

Learning Anomalies with Normality Prior for Unsupervised Video Anomaly Detection

Haoyue Shi^{1,3†}, Le Wang^{1*}, Sanping Zhou¹, Gang Hua², and Wei Tang³

¹ National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

² Multimodal Experiences Research Lab, Dolby Laboratories

³ University of Illinois Chicago

Abstract. Unsupervised video anomaly detection (UVAD) aims to detect abnormal events in videos without any annotations. It remains challenging because anomalies are rare, diverse, and usually not well-defined. Existing UVAD methods are purely data-driven and perform unsupervised learning by identifying various abnormal patterns in videos. Since these methods largely rely on the feature representation and data distribution, they can only learn salient anomalies that are substantially different from normal events but ignore the less distinct ones. To address this challenge, this paper pursues a different approach that leverages data-irrelevant prior knowledge about normal and abnormal events for UVAD. We first propose a new normality prior for UVAD, suggesting that the start and end of a video are predominantly normal. We then propose normality propagation, which propagates normal knowledge based on relationships between video snippets to estimate the normal magnitudes of unlabeled snippets. Finally, unsupervised learning of abnormal detection is performed based on the propagated labels and a new loss re-weighting method. These components are complementary to normality propagation and mitigate the negative impact of incorrectly propagated labels. Extensive experiments on the ShanghaiTech and UCF-Crime benchmarks demonstrate the superior performance of our method. The code is available at <https://github.com/shyern/LANP-UVAD.git>.

Keywords: Video anomaly detection · Self-training · Prior Knowledge

1 Introduction

Video anomaly detection aims to detect abnormal events in video sequences along the temporal dimension. This task is essential for intelligent surveillance [55] and crime detection [5]. Most existing methods are trained with partial supervision that requires manual annotations. For instance, one-class methods [23, 40, 41] use normal videos as the training set, while weakly-supervised methods [5, 33, 36]

*Corresponding author. [†] Part of this work was done while Haoyue Shi was a visiting scholar at University of Illinois Chicago.

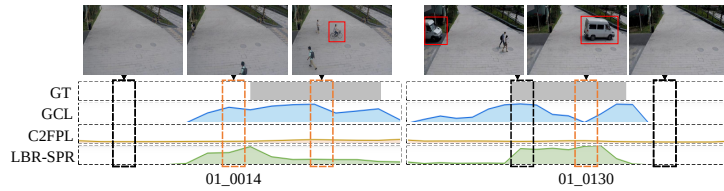


Fig. 1: Two examples of abnormal videos on ShanghaiTech [20]. The gray barcode is the ground truth. The line charts are anomaly scores of three methods [1, 46, 47]. The instances in the orange boxes are easily predicted incorrectly. This is because these methods are data-driven and highly rely on feature representation and data distribution, often struggling to capture less salient anomalies.

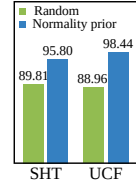


Fig. 2: Accuracy of normal snippets selected by the normality prior and random.

rely on video-level annotations during training. In contrast, unsupervised video anomaly detection (UVAD) methods [28, 37, 46, 47] attempt to detect anomalies without any annotations. It remains challenging because abnormal events in UVAD are rare, diverse, and usually not well-defined.

Existing UVAD methods [1, 28, 35, 37, 38, 46, 47] are based on self-training or reconstruction. Reconstruction-based methods [37, 38, 46] learn normal patterns from all training videos by gradually filtering out anomalies with large reconstruction errors. Self-training methods [1, 27, 35, 47] first generate pseudo-labels, which are then used as supervision for training, achieving state-of-the-art performance recently.

All current UVAD methods are purely data-driven, performing unsupervised learning by identifying abnormal patterns in videos. However, these methods heavily rely on the feature representation and data distribution, often struggling to capture less salient anomalies. In reconstruction-based methods [37, 38, 46], where normal and abnormal patterns are easily memorized by the autoencoder, imperceptible abnormal events may be easily overlooked. Self-training methods use data-driven strategies to generate pseudo-labels. For instance, GCL identifies anomalies based on local contrast in video snippets. They focus on high-contrast anomalies, potentially leading to the problem of anomalies attenuation, *i.e.*, overlooking less salient anomalies with low contrast. Al *et al.* [1] separate normal and anomalies by clustering over the entire dataset, making it challenging to distinguish them without pre-defined normal events. This challenge is illustrated in Figure 1, showing results from previous representative methods [1, 46, 47]. We can see that results from different methods vary significantly, and they can easily miss less salient anomalies.

In essence, UVAD is highly ill-posed. There still lacks a common definition of “what an anomaly is” in the community, and simply relying on data to detect anomalies is prone to failure. We tackle this problem from a different perspective. Unlike previous methods driven by data, our idea is to leverage *data-irrelevant* prior knowledge about normal and abnormal events to aid in identifying anomalies. By characterizing what normal and abnormal events look like beyond the

data, this prior knowledge helps address ambiguities between normal and abnormal events that cannot be resolved from a purely data-driven perspective, leading to more effective anomaly detection.

What is good prior knowledge for UVAD? It should be both informative and of high quality. We draw inspiration from the area of saliency detection [7, 16, 17, 45]. Concretely, inspired by the boundary prior [7, 45] commonly used in traditional saliency detection methods, we introduce a normality prior, that is, the start and end of a video are mostly normal. It is more robust than the center prior [14, 19] for anomalies because anomalies can be located far away from the temporal center but rarely touch the temporal boundary. We validate the normality prior on several datasets, as shown in Figure 2, which shows that compared to random selection, using our normality prior to select normal snippets is significantly more accurate. A more detailed experimental discussion is presented in Section 4.3.

