Localizing Anomalies From Weakly-Labeled Videos

Hui Lv, Chuanwei Zhou, Zhen Cui, *Member, IEEE*, Chunyan Xu, Yong Li, and Jian Yang

***Abstract*— Video anomaly detection under video-level labels is currently a challenging task. Previous works have made progresses on discriminating whether a video sequence contains anomalies. However, most of them fail to accurately localize the anomalous events within videos in the temporal domain. In this paper, we propose a Weakly Supervised Anomaly Localization (WSAL) method focusing on temporally localizing anomalous segments within anomalous videos. Inspired by the appearance difference in anomalous videos, the evolution of adjacent temporal segments is evaluated for the localization of anomalous segments. To this end, a high-order context encoding model is proposed to not only extract semantic representations but also measure the dynamic variations so that the temporal context could be effectively utilized. In addition, in order to fully utilize the spatial context information, the immediate semantics are directly derived from the segment representations. The dynamic variations as well as the immediate semantics, are efficiently aggregated to obtain the final anomaly scores. An enhancement strategy is further proposed to deal with noise interference and the absence of localization guidance in anomaly detection. Moreover, to facilitate the diversity requirement for anomaly detection benchmarks, we also collect a new traffic anomaly (TAD) dataset which specifies in the traffic conditions, differing greatly from the current popular anomaly detection evaluation benchmarks. The dataset and the benchmark test codes, as well as experimental results, are made public on http://vgg-ai.cn/pages/Resource/ and https://github.com/ktr-hubrt/WSAL. Extensive experiments are conducted to verify the effectiveness of different components, and our proposed method achieves new state-of-the-art performance on the UCF-Crime and TAD datasets.**

***Index Terms*— Anomaly detection, anomaly localization, weak supervision, traffic anomaly dataset.**

I. Introduction

ANOMALY detection, which aims to recognize those behaviors or appearance patterns that do not conform to usual patterns [1]–[3], is of great importance for the alarm of potential risks or dangers. With the large-scale deployment of surveillance, an urgent requirement of intelligent systems is to automatically filter out possibly abnormal events.

Manuscript received August 16, 2020; revised February 7, 2021 and March 8, 2021; accepted April 1, 2021. Date of publication April 19, 2021; date of current version April 26, 2021. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20190019, in part by the National Natural Science Foundation of China under Grant 62072244 and Grant 61972204, in part by the Natural Science Foundation of Shandong Province under Grant ZR2020LZH008, and in part by the State Key Laboratory of High-end Server & Storage Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhao Zhang. *(Corresponding author: Zhen Cui.)*

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: hubrthui@njust.edu.cn; cwzhou@njust.edu.cn; zhen.cui@njust.edu.cn; cyx@njust.edu.cn; yong.li@njust.edu.cn; csjyang@njust.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3072863

Anomaly detection is typically tackled under constrained supervision that only normal data or limited annotations are provided in the training phase [4]–[9]. As anomalous events rarely happen in real-life situations, which brings in the scarcity of annotations, several methods [7]–[9] have been proposed to model the shared pattern among normal videos in the training phase and detect the outliers as anomalies during testing. However, these methods often fail in identifying anomalies when facing complicated or unseen scenes. Recently, researchers [4] select to leverage the video-level labels for developing robust anomaly detectors. The release of UCF-Crime dataset [4] activates this direction which encourages the detectors to take the best of the weak signals of video-level. Although a large gain has been observed in this domain [5], [6], it still lacks an efficient way to temporally localize anomalous frames.

In previous methods, the performances on the overall test set are calculated and reported as the evaluation results. However, in this case the temporal anomaly localization capability of detectors is somewhat unrevealed. Since the whole test set contains both the normal and anomaly videos, the superior performance on normal videos conceals the poor accuracy of anomaly localization within anomalous videos. To reveal the problem therein, we conduct statistic analysis on the anomaly data of UCF-Crime test set. ROC curves of two state-of-the-art (SOTA) methods, as well as ours, are plotted in Figure 1. The details of corresponding metrics can be found in Section IV-B. A test sample (video name: *Burglary079*) is also shown in the left part of the figure. We can find that the localization accuracy of the two methods on anomalous videos are 54.25% and 59.02% respectively, in term of AUC. It is worth mentioning that an AUC of 50% can be obtained by random binary prediction of anomalies. To sum up, there exists a large space for improving the temporal localization of anomalies.

To facilitate the localization property of anomaly detection, we propose a Weak-Supervised Anomaly Localization (WSAL) method to detect anomalies with video-level labels. In our WSAL model, we investigate into two aspects of the anomaly, which are the semantic and context. The anomalies are defined as the uncommon activities that differ from the usual pattern. Thus, the extracted semantics can act as a direct cue to infer anomalies. Based on this point, existing methods [4], [5] treat each video as frame-by-frame images or direct optical flows and extract fine-grained semantic representations for further anomaly detection. While in this manner, the temporal evolution across consecutive frames is not adequately exploited. For example, in the long temporal domain, a sudden change of the dynamic variation uncovers the anomaly itself.

On the other hand, owing to rough supervisory signals of video level, anomaly detectors are prone to false alarms or missed detection. For instances, the drastic environment changes as well as noise interruptions caused by hardware failure may lead to unwanted high probabilities from anomaly detectors. These influences in long and untrimmed videos ought to be suppressed or excluded from the anomalies. Toward this end, we put forward a noise stimulation strategy to tackle inevitable interference lying in untrimmed videos, whose quality can not be guaranteed. Moreover, we introduce hand-crafted anomalies, similar to actual anomalies, to provide pseudo location signals as guidance for the model learning process. Above two strategies make up for our enhancement strategy to boost the weakly-supervised learning and strengthen the robustness of anomaly detection. Thoroughly, we equip raw video data with the augments of video noises and hand-crafted anomalies. As a consequence, the weak labels are expanded with pseudo location signals as auxiliary.

