Deep Anomaly Discovery from Unlabeled Videos via Normality Advantage and Self-Paced Refinement

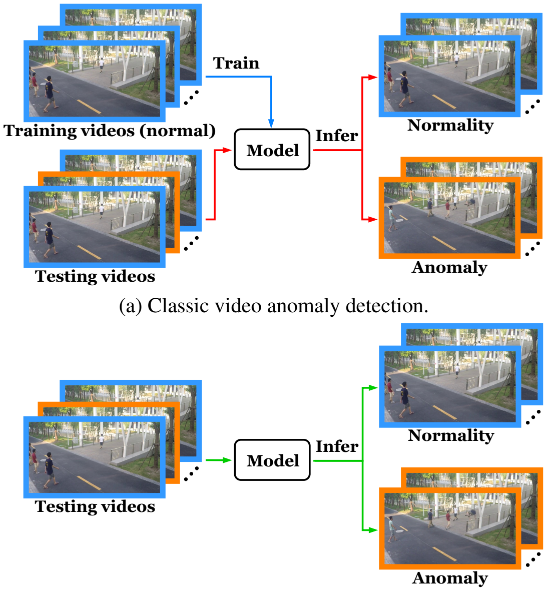
Guang Yu[*] Siqi Wang[*†] Zhiping Cai[†] Xinwang Liu, Chuanfu Xu, Chengkun Wu
National University of Defense Technology, China
{guangyu, wangsiqi10c, zpcai, xinwangliu, xuchuanfu, chengkun_wu}@nudt.edu.cn

Abstract

While classic video anomaly detection (VAD) requires labeled normal videos for training, emerging unsupervised VAD (UVAD) aims to discover anomalies directly from fully unlabeled videos. However, existing UVAD methods still rely on shallow models to perform detection or initialization, and they are evidently inferior to classic VAD methods. This paper proposes a full deep neural network (DNN) based solution that can realize highly effective UVAD. First, we, for the first time, point out that deep reconstruction can be surprisingly effective for UVAD, which inspires us to unveil a property named "normality advantage", i.e., normal events will enjoy lower reconstruction loss when DNN learns to reconstruct unlabeled videos. With this property, we propose Localization based Reconstruction (LBR) as a strong UVAD baseline and a solid foundation of our solution. Second, we propose a novel self-paced refinement (SPR) scheme, which is synthesized into LBR to conduct UVAD. Unlike ordinary self-paced learning that injects more samples in an easy-to-hard manner, the proposed SPR scheme gradually drops samples so that suspicious anomalies can be removed from the learning process. In this way, SPR consolidates normality advantage and enables better UVAD in a more proactive way. Finally, we further design a variant solution that explicitly takes the motion cues into account. The solution evidently enhances the UVAD performance, and it sometimes even surpasses the best classic VAD methods. Experiments show that our solution not only significantly outperforms existing UVAD methods by a wide margin (5% to 9% AUROC), but also enables UVAD to catch up with the mainstream performance of classic VAD.

1. Introduction

Video anomaly detection (VAD) [29, 58] has constantly been a valuable topic in computer vision, as it aims to automatically discover abnormal events (i.e., anomalies) that deviate from frequently-seen normal routine in surveillance videos. With a great potential to be applied to realms like public security and city management [53,90], VAD enjoys a continuous interest from both academia and industry. However, VAD remains open and unsolved. The underlying reason is that anomalies are typically rare and novel, and such characteristics make anomalies hard to be foreseen or enumerated in practice. As a result, a sufficient and comprehensive collection of anomaly data can be particularly difficult or even impossible, which makes the fully supervised classification paradigm not directly applicable to VAD.



(a) Classic video anomaly detection.

(b) Unsupervised video anomaly detection (UVAD).

Figure 1. Comparison of classic VAD and UVAD.

Thus, ***classic VAD*** follows a *semi-supervised* setup, which labels a training set that contains only normal videos to train a normality model (see Fig. 1a). During inference, video events that do not fit this normality model are viewed as anomalies. Although such a classic semi-supervised

[*] Equal contribution.
[†] Corresponding author.

VAD paradigm avoids the thorny issue to collect anomaly data, it still requires human efforts to label a training set with pure normal events. The labeling process can also be particularly tedious and labor-intensive, especially when faced with surging surveillance videos. To alleviate this problem, a natural idea is to perform ***unsupervised VAD*** (UVAD), which aims to discover anomalies directly from fully unlabeled videos in an unsupervised manner (see Fig. 1b). In this way, UVAD no longer requires labeling normal videos to build a training set, which can significantly reduce the cost of time and labor. Therefore, several recent works [8, 39, 53, 74] have explored this topic as a promising alternative to classic VAD (reviewed in Sec. 2.1).

Despite some progress, we notice that existing UVAD solutions suffer from two prominent limitations: **(1)** *Existing UVAD methods typically rely on shallow models to perform detection or initialization, and most of them still involve hand-crafted feature descriptors.* To be more specific, the core idea of representative UVAD methods [8, 39, 74] is to detect drastic changes as anomalies, which often involves learning a shallow detection model (e.g., logistic regression) with descriptor (e.g., 3D gradients) based video representations. However, the expressive power of both the shallow model and hand-crafted descriptors can be limited. The latest work [53] for the first time introduces deep neural networks (DNNs) to avoid hand-crafted descriptors, but it must resort to an initialization step that involves an isolation forest [36] model to obtain initial results. **(2)** *The performance of existing UVAD methods is evidently inferior to classic VAD methods.* Taking the commonly-used UCSDped1 and UCSDped2 dataset for an example, recent classic VAD methods usually lead existing UVAD methods by about 10% AUROC. Meanwhile, existing UVAD methods typically report their performance on earlier datasets, while their applicability and effectiveness on recent benchmark dataset like ShanghaiTech [38] are also unknown.

To move beyond the above limitations, we propose a novel DNN based solution that can perform UVAD in a highly effective and fully end-to-end manner. Specifically, this paper contributes to UVAD in terms of three aspects:

- We, for the first time, point out that deep reconstruction is actually surprisingly effective for UVAD, while such effectiveness further motivates us to unveil the property named *"normality advantage"*. Based on such a property, we design Localization based Reconstruction (LBR), which serves as a strong deep UVAD baseline and the solid foundation of our deep UVAD solution.

- We design a novel self-paced refinement (SPR) scheme, which is synthesized into LBR to consolidate normality advantage and enable more proactive UVAD. Unlike ordinary self-paced learning (SPL) that gradually injects training samples from easy to hard, the proposed SPR scheme aims to drop suspicious samples, so as to remove anomalies and focus on learning with normality. To our best knowledge, this is also the first attempt to tailor SPL for addressing VAD.