How should the prior knowledge be used? A straightforward method is to directly compare other snippets with the normal prior, *i.e.*, start and end snippets of a video. However, since normal frames vary over time, the similarities between the normal prior and distant normal frames can be very large. They can be easily mislabeled without considering the temporal and semantic consistency of video snippets. In this paper, we propose Normality Propagation, which aims to propagate the normal information based on relationships between video snippets to estimate the normal magnitudes of unlabeled snippets, which represent the normal degree they received. Different from traditional label propagation in semi-supervised learning, our normality propagation features several innovations: 1) To overcome the limitation of no labeled snippets, we propose to use the normality prior to specify normal snippets. 2) We propose a temporally-modulated feature-based similarity matrix to model pairwise similarities. 3) We apply the propagation in a more efficient way, *i.e.*, over snippets in a video instead of the whole dataset used in traditional label propagation. We then perform unsupervised learning of abnormal detection based on the propagated labels and a new loss re-weighting method. They are complementary to normality propagation and mitigate the negative impact of incorrectly propagated labels.

The main contributions are summarized as follows:

- Unlike previous UVAD methods that are purely data-driven, we propose to use the data-irrelevant normality prior to identify abnormal events. To the best of our knowledge, such prior has never been studied before in the area of UVAD.
- We introduce normality propagation to effectively propagate the normality prior to unlabeled snippets for pseudo label generation.
- We perform unsupervised learning of abnormal detection based on the propagated labels and a new loss re-weighting method. They are complementary to normality propagation and mitigate the negative impact of incorrectly propagated labels.
- Extensive experiments on ShanghaiTech [20] and UCF-Crime [33] demonstrate the effectiveness of the proposed method.

2 Related Work

The literature on video anomaly detection (VAD) is rich. We mostly restrict the discussion to approaches for unsupervised video anomaly detection, label propagation, self-training and pseudo labeling.

2.1 Unsupervised Video Anomaly Detection

Video Anomaly Detection (VAD) aims to detect abnormal events in video sequences. Full-supervised methods [18, 39] tackle this task with precise annotations. Most existing methods are trained in partial supervision that needs manual annotations. For instance, one-class methods [6, 9, 20, 22, 23, 29, 34, 40, 41, 54] use only normal videos to train the detection model, weakly-supervised methods [5, 24, 25, 31–33, 36, 42, 49, 51, 55] need video-level annotations during training. In contrast, unsupervised methods [28, 37, 38, 46, 47] attempt to detect anomalies without any manual annotations, which are laborious, expensive, and prone to large variations. It remains challenging because abnormal events in UVAD are rare, diverse, and usually not well-defined. In the current work, we explore unsupervised mode for video anomaly detection.

Existing UVAD methods [28, 37, 38, 46, 47] can be categorized into reconstruction based methods and self-training based methods. These methods are purely data-driven, performing unsupervised learning by identifying abnormal patterns in videos. Reconstruction-based methods [37, 38, 46] learn normal patterns from all training videos by gradually filtering out anomalies with large reconstruction errors. For instance, Yu *et al.* [46] design a novel self-paced refinement scheme to remove anomalies with the reconstruction model. Tur *et al.* [37] leverage the reconstruction capability of diffusion models to detect anomalies with larger reconstruction errors. They [38] further employ conditional diffusion models to improve detection performance conditioned on compact motion representations. However, as normal and abnormal patterns are easily memorized by the autoencoder, they easily overlook imperceptible abnormal events.

Self-training-based methods [1, 35, 47] have achieved good performance recently. They generate pseudo labels first, then use pseudo labels as self-supervision. Pseudo labels are generated relying on data-driven strategies. For instance, Zaheer [47] propose a generative cooperative learning method. It identifies anomalies based on the local contrast in video snippets *i.e.*, the difference between consecutive snippets. They may label the boundaries of abnormal events well but can attenuate interior anomalies. Al *et al.* [1] cluster over the entire dataset to detect anomalies belonging to a smaller cluster. Normal and abnormal videos may appear in the same scene, but it is hard to separate them without a pre-defined normal. Thakare *et al.* [35] use OneClassSVM and iForest to find anomalies that lie outside the constructed hypersphere. However, previous UVAD methods are driven by data. They largely rely on feature representation and data distribution. Thus they easily overlook less salient anomalies. Our idea is to leverage prior knowledge about normal events that are irrelevant to the data to help identify anomalies.

2.2 Label propagation

Variants of label propagation [56] have been applied to various computer vision tasks, including learning with noisy labels [11, 26], few-shot learning [3, 21, 30], and semi-supervised learning [10, 57]. Method [11] for learning with noisy labels proposes a neighbour consistency regularization loss that encourages examples with similar feature representations to have similar predictions. Multi-Objective Interpolation Training (MOIT) [26] identifies and refines noisy examples using neighbours’ predictions. Liu *et al.* [21] consider label propagation with stochastic gradient descent for episodic few-shot learning. For semi-supervised learning, Iscen *et al.* [10] use label propagation to obtain labels for unsupervised data based on their neighbours in the feature space. Label propagation is not directly applicable to UVAD as both the initial labels and the similarity matrix are unavailable. We propose normality propagation to address these challenges. On the one hand, we propose the data-irrelevant normality prior to specifying the initial labels. On the other hand, we design a temporally-modulated similarity matrix to effectively propagate the normality prior across the video.