So far, there are few datasets available for anomaly detection, most of them are with small-scale or constrained scenarios, like UCSD Peds [10], Avenue [11], ShanghaiTech [8], and Street Scene [12]. Also, these datasets are initially used for semi-supervised anomaly detection with normal training samples. For the problem under video-level scenario, only UCF-Crime [4] dataset is now available publicly to our knowledge. Thus, we build a new large-scale traffic anomaly detection (TAD) dataset with long surveillance videos under traffic scene. The proposed dataset consists of realistic anomalies on roads with various appearance and motion pattern, which facilitates the diversity requirement for anomaly detection benchmarks. In addition, we implement and compare different SOTA anomaly detection approaches on the UCF-Crime and our TAD dataset. We hope the newly collected benchmark will boost the development of anomaly detection in research domain and real-life application. The main contributions of this paper are as follows:

1) Deeply delving into anomaly detection, we propose a weak-labeled anomaly localization method, in which we employ a high-order context encoding model to encode temporal variations as well as high-level semantic information for weak-supervised anomaly detection;
2) We introduce a weak-supervision enhancement strategy by stimulating video noises and building virtual indicative locations to suppress or exclude those interruption of false-anomaly signals;
3) We build a new weak-labeled traffic anomaly detection dataset with extensive benchmark tests, and report the new SOTA results on the proposed TADdataset as well as the UCF-Crime dataset.

The rest parts of the paper are organized as follows: In Section II, we review the literature of anomaly detection in surveillance videos. In Section III we introduce the proposed WSAL method in details. In Section IV, we conduct experiments to compare our proposed method with other SOTA methods as well as elaborated ablation studies to fully analyze different components. Finally, we conclude the paper in Section V.

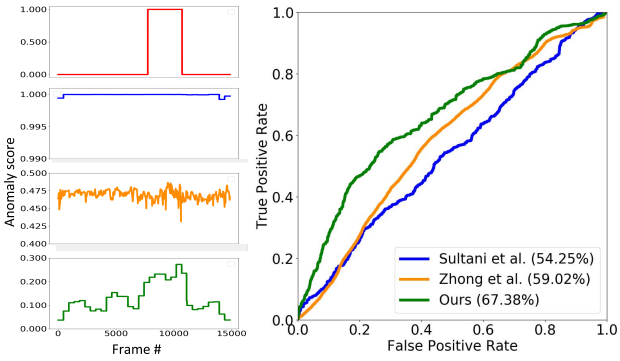


Fig. 1. Anomaly localization comparisons. **Left:** A comparison of *Burglary* case on UCF-Crime (x-axis corresponds to frames and y-axis corresponds to the anomaly score.). Groundtruth is shown in the top-left, following by three methods: Sultani *et al.* [4], Zhong *et al.* [5] and ours. **Right:** ROC curves of frame-level anomaly localization on all anomaly videos.

II. Related Work

The techniques of anomaly detection in surveillance videos have long been developed as a tool for mining unusual patterns in videos [10], [13]–[16]. The family can be divided into two categories, based on how and how much supervision is accessible. The details are discussed in the following.

Video anomaly detectors are originally designed in an unsupervised manner [9], [17]–[20] that only normal samples are available in the training phase without any labels. They first involve modeling normal behavior and then detecting samples that deviate from it. Motion trajectory, as one of common basic factors, has been utilized to detect anomalies in [15], [21], [22]. Although such methods can be easily implemented and have a fast execution speed, tracking is prone to failure in crowded or cluttered scenes. An alternative approach is to tackle the original task as a problem of novelty detection, e.g., sparse coding [11], [13], [23], distance-based methods [24], the mixture of dynamic models on texture [25] and the mixture of probabilistic PCA [26]. These models are generally built on the low-level features (e.g., a histogram of oriented gradients (HOG) and the histogram of oriented flows (HOF)) extracted from densely sampled image patches. There are also works that improve the tradition approach into VAD, such as in [27], the authors proposes a spatial localization constrained sparse coding approach for anomaly detection in traffic scenes, which fuses these two aspects of motion orientation and magnitude to obtain a robust detection result. Several recent approaches have investigated the learning-based features using autoencoders [28], [29], which minimize reconstruction errors on the normal patterns in the training process. Shi *et al.* [30] have proposed to modify original LSTM with ConvLSTM and used it for precipitation forecasting. Liu *et al.* [7] have designed a future prediction network to infer the coming frames and detect anomalies according to the quality of predicted frames. Despite the advances in developing unsupervised anomaly detection approaches, these detectors are easily to fall down when dealing with complicated or unseen environments.

Recently, great advance has been witnessed in weak supervision, for example in [31], the authors introduce weak supervision into adversarial domain adaptation for improving the

segmentation performance from synthetic data to real scenes. Inspired by them, various methods based on weak supervision situation have been introduced, they employ both normal and abnormal data along with video-level annotations for building robust anomaly detection model [4]–[6], [32], [33]. Among them, Multiple Instance Learning (MIL) is introduced for pattern modeling under weak supervision [4], [6], [33]. Sultani *et al.* [4] consider anomaly detection as a MIL problem with a novel ranking loss function. Later, by extending it, Zhu *et al.* [6] introduce the attention mechanism for better localizing anomalies. Due to the absence of anomaly positions in training phase, these two methods cannot predict anomaly frames well. For this, Zhong *et al.* [5] attempt to construct supervised signals of anomaly positions through iteratively refining them. However these methods focus on predicting segment labels while neglecting modeling hidden temporal context information. Temporal or context aggregation technology has been widely adopted as in [34] the authors adopt a volumetric structure to effectively synthesize spatiotemporal information of the same target from the current time and history frames to enhance detection. In [35], the authors propose two nuclear- and L2,1-norm regularized neighborhood preserving projection methods for extracting representative 2D image features. While in our work, we not only propose a high-order context encoding structure for temporal context aggregation but also modeling variations through the video sequences as a dynamic cue and incorporate it with semantic cues to better localize anomalies. Besides, we introduce a weak-supervision enhancement strategy to suppress false-anomaly signals.

III. The Proposed Method

In this section, we will introduce our Weak-Supervised Anomaly Localization (WSAL) method in details. We first give the basic formulation for the anomaly localization and core modules of our WSAL are elaborated thoroughly then.

A. Formulation

The purpose of anomaly detection is to estimate the anomaly status of a video and localize the anomalies in the video sequence if exist. In the weakly supervised scenario, a video sequence $\mathcal{X}$ and its corresponding video-level annotation y $\in \{0, 1\}$ are given, where the case 'y=1' means there exists anomaly in this sequence otherwise 'y=0' indicates that there is no anomaly in $\mathcal{X}$. We start with dividing the entire video into several segments with equal lengths, denoted as $\mathcal{X} = (\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_m)$. The goal of video segmenting is to alleviate computation burden resulting from almost-repetitive video frames. For the i-th segment $\mathcal{X}_i$, we first use a classical convolutional network to extract features for each frame, and the segment feature x_i is obtained by aggregating the features of all frames within the segment. As a consequence, the sequence $\mathcal{X}$ could now be represented by the m-tuple features $(\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_m)$. We can now use this m-tuple to determine whether the current video contains any anomaly or not, in the manner that assigning each segment in the video with an anomaly score, indicating the probability of being anomalous.

