to perform this estimation task based on the following facts. First, since adjacent frames and missing frames are temporally connected, adjacent frames and missing frames contain more similar motions compared to non-adjacent frames from the same scene but farther away. Second, video frames from the same scene share more similar appearance features with missing frames than video frames from other scenes.

Accordingly, we learn the motion information in the video by triplet loss [34] taking the discontinuous video clips v_d^i as the anchor point, the missing frames I_m as the positive samples, and the continuous video clips v_c^i as the negative samples. In addition, we take discontinuous clips

v_d^i as anchor points, continuous clips v_c^i from the same scene as positive samples, and video clips $\{v^j\}_{i \neq j}$ from different scenes as negative samples, to learn the appearance information in the video by contrastive loss [19]. Defining the cosine similarity computation operation as $sim(\cdot, \cdot)$, and the hyperparameter that balances the relative contributions of the triplet loss and the contrastive loss as $\omega \in [0, 1]$, the missing frame estimation loss is denoted as:

$$\mathcal{L}_3 = \frac{1}{N}\sum_{i=1}^{N}(\omega \times \max(0, \gamma - (p_i^+ - p_i^-)) - (1-\omega)\log\left(\frac{q_i^+}{q_i^+ + \sum_{j=1, j\neq i}^{N} q_{i,j^-}}\right)), \tag{3}$$

$$p_i^+ = \text{sim}\left(P_3(V_d^i, I_i)\right), \tag{4}$$

$$p_i^- = \text{sim}\left(P_3(V_d^i, V_c^i)\right), \tag{5}$$

$$q_i^+ = \exp\left(\text{sim}\left(P_3(V_d^i, V_c^i)\right)/\tau\right), \tag{6}$$

$$q_{i,j}^- = \exp\left(\text{sim}\left(P_3(v_d^i, v_d^j)\right)/\tau\right) + \exp\left(\text{sim}\left(P_3(v_d^i, v_c^j)\right)/\tau\right), \tag{7}$$

where p_i^+ and p_i^- denote the similarity between positive sample pairs and negative sample pairs in the triplet loss, respectively, q_i^+ denotes a single positive sample pair in the contrastive loss, ${q_{i,j}}^-$ denotes the similarity of negative sample pairs in the contrastive loss, and τ is the temperature factor in the contrastive loss. Increasing ω allows for more focus on motion information, while decreasing ω allows for more focus on appearance information. Overall, the model is incentivized to learn fine-grained motion changes and contextual features, including background information and object appearance.

Joint Optimization. The above three proxy tasks learn spatio-temporal features in videos from different levels, and we jointly optimize our shared video encoder F with the multi-loss function.

$$\mathcal{L}_F = w_1\mathcal{L}_1 + w_2\mathcal{L}_2 + w_3\mathcal{L}_3, \tag{8}$$

where $w_1, w_2, w_3 \in [0, 1]$ are the individual weights on the $\mathcal{L}_1, \mathcal{L}_2$ and $\mathcal{L}_3$. Joint optimization helps the model to learn video spatio-temporal features from both coarse- and fine-grained scales to perceive anomalies of various scales.

3.5. VAD Inference

In the inference, we use T frames as the processing unit. Following the setup of the SOTA method [20, 41, 46], we choose not to compute anomaly scores using the error between each estimated frame $\hat{I}_t$ $(1 \leq t \leq T)$ and the corresponding ground truth I_t. Instead, we opt to employ the Peak Signal-to-Noise Ratio (PSNR) corresponding to the frame with the largest Mean-Square Error (MSE) among the T frames and the ground truth. This approach serves as the detection metric for the video sequence, as follow:

$$\text{MSE}\left(\boldsymbol{I}_t, \hat{\boldsymbol{I}}_t\right) = \frac{1}{H \cdot W}\sum_{i,j}\left\|\boldsymbol{I}_t(i,j) - \hat{\boldsymbol{I}}_t(i,j)\right\|_2^2, \tag{9}$$