2.3 Self-training and Pseudo labeling

Self-training is a simple but effective technique used in unsupervised learning [12, 13, 52, 53], semi-supervised learning [10, 43], and weakly-supervised video anomaly detection [5, 55]. It initializes pseudo labels first, then uses labels predicted by the model as self-supervision. Zhang *et al.* [52] first generate saliency pseudo labels by using contrast-based SOD methods [4, 44], then designed a noise modelling module to deal with noises in saliency cues. The unsupervised person re-identification method [53] generates pseudo labels based on clustering results, then refines pseudo labels with clustering consensus over training epochs and temporal ensembling techniques. Lee *et al.* [10] uses the current network to infer pseudo labels of unlabeled data, and then re-trains the model on both labeled and unlabeled data. Zhong *et al.* [55] propose an alternate training framework for weakly supervised VAD, which generates frame-level pseudo labels for abnormal videos according to an action classifier [2]. In this work, we use self-training as an approach for UVAD. Our method differs from all prior work in that we propose a normality prior and introduce normality propagation, which effectively propagates the normality prior to unlabeled snippets for pseudo label generation. We also introduce a loss re-weighting strategy for robust abnormal detection.

3 Method

Problem Statement. In Unsupervised Video Anomaly Detection (UVAD), both normal and abnormal videos are included in the training set, while annotations are not provided. Assume that we have a set of N training videos, and each video is divided into a series of non-overlapping snippets. The goal of UVAD is to learn a snippet-level anomaly classifier $f_\theta(\cdot)$ that predicts the anomaly score of the snippet. A higher score indicates the snippet is more likely to be abnormal.

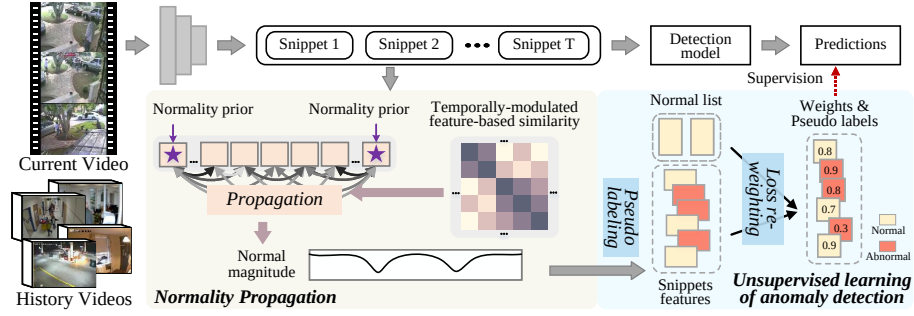


Fig. 3: Overview of our method. We first use the normality prior to specify that the start and end of a video are normal. Then we propose a new normality propagation method to propagate the normal information based on the temporally-modulated feature-based similarity for estimating normal magnitudes. After that, We perform unsupervised learning of abnormal detection based on the propagated labels and a new loss re-weighting method.

Overview. As previous methods are purely data-driven, they heavily rely on feature representation and data distribution, thereby often struggling to capture less distinct anomalies. To address this challenge, our key idea is to leverage data-irrelevant prior knowledge about normal/abnormal events to aid in identifying anomalies in videos. By characterizing what normal and abnormal events look like beyond the data, this prior knowledge helps address ambiguities between normal and abnormal events that cannot be resolved from a purely data-driven perspective, leading to more effective anomaly detection. Figure 3 illustrates an overview of our method. We first describe the normality prior in natural videos, and use it to specify normal snippets in a video. Then we propose normality propagation to propagate normal knowledge based on the temporally-modulated feature-based similarity matrix to estimate the normal magnitudes of unlabeled snippets. After propagation, unsupervised learning of abnormal detection is performed based on the propagated labels and a new re-weighting method.

Feature Extraction. Following [33, 47], we utilize a fixed-weight backbone network to extract features for each snippet. Formally, we denote features in a video as $\mathbf{X} \in \mathbb{R}^{D \times L}$, where L and D are the number of video snippets of a video and the feature dimension.

3.1 Normality Prior in Our Method

We propose to use a prior about normal events in natural videos, namely normality prior. Such prior has never been studied in previous methods, but it can help address challenges faced by purely data-driven methods.

Our normality prior is that the start and end of a video are mostly normal. It is inspired by the boundary prior [7, 16, 17, 45] widely used in traditional saliency detection methods: the image boundary is mostly background. The boundary

prior is more general than the center prior [14, 19] to identify anomalies because anomalies can be located far away from the temporal center, but they rarely touch the temporal boundary. This prior is validated on several anomaly detection datasets, as statistical analysis in Figure 2 and experimental discussion in Section 4.3.

3.2 Normality Propagation

Normality propagation aims to effectively propagate the normality prior to the unlabeled snippets for pseudo label generation. Given features of a video with L snippets $\mathbf{X}$, we propagate normal knowledge over video snippets based on pairwise similarities for estimating their normal magnitudes $\mathbf{z}$, which represent the normal degree they received.

As normal videos contain normal snippets only, we only specify the first and end snippets in a video as normal and do not mark any abnormal ones. We first define a label vector $\mathbf{y} \in \mathbb{R}^L$ with $y_i = 1$ if $i = 1$ or $i = L$ and $y_i = 0$ otherwise. Clearly, $\mathbf{y}$ is consistent with the initial normal magnitudes of video snippets.