- We further design a motion enhanced solution that explicitly takes the motion cues into account. The variant solution can consistently enhance the detection capability, and sometimes even allows our UVAD solution to outperform state-of-the-art classic VAD methods.

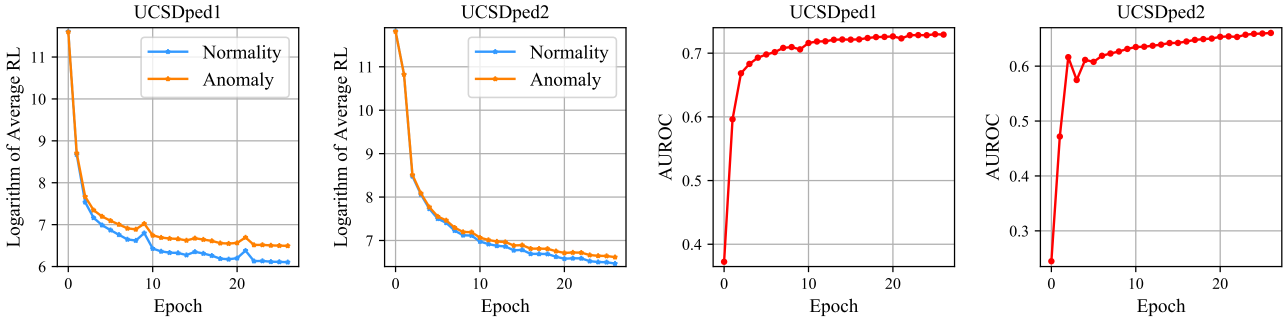
Experiments demonstrate the remarkable advantage of our solution against its UVAD counterparts. Furthermore, it for the first time achieves readily comparable performance to recent classic VAD methods on mainstream benchmarks.

2. Related Work

2.1. Video Anomaly Detection (VAD)

Classic VAD. Early classic VAD methods usually consist of two steps: First, they utilize hand-crafted feature descriptors (e.g., trajectory [56], dynamic texture [48], histogram of optical flow [7], 3D gradients [41]) to represent original training videos. Then, the extracted features are fed into a shallow normality model for training and inference, such as sparse reconstruction models [7, 41, 98], probabilistic models [6, 48], one-class classifiers [77] and nature inspired models [50, 70]. As manual descriptor design can be troublesome and inflexible, recent works are naturally motivated to introduce DNNs for automatic representation learning and end-to-end VAD. Thus, DNN based classic VAD methods have enjoyed a surging interest and explosive development [26, 31, 49, 51, 59–62, 66, 91]. Due to the absence of anomalies in training, they usually build a DNN normality model by training the DNN to perform some surrogate learning tasks, such as reconstruction [73, 82, 87, 88] and prediction [5, 9, 35, 42, 63, 97]. To improve representation learning and normality modeling, various DNN models have been explored such as recurrent neural networks [45, 46] and generative adversarial network [62, 64, 92]. A more detailed review on classic VAD can be found in [58]. Besides, note that deep VAD in this paper refers to directly learning from pixel-level video data by DNNs for VAD.

UVAD. Compared with thoroughly-studied classic VAD, only limited works have explored this emerging topic: Del et al. [8] pioneer the exploration of UVAD by detecting drastic changes as anomalies. Specifically, they describe each video frame by hand-crafted descriptors, and then train a shallow classifier to differentiate two temporally consecutive set of features. Afterwards, an easy classification indicates a drastic change, while shuffling is used to make the classification order-independent; Ionescu et al. [74] follow the direction of [8], but improve change detection by a



(a) Average reconstruction loss (RL) of normal and abnormal frames.

(b) Frame-level Area Under ROC Curve (AUROC) during training.

Figure 2. A demonstration of normality advantage by FBR on the testing set of UCSDped1 and UCSDped2 dataset.

more sophisticated unmasking scheme: With features calculated by hand-crafted descriptor and pre-trained DNN, they iteratively remove the most discriminative feature in classification. Frames that are still easy to classify after several rounds of removal are viewed as anomalies; Liu et al. [39] study the connection between unmasking and statistical learning, and further enhance the performance by a history sampling method and a new frame-level motion feature. Unlike above methods that are essentially based on the change detection paradigm, the latest work from Pang et al. [53] first obtain the preliminary detection results by leveraging a pre-trained DNN and an isolation forest [36]. Results are then refined by performing a two-class ordinal regression in a self-trained fashion.

Weakly-supervised VAD (WVAD). WVAD has been another heated topic [12, 32, 37, 57, 68, 72, 81, 93, 101] in current research. Unlike classic VAD or UVAD, WVAD utilizes video-level annotations for training, so as to reduce the cost of labeling [68]. Since WVAD usually adopts a different setup and benchmarks from most classic VAD and UVAD works, we will not discuss WVAD in this paper.

2.2. Self-Paced Learning

Self-paced learning (SPL) is a branch of curriculum learning (CL) [67, 79]. Motivated by the beneficial learning order in human curricula, CL introduces a learning strategy that trains the model with samples in an easy-to-hard manner [2]. To avoid the manual design of difficulty measures in classic CL, SPL is proposed to automatically measure the difficulty of samples based on the training losses [30]. Specifically, given a leaning objective, SPL embeds learnable sample weights and a self-paced (SP) regularizer into the objective. The SP regularizer enables SPL to learn a proper weight for each sample, so as to control the curriculum of learning. As a center issue of SPL, the design of SP regularizer has been extensively studied [10, 20, 27, 28, 30, 34, 86, 99], and the plug-and-play nature of SPL enables it to be widely applied to various tasks, such as classification [71, 85], object segmentation [95], domain adaptation [96], object detection [65, 94], clustering [18, 21], object re-identification [15]. However, to our best knowledge, none of existing works has explored SPL for VAD.

3. The Proposed UVAD Solution

3.1. Reconstruction in Classic VAD

Although our goal is to develop a deep UVAD solution, it will be helpful to recall how DNN addresses classic VAD in the first place. Owing to the lack of anomalies in training, DNN cannot learn representations directly by supervised classification. Instead, reconstruction has been a frequently-used deep learning paradigm for classic VAD. Typically, the reconstruction paradigm learns to embed the normal training video $\mathbf{x}$ into a low-dimensional embedding by an encoder network $f_e(\cdot)$, and then reconstruct the input video from the embedding by a decoder network $f_d(\cdot)$. This goal is often realized by solving the objective below:

$$\min_{\boldsymbol{\theta}} \sum_{\mathbf{x}} L_R(f_d(f_e(\mathbf{x})), \mathbf{x}|\boldsymbol{\theta}) + R(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ denotes all learnable parameters of the encoder and decoder, and $L_R(\cdot, \cdot|\boldsymbol{\theta})$ is a loss function that measures the reconstruction loss (RL) under parameters $\boldsymbol{\theta}$. $R(\boldsymbol{\theta})$ is a regularization term that prevents overfitting. By Eq. (1), DNN is expected to learn normality patterns and reconstruct normal events well, while large RL is produced for unseen anomalies. As a straightforward deep learning paradigm, reconstruction is extensively applied to classic VAD [58].