$$PSNR_t = 10\log_{10}\frac{\text{MAX}_{\boldsymbol{I}_t}^2}{\text{MSE}\left(\boldsymbol{I}_t, \hat{\boldsymbol{I}}_t\right)}, \tag{10}$$

where H and W are the height and width of the frames, respectively, $\boldsymbol{I}_t(i,j)/\hat{\boldsymbol{I}}_t(i,j)$ represents the RGB value of the (i,j)-th pixel in $I_t/\hat{I}_t$, and $\text{MAX}_{\boldsymbol{I}_t}$ represents the maximum value of the image pixel in I_t. Finally, we calculate the anomaly score S_t for the t-th frame I_t by normalizing the PSNR values over all T frames:

$$S_t = 1 - \frac{PSNR_t - \min_{t'} PSNR_{t'}}{\max_{t'} PSNR_{t'} - \min_{t'} PSNR_{t'}}, \tag{11}$$

where $1 \leq t' \leq T$. A higher score S_t indicates that the frame I_t is more likely to be anomalous.

4. Experiment

4.1. Dataset and Evaluation Metrics

Cross-Scene Anomaly Dataset Avenue [25] contains 16 training videos and 21 test videos with 47 abnormal events, including running and throwing. The scenes in this dataset are single and the anomalies are mainly large-scale. ShanghaiTech [27] has 13 scenes with complex lighting conditions and different perspectives. In addition, the dataset includes anomalies caused by sudden movements, such as chasing and arguing. The different perspectives and the unfixed position of the camera lead to a large variation of both the object scale and the anomaly scale in the scene. This is a significant challenge to the robustness of anomaly detection methods to scale variations. UCF-Crime [36] comprises 13 anomaly types, spanning a total of 128 hours of video footage. This dataset is complex because of the unconstrained backgrounds and anomalies of various scales. The training set contains video-level annotations, whereas the testing set includes frame-level annotations.

Scene-Dependent Anomaly Dataset. Campus [2] is currently the most challenging dataset in its field with 43 scenes, 28 classes of anomalous events and 16 hours of videos. Especially, it contains scene-dependent anomalies, which means an normal event may be abnormal in another scene. Detecting scene-dependent anomalies requires the model to understand the scene and learn highly discriminative features rather than overfit.

Method	Length T	ShanghaiTech		Campus	
		AUC		AUC	
		Micro	Macro	Micro	Macro
Continuity Judgment	8	**78.1**	**82.3**	60.1	63.2
	12	77.2	81.1	**60.3**	**63.9**
	16	75.1	79.2	58.2	60.9
Discontinuity Localization	8	82.6	85.5	62.1	65.4
	12	**83.4**	**87.2**	**62.6**	**65.1**
	16	82.1	85.5	61.9	65.0
Missing Frame Estimation	8	79.8	83.2	63.5	66.8
	12	**81.2**	**84.9**	**65.2**	**68.9**
	16	80.5	84.0	64.9	68.0
Overall Framework	8	84.2	88.8	66.3	69.2
	12	**85.1**	**89.8**	**67.3**	**70.9**
	16	83.9	88.3	66.0	69.2

Table 1. Comparison of proposed framework and individual proxy tasks between different lengths T of input clips. We report Micro and Macro AUC (%) on ShanghaiTech and Campus datasets. The best performing results are marked in bold.

Evaluation Metrics The area under the ROC curve (AUC) serves as a commonly used metric for evaluation and comparison. A higher AUC score indicates a better anomaly detection capability. Following previous research [12], we evaluate both the Micro and Macro versions of AUC.