Considering that normal and abnormal events are temporally consistent, and they are similar by themselves but frames between them vary greatly, we propose a temporally-modulated feature-based similarity matrix to model pairwise similarities, as shown in Figure 4. We define a matrix $\mathbf{W} \in \mathbb{R}^{L \times L}$ to describe the proximity of snippets in the feature space and temporal domain. We construct $\mathbf{W}$ by computing the snippets' feature space similarities and then modulating them by their respective temporal positions. Specifically, we define a similarity matrix $\mathbf{W}^f$ computed in the feature-space as:

$$\mathbf{W}_{i,j}^f = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathbf{x}_i$ denotes the i -th snippet feature in a video, $\sigma = 0.1$ is a hyper-parameter to control the strength of the similarity. A similar matrix $\mathbf{W}^t$ is defined in the time-space, and elements are computed from the time stamps. For L snippets, the time stamps are defined as $\mathbf{t} = \{1, 2, \dots, L\}$, and each element in $\mathbf{W}^t$ is computed as:

$$\mathbf{W}_{i,j}^t = \begin{cases} \exp\left(-\frac{|t_i - t_j|}{L}\right), & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where t_i denotes the i -th snippet time stamp in a video. We then compute temporally-modulated feature-based similarity matrix $\mathbf{W}$ as:

$$\mathbf{W}_{i,j} = \mathbf{W}_{i,j}^f \cdot \mathbf{W}_{i,j}^t. \quad (3)$$

$\mathbf{W}_{i,j}$ therefore specifies the temporally weighted similarity between snippets i and j . Its symmetrically normalized counterpart $\mathbf{S} = \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$, in which $\mathbf{D}$ is a diagonal matrix with its (i, i) -element equal to the sum of the i -th row of $\mathbf{W}$.

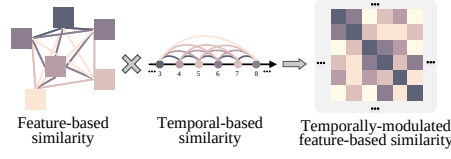


Fig. 4: Temporally-modulated feature-based similarity.

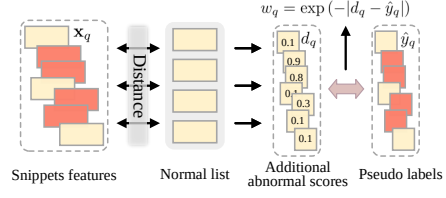


Fig. 5: Loss re-weighting strategy.

Normality propagation is to let every snippet iteratively spread its normal knowledge to its neighbours until a global stable state is achieved, that is, iterate $\mathbf{z}(n+1) = \alpha \mathbf{S} \mathbf{z}(n) + (1-\alpha) \mathbf{y}$ until convergence, where α is a hyper-parameter in the range $(0, 1)$, n is the index of an iteration, and $\mathbf{z}(0) = \mathbf{y}$. The iterative process converges to a simple solution: $\mathbf{z}^* = (1-\alpha)(\mathbf{I} - \alpha \mathbf{S})^{-1} \mathbf{y}$, which is clearly equivalent to

$$\mathbf{z}^* = (\mathbf{I} - \alpha \mathbf{S})^{-1} \mathbf{y}, \quad (4)$$

where $\mathbf{I}$ is the identity matrix. $\mathbf{z}^*$ can be efficiently obtained by using the conjugate gradient from the linear system $(\mathbf{I} - \alpha \mathbf{S}) \mathbf{z}^* = \mathbf{y}$. Derivations refer to the supplementary material.

It is interesting to observe that the vector $\mathbf{z}^*$ as defined by Eq.(4) is equivalent to the solver of the following objective function:

$$\mathcal{Q}(\mathbf{z}) = \alpha \mathbf{z}^T \mathbf{S} \mathbf{z} + (1-\alpha)(\mathbf{z} - \mathbf{y})^T (\mathbf{z} - \mathbf{y}), \quad (5)$$

The first term encourages similar snippets to get the same predictions, while the second term attempts to maintain predictions for the specified normal examples [56].

The optimized normal magnitudes of snippets in a video $\mathbf{z}^*$ indicate different degrees of normal knowledge that they received from the normality prior and their neighbours. A higher value means that this snippet is more likely to be normal, and vice versa.

Discussion. The normality prior specifies normal snippets effectively. By characterizing what normal events look like beyond the data, this prior knowledge helps address ambiguities between normal and abnormal events that cannot be resolved from reconstruction-based methods [37, 38, 46] and the global cluster-based methods [1]. Besides, normality propagation leverages the normality prior in a more effective way than directly comparing the video snippets with the normal prior. It estimates the normal magnitudes of each snippet not only based on the labeled normal snippets but also takes into account their neighbours' normal magnitudes. Thus normal snippets that are far away from specified ones will be affected by their normal neighbours and then can be labeled correctly. Besides, the previous ‘‘anomalies attenuation’’ problem is significantly alleviated because anomalies in a video are usually different from normal. In addition, normality propagation is a transductive method. It is easy to implement, fast and suitable for generating pseudo labels for UVDA.

3.3 Unsupervised Learning of Abnormal Detection

We perform unsupervised learning of abnormal detection based on the propagated labels and a new loss re-weighting method. They are complementary to normality propagation and mitigate the negative impact of incorrectly propagated labels.