3.2. Normality Advantage in UVAD

Despite the popularity of DNN based reconstruction in classic VAD, it has not been explored as a deep solution to UVAD. Seemingly, learning by unlabeled videos mixed with anomalies also enables DNN to reconstruct anomalies, which disables it from discriminating anomalies. However, we argue that it may not be true: In most cases, anomalies are unusual events that occur at a low probability, while the majority of events in videos are still normal. When DNN

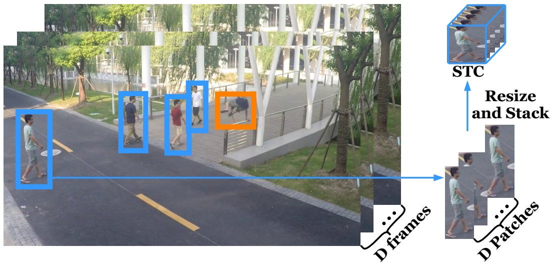


Figure 3. Localizing foreground to build spatio-temporal cube.

learns to reconstruct unlabeled videos that contain anomalies, *the imbalanced nature of normality/anomaly tends to bias the DNN model towards the majority class (normality), which offers us a chance to differentiate normality and anomalies*. Besides, we also notice that such bias is reported in simulated outlier image removal experiments [76, 84].

Motivated by such an intuition, we conduct some basic experiments to test whether DNN based reconstruction can be a feasible deep solution to UVAD: Following most UVAD works [8, 39, 74], we directly use the testing set of a VAD benchmark dataset as unlabeled videos with anomalies, while both the training set and testing set labels are strictly unused when training the DNN. To perform reconstruction, we train a multi-layer fully convolutional autoencoder (CAE) network to reconstruct the frames of unlabeled videos. To evaluate the reconstruction of normal and abnormal frames, we compute the average RL of normal frames and abnormal frames respectively. As an example, we visualize the logarithm of the average RL on UCSDped1 and UCSDped2 dataset in Fig. 2a, and some interesting observations can be drawn: Initially, the averaged RL of normal and abnormal events are very close. Afterwards, a loss gap gradually appears between normal and abnormal frames, which suggests that DNN prioritizes the reconstruction of normality. Moreover, the gap persists to exist as the training continues. Such observations lead to an interesting conclusion: ***Normality tends to play a more advantageous role (i.e., enjoys a lower reconstruction loss) when DNN learns to reconstruct both normality and anomalies in unlabeled videos***, which is named as ***normality advantage*** of UVAD.

To further validate whether normality advantage can be utilized to discriminate anomalies, we simply use RL as the anomaly score of each video frame, and calculate frame-level AUROC [48] to quantitatively evaluate the VAD performance during the learning process: As shown in Fig. 2b, whilst the VAD performance is poor at the beginning, it will be rapidly improved in 3-5 starting epochs. Afterwards, the AUROC tends to increase slowly and gradually levels off. As a consequence, those observations demonstrate the possibility to exploit normality advantage for deep UVAD. In addition, we would like to make the following remarks: **(1)** Normality advantage stems from the dominant role of normal events in videos. This role is essentially decided by the nature of anomalies, which are supposed to be rare events that divert from the majority. Actually, when a certain anomaly becomes frequent, they should be viewed as the new normality. Thus, we simply assume that normality advantage usually holds in the context of UVAD. **(2)** In Sec. 4.3, we will show that other deep learning paradigms (e.g., prediction) can also exploit this property to perform UVAD. This paper will focus on reconstruction as it is one of the most frequently-used deep paradigms in VAD.

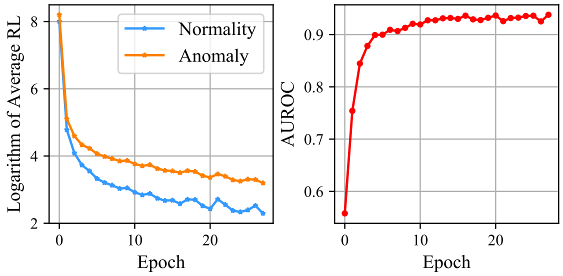


Figure 4. Average RL of normal/abnormal STCs (left) and frame-level AUROC (right) of LBR on UCSDped2 dataset.

3.3. Localization based Reconstruction (LBR)

Normality advantage renders frame based reconstruction (FBR) a feasible deep solution to UVAD, but its performance is still inferior to existing UVAD methods. For example, the RL gap of FBR on UCSDped2 dataset is relatively small (see Fig. 2a), while its AUROC is also unsatisfactory. Actually, there is an important reason for its unsatisfactory performance: In many cases, only a small region of a video frame is anomalous, while the remaining part is still normal. Thus, FBR is obviously not the optimal way to manifest normality advantage, since video events cannot be precisely represented on a per-frame basis. Inspired by recent works [23, 25, 90] that explore localization for classic VAD, we propose to introduce localization as a remedy to the drawback of FBR. Although localization is first introduced by classic VAD, we must point out that localization brings one unique benefit to UVAD: *Localization is able to magnify the normality advantage when performing UVAD*. An example is shown in Fig. 3: Consider a video frame with four walking pedestrians (normality) and one fence jumper (anomaly). For frame based analysis, the entire frame will be viewed as one abnormal event. By contrast, localization enables us to extract four normal events and one abnormal event. In this way, more normal events will exhibit a larger advantage against the anomalies in reconstruction.