To predict the state (normal or abnormal) of a video, we derive a novel function to describe the video by estimating the anomalous margin among a video, formally,

$$\mathcal{S}(\mathcal{X}) = \max_{i,j=1,\ldots,m} f(\psi(\mathrm{x}_{i-k}, \ldots, \mathrm{x}_i, \ldots, \mathrm{x}_{i+k}), \psi(\mathrm{x}_{j-k}, \ldots, \mathrm{x}_j, \ldots, \mathrm{x}_{j+k})), \quad (1)$$

where

- ψ is a high-order function that encodes an anchored segment as well as its adjacent $2k$ segments in the temporal context. To mine the anomalies, we consider two aspects of information: spatial semantics and temporal variations. The function ψ is modeled with a high-order dynamic regression to generate semantic features and predict variations within local window $[-k, k]$. Please see Section III-B for more details.
- f is a margin distance metric measuring the anomaly score margin between the segment position i and j. The more close the predicted anomaly scores are, the smaller the distance is.
- $\mathcal{S}(\cdot)$ is the score of a video that computes the maximum relative distance of pairwise positions. The scores of normal videos are expected to be smaller than anomalous videos. Thus, the maximum-distance strategy constrains entire normal videos more smooth than anomaly videos, which complies with the conventional assumption.
- max function is chosen to capture the largest score margin, which can represent the extent of abnormalities in a video. Since anomaly scores are all close to zero in a normal video, leading to the score margin with a small value. While in an anomalous video, the anomalies, lying in normal background, will lead to a large score margin.

Given a batch of training data $\{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \cdots, \mathcal{X}^{(n)}\}$ and the corresponding video labels $\{y^{(1)}, y^{(2)}, \cdots, y^{(n)}\}$, we define a margin loss function as:

$$\zeta_1(\{\mathcal{X}^{(i)}\}|_{i=1}^{n}) = \max\ \{0,\ 1 - \frac{1}{n_1}\sum_{i=1}^{n}[\mathcal{S}(\mathcal{X}^{(i)})|y^{(i)} = 1] + \frac{1}{n_0}\sum_{j=1}^{n}[\mathcal{S}(\mathcal{X}^{(j)})|y^{(j)} = 0]\ \}, \quad (2)$$

here n_1, n_0 are the total amounts of anomaly and normal samples. As the function only depends on video-level labels, the learning process belongs to the case of weak supervision.

In addition, we augment training samples to generate two types of data: noise data $\{\ddot{\mathcal{X}}^{(i)}\}|_{i=1}^{\ddot{n}}$ and pseudo-location data $\{\check{\mathcal{X}}^{(i)}\}|_{i=1}^{\check{n}}$, where $\ddot{n}$ and $\check{n}$ are the amounts of pseudo samples. The former could help the detector reduce mis-judgement where some noised normal videos are predicted as anomaly labels, whilst the latter provides direct guidance to localize anomalous frames. Let $\{\mathcal{X}'\}|_{i=1}^{n'} = \{\ddot{\mathcal{X}}^{(i)}\}|_{i=1}^{\ddot{n}} \cup \{\check{\mathcal{X}}^{(i)}\}|_{i=1}^{\check{n}}$ denote all augmentation samples, where $n' = \ddot{n} + \check{n}$. Finally, we derive the objective function to optimize as:

$$\zeta = \zeta^{\mathcal{O}}(\{\mathcal{X}^{(i)}\}|_{i=1}^{n}) + \lambda\zeta^{\mathcal{A}}(\{\mathcal{X}'^{(i)}\}|_{i=1}^{n'}), \quad (3)$$

where λ is the balance factor between the original and the augmented data. Loss function $\zeta^{\mathcal{O}}$, defined on the original weak-labeled data, uses the margin loss ζ_1 in Equation (2), and the details will be listed in Section III-B. Loss $\zeta^{\mathcal{A}}$ is imposed on noise data as well as pseudo-location data, which will be introduced in Section III-C.

In testing process, given a video, we obtain the anomaly status of each segment by aggregating the consensus of spatial semantics and dynamic variations defined in the following.

B. High-Order Context Encoding

Previous approaches [4]–[6] directly infer the anomaly scores from input visual features in an intuitive way, while neglecting the guidance of the temporal context for anomaly localization. Intuitively, the rarely occurred anomalies among the normal patterns will lead to significant changes in the time domain. Therefore, the dynamic variations in the time series are able to indicate the existence of anomalies. Inspired by this, we propose to leverage the temporal context information for the immediate spatial semantics and dynamic temporal variations, and aggregate both cues for accurately locating anomalies.

In the beginning, we design a High-order Context Encoding (HCE) model to extract high-level semantic features and encode the variations in time series. The input is the feature vectors $(\mathrm{x}_1, \cdots, \mathrm{x}_m)$ extracted from consecutive segments. The regression process is formulated as:

$$\tilde{\mathrm{x}}_t = \sum_{j=-k,\cdots,k,\ j\neq 0} \mathbf{W}_j \mathrm{x}_{t+j} + \mathbf{W}_0 \mathrm{x}_t + \mathbf{b}, \tag{4}$$

where $\mathbf{W}_j$ is a projection function on the j-th segment, $\mathbf{b}$ is a bias term. The output encodes the context information of the anchored segment and adjacent segments, i.e., $(\tilde{\mathrm{x}}_{t-k}, \cdots, \tilde{\mathrm{x}}_{t-1}, \mathrm{x}_t, \tilde{\mathrm{x}}_{t+1}, \cdots, \tilde{\mathrm{x}}_{t+k})$. The intuition is that t-th high-order feature vector collects the fruitful information from its $2k$ neighbors, which can facilitate both the mining of immediate spatial semantics and local dynamic variations. Actually the regression can be stacked as a hierarchical structure by taking the output $\tilde{\mathrm{x}}$ as the input in a recursive manner. In practice, we find the simple one-layer regression can perform well.

The neighbor size k controls the temporal context modeled in each local segment $\tilde{\mathrm{x}}_t$. Then to exploit the immediate semantic information of the anchored segment, we use a fully connected layer, activated by a Sigmoid function, to obtain an anomaly score. Formally:

$$\psi^{sem}(\tilde{\mathrm{x}}_t) = \sigma(\mathbf{w}_{sem}\tilde{\mathrm{x}}_t + b_{sem}), \tag{5}$$

where $\psi^{sem}(\tilde{\mathrm{x}}_t)$ represents the semantics score, w_{sem} and b_{sem} are the weight and bias of the fully connected layer and σ stands for the sigmoid function.