Pseudo Labeling. We generate pseudo-labels based on normal magnitudes. Video-level pseudo labels are generated first, and then snippet-level pseudo labels are generated. Specifically, we select N^{nor} videos that have lower video scores than normal videos, where the video score is defined as the standard deviation of normal magnitudes of snippets in a video. Standard deviations of abnormal videos will be large as the abnormal video contains both normal and abnormal snippets. After that, we select snippets that have lower $r\%$ normality scores in the abnormal video as abnormal snippets. Finally, we formulate the pseudo label of the t -th snippet in the i -th video as

$$\hat{y}_t^i = \begin{cases} 1, & \text{if } i \in \mathcal{I}^{\text{abn}} \wedge t \in \mathcal{I}_i^{\text{abn}} \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $\mathcal{I}^{\text{abn}}$ is the set of indexes of abnormal videos, $\mathcal{I}_i^{\text{abn}}$ is the set of indexes of abnormal snippets in the i -th video.

Loss Re-weighting. Because a few videos may not follow the normality prior and the normality propagation is imperfect, we will have incorrect pseudo labels. We propose a loss re-weighting strategy to mitigate the negative impact of the noisy pseudo labels. As shown in Figure 5, we first select N^{hnor} high-confident normal videos and use the mean feature of each selected video to construct the normal list $\mathcal{M} = \{\mathbf{m}_0, \mathbf{m}_1, \dots, \mathbf{m}_{N^{\text{hnor}}}\}$. We then estimate additional abnormal scores based on the global normal list. For a snippet $\mathbf{x}_q$, its additional score is defined as

$$d_q = \min_{\mathbf{m}_i \in \mathcal{M}} 1 - \langle \mathbf{x}_q, \mathbf{m}_i \rangle, \quad (7)$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity between two vectors. The additional abnormal score estimates the probability of a snippet being abnormal in a global view. It is complementary to the pseudo labels generated by normality propagation in a local view. Then, the reliability is measured by the discrepancy between pseudo labels and additional normal-based anomaly scores. We further formulate the loss weight of the snippet $\mathbf{x}_q$ as

$$w_q = \exp(-|d_q - \hat{y}_q|). \quad (8)$$

The loss re-weighting strategy penalizes pseudo-labels with high discrepancy to guide the learning through reliable pseudo-labels.

After pseudo labels and their loss weights are generated, we train a classification model with the weighted binary cross-entropy loss as

$$\mathcal{L} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x}_i \in \mathcal{X}} -w_i \hat{y}_i \log(s_i) - w_i (1 - \hat{y}_i) \log(1 - s_i), \quad (9)$$

where $\mathcal{X}$ is the set of video snippets in a mini-batch, $s_i = f_\theta(\mathbf{x}_i)$ is the anomaly prediction of the i -th snippet.

4 Experiment

In this section, we first provide experimental details, then draw comparisons with the existing UVAD methods, and finally study different components of our method.

4.1 Experimental details

Evaluation Datasets. We evaluate our method on two popular video anomaly detection datasets: ShanghaiTech [20] and UCF-Crime [33].

ShanghaiTech [20] is a popular dataset used in video anomaly detection. It contains 437 campus surveillance videos (330 normal videos, 107 abnormal videos) with different locations spanning and camera angles. Recent UVAD methods follow the data organization of [55] but do not use annotations in the training set. Specifically, the training set contains 63 abnormal videos and 175 normal videos, and the testing set contains 44 abnormal videos and 155 normal videos.

UCF-Crime [33] is a large-scale video anomaly detection dataset. It consists of 1900 real-world surveillance videos (950 normal videos, 950 abnormal videos) with 13 different types of realistic abnormal events. It is a complex dataset due to videos containing diverse backgrounds and durations. Recent UVAD methods follow the data organization of [33] but without using training video labels. Specifically, the training split has 810 abnormal and 800 normal videos, while the testing split has 140 abnormal and 150 normal videos.

Evaluation Metrics. Following prior work [37, 47], we use the frame-level area under the ROC curve (AUC) for evaluation and comparisons. Larger AUC values indicate better performance.

Implementation Details. We use two backbones, *i.e.*, ResNext3d [8] and I3D [2], to extract features for each snippet receptively. Our detection model is composed of a temporal convolution and two linear layers. During our training procedure, each mini-batch consists of 30 randomly selected videos, and each video is sampled into 32 snippets. We train the model for 300 epochs with the RMSprop optimizer using a learning rate of 0.0001 and a momentum of 0.6. We set $\alpha = 0.99$ for normality propagation, N^{nor} as the number of normal videos in the training set, and the abnormal ratio $r\% = 40\%$. Besides, we set the number of high confident normal videos $N^{\text{hnor}} = h * N^{\text{nor}}$, and $h = 0.1$.

4.2 Comparison with State-of-the-art Methods

In Table 1, we compare the proposed method with existing unsupervised video anomaly detection methods [37, 38, 47] on two different datasets, *i.e.*, ShanghaiTech [20] and UCF-Crime [33]. Selected weakly-supervised and one-class

Table 1: Comparison with state-of-the-art methods in AUC (%) on ShanghaiTech and UCF-Crime. We divide the methods into weakly-supervised, one-class, and unsupervised. Best results are in **bold**. We implemented [46] and [1] and computed their AUC scores.