4.2. Implementation Details

We train and evaluate our method with an NVIDIA RTX 3090 GPU. In the training phase, we resize the resolution of all input video clips to 256×256 pixels, while the values of the pixels in all frames are normalized to the range [0, 1]. For the pre-training of the three proxy tasks, we utilize AdamW as the optimizer while the length of the continuous video clips is set to $T = 12$ frames. The initial learning rate is set to 0.0003 and is gradually decayed following the scheme of cosine annealing. For the hyperparameters, we set $\omega = 0.5$ in Equation 3, $w_1 = w_2 = w_3 = 1$ in Equation 8, and the temperature factor $\tau = 0.1$ in the contrastive loss. In our reported experimental performance, the shared backbone network F is implemented as I3D-RGB [3]. In addition, we incorporate the designed proxy tasks into the backbone of the SOTA methods [2, 11], the experimental results are also reported. Following the existing method setup [11, 24, 33], we perform the proxy task at the object-level with the multi-task backbone [11] and at the frame-level with the other backbones.

4.3. Ablation Study

Considering the ShanghaiTech and Campus datasets contain diverse perspectives and scenes, and large variations in scales of anomaly, we conduct exhaustive ablation experiments on the ShanghaiTech and Campus datasets.

Sensitivity to the Length of Video Clips. The length T of the input clips is an important setting in our proposed framework. With a small T, both continuity judgments and discontinuity localization will be so ambiguous as to be difficult. While with a larger T, the model may not need to learn high-quality spatio-temporal features to solve the task due to enough spatio-temporal information in the feed. Table 1 shows the sensitivity of each proxy task to T. For the discontinuous localization and missing frame estimation tasks, $T = 12$ presents better performance. Although at $T = 8$, the continuity judgment yields competitive results on the ShanghaiTech dataset, the overall framework achieves the best outcomes with jointly optimized at $T = 12$.

ID	CJ	DL	MFE	FP	ShanghaiTech		Campus	
					AUC		AUC	
					Micro	Macro	Micro	Macro
1	-	-	-	✓	70.2	71.2	58.1	58.6
2	✓	-	-	-	77.2	81.1	60.3	63.9
3	-	✓	-	-	83.4	87.2	62.6	65.1
4	-	-	✓	-	81.2	84.9	65.2	68.9
5	✓	✓	-	-	84.1	87.8	62.8	65.4
6	✓	-	✓	-	82.3	85.6	65.3	69.2
7	-	✓	✓	-	84.5	89.5	66.2	70.1
8	✓	✓	✓	-	**85.1**	**89.8**	**67.3**	**70.9**
9	✓	✓	-	✓	84.2	87.6	62.8	65.6

Table 2. Ablation experiments on the contributions of each proxy task. We report the AUC (%) scores on ShanghaiTech and Campus datasets. 'CJ', 'DL' and 'MFE' stand for the three proxy tasks of continuity judgment, discontinuity localization and missing frame estimation, respectively. In addition we add experiments on frame prediction for comparison, 'FP' represents frame prediction.

Effect of Continuity Judgment. The results of the ablation experiments for each proxy task are shown in Table 2, where we take the frame prediction [20] performed on the same backbone F as the baseline and train it with gradient loss and intensity loss (ID 1). Performing the continuity judgment task alone improves the performance of the model compared to frame prediction (ID 2). This suggests that continuity judgment task motivates the model to generate higher quality representations than frame prediction.

Effect of Discontinuity Localization. Compared to other proxy tasks and baseline, the performance of the discontinuous localization task alone can be significantly improved, especially on ShanghaiTech which contains anomalies of various scales (ID 3, ID 5 and ID 7). As this task samples discontinuous positions uniformly along the time axis, it motivates the model to capture subtle variations between frames, obtaining fine-grained representations.