Following this idea, we propose *localization based Reconstruction* (LBR) as a new deep baseline for UVAD: As to localization, we follow the localization scheme proposed in [90], which is shown to achieve both precise and com-

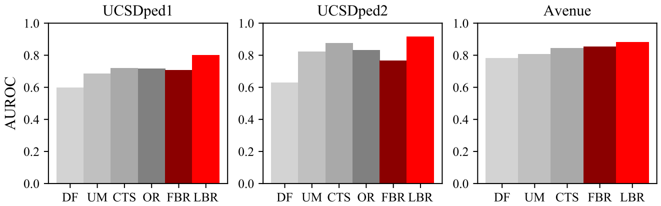


Figure 5. AUROC comparison between FBR/LBR and existing UVAD methods (DF [8], UM [74], CTS [39], OR [53]).

prehensive localization (the procedure is detailed in supplementary material). For each localized object on the frame, we extract D patches from the current and adjacent $(D-1)$ frames. Extracted patches are resized into $H \times W$, and then stacked into a $H \times W \times D$ spatio-temporal cube (STC), which is used to represent a video event (illustrated in Fig. 3). The DNN is then trained to reconstruct extracted STCs, while RL of STCs are also used as anomaly scores. To perform frame-level evaluation, the maximum of all STCs' scores on a frame is considered as the score of this frame. To illustrate how LBR magnifies normality advantage, we visualize LBR's average RL of normality/anomaly and AUROC in training on UCSDped2 dataset, on which FBR performs poorly. As shown in Fig. 4, LBR enjoys a remarkably larger RL gap than FBR, while the frame-level AUROC also grows to over 90%. In Fig. 5, we further compare frame-level AUROC of FBR and LBR (detailed in Sec. 4.1) with existing UVAD methods on several commonly-used VAD benchmarks, and find that LBR is surprisingly effective: As a straightforward baseline, LBR has already been able to outperform all existing UVAD methods on those benchmarks. Meanwhile, LBR achieves a large performance gain when compared with FBR, which verifies the importance of localization for UVAD. Consequently, the proposed LBR is able to lay a solid foundation for our deep UVAD solution.

3.4. Self-Paced Refinement (SPR)

Although LBR is shown to be a strong UVAD baseline, it passively relies on normality advantage to detect anomalies, and anomalies are constantly reserved in training. Nevertheless, the proactive removal of anomalies is obviously more preferable. To be more specific, we intend to sort out suspicious anomalies by RL and actively reduce anomalies' influence on DNN, so as to refine the DNN model and consolidate the normality advantage. To this end, we notice that self-paced learning (SPL) [30] provides an elegant strategy to adjust the influence of each individual sample in learning. However, traditional SPL usually injects harder samples to training in an incremental manner, but our goal is to gradually remove suspicious anomalies from the given data. To bridge this gap, we design a novel *Self-Paced Refinement* (SPR) scheme for UVAD, which is detailed below:

We first review the ordinary SPL as preliminaries. Formally, let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denote the training set, where $\mathbf{x}_i$ and y_i represent i-th sample and its learning target respectively. A model f parameterized by $\boldsymbol{\theta}$ maps a sample $\mathbf{x}_i$ to a prediction $f(\mathbf{x}_i)$, while the training loss $L(f(\mathbf{x}_i), y_i)$ is calculated by some loss function L. The learning goal is usually written as the following objective:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} L(f(\mathbf{x}_i), y_i|\boldsymbol{\theta}) \tag{2}$$

Note that we omit the regularization term R for simplicity. As to SPL, it embeds the learnable sample weights $\mathbf{v} = [v_i, \ldots, v_N] \in [0,1]^N$ and a self-paced (SP) regularizer $g(\mathbf{v}|\lambda)$ into the above learning objective, where λ is an age parameter to control the learning pace. Specifically, the goal of SPL is to solve the optimization problem below:

$$\min_{\boldsymbol{\theta}, \mathbf{v}} \sum_{i=1}^{N} v_i L(f(\mathbf{x}_i), y_i|\boldsymbol{\theta}) + g(\mathbf{v}|\lambda) \tag{3}$$

Eq. (3) can be solved by an alternative search strategy (ASS) [30], which alternatively optimizes $\boldsymbol{\theta}$ or $\mathbf{v}$ while keeping the other fixed. To facilitate optimization of $\mathbf{v}$, the SP regularizer $g(\mathbf{v}|\lambda)$ is usually designed to be convex, so when fixing $\boldsymbol{\theta}$ the global minimum $\mathbf{v}^*$ can be easily yielded by setting the partial derivative to be 0. It can be shown that $\mathbf{v}^*$ is usually determined by the training loss $L(f(\mathbf{x}_i), y_i)$ and age parameter λ. To enable SPL, λ is usually initialized by a small value, which produces $\mathbf{v}^*$ that only involves a few easy samples with small loss at the early training stage. Then, λ is gradually increased to introduce harder samples into training until all samples are considered in the end.

As shown above, SPL can adjust the weights of samples by considering their hardness and the current learning stage. Such desirable abilities make SPL perfectly eligible for enlarging normality advantage, which can be realized by assigning smaller weights to suspicious anomalies with large RL. Thus, we develop SPR from SPL: Concretely, given a sampled batch of STCs $\{\mathbf{c}_i\}_{i=1}^n$ ($\mathbf{c}_i$ denotes the i-th STC), SPR minimizes an objective $\mathcal{L}_{SPR}$ w.r.t. the DNN parameters $\boldsymbol{\theta}$ and sample weights $\mathbf{v}$, while $\mathcal{L}_{SPR}$ is defined by:

$$\mathcal{L}_{SPR} = \sum_{i=1}^{n} v_i L_i(\boldsymbol{\theta}) + g(\mathbf{v}|\lambda) \tag{4}$$

where $L_i(\boldsymbol{\theta}) = L_R(f_d(f_e(\mathbf{c}_i)), \mathbf{c}_i|\boldsymbol{\theta})$ represents the RL of $\mathbf{c}_i$, and the regularization term $R(\boldsymbol{\theta})$ is also omitted for simplicity. As mentioned above, Eq. (4) is optimized by ASS: When $\mathbf{v}$ is fixed, the objective can be transformed into:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} v_i L_i(\boldsymbol{\theta}) \tag{5}$$

The goal in Eq. (5) can be optimized by gradient descent. In fact, it assigns a weight to each STC when DNN learns to reconstruct STCs, which encourages DNN to place more emphasis on reconstructing the STC with larger weight v_i. When $\boldsymbol{\theta}$ is fixed, the optimal v_i^* can be obtained by solving:

$$\min_{v_i \in [0,1]} \sum_{i=1}^{n} v_i L_i(\boldsymbol{\theta}) + g(\mathbf{v}|\lambda) \tag{6}$$

Qualitatively, our SPR expects the optimal sample weight v_i^* yielded by Eq. (6) to meet the following requirements: When the loss of a STC is very large/small among its peers, it is highly likely to be abnormal/normal. Accordingly, its sample weight v_i should be directly set to 0/1. Otherwise, the sample weight should be negatively correlated with its likelihood to be abnormal, which is embodied by its RL. Such requirements motivate us to leverage a mixture SP regularizer [27] for SPR, which is in the following form:

$$g(\mathbf{v}|\lambda, \lambda') = -\rho \sum_{i=1}^{n} \ln(v_i + \frac{\rho}{\lambda}) \tag{7}$$

where λ' is an additional parameter that satifies $\lambda > \lambda' > 0$, and $\rho = \frac{\lambda\lambda'}{\lambda - \lambda'}$. As the mixture SP regularizer is convex, v_i^* to Eq. (6) can be derived by setting the partial derivative of $\mathcal{L}_{SPR}$ w.r.t v_i to zero, which yields:

$$\frac{\partial \mathcal{L}_{SPR}}{\partial v_i} = L_i(\boldsymbol{\theta}) - \frac{\rho}{v_i + \frac{\rho}{\lambda}} = 0, \quad i = 1, \cdots, n \tag{8}$$