To measure the variation between two adjacent segments, we take the cosine similarity measurement: $\cos(\tilde{\mathrm{x}}_{t-1}, \tilde{\mathrm{x}}_t) = \tilde{\mathrm{x}}_{t-1}^{\top}\tilde{\mathrm{x}}_t / (\|\tilde{\mathrm{x}}_{t-1}\|^2 \|\tilde{\mathrm{x}}_t\|^2)$. The corresponding distance metric is $1 - \cos(\tilde{\mathrm{x}}_{t-1}, \tilde{\mathrm{x}}_t)$, which has a large value for dramatic variations. Then the second-order discrepancy of local variations is computed as an indicator of anomaly, which becomes:

$$\psi^{var}(\tilde{\mathrm{x}}_t) = (2 - \cos(\tilde{\mathrm{x}}_{t-1}, \tilde{\mathrm{x}}_t) - \cos(\tilde{\mathrm{x}}_t, \tilde{\mathrm{x}}_{t+1})))/4, \tag{6}$$

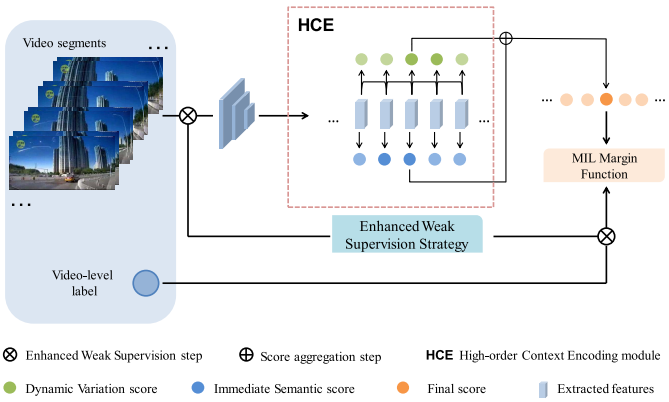


Fig. 2. Framework of WSAL. Video clips are organized in segment level and inputted into the backbone model. These extracted features are processed in HCE module to generate anomaly scores from the cues of immediate semantics and dynamic variations. Then the predicted scores are aggregated and supervised in a novel MIL Margin objective function using the video-level labels. In weak supervision, only video-level annotations are available, which lacks accurate temporal location guidance. Motivated by this, we introduce Enhanced Weak Supervision strategy for data augmentation and generating pseudo anomaly signals. Better viewed in color.

where we make the score value divided by four to normalize the scalar into [0, 1].

Then, we obtain the singularity of a sequence from the dual context cues, with the margin measurement f as L1-distance:

$$\mathcal{S}^{sem}(\mathcal{X}) = \max_{i,j=1,\cdots,m} |\psi^{sem}(\tilde{\mathrm{x}}_i) - \psi^{sem}(\tilde{\mathrm{x}}_j)|, \tag{7}$$

$$\mathcal{S}^{var}(\mathcal{X}) = \max_{i,j=1,\cdots,m} |\psi^{var}(\tilde{\mathrm{x}}_i) - \psi^{var}(\tilde{\mathrm{x}}_j)|. \tag{8}$$

By plugging above singularity tuple into Equation (2), the acquired margin losses of the dual context are denoted as ζ_1^{sem} and ζ_1^{var}. Since the scores of normal events are targeted to 0, and those of anomalous are sparse (scarce of anomalies), we place a sparsity constraint on the loss function. Added with the sparsity constraint of weight β, the margin loss of dual context becomes:

$$\zeta^{\mathcal{O}} = \zeta_1^{sem}(\{\mathcal{X}^{(i)}\}|_{i=1}^n) + \zeta_1^{var}(\{\mathcal{X}^{(i)}\}|_{i=1}^n) + \frac{\beta}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}(|s_t^{sem}| + |s_t^{var}|). \tag{9}$$

C. Enhanced Weak Supervision

1) Noise Simulation: As is mentioned in Section I, noises in videos lead to serious interference for anomaly detection, especially localization. Due to the unavoidable external factors, it tends to exist noisy artifacts such as lens jitter in the videos which is going to result in misjudgments. To mitigate this issue, we introduce a noise simulation strategy in which we fuse the raw videos with varying degrees of video noises, such as blur, picture interruption as well as lens jitter. Specifically, we augment the normal video sequences with three kinds of video noise simulations, which are motion blur (kernel size: 5, angle: [−45°, 45°]), black/blue/purple blocks ([1/4, 1] of raw image size) and random scale (−20 to +20% on x- and y-axis independently)). We randomly choose m segments in a normal video sequence to augment and the augmented data are still treated normal.

Given the simulating noise data $\{\ddot{\mathcal{X}}^{(i)}\}|_{i=1}^{\ddot{n}}$ and the corresponding label set $\{\ddot{y}^{(i)}\}|_{i=1}^{\ddot{n}}$, we apply a supervised constraint on the predicted anomaly states $\{\ddot{s}_t^{(i)}\}|_{i=1}^{\ddot{n}}$:

$$\zeta^{nse}(\{\ddot{\mathcal{X}}^{(i)}\}|_{i=1}^{\ddot{n}}) = \frac{1}{\ddot{n}}\sum_{i=1}^{\ddot{n}}\sum_{t=1}^{m}(\ddot{s}_t^{(i)} - \ddot{y}^{(i)})^2, \tag{10}$$

$$\text{s.t. ,}\ \ddot{s}_t^{(i)} = \frac{1}{2}([\ddot{s}_t^{(i)}]^{sem} + [\ddot{s}_t^{(i)}]^{var}). \tag{11}$$

2) Hand-Crafted Anomaly: The noise simulation strategy introduced above is able to alleviate the false alarm for normal videos. However, for anomaly videos, there still lacks enough data for model training. In particular, there exists no explicit location supervision in those anomalous videos which brings in great challenge for effective anomaly localization. To mitigate this issue, we then introduce hand-crafted anomaly to boost the anomaly localization performance via creating explicit location instructions for anomaly localization. We name hand-crafted anomalies as pseudo-location data $\{\check{\mathcal{X}}^{(i)}\}|_{i=1}^{\check{n}}$. Specifically, we first randomly choose a pair of normal and abnormal videos. Then several random segments of the normal video are selected and fused with several segments from the abnormal video. Finally, the obtained segments are combined with the remaining normal video segments to form a pseudo anomalous sequence. Those fused segments are targeted to be abrupt in the obtained sequence to create a pseudo anomalous sample since the substitutes differ from the distribution of the original normal video due to different scenes.