Effect of Missing Frame Estimation. Performing missing frame estimation alone is effective in improving performance on Campus compared to baseline and other proxy tasks, which contains scene-dependent anomalies (ID 4). This improvement illustrates that our proposed contrastive learning scheme can help the model understand the scene. Compared to performing the frame prediction task to learn RGB features (ID 1 and ID 9), our proposed scheme learns highly discriminative motion features and semantically ef-

year	Method	Avenue AUC		ShanghaiTech AUC	
		Micro	Macro	Micro	Macro
2018	Liu *et al.* [20]	85.1	-	72.8	-
	Liu *et al.* [22]	84.4	-	-	-
	Sultani *et al.* [36]	-	-	76.5	-
2019	Gong *et al.* [14]	83.8	-	71.2	-
	Lee *et al.* [18]	90.0	-	76.2	-
	Ionescu *et al.* [17]	87.4	90.4	78.7	84.9
2020	Park *et al.* [31]	88.5	-	70.5	-
	Sun *et al.* [37]	89.6	-	74.7	-
	Lu *et al.* [26]	85.8	-	75.9	-
	Wang *et al.* [42]	87.0	-	79.3	-
	Yu *et al.* [47]	89.6	-	74.8	-
2021	Liu *et al.* [23]	91.1	-	76.2	-
	Lv *et al.* [28]	89.5	-	73.8	-
	Georgescu *et al.* [11]*	91.5	91.9	82.4	89.3
	Georgescu *et al.* [12]*	92.3	90.4	82.7	89.3
2022	Wang *et al.* [41]	88.3	-	76.6	-
	Zaheer *et al.* [49]	74.2	-	79.6	-
	Chen *et al.* [5]	90.3	-	78.1	-
	Zhong *et al.* [52]	89.0	-	74.5	-
	Cho *et al.* [6]	88.0	-	76.3	-
	Yang *et al.* [45]	89.9	-	74.7	-
	Ristea *et al.* [33]*	92.9	91.9	83.6	89.5
	Acsintoae *et al.* [1]▽*	93.0	93.2	83.7	90.5
2023	Yang *et al.* [46]	89.9	-	73.8	-
	Cao *et al.* [2]	86.8	-	79.2	-
	Liu *et al.* [21]	92.8	-	78.8	-
	Singh *et al.* [35]	86.0	-	76.6	-
	Sun *et al.* [39]	92.4	-	83.0	-
	Sun *et al.* [38]	91.5	-	78.6	-
	Liu *et al.* [24]▽	91.8	92.3	83.8	87.8
	Liu *et al.* [24] ▽*	93.6	93.9	85.0	91.4
	Ours	92.4	92.9	85.1	89.8
	Ours▽	93.2	92.5	86.2	91.0
	Ours*(Estimation Task)	92.6	93.0	85.5	91.8
	Ours*(Three Tasks)	93.6	93.8	86.8	92.4
	Ours▽*(Three Tasks)	**94.3**	**94.5**	**87.5**	**93.0**

Table 3. Comparison with SOTA methods of the Micro and Macro AUC (%) on Avenue and ShanghaiTech datasets. The best performing results are marked in bold.
▽: Methods apply virtual dataset for training.
∗: Methods utilize multi-task model ([11] or [12]) as backbone.

Method	Reference	Micro AUC	Macro AUC
Park *et al.* [31]	CVPR20	68.9	72.4
Georgescu *et al.* [11]	CVPR21	74.6	78.2
Wang *et al.* [41]	TNNLS22	72.9	76.8
Sun *et al.* [39]	CVPR23	75.5	78.3
Ours	-	**80.6**	**83.9**

Table 4. Results of Micro and Macro AUC(%) on UCF-Crime dataset. The best performing results are marked in bold.

fective contexts, including the appearance of objects and background (ID 4 and ID 8).

Effect of Joint Optimization. Although a single task may be able to achieve competitive performance. Experiments have shown that when proxy tasks are executed in pairs (ID 5, ID 6 and ID 7) and all tasks are executed together (ID 8), the performance of the model continues to improve to an optimum. This joint improvement is attributed to the fact that different proxy tasks learn spatio-temporal features at different levels, and they complement each other in motion and contextual feature learning.