Based on Eq. (8) and the constraint $v_i \in [0,1]$, a closed-formed solution to Eq. (6) can be derived as follows:

$$v_i^* = \begin{cases} 0, \quad L_i(\boldsymbol{\theta}) \geq \lambda \\ \dfrac{\rho}{L_i(\boldsymbol{\theta})} - \dfrac{\lambda'}{\lambda - \lambda'}, \quad \lambda' < L_i(\boldsymbol{\theta}) < \lambda \\ 1, \quad L_i(\boldsymbol{\theta}) \leq \lambda' \end{cases} \tag{9}$$

From Eq. (9), we can see how SPR consolidates normality advantage and excludes anomalies in an active way: When the RL of a STC $L_i(\boldsymbol{\theta})$ is larger than a upper threshold λ, its weight v_i will be directly set to 0, which suggests that this STC will be directly dropped from the current iteration. Similarly, the weight of STC will be directly set to 1 when its RL is smaller than a lower threshold λ', which enables it to fully participate in learning. For those STCs that are less certain ($\lambda' < L_i(\boldsymbol{\theta}) < \lambda$), their weights are inversely proportional to their RL. Next, the most important issue is to determine λ and λ', and we propose a self-adaptive strategy to calculate them by the statistics of RL: At the t-th iteration of model updating, the lower threshold $\lambda' = \mu(t) + \sigma(t)$, where $\mu(t)$ and $\sigma(t)$ denote the mean and standard deviation of STCs' RL in the current batch. The design of $\lambda'(t)$ indicates that we expect the majority of events to be normal. As to the upper threshold λ, we set it as follows:

$$\lambda = \max\{\mu(t) + (4 - t \cdot r) \cdot \sigma(t), \lambda'\} \tag{10}$$

where r is the shrink rate that usually takes a small value. The intuition behind λ is also straightforward: At the beginning, we only view STCs with very high RL ($L_i(\boldsymbol{\theta}) \geq \mu(t) + 4\sigma(t)$) as certain anomalies. As the learning continues, the normality advantage becomes more evident and allows us to exclude more anomalies. Thus, as t increases, Eq. (10) enables us to gradually shrink the coefficient of $\sigma(t)$ until λ decreases to λ', so as to exclude a larger portion of suspicious anomalies. Since the initial RL is not informative, SPR is introduced after a few warm-up epochs, which allows normality to establish the preliminary advantage. The whole SPR scheme is presented in Algorithm 1.

Algorithm 1 Self-Paced Refinement

```
Input: A DNN $f$ with parameters $\boldsymbol{\theta}$, the set $\mathcal{C}$ of $N$ STCs collected from
    unlabeled videos, batch size $n$, training epoch $T$, warm-up epoch $T'$
Output: The updated parameters $\boldsymbol{\theta}$
Initialize $\boldsymbol{\theta}$, $t = 0$
for $i = 1 \to T$ do
    for $j = 1 \to \lceil \frac{N}{n} \rceil$ do
        Randomly sample a batch of $n$ data from $\mathcal{C}$
        if $i \leq T'$ then
            Update $\boldsymbol{\theta}$ by Eq. (1)
        else
            Compute $\lambda' = \mu(t) + \sigma(t)$ and $\lambda$ by Eq. (10)
            $t = t + 1$
            Update $\mathbf{v}$ by Eq. (9)
            Updata $\boldsymbol{\theta}$ by Eq. (5)
        end if
    end for
end for
```

3.5. Motion Enhanced UVAD Solution

Many VAD works have pointed out the importance of motion cues [40, 52, 90]. Hence, we also design a motion enhanced UVAD solution, which consists of the following steps: First, to represent motion in videos, we adopt the dense optical flow [14], which depicts pixel-wise motion by estimating the correspondence between two frames. The optical flow map of each video frame can be computed efficiently by a pre-trained DNN model (e.g., FlowNet v2 [24]). Then, based on the location of each foreground object, we extract D optical flow patches from the optical flow maps that correspond to the current and $(D-1)$ neighboring frames. Similar to the construction of STC, D optical flow patches are resized and stacked into a $H \times W \times D$ optical flow cube (OFC). Then, we introduce a separated motion encoder $f_e^{(m)}$ and decoder $f_d^{(m)}$, which are trained to reconstruct the OFC by taking its corresponding STC as input:

$$\min_{\boldsymbol{\theta}'} \sum_{i=1}^{n} L_R(f_d^{(m)}(f_e^{(m)}(\mathbf{c}_i)), \mathbf{c}_i^{(o)} | \boldsymbol{\theta}') + R(\boldsymbol{\theta}') \tag{11}$$

where $\mathbf{c}_i^{(o)}$ represents the OFC for input STC $\mathbf{c}_i$. $\boldsymbol{\theta}'$ is the set of parameters for $f_e^{(m)}$ and $f_d^{(m)}$. After training, the RL of an OFC is computed as the motion anomaly scores $S^{(m)}$. The final anomaly score S is computed as follows:

$$S(\mathbf{c}_i) = \omega_a \frac{S^{(a)}(\mathbf{c}_i) - \mu^{(a)}}{\sigma^{(a)}} + \omega_m \frac{S^{(m)}(\mathbf{c}_i) - \mu^{(m)}}{\sigma^{(m)}} \quad (12)$$

where $S^{(a)}$ is the appearance anomaly score obtained by RL of STCs, and $\mu^{(a)}$, $\sigma^{(a)}$, $\mu^{(m)}$, $\sigma^{(m)}$ are means and standard deviations of appearance/motion anomaly scores for all STCs/OFCs. $\mu^{(a)}$, $\sigma^{(a)}$, $\mu^{(m)}$, $\sigma^{(m)}$ are computable as UVAD handles all testing videos in a transductive manner.