To decrease the abrupt in the fused sequence, we fused the features of abnormal segments with the normal ones. The coefficient of the anomaly feature ranges in [0.2,0.5] and were randomly generated during the training process. Rather than simply assigning a fixed score (e.g., 1) for the simulated abnormal video will bring in a degenerate solution because the signal can encourage the remaining normal segments to have a high anomaly score along with pseudo-location data. To mitigate the issue above, we propose a simple yet effective skill by barely pushing the fused segments $\{\check{\mathcal{X}}_i\}|_{i=1}^{\check{n}}$ to have a higher score than the others. The supervision constraint is derived as:

$$\zeta^{loc}(\{\check{\mathcal{X}}_i\}|_{i=1}^{\check{n}}) = \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\sum_{t\in\mathcal{I}} \max(0, \check{s}_t^{(i)} - \max_{j\notin\mathcal{I}}\{\check{s}_j^{(i)}\}), \tag{12}$$

where $\check{s}_t^{(i)}$ denotes the anomaly score estimated by HCE module and $\mathcal{I}$ is a collection of the indexes for those pseudo location segments of the hand-crafted anomalies. Integrating the above two augmentation techniques, the objective function of weak supervision enhancement strategy becomes:

$$\zeta^{\mathcal{A}} = \zeta^{nse}(\{\ddot{\mathcal{X}}^{(i)}\}|_{i=1}^{\ddot{n}}) + \zeta^{loc}(\{\check{\mathcal{X}}_i\}|_{i=1}^{\check{n}}). \tag{13}$$

Combining Eqn. 9 and Eqn. 13, we finally arrive at the overall objective function which is denoted by Eqn. 3.

D. Traffic Anomaly Detection (TAD) Dataset

So far, most existing video anomaly datasets are prepared for unsupervised case, e.g., UCSD Pedestrian 1&2 [10], Subway Entance & Exit [2], Avenue [11], etc. These unsupervised datasets are either small in scale or under the constraint of limited scenes. For example, videos in Avenue are short and some of the anomalies are performed by actors (e.g., throwing paper), which are unrealistic. Different from them, UCF-Crime [4] dataset is a newly released large-scale dataset proposed for weak supervision case. Long untrimmed surveillance videos, covering 13 real-world anomalies, are collected in the dataset. It has a total of $1,900$ surveillance videos, which consists of $1,610$ training videos and 290 test videos. Note that only video-level annotations are provided in the training set, and frame-level annotations are available for evaluation on the test set. The comparison of video anomaly detection datasets are shown in Table I.

TABLE I
A COMPARISON OF ANOMALY DETECTION DATASETS

Dataset	Target Domain	# Videos	Total Frames	View	Supervision
UCSD Ped 1/2 [10]	Campus	98	18,560	3rd-person	Unsupervised
CUHK Avenue [11]	Campus	37	30,652	3rd-person	Unsupervised
Street scene [12]	Street	81	203,251	3rd-person	Unsupervised
Shanghai Tech [8]	Campus	81	317,398	3rd-person	Unsupervised
UCF-Crime	General	1900	13,769,300	3rd-person	Weakly-supervised
Aadv [37]	Traffic	1750	175,000	1st-person	Fully-supervised
A3D [36]	Traffic	1500	208,166	1st-person	Unsupervised
Ours	**Traffic**	**500**	**540,272**	**1st&3rd-person**	**Weakly-supervised**

Although the datasets mentioned above have greatly promoted the development of anomaly detection methods, there still lacks benchmarks of enough diversity for evaluation. To further meet the benchmark diversity requirement, we here propose a new anomaly detection dataset which specifies in the traffic scenes, differing greatly from the datasets mentioned before. Traffic video monitoring plays an essential role in early warning and emergency assistance for car accidents. It is an urgent need to design effective anomaly detection systems for surveillance videos on roads. In traffic scenes, many factors, such as the vehicles moving at a high speed and various road conditions, add up to the hardness of anomaly detection. So far, there is not any specific dataset for traffic anomaly detection. Although UCF-Crime contains road accidents videos, most of anomalies in traffic scenarios are not covered in this dataset. Basically, a large-scale and complex dataset is of great importance for devising and evaluating various methods. It is out desire to push the study of anomaly detection towards the usage in real traffic application.

To date, there are also public-available anomaly datasets on traffic scenes, however they are with limited scenarios. For example, in [36], a video dataset of unsupervised traffic accident detection is released and authors in [37] propose a fully supervised traffic anomaly detection benchmark. Both the above datasets are made up of samples of short clips, which cannot represent the real situations that the anomalies rarely exist among a large quantity of normal data, and these datasets are not target for weak supervision. Moreover, these datasets only consist of first-person or dashboard videos, which lacks the videos captured by surveillance cameras on roads.

Hence, we are motivated to construct a new large-scale dataset under the traffic scenes for video anomaly detection under weak supervision. The collected TAD dataset consists of long untrimmed videos which cover 7 real-world anomalies

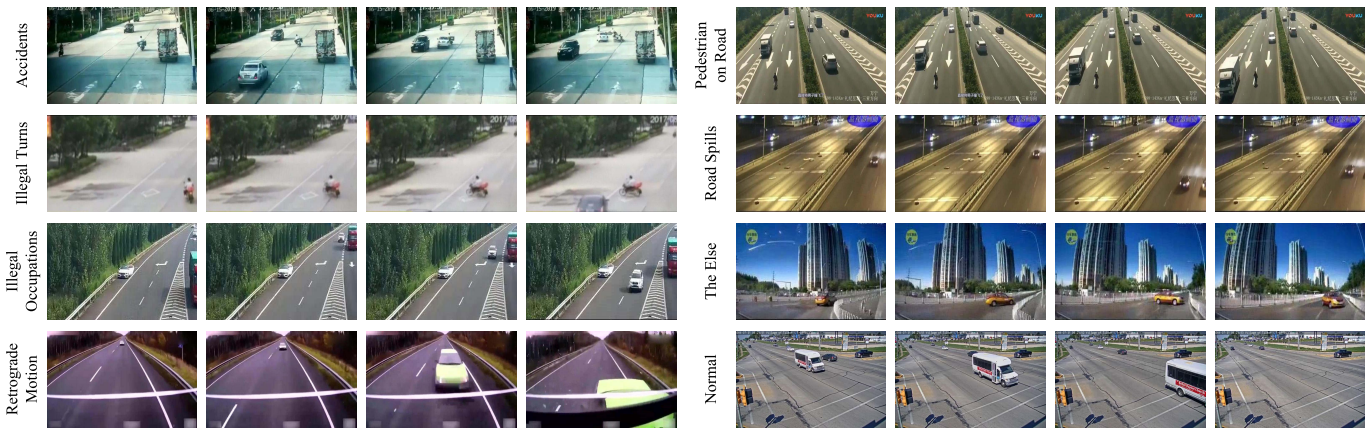


Fig. 3. Examples of different anomalies in the collected TAD dataset. Best viewed in color.