	Method	Features	ShanghaiTech	UCF-Crime
Weakly-supervised	Sultani <i>et al.</i> [33]	C3D	-	75.41
	CLAWS [49]	C3D	89.67	83.03
	CLAWS Net+ [50]	C3D	90.12	83.37
	MIST [5]	C3D	93.13	81.40
	RTFM [36]	C3D	91.51	83.28
	MIST [5]	I3D	94.83	82.30
	RTFM [36]	I3D	97.21	84.30
	Zhang <i>et al.</i> [51]	I3D	-	86.22
	CLAWS [49]	ResNext	-	82.61
	CLAWS Net+ [50]	ResNext	91.46	84.16
	Zaheer <i>et al.</i> [47]	ResNext	86.21	79.84
One-class	Lu <i>et al.</i> [23]	-	68.00	65.51
	BODS [40]	I3D	-	68.26
	GODS [40]	I3D	-	70.46
	OGNet [48]	ResNext	69.90	69.47
	Zaheer <i>et al.</i> [47]	ResNext	79.62	74.20
Unsupervised	DyAnNet [35]	I3D	-	79.76
	C2FPL [1]	I3D	-	80.65
	Ours	I3D	88.32	80.02
	Kim <i>et al.</i> [15]	ResNext	56.47	52.00
	LBR-SPR [46]	ResNext	77.12	57.18
	GCL [47]	ResNext	78.93	71.04
	Tur <i>et al.</i> [37]	ResNext	68.88	62.91
	Tur <i>et al.</i> [38]	ResNext	66.36	63.52
	C2FPL [1]	ResNext	67.36	74.71
	Ours	ResNext	86.46	76.64

methods [36,40,48,51] are presented for reference. We reimplement LBR-SPR [46] on our dataset splits and C2FPL [1] on ResNext features for fair comparisons. As we can see, our method outperforms almost all previous methods on different features. We establish a new state-of-the-art on ShanghaiTech with 86.46% AUC and UCF-Crime with 76.64% AUC with ResNext as the backbone. We also achieve comparable performance using I3D as the backbone.

In addition, Unsupervised methods [1,37,47] still perform inferior to weakly-supervised ones [5,33] because no annotations are provided in UVAD. But our method achieves better performance than one-class methods [40,50], verifying the effectiveness of the unsupervised setting in video anomaly detection.

4.3 Ablation Study and Analysis

We conduct a series of ablation studies to understand better how the proposed method works, where we use ResNext [8] as the backbone to extract features.

Validation of Normality Prior. In this experiment, we validate the normality prior to the experimental performance. Table 2 shows the precision and

Table 2: Precision (%) and Recall (%) of pseudo labels generated by normality propagation with Random, Data-driven Prior, and Normality Prior as labeled normal snippets, as well as corresponding testing performance (AUC %).

	ShanghaiTech			UCF-Crime		
	Precision	Recall	TestAUC	Precision	Recall	TestAUC
Random	17.90	29.27	81.69	13.28	42.41	60.33
Data-driven Prior	19.74	24.64	83.59	15.19	48.23	68.51
Normality Prior	34.39	42.92	85.96	20.37	54.28	75.99

Table 3: Ablation study of normality propagation. T, F, and T&F mean Temporal-based, Feature-based, and Temporally-modulated Feature-based pairwise similarities. The metric is AUC (%).

	Pairwise Similarity	ShanghaiTech	UCF-Crime
Direct Comparison	-	72.33	71.22
Normality Propagation	T	79.62	57.99
	F	79.73	63.01
	T&F	85.96	75.99

recall of pseudo labels generated by using normality propagation with random, the data-driven prior, and the normality prior as labeled normal snippets. It also shows the corresponding testing performance. As we can see, specified normal snippets with the normality prior generate better pseudo labels and testing performance. It is more informative than data-driven prior because the latter often favors simple, similar normal snippets, e.g., assuming snippets exhibiting minimal contrast are considered normal. We also validate its robustness beyond these standard benchmarks in the supplementary material.

Ablation Study of Normality Propagation. How to use the normality prior is very important. We compare different strategies for generating pseudo labels using the specified normal snippets, including making direct comparisons with specified normal snippets, normality propagation with temporal-based (T), feature-based (F), and temporally-modulated feature-based (T&F) pairwise similarities. The testing performances are presented in Table 3. Direct comparison has inferior performance than normality propagation with T&F. This is because direct comparison overlooks relationships between unlabeled snippets. In addition, our T&F pairwise similarity outperforms T or F pairwise similarity that only uses the similarity in the time-space or feature-space because we take into account temporal and feature similarities simultaneously. In summary, the proposed normality propagation effectively propagates the normality prior to unlabeled snippets based on the T&F pairwise similarities matrix. It mostly benefits from knowledge in labeled normal snippets as well as temporal and semantic consistency in unlabeled snippets to generate pseudo labels.

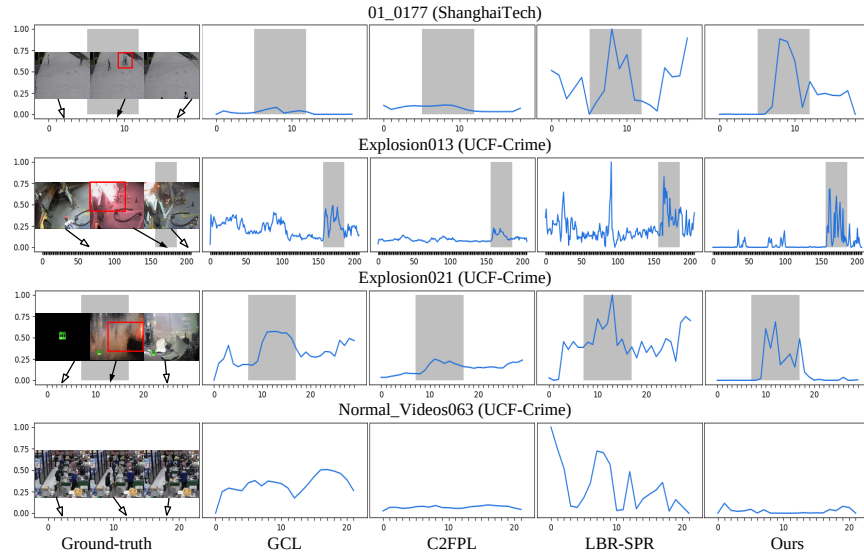
Ablation Study on Loss Re-weighting. In this experiment, we validate the effectiveness of the proposed loss re-weighting strategy. The results are shown

Table 4: Ablation study on the effectiveness of the loss re-weighting (LR) in our method.