4. Empirical Evaluations

4.1. Experimental Settings

We evaluate the proposed UVAD solution (LBR-SPR) on the following commonly-used public VAD datasets: UCSDped1/UCSDped2 [48], Avenue [41] and ShanghaiTech [38]. To perform UVAD, we adopt two types of UVAD setups in previous UVAD works: **(1)** ***Partial mode*** [8, 39, 74]: Only the original testing set of a dataset is used for learning, while the original training set is discarded. **(2)** ***Merge mode*** [53]: The original training set and testing set are merged into one unlabeled set for learning. For both modes, labels are strictly unused in learning. Note that performance evaluation is only conducted on the original testing set of each benchmark, so as to enable comparison with existing VAD methods in the literature. For quantitative evaluation, we adopt the most commonly-used frame-level AUROC [48] in recent VAD works, while we also introduce and report other metrics like equal error rate (EER) and pixel-level AUROC in supplementary material. To construct STCs and OFCs, we adopt the localization scheme in [90] and set $H = W = 32$ and $D = 5$. The reconstruction is performed by a 7-layer fully convolutional autoencoder network, which is optimized by the default Adam optimizer in PyTorch toolbox [55]. The batch size in training is 256, while RL is computed by mean square error (MSE). For the shrink rate r, we adopt 0.0001 for UCSDped1/Avenue and 0.005 for UCSDped2/ShanghaiTech. The number of training epochs is set by $T = 30$, while $T' = 5$ epochs are typically used for warm-up. As to motion enhanced solution, we set (ω_a, ω_m) to be $(0.5, 1)$ for UCSDped1/UCSDped2/Avenue, and $(0.1, 1)$ for ShanghaiTech. Note that more details are provided in supplementary material due to page limit.

Table 1. Frame-level AUROC comparison. Note that LBR-SPR* indicates the performance of LBR-SPR under partial mode, while LBR-SPR$^+$ indicates the performance under merge mode (explained in Sec. 4.1). ME denotes motion enhancement.

Setup	Method	Ped1	Ped2	Avenue	SHTech
Classic VAD	CAE [22]	81.0%	90.0%	70.2%	-
	ST-CAE [100]	92.3%	91.2%	80.9%	-
	sRNN [45]	-	92.2%	81.7%	68.0%
	WTA-CAE [73]	91.9%	96.6%	82.1%	-
	LSTM-AE [44]	75.5%	88.1%	77.0%	-
	AM-GAN [62]	97.4%	93.5%	-	-
	Recounting [23]	-	92.2%	-	-
	FFP [38]	83.1%	95.4%	85.1%	72.8%
	AnoPCN [89]	-	96.8%	86.2%	73.6%
	Attention [103]	83.9%	96.0%	86.0%	-
	PDE-AE [1]	-	95.4%	-	72.5%
	Mem-AE [19]	-	94.1%	83.3%	71.2%
	AM-Corr. [52]	-	96.2%	86.9%	-
	AnomalyNet [102]	83.5%	94.9%	86.1%	-
	Object-Centric [25]	-	97.8%	90.4%	84.9%
	MLAD [75]	82.3%	99.2%	71.5%	-
	BMAN [33]	-	96.6%	90.0%	76.2%
	Clustering-AE [4]	-	96.5%	86.0%	73.3%
	r-GAN [43]	86.3%	96.2%	85.8%	77.9%
	DeepOC [83]	83.5%	96.9%	86.6%	-
	SIGNet [11]	86.0%	96.2%	86.8%	-
	Multipath-Pred. [78]	83.4%	96.3%	88.3%	76.6%
	Mem-Guided [54]	-	97.0%	88.5%	70.5%
	CAC [80]	-	-	87.0%	79.3%
	Scene-Aware [69]	-	-	89.6%	74.7%
	VEC [90]	-	97.3%	90.2%	74.8%
	BAF [17]	-	98.7%	92.3%	82.7%
	AMMCN [3]	-	96.6%	86.6%	73.7%
	SSMTL[‡] [16]	-	97.5%	91.5%	82.4%
	MPN [47]	85.1%	96.9%	89.5%	73.8%
	HF^2 [40]	-	99.3%	91.1%	76.2%
	CT-D2GAN [13]	-	97.2%	85.9%	77.7%
UVAD	DF [8]	59.6%	63.0%	78.3%	-
	UM [74]	68.4%	82.2%	80.6%	-
	CTS [39]	71.8%	87.5%	84.4%	-
	OR [53]	71.7%	83.2%	-	-
	LBR-SPR* (w/o ME)	**81.1%**	93.3%	88.5%	71.1%
	LBR-SPR* (w/ ME)	**81.1%**	**95.7%**	**92.8%**	**72.1%**
	LBR-SPR$^+$ (w/o ME)	79.4%	97.0%	89.7%	71.9%
	LBR-SPR$^+$ (w/ ME)	**80.9%**	**97.2%**	**90.7%**	**72.6%**

4.2. Comparison with State-of-the-art Methods

In Table 1, we compare the performance of LBR-SPR with state-of-the-art UVAD solutions. The performance of recent classic VAD methods is also included, while they are listed here as a reference. From Table 1, we can draw the following conclusions: **(1)** The proposed LBR-SPR solution consistently outperforms state-of-the-art UVAD methods by a notable margin under all configurations. Even for the basic LBR-SPR without motion enhancement (ME), it is able to achieve 4%-9% AUROC gain when only videos from the testing set are used (partial mode). **(2)** Meanwhile, LBR-SPR successfully bridges the gap between UVAD and classic VAD. On these benchmarks, LBR-SPR can achieve comparable or even superior performance to latest classic VAD methods in most cases. **(3)** Taking motion cues into consideration typically strengthens the performance of the proposed method. In particular, motion enhancement (ME) brings about 4.3% AUROC gain on Avenue dataset under

[‡] As micro AUROC is used, we reported results from the official page of [16] (https://github.com/lilygeorgescu/AED-SSMTL).

Table 2. Influence of SPR on frame-level AUROC.

Mode	Method	Ped1	Ped2	Avenue	SHTech
Partial	LBR	79.7%	90.9%	90.4%	71.7%
	LBR-SPR	**81.1%**	**95.7%**	**92.8%**	**72.1%**
Merge	LBR	80.1%	91.8%	89.5%	71.7%
	LBR-SPR	**80.9%**	**97.2%**	**90.7%**	**72.6%**

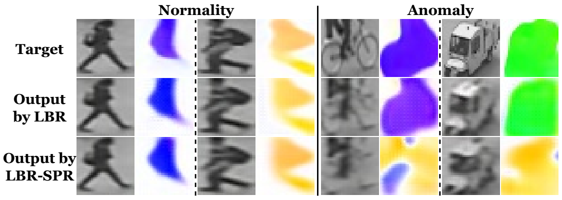


Figure 6. Reconstruction results for LBR and LBR-SPR.

partial mode, which even enables LBR-SPR to yield superior performance (92.8% AUROC) to state-of-the-art classic VAD methods. **(4)** The merge mode does not necessarily produce better performance than the partial mode, e.g., on UCSDped1 and Avenue. A possible reason is that normal events in the training set are slightly different from those of the testing set. Such a distribution shift distracts DNN from reconstructing normality in the testing set, which may undermine the normality advantage and UVAD performance.