Method	Reference	Miceo AUC	Macro AUC
Liu *et al.* [20]	CVPR18	57.9	60.2
Gong *et al.* [14]	CVPR19	61.9	62.5
Ionescu *et al.* [17]	CVPR19	59.3	63.4
Park *et al.* [31]	CVPR20	62.5	63.6
Liu *et al.* [23]	ICCV21	63.7	-
Lv *et al.* [28]	CVPR21	64.4	-
Wang *et al.* [41]	TNNLS22	61.9	64.2
Cao *et al.* [2]△	CVPR23	68.2	-
Ours	-	67.3	70.9
Ours△	-	**70.1**	**72.2**

Table 5. Comparison with SOTA methods of the Micro and Macro AUC(%) on Campus dataset. The best performing results are marked in bold.
△:Version that utilize scene-conditioned model [2] as backbone.

4.4. Comparisons with State-Of-The-Arts

We compare the proposed framework with SOTA methods in terms of Micro and Macro AUC(%). It is noteworthy that the current state-of-the-art methods [1, 24, 33] utilize the multi-task framework proposed by Georgescu *et al.* ([11] and [12], respectively) as the backbone, where [1] and [24] employ virtual data for training. Therefore, we evaluate different configurations of the proposed method on the Avenue and ShanghaiTech datasets, as shown in Table 3. With the multi-task backbone [11] (marked with ∗ in Table 3), we assess two versions: one involves replacing only the prediction component with our designed missing estimation task, while keeping the other proxy tasks unchanged. The other version replaces both proxy tasks in the motion branch with continuity judgment and discontinuity localization tasks in addition to incorporating the missing estimation task. In addition, we report the performance of applying virtual data for training (marked with ▽ in Table 3).

Results on Avenue. As shown in Table 3, our method achieved the highest AUC scores, obtaining a Micro AUC of 94.3% and a Macro AUC of 94.5%. Without the backbone of multi-task [11, 12] and virtual data [24], our method still scored the best performance with 92.9% Macro AUC.

Results on ShanghaiTech. Our proposed method yields a Micro AUC of 87.5% and a Macro AUC of 93.0% on the ShanghaiTech dataset. With the backbone of multi-task [11, 12], we improve the Micro AUC by 3.1% compared to Georgescu *et al.* [11] in the setup with only the missing estimation task, and by 4.4% when incorporating three designed proxy tasks. Without the backbone of multi-task

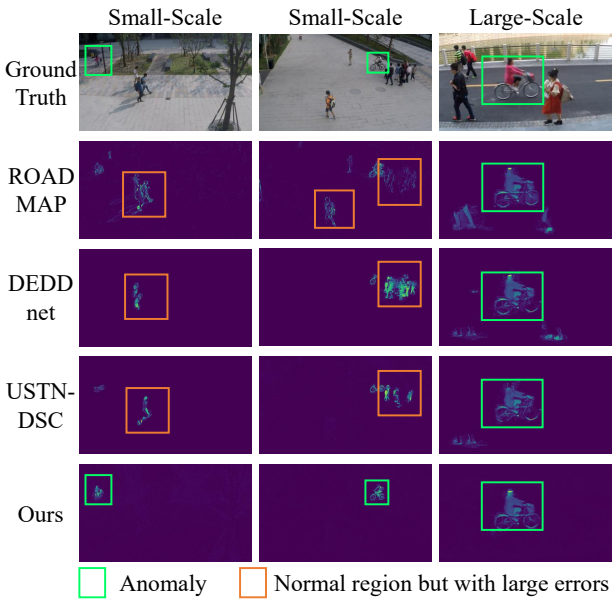

Figure 3. Examples of detection for anomalies of different scales from the ShanghaiTech dataset, including ground truth and prediction or reconstruction error maps from different methods (from top to bottom: ROADMAP[41], DEDDnet[52], USTN-DSC[46], and our method). Best viewed in color.

[11, 12] and virtual data [24], our method still achieves a Micro AUC score of 85.1%,which is comparable to the method [24] that employed the multi-task backbone and virtual training data. In comparison with the multi-scale VAD methods ([41] and [52]), our method achieves a significant improvement of 8.5% and 10.6%, respectively.