on roads, including *Vehicle Accidents, Illegal Turns, Illegal Occupations, Retrograde Motion, Pedestrian on Road, Road Spills* and *The Else* (i.e., the remaining anomalies with fewer quantity are put together as one category). Some cases of the anomalies are shown in Figure 3. The proposed dataset is comprehensive that includes realistic videos from various scenarios, weather conditions and daytime periods.

1) Data Collection: Traffic videos from various countries are collected and annotated under a detailed and unified plan. Raw videos are downloaded from YouTube or Google website. The collected videos are mostly recorded by CCTV cameras mounted on the roads. We remove videos which fall into any of the following cases: manually edited, prank videos, and containing compilation. Videos with ambiguous anomalies are also excluded.

2) Data Partition and Annotation: Our TAD dataset contains a total of about 25 hours videos, average 1075 frames per clips. The anomalies randomly occur in each clip, about 80 frames average and there are one to two random anomalies in a video sequence. Finally, 500 traffic surveillance videos are saved and annotated for anomaly detection, with 250 abnormal and normal videos respectively. The whole dataset is randomly partitioned into two parts: training set with 400 videos, and test set with 100 videos. Both training and test sets contain normal and abnormal videos and all seven kinds of anomalies at various temporal locations in anomalous videos. Following the setting of weak supervision as [4], the training set is equipped with video-level annotations, and frame-level annotations are provided for the inference set.

Our proposed TAD dataset contains totally different abnormal scenarios than the current benchmarks. We believe that it could be used to better evaluate the effects of different anomaly detection algorithms from another perspective. We hope our TAD dataset could serve as a standard benchmark for better promoting the development of anomaly detection methods.

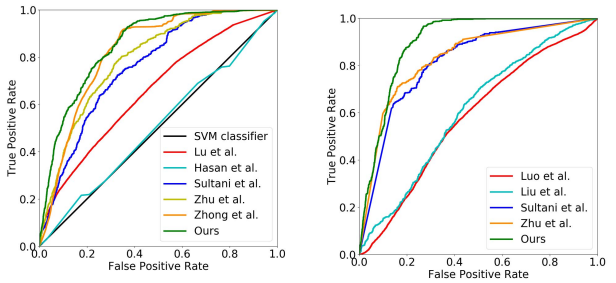
IV. Experiments

A. Implementation Details

As in [5], we adopt the Temporal Segment Network (TSN) [38], which is a powerful action feature extractor, as our backbone net. We use the BN-Inception version of TSN to extract features for our proposed WSAL method. We extract features from the global average pooling layer (1024-dim). For the UCF-Crime dataset, we use the model weights finetuned on UCF-Crime as in [5] to extract features. While on our TAD dataset we only use the model weights pretrained on Kinetics-400 dataset. We first divide each video into 32 non-overlapping segments empirically as in previous works [4]–[6] for a fair comparison. Hence, for each video, we have a 32×1024 feature matrix. In our VAD model, the input features are first run through 2 fully connect layers with 512 and 128 dimension, respectively. Then, the dimensions of fully connect layers in HCE module and the immediate semantic score are 128 and 1, respectively. Dropout operations of 60% rate are implemented after each fully connect layer, except the fully connect layer for immediate semantic score. During the training phase, we randomly select 30 positive and 30 negative bags as a mini-batch. We employ Adagrad [39] optimizer with the initial learning rate of 0.001. The parameter of sparsity constraint in the margin loss is set to $\beta = 0.00008$ as in [4], [6] and the weight of strategy for weak supervision enhancement is set to $\lambda = 1.0$ for the best performance. We train the model for a total of 3K iterations, decrease the learning rate by half at 1.2K, 2.4K and stop at 3K. All hyper-parameters are the same for both UCF-Crime and TAD datasets.

B. Evaluation Metrics

For anomaly detection [7], [25], Receiver Operation Characteristic (ROC) is used as a standard evaluation metric. It is calculated by gradually changing the threshold of regular scores on the predicted anomaly scores. Then the Area Under Curve (AUC) is accumulated to a score for the performance evaluation. A higher value indicates a better anomaly detection performance. Following the previous works [4]–[6], we apply ROC curves and frame-level AUC for anomaly detection performance comparison. Due to the lack of frame-level annotations on the training split and verification split for ablation studies, we use video-level AUC as the measurement for tuning the hyper-parameters. In addition, we also use the ROC



(a) ROC curves on UCF-Crime (b) ROC curves on our TAD

Fig. 4. ROC curves with various anomaly detection methods on the UCF-Crime dataset and TAD dataset.

and AUC on the anomaly subset to serve as the evaluation metric for anomaly localization ability.

C. Comparison With SOTA Methods

1) On UCF-Crime Dataset: For fair comparison, we reproduce the methods of [4] and [5] by running their publicly released codes. Other statistical results are drawn from the work [4]. We compare our WSAL with several anomaly detection methods. Specifically, a binary SVM classifier is set as the baseline method. In this case, the anomalous and normal videos are treated as two separate class. Models from Lu *et al.* [11] and Hasan *et al.* [17] are two unsupervised methods, training with the normal videos in UCF-Crime training set. The remaining Sultani *et al.* [4], Zhu *et al.* [6] and Zhong *et al.* [5] are SOTA weakly-supervised methods. As shown in Table II, on the whole test set which contains both the normal and abnormal videos, we boost the best performance of overall AUC from the 82.12% to 85.38% by a large margin. In Figure 5(a), we plot the ROC curves of SOTA methods on the whole UCF-Crime dataset and it vividly shows the superoirity of our proposed WSAL method over other SOTA methods. As for the Anomaly subset, our proposed method exceeds the SOTA detectors by 9% over [5] and 13% over [4], achieving a significant progress on the anomaly localization perspective.