4.3. Discussion

Role of Self-Paced Refinement. To demonstrate the importance of SPR to our UVAD solution, we conduct an ablation study that compares LBR and LBR-SPR under both partial and merge mode. As suggested by results in Table 2, SPR constantly brings tangible performance improvement to the LBR baseline. In particular, SPR improves LBR by 4% to 5% AUROC on UCSDped2, as it contains a relatively high proportion of anomalies. To provide a more intuitive illustration, we further visualize some reconstruction results of LBR and LBR-SPR in Fig. 6. As shown by the figure, while LBR and LBR-SPR both reconstruct normal foreground object and its optical flow satisfactorily, LBR-SPR reconstructs anomalies in an obviously worse manner than LBR, which makes anomalies more discriminative.

Sensitivity Analysis. In Fig. 7, we also conduct sensitivity analysis on key parameters in our solution: **(1)** The shrink rate r. For demonstration, we evaluate the performance of LBR-SPR on UCSDped2 and ShanghaiTech when r is varied between 0.001 and 0.01. As shown in Fig. 7, variation of r produces up to 0.5% AUROC fluctuation, which shows that the performance is not sensitive to r. **(2)** Weights of anomaly scores (ω_a, ω_m). To facilitate analysis, we simply fix $\omega_m = 1$ and vary ω_a between 0.1 and 1 in our experiments: On UCSDped2, LBR-SPR enjoys a stable performance, while the AUROC drops by at most 1.1% on ShanghaiTech when ω_a increases. However, it is noted that LBR-SPR without ME still yields satisfactory performance.

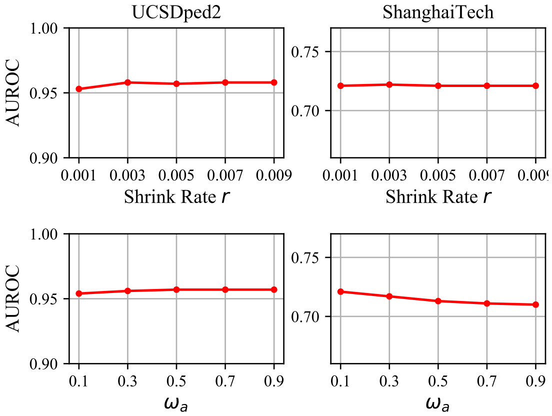


Figure 7. Parameter sensitivity analysis.

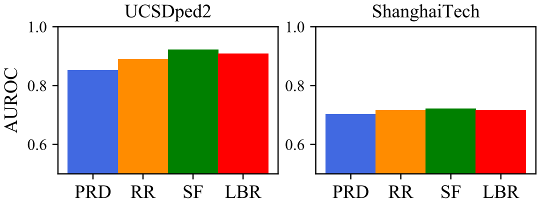


Figure 8. Other learning paradigms for UVAD.

Other Learning Paradigms. As we discussed in Sec. 3.2, normality advantage should also be observed in other learning paradigms. To verify this, we test three additional paradigms: Prediction (PRD), reverse reconstruction (RR) and shuffling (SF). PRD aims to predict the final patch of a STC/OFC by the remaining patches; RR aims to reconstruct a STC/OFC from its reversed patch sequence; SF aims to recover a STC/OFC with randomly shuffled patches. We compare the performance of PRD/RR/SF with raw LBR on UCSDped2 and ShanghaiTech. As shown in Fig. 8, other paradigms also yield close or better AUROC, which unveils the possibility to explore diverse paradigms for UVAD. More discussion are presented in supplementary material.

5. Conclusion

In this paper, we first reveal the advantageous role of normality in DNN based reconstruction, which enables us to propose LBR as a strong UVAD baseline. Based on LBR, we design a novel SPR scheme to remove anomalies actively, while motion cues are also exploited to further boost our solution. Our deep solution not only outperforms previous UVAD methods by a large margin, but also bridges the performance gap between UVAD and classic VAD.

Acknowledgements. The work is supported by National Natural Science Foundation of China (62006236), NUDT Research Project (ZK20-10), HPCL Autonomous Project (202101-15).

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2019. 7

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML '09*, 2009. 3

[3] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *AAAI*, 2021. 7

[4] Y. Chang, Z. Tu, Wei Xie, and J. Yuan. Clustering driven deep autoencoder for video anomaly detection. In *ECCV*, 2020. 7

[5] Dongyue Chen, P. Wang, Lingyi Yue, Yu xin Zhang, and Tong Jia. Anomaly detection in surveillance video based on bidirectional prediction. *Image Vis. Comput.*, 98:103915, 2020. 2

[6] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2917, 2015. 2

[7] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3449–3456. IEEE Computer Society, 2011. 2

[8] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision*, pages 334–349. Springer, 2016. 2, 4, 5, 7

[9] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021. 2

[10] Yanbo Fan, Ran He, Jian Liang, and Bao-Gang Hu. Self-paced learning: An implicit regularization perspective. In *AAAI*, 2017. 3

[11] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and F. Yang. Anomaly detection with bidirectional consistency in videos. *IEEE transactions on neural networks and learning systems*, PP, 2020. 7

[12] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14009–14018, June 2021. 3

[13] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 7

[14] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. 6

[15] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11309–11321. Curran Associates, Inc., 2020. 3

[16] Mariana-Iuliana Georgescu, Antonio Bărbălău, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12737–12747, 2021. 7

[17] 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 transactions on pattern analysis and machine intelligence*, PP, 2021. 7

[18] Kamran Ghasedi, Xiaoqian Wang, Cheng Deng, and Heng Huang. Balanced self-paced learning for generative adversarial clustering network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4386–4395, 2019. 3

[19] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 7

[20] Maoguo Gong, Hao Li, Deyu Meng, Qiguang Miao, and Jia Liu. Decomposition-based evolutionary multiobjective optimization to self-paced learning. *IEEE Transactions on Evolutionary Computation*, 23:288–302, 2019. 3

[21] Xifeng Guo, Xinwang Liu, En Zhu, Xinzhong Zhu, Miaomiao Li, Xin Xu, and Jianping Yin. Adaptive self-paced deep clustering with data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 32:1680–1693, 2020. 3

[22] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 7

[23] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017. 4, 7

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 6

[25] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric autoencoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019. 4, 7

[26] Radu Tudor Ionescu, Sorina Smeureanu, M. Popescu, and B. Alexe. Detecting abnormal events in video using nar-

rowed normality clusters. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1951–1960, 2019. 2