Results on UCF-Crime. Due to the absence of published results (methods only learning from normal data) on the UCF-Crime dataset, we implement the code from the existing literature [11, 31, 39, 41]. As shown in Table 4, our proposed method achieves a significant improvement compared to the second best method by 5.1% in terms of Micro AUC and 5.6% in terms of Macro AUC.

Results on Campus. We evaluate an additional version on the Campus dataset with the scene-conditioned model (provided by Cao *et al.* [2]) as the backbone, as it is necessary to incorporate scene features for detecting scene-dependent anomalies. As shown in Table 5, our method achieves the highest Micro AUC score of 70.1% and Macro AUC score of 72.2%. Without the scene-conditioned model, our method still obtain a Micro AUC of 67.3%, surpassing other RGB reconstruction-based [14, 28, 31] or RGB prediction-based methods [20, 23, 41].

4.5. Qualitative Results

Fig. 1 shows the anomaly curves of the test video and the comparison of our method with PMem [31] (reconstruction-based method) and RoadMap [41] (prediction-based method). When there are small-scale anomalies in the sampled videos, the anomaly score curves of the other two methods have obvious missed detections, while our method can accurately perceive the small-scale anomalies. In addition, Fig. 3 illustrates more examples, including the comparison with the multi-scale VAD methods [41, 52] and the comparison with self-supervised VAD method [46]. It can be observed that for normal regions in video frames, our method is able to estimate them well, while for abnormal event regions large errors occur. While other methods generate large errors for some normal regions, especially for frames with small-scale anomalies.

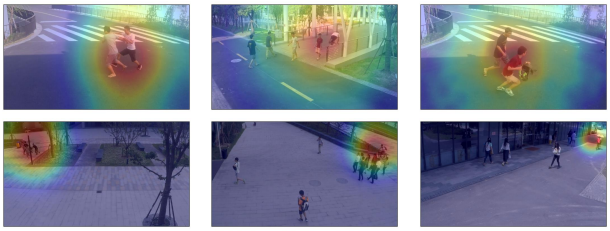

Figure 4. Visualization of salient regions of test frames by the proposed method. Best viewed in color.

Visualizations of the salient regions of the frames by our method are illustrated in Fig. 4. For both large-scale anomalies (1-th row) and small-scale anomalies (2-th row), our model can accurately focus on the anomalous regions.

5. Conclusion

In this paper, we observe that anomalous spatial scales affect the feature learning of the model. Small-scale anomalies require the model to learn highly discriminative features in a fine-grained manner rather than background noise. To detect multi-scale anomalies, we design proxy tasks supervised by video continuity that motivate the model to learn spatio-temporal features in both coarse-grained and fine-grained manners. Furthermore, instead of performing RGB reconstruction and prediction, we estimation frames in feature space through contrastive learning to learn highly discriminative features of appearance and motion. Experiments conducted on four challenging benchmark datasets validate the effectiveness of our proposed method.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grants (62101064, 62171057, 62201072, U23B2001, 62001054, 62071067), the National Postdoctoral Program for Innovative Talents under Grant BX20230052, China Postdoctoral Science Foundation (2023TQ0039), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *CVPR*, pages 20111–20121, 2022. 2, 7

[2] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *CVPR*, pages 20392–20401, 2023. 2, 5, 6, 7, 8

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 6

[4] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *ECCV*, pages 329–345, 2020. 2

[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*, pages 230–238, 2022. 1, 2, 7

[6] MyeongAh Cho, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognit.*, 129:108703, 2022. 7

[7] Romero F. A. B. de Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Reda Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *CVPR*, pages 11996–12004, 2019. 2

[8] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and Feng Yang. Anomaly detection with bidirectional consistency in videos. *IEEE Trans. Neural Networks Learn. Syst.*, 33(3):1079–1092, 2022. 2

[9] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. MIST: multiple instance self-training framework for video anomaly detection. In *CVPR*, pages 14009–14018, 2021. 1