	ShanghaiTech UCF-Crime	
Ours w/o LR	85.96	75.99
Ours	86.46	76.64

Table 5: Ablation study on different values of high confident normal videos, where we set different h .

	0.1	0.3	0.5
ShanghaiTech	86.46	86.18	86.24
UCF-Crime	76.64	76.63	76.42

**Fig. 6:** Qualitative results. Curves represent the predicted anomaly scores. The grey background corresponds to the ground truth. The white and black arrows denote the locations of normal and abnormal frames displayed on the left.

in Table 4. As we can see, our method with normality propagation achieves good performance over two datasets. By adding the normal-based loss re-weight, we obtain better performance. This is because it provides the global normal information for re-weighting pseudo labels, which is complementary to the normality propagation that generates pseudo labels in a local view. This also verified that our loss re-weighting strategy can mitigate the negative effects of incorrectly propagated labels.

Ablation Study of hyperparameters. The number of highly confident normal videos N^{hnor} is an important hyperparameter in our normal-based loss re-weighting. We set it as $h * N^{\text{nor}}$, and N^{nor} is set as the number of normal videos of the corresponding dataset. Here we test its sensitivity on two datasets. Results are presented in Table 5. Our method consistently achieves AUC higher than 86% on ShanghaiTech and 76% on UCF-Crime. Thus, our method is insensitive to N^{hnor} . More experimental results refer to the supplementary material.

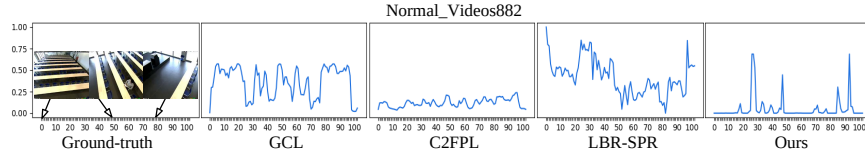


Fig. 7: A failure case on UCF-Crime.

4.4 Qualitative Analysis

Figure 6 shows some examples of detection results of our method and three previous representative methods, *i.e.*, one reconstruction-based method LBR-SPR [46], and two self-training methods GCL [47] and C2FPL [1]. Our method can detect the abnormal event well and predict abnormal scores of the normal frames very close to zero. In abnormal examples of *Explosion013* (UCF-Crime) and *01_0177* (ShanghaiTech), compared with other data-driven methods, our method locates anomalies more accurately. This is because our method uses the normality prior effectively. It helps address ambiguities between normal and abnormal events that cannot be resolved from a purely data-driven perspective, leading to more effective anomaly detection.

Limitation. As our method is built on the semantic consistency of normal events, inevitably, it fails when there are multiple types of normal events in a video, as shown in Figure 7 shows a typical failure case: A normal video captured from different angles. Nevertheless, our method performs well in most videos, as shown in Figure 6. In the future, we will explore combining with other methods that focus on detecting anomalies in multiple scenes to overcome this limitation.

5 Conclusion

Unlike existing methods driven by data, this paper leverages data-irrelevant prior knowledge for UVAD. We first propose a new normality prior, suggesting that the start and end of a video are predominantly normal. We then introduce normality propagation, which propagates normal knowledge based on relationships between video snippets to estimate normal magnitudes of unlabeled snippets. Finally, we perform unsupervised learning of abnormal detection based on the propagated labels and a new loss re-weighting method. Extensive experiments indicate that our method outperforms existing state-of-the-art methods.

Acknowledgements

This work was supported in part by National Science and Technology Major Project under Grant 2023ZD0121300, National Natural Science Foundation of China under Grants 62088102, 12326608 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

1. Al-Lahham, A., Tastan, N., Zaheer, M.Z., Nandakumar, K.: A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection. In: WACV. pp. 6793–6802 (2024)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
3. Chen, C., Yang, X., Xu, C., Huang, X., Ma, Z.: Eckpn: Explicit class knowledge propagation network for transductive few-shot learning. In: CVPR. pp. 6596–6605 (2021)
4. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. *IEEE TPAMI* **37**(3), 569–582 (2014)
5. Feng, J.C., Hong, F.T., Zheng, W.S.: Mist: Multiple instance self-training framework for video anomaly detection. In: CVPR. pp. 14009–14018 (2021)
6. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV. pp. 1705–1714 (2019)
7. Grady, L., Jolly, M.P., Seitz, A.: Segmentation from a box. In: ICCV. pp. 367–374 (2011)
8. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: CVPR. pp. 6546–6555 (2018)
9. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR. pp. 733–742 (2016)
10. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: CVPR. pp. 5070–5079 (2019)
11. Iscen, A., Valmadre, J., Arnab, A., Schmid, C.: Learning with neighbor consistency for noisy labels. In: CVPR. pp. 4672–4681 (2022)
12. Ji, H., Wang, L., Zhou, S., Tang, W., Zheng, N., Hua, G.: Meta pairwise relationship distillation for unsupervised person re-identification. In: ICCV. pp. 3661–3670 (2021)
13. Ji, W., Li, J., Bi, Q., Guo, C., Liu, J., Cheng, L.: Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. In: ICLR (2022)
14. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. pp. 2106–2113 (2009)
15. Kim, J.H., Kim, D.H., Yi, S., Lee, T.: Semi-orthogonal embedding for efficient unsupervised anomaly segmentation. *arXiv preprint arXiv:2105.14737* (2021)
16. Lempitsky, V., Kohli, P., Rother, C., Sharp, T.: Image segmentation with a bounding box prior. In: ICCV. pp. 277–284 (2009)
17. Li, H., Lu, H., Lin, Z., Shen, X., Price, B.: Inner and inter label propagation: salient object detection in the wild. *IEEE TIP* **24**(10), 3176–3186 (2015)
18. Liu, K., Ma, H.: Exploring background-bias for anomaly detection in surveillance videos. In: ACM MM. pp. 1490–1499 (2019)
19. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. *IEEE TPAMI* **33**(2), 353–367 (2010)
20. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection—a new baseline. In: CVPR. pp. 6536–6545 (2018)
21. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. In: ICLR (2019)

22. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: ICCV. pp. 13588–13597 (2021)
23. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: ICCV. pp. 2720–2727 (2013)
24. Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., Zhang, H.: Unbiased multiple instance learning for weakly supervised video anomaly detection. In: CVPR. pp. 8022–8031 (2023)
25. Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J.: Localizing anomalies from weakly-labeled videos. *IEEE TIP* **30**, 4505–4515 (2021)
26. Ortego, D., Arazo, E., Albert, P., O’Connor, N.E., McGuinness, K.: Multi-objective interpolation training for robustness to label noise. In: CVPR. pp. 6606–6615 (2021)
27. Pang, G., Shen, C., van den Hengel, A.: Deep anomaly detection with deviation networks. In: ACM SIGKDD. pp. 353–362 (2019)
28. Pang, G., Yan, C., Shen, C., Hengel, A.v.d., Bai, X.: Self-trained deep ordinal regression for end-to-end video anomaly detection. In: CVPR. pp. 12173–12182 (2020)
29. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: CVPR. pp. 14372–14381 (2020)
30. Rodríguez, P., Laradji, I., Drouin, A., Lacoste, A.: Embedding propagation: Smoother manifold for few-shot classification. In: ECCV. pp. 121–138 (2020)
31. Sapkota, H., Yu, Q.: Bayesian nonparametric submodular video partition for robust anomaly detection. In: CVPR. pp. 3212–3221 (2022)
32. Shi, H., Wang, L., Zhou, S., Hua, G., Tang, W.: Abnormal ratios guided multi-phase self-training for weakly-supervised video anomaly detection. *IEEE TMM* **26**, 5575–5587 (2023)
33. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488 (2018)
34. Sun, S., Gong, X.: Hierarchical semantic contrast for scene-aware video anomaly detection. In: CVPR. pp. 22846–22856 (2023)
35. Thakare, K.V., Raghuwanshi, Y., Dogra, D.P., Choi, H., Kim, I.J.: Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In: WACV. pp. 5541–5550 (2023)
36. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV. pp. 4975–4986 (2021)
37. Tur, A.O., Dall’Asen, N., Beyan, C., Ricci, E.: Exploring diffusion models for unsupervised video anomaly detection. In: ICIP. pp. 2540–2544 (2023)
38. Tur, A.O., Dall’Asen, N., Beyan, C., Ricci, E.: Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In: ICIAP. pp. 49–62 (2023)
39. Wan, B., Jiang, W., Fang, Y., Luo, Z., Ding, G.: Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing* **15**(14), 3454–3465 (2021)
40. Wang, J., Cherian, A.: GODS: Generalized one-class discriminative subspaces for anomaly detection. In: ICCV. pp. 8201–8211 (2019)
41. Wang, L., Tian, J., Zhou, S., Shi, H., Hua, G.: Memory-augmented appearance-motion network for video anomaly detection. *PR* **138**, 109335 (2023)

42. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV. pp. 322–339 (2020)
43. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR. pp. 10687–10698 (2020)
44. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR. pp. 1155–1162 (2013)
45. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR. pp. 3166–3173 (2013)
46. Yu, G., Wang, S., Cai, Z., Liu, X., Xu, C., Wu, C.: Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In: CVPR. pp. 13987–13998 (2022)
47. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: CVPR. pp. 14744–14754 (2022)
48. Zaheer, M.Z., Lee, J.h., Astrid, M., Lee, S.I.: Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In: CVPR. pp. 14183–14193 (2020)
49. Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.I.: CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: ECCV. pp. 358–376 (2020)
50. Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.I.: Clustering aided weakly supervised training to detect anomalous events in surveillance videos. IEEE TNNLS (2022)
51. Zhang, C., Li, G., Qi, Y., Wang, S., Qing, L., Huang, Q., Yang, M.H.: Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In: CVPR. pp. 16271–16280 (2023)
52. Zhang, J., Zhang, T., Dai, Y., Harandi, M., Hartley, R.: Deep unsupervised saliency detection: A multiple noisy labeling perspective. In: CVPR. pp. 9029–9038 (2018)
53. Zhang, X., Ge, Y., Qiao, Y., Li, H.: Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: CVPR. pp. 3436–3445 (2021)
54. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: ACM MM. pp. 1933–1941 (2017)
55. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: CVPR. pp. 1237–1246 (2019)
56. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. *NeurIPS* **16**, 321–328 (2003)
57. Zhuang, F., Moulin, P.: Deep semi-supervised metric learning with mixed label propagation. In: CVPR. pp. 3429–3438 (2023)