[27] Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. *Proceedings of the 22nd ACM international conference on Multimedia*, 2014. 3, 6

[28] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, S. Shan, and Alexander Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014. 3

[29] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2), 2018. 1

[30] M. Pawan Kumar, Ben Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 3, 5

[31] Yuandu Lai, R. Liu, and Yahong Han. Video anomaly detection via predictive autoencoder with gradient-based attention. *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020. 2

[32] Dongha Lee, Sehun Yu, Hyunjun Ju, and Hwanjo Yu. Weakly supervised temporal anomaly segmentation with dynamic time warping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7355–7364, October 2021. 3

[33] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2020. 7

[34] Hao Li and Maoguo Gong. Self-paced convolutional neural networks. In *IJCAI*, 2017. 3

[35] Sen Li, Jianwu Fang, Hongke Xu, and J. Xue. Video frame prediction by deep multi-branch mask network. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2

[36] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. 2, 3

[37] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, and Shenghua Gao. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, 2019. 3

[38] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 2, 7

[39] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *BMVC*, page 71, 2018. 2, 3, 4, 5, 7

[40] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13588–13597, October 2021. 6, 7

[41] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 2, 7

[42] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2

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

[44] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017. 7

[45] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 7

[46] Weixin Luo, W. Liu, Dongze Lian, J. Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1070–1084, 2021. 2

[47] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15425–15434, June 2021. 7

[48] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010. 2, 4, 7

[49] Amir Markovitz, Gilad Sharir, Itamar Friedman, L. Zelnik-Manor, and S. Avidan. Graph embedded pose clustering for anomaly detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10536–10544, 2020. 2

[50] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009. 2

[51] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019. 2

[52] Trong-Nguyen Nguyen and Jean and Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283, 2019. 6, 7

[53] Guansong Pang, C. Yan, Chunhua Shen, A. V. D. Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-

to-end video anomaly detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12170–12179, 2020. 1, 2, 3, 5, 7

[54] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14360–14369, 2020. 7

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 7

[56] Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, 18(11):1544–1554, 2008. 2

[57] Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 173–183, October 2021. 3

[58] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3

[59] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. Perceptual metric learning for video anomaly detection. *Mach. Vis. Appl.*, 32:63, 2021. 2

[60] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2587–2596, 2020. 2

[61] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, E. Sangineto, and N. Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698, 2018. 2

[62] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017. 2, 7

[63] R. Rodrigues, Neha Bhargava, R. Velmurugan, and S. Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2615–2623, 2020. 2

[64] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018. 2

[65] E. Sangineto, Moin Nabi, Dubravko Culibrk, and N. Sebe. Self paced deep learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:712–725, 2019. 3

[66] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. Deep appearance features for abnormal behavior detection in video. In Sebastiano Battiato, Giovanni Gallo, Raimondo Schettini, and Filippo Stanco, editors, *Image Analysis and Processing - ICIAP 2017*, pages 779–789, Cham, 2017. Springer International Publishing. 2

[67] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and N. Sebe. Curriculum learning: A survey. *ArXiv*, abs/2101.10382, 2021. 3

[68] Waqas Sultani, C. Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 3

[69] Che Sun, Y. Jia, Yao Hu, and Y. Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 7

[70] Q. Sun, H. Liu, and T. Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, page S0031320316302771, 2016. 2

[71] Ye Tang, Yu-Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. *Proceedings of the 20th ACM international conference on Multimedia*, 2012. 3

[72] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4975–4986, October 2021. 3

[73] Hanh TM Tran and David Hogg. Anomaly detection using a convolutional winner-take-all autoencoder. In *Proceedings of the British Machine Vision Conference 2017*. British Machine Vision Association, 2017. 2, 7

[74] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2895–2903, 2017. 2, 4, 5, 7

[75] Hung Thanh Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Q. Phung. Robust anomaly detection in videos using multilevel representations. In *AAAI*, 2019. 7

[76] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and M. Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *NeurIPS*, 2019. 4

[77] S. Wang, E. Zhu, J. Yin, and F. Porikli. Video anomaly detection and localization by local motion based joint video representation and ocelm. *Neurocomputing*, 277(FEB.14):161–175, 2017. 2

[78] X. Wang, Zhengping Che, Ke Yang, Bo Jiang, Jian-Bo Tang, Jieping Ye, Jingyu Wang, and Q. Qi. Robust unsupervised video anomaly detection by multi-path frame prediction. *IEEE transactions on neural networks and learning systems*, PP, 2021. 7

[79] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021. 3

[80] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 7

[81] 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*, 2021. 3

[82] P. Wu, Jing Liu, M. Li, Yujia Sun, and Fang Shen. Fast sparse coding networks for anomaly detection in videos. *Pattern Recognit.*, 107:107515, 2020. 2

[83] P. Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE Transactions on Neural Networks and Learning Systems*, 31:2609–2622, 2020. 7

[84] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1511–1519, 2015. 4

[85] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 247–263, Cham, 2020. Springer International Publishing. 3

[86] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view self-paced learning for clustering. In *IJCAI*, 2015. 3

[87] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 2

[88] Shiyang Yan, Jeremy S Smith, Wenjin Lu, and Bailing Zhang. Abnormal event detection from videos using a two-stream recurrent variational autoencoder. *IEEE Transactions on Cognitive and Developmental Systems*, 2018. 2

[89] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813. ACM, 2019. 7

[90] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020. 1, 4, 6, 7

[91] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and W. Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE transactions on neural networks and learning systems*, PP, 2021. 2

[92] M. Zaheer, Jin ha Lee, M. Astrid, and Seungik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14171–14181, 2020. 2

[93] M. Zaheer, Arif Mahmood, M. Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *ECCV*, 2020. 3

[94] Dingwen Zhang, Junwei Han, Long Zhao, and Deyu Meng. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision*, 127:363–380, 2018. 3

[95] Dingwen Zhang, Le Yang, Deyu Meng, Dong Xu, and Junwei Han. Spftn: A self-paced fine-tuning network for segmenting objects in weakly labelled videos. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5340–5348, 2017. 3

[96] Weichen Zhang, Dong Xu, Wanli Ouyang, and Wen Li. Self-paced collaborative and adversarial network for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:2047–2061, 2021. 3

[97] Y. Zhang, Xiushan Nie, Rundong He, Meng Chen, and Y. Yin. Normality learning in multispace for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2

[98] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. IEEE, 2011. 2

[99] Qian Zhao, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, 2015. 3

[100] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941. ACM, 2017. 7

[101] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[102] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 2019. 7

[103] Joey Tianyi Zhou, Le Zhang, Zhiwen Fang, Jiawei Du, Xi Peng, and Xiao Yang. Attention-driven loss for anomaly detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 7