We draw the following conclusions upon above experimental results: 1) SVM classifier fails to distinguish the anomalous and normal videos, mainly because the normal patterns take the dominate position in both normal and anomalous videos, and make the classifier difficult to capture the rare anomalies; 2) By encoding the normal patterns and building the corresponding semantic boundary, unsupervised methods [17] and [11] achieve better results than SVM classifier; 3) Owing to the benefits of weak labels, the performances of weakly-supervised methods [4], [6] and [5] are superior than above approaches. Nevertheless, previous weakly-supervised methods infer the anomaly status from high-level semantic features intuitively, while neglecting an important property of the anomaly, which is the dynamic evolution lying in time series. The considerable gain in anomaly localization promotes the improvement of overall anomaly detection accuracy. Some visual results of our method on test cases are shown in Figure 5. Compared with the other two SOTA methods, a large degree of distinction between the normal and anomalous can be achieved by our approach. As a result, a superior performance is achieved with the better anomaly-discriminating capability.

TABLE II

QUANTITATIVE COMPARISON ON THE UCF-CRIME DATASET. * SYMBOL INDICATES THE METHOD IS TRAINED WITH NORMAL VIDEOS ONLY

Method	Overall AUC(%)	Anomaly Subset AUC(%)
SVM	50	50
Hasan et al.* [17]	50.60	-
Lu et al.* [11]	65.51	-
Sultani et al. [4]	75.41	54.25
Zhu et al. [6]	79.10	62.18
Zhong et al. [5]	82.12	59.02
Ours	**85.38**	**67.38**

2) On the Proposed TAD Dataset: To compare the performance of different methods under other circumstances, we conduct comparison experiments on the TAD dataset. We compare our WSAL model with four SOTA anomaly detection methods, including two unsupervised methods (Luo *et al.* [8] and Liu *et al.* [7]) and two weakly-supervised methods (Sultani *et al.* [4] and Zhu *et al.* [6]) For unsupervised models, we follow their implementation and train the models on the training subset where only normal videos are provided. All models are re-trained with the same features extracted using TSN, except [7] which takes the RGB frames as inputs.

The quantitative comparisons of AUC are revealed in Table III and the corresponding ROC curves are drawn in Figure 5(b). Similar as upon UCF-Crime, weakly-supervised methods are able to obtain much better performance than unsupervised ones. As both normal and abnormal training samples are provided, the weakly-supervised methods own much better understanding of the intrinsic nature of anomaly. It demonstrates that 1) weak annotations with a little cost of time and work can greatly benefit the performance of VAD methods, and 2) the collected dataset with various scenes and different kinds of anomalies are comlex and extremely challenging for unsupervised methods, compared with current unsupervised benchmark, such as Ped 1/2, Avenue, etc., which could activate future research direction for unsupervised methods. In addition, our WSAL method also achieves better performance with a gain of 6% AUC over previous SOTA [6]. The prominent advances on the two large-scale and comprehensive benchmarks prove the superiority of our method on detecting and localizing anomalies.

D. Ablation Studies

To comprehensively study the impact of different components we proposed, we conduct various ablation studies in this part. All experiments are conducted on the UCF-Crime dataset, all hyper-parameters are kept the same as the WSAL method if not otherwise claimed.

1) Analysis of the Dual Context Ensemble: To verify the effectiveness of the proposed ensemble mechanism of immediate semantics and dynamic variations, we construct three variants of the proposed WSAL method where only the immediate semantics or the dynamic variations or both of them are adopted. Detailed comparisons are presented in the first three lines of Table IV. When only immediate semantics

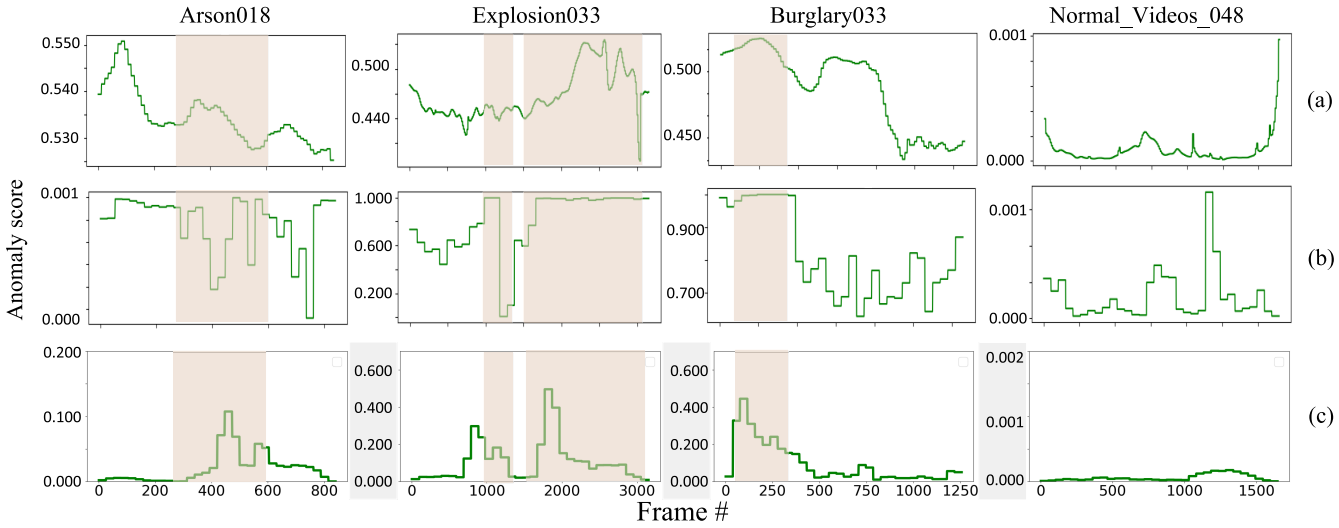


Fig. 5. Visualization of predictions on the UCF-Crime test cases. The x-axis denotes the video frame # and y-axis is corresponding to the anomaly score. In the figure, (a), (b), and (c) denotes the results of [4], [5] and our model, respectively. The green curves are predictions of various approaches. The light orange regions are ground truth anomalies. Video names are labeled in the blank.