[10] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *(ICCV)*, pages 10318–10329, 2023. 2

[11] 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*, pages 12742–12752, 2021. 2, 6, 7, 8

[12] 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, 2022. 2, 6, 7, 8

[13] 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 *ICCV*, pages 1705–1714, 2019. 2

[14] 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 *ICCV*, pages 1705–1714, 2019. 2, 7, 8

[15] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *CVPR*, pages 733–742, 2016. 2

[16] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *ICCV*, pages 8771–8780, 2021. 2

[17] 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*, pages 7842–7851, 2019. 2, 7

[18] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. BMAN: bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Trans. Image Process.*, 29:2395–2408, 2019. 2, 7

[19] Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *AAAI*, pages 1564–1573, 2022. 2, 3, 5

[20] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *CVPR*, pages 6536–6545, 2018. 2, 5, 6, 7, 8

[21] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *CVPR*, pages 12147–12156, 2023. 7

[22] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *BMVC*, page 71. BMVA Press, 2018. 7

[23] 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*, pages 13568–13577, 2021. 1, 2, 7, 8

[24] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *CVPR*, pages 24500–24510. IEEE, 2023. 1, 2, 6, 7, 8

[25] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 FPS in MATLAB. In *ICCV*, pages 2720–2727, 2013. 2, 5

[26] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *ECCV*, pages 125–141, 2020. 7

[27] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *ICCV*, pages 341–349, 2017. 2, 5

[28] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021. 2, 7, 8

[29] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, pages 8022–8031, 2023. 1

[30] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *ICCV*, pages 1273–1283, 2019. 2

[31] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, pages 14360–14369, 2020. 1, 2, 7, 8

[32] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *WACV*, pages 1896–1904, 2019. 2

[33] Nicolae-Catalin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *CVPR*, pages 13566–13576, 2022. 2, 6, 7

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 4

[35] Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *CVPR*, pages 18717–18726, 2023. 7

[36] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 2, 5, 7

[37] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *ACM Multimedia*, pages 184–192, 2020. 7

[38] Che Sun, Chenrui Shi, Yunde Jia, and Yuwei Wu. Learning event-relevant factors for video anomaly detection. In *AAAI*, pages 2384–2392, 2023. 7

[39] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *CVPR*, pages 22846–22856, 2023. 1, 2, 7, 8

[40] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, pages 4955–4966, 2021. 1

[41] 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 Networks Learn. Syst.*, 33 (6):2301–2312, 2022. 1, 2, 3, 5, 7, 8

[42] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. In *ACM Multimedia*, pages 2463–2471, 2020. 2, 7

[43] Jie Wu, Wei Zhang, Guanbin Li, Wenhao Wu, Xiao Tan, Yingying Li, Errui Ding, and Liang Lin. Weakly-supervised spatio-temporal anomaly detection in surveillance video. In *IJCAI*, pages 1172–1178, 2021. 1

[44] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. Hierarchical self-supervised representation learning for movie understanding. In *CVPR*, pages 9717–9726, 2022. 3

[45] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *ECCV*, pages 404–421, 2022. 2, 7

[46] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, pages 14592–14601, 2023. 1, 2, 5, 7, 8

[47] 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 Multimedia*, pages 583–591, 2020. 7

[48] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE Trans. Neural Networks Learn. Syst.*, 33(8):3572–3586, 2022. 2

[49] Muhammad Zaigham Zaheer, Arif Mahmood, Muhammad Haris Khan, Mattia Segù, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, pages 14724–14734, 2022. 7

[50] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *CVPR*, pages 16271–16280, 2023. 1

[51] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *ACM Multimedia*, pages 1933–1941, 2017. 2

[52] Yuanhong Zhong, Xia Chen, Yongting Hu, Panliang Tang, and Fan Ren. Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 32(12):8285–8296, 2022. 2, 3, 7, 8

[53] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI*, pages 3769–3777. AAAI, 2023. 1