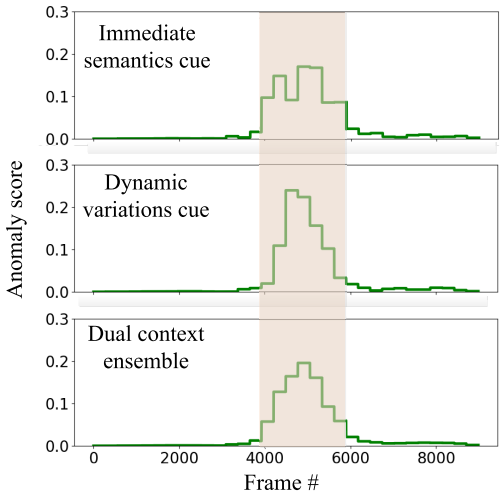


Fig. 6. An visualization case of the dual context cues on the UCF-Crime dataset. The light orange region denotes the groundtruth anomaly. From the top to the bottom, the curves represent the anomaly scores of immediate semantics cue, dynamic variations cue and consensus of above two cues, respectively. A more robust and smooth prediction is observed from the dual context model.

TABLE III

QUANTITATIVE COMPARISON ON OUR TAD DATASET. * SYMBOL INDICATES THE METHOD IS TRAINED WITH NORMAL VIDEOS ONLY

Method	Overall AUC(%)	Anomaly Subset AUC(%)
Luo et al.* [8]	57.89	55.84
Liu et al.* [7]	69.13	55.38
Sultani et al. [4]	81.42	55.97
Zhu et al. [6]	83.08	56.89
Ours	**89.64**	**61.66**

TABLE IV

ABLATION STUDIES OF OUR WSAL METHOD ON THE UCF-CRIME DATASET. THE MEANINGS OF THE ABBREVIATIONS IN THE TABLE ARE AS FOLLOWS: IS: IMMEDIATE SEMANTICS; DV: DYNAMIC VARIATIONS; HCE: HIGH-ORDER CONTEXT ENCODING; NS: NOISE SUPPRESSION; HA: HAND-CRAFTED ANOMALY

IS	DV	HCE	NS	HA	Overall AUC(%)	Anomaly Subset AUC(%)
✓					81.44	61.13
	✓				82.52	62.38
✓	✓				82.95	63.65
✓	✓	✓			84.44	64.95
✓	✓	✓	✓		84.86	66.28
✓	✓	✓		✓	84.95	66.55
✓	✓	✓	✓	✓	**85.38**	**67.38**

are exploited, the algorithm is able to achieve satisfactory accuracies of 81.44% and 61.13% w.r.t anomaly detection and anomaly localization. It means that the immediate semantics can provide useful information for the tasks, and if dynamic variations are adopted, the performances are boosted by 1.12% and 1.25% separately. One case for visualization is shown in Figure 6. There is only one clear and sharp peak in the prediction of dynamic variations cue, compared with the results of immediate semantics cue. It demonstrates that the dynamic variations are good at capturing the sudden occurrence of the anomaly even when the immediate semantics cue may bring in uncertainty. By aggregating the immediate semantics and dynamic variations cues, the detection performance is more robust under various circumstances, with a higher detection and localization accuracy. Thanks to the complementary characteristic of the immediate semantics and dynamic variations cue, which stand for different aspects of the anomaly.

2) Analysis of HCE Model: We study the influence of High-order Context Encoding on the new training and verification splits of UCF-Crime. Since only video-level labels are available in verification split, video-level AUC is measured by aggregating the segment-level model predictions as in Formula 1 and then calculating the AUC results. As to the incorporation of temporal context, an appropriate temporal window size k is critical for the final performance. We slowly increase the window size k from 0 to 3 and the results are listed in Table V. When the window size k grows, the accuracy of video-level predictions improves drastically from 0 to 1, with a performance gain of 1.6%. It means that appropriate

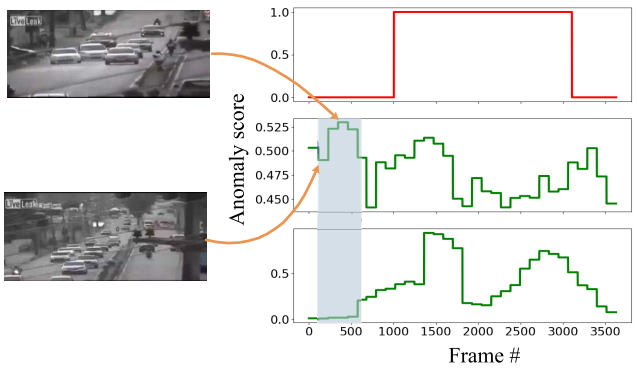


Fig. 7. An visualization case of video noises on the UCF-Crime dataset. The light blue region in the video sequence contains video noises. The red curves in the top-right denotes the groundtruth anomalies. The curve in the middle-right represents the result of baseline method [5]. The bottom-right curve belongs to the result of our method. In this case, the noise comes from the lens jitters. The drastic view change easily leads to the false detection of the basic model.

TABLE V
ANALYSIS OF THE WINDOW SIZE IN HCE MODEL ON THE UCF-CRIME DATASET

Window Size	0	1	2	3
Video-level AUC(%)	93.39	95.01	95.65	95.73

aggregation of the temporal context possesses great potential for anomaly detection. The fruitful information in the temporal neighborhood facilitates the learning of anomaly semantics, as well as the encoding of the temporal evolution. Finally, we choose the size $k = 2$ for trading off between model size and performance, since the accuracy gain slows down when the window size further increases.

Further, We conduct ablation studies for analysing impacts of high-order context encoding on various feature sources as in Table. VI. For feature extraction on I3D [40], we implement official released model; On R(2+1)D [41], we use the official 18 layer model, both the models are pre-trained on kinetics-400. As is shown, when the high-order context encoding module is implemented with input feature of I3D and R(2+1)D, the anomaly location accuracy is increased by over 3% (from 51.23% to 55.04% with I3D inputs and from 47.96% to 51.25% with R(2+1)D inputs), together with about 2% on overall accuracy. Moreover, based on the TSN feature pre-trained on UCF-Crime (by tiling video-level label for each video segment as in [5]), the proposed HCE module can still achieve remarkable progress with overall AUC from 82.95% to 84.44%. It demonstrates that although temporal aggregation has been done in the feature extraction model, our high-order context encoding is still beneficial for highlighting and capturing transient anomalous information in long-time sequences. It boosts the performance of video anomaly detection.

3) Enhanced Weak Supervision: We conduct studies to verify the effectiveness of the proposed video noise augmentation and hand-crafted anomaly separately. The detailed results are reported in Table IV. We first augment the training process by adding video noise simulations. The anomaly detection AUC is boosted from 84.44% to 84.86% and the anomaly localization accuracy is further boosted by 1.33%. One case is plotted in Figure 7, our noise stimulation strategy contributes to alleviate the interference caused by lens jitters. Note that the anomaly localization performance improvement is non-trivial. It clearly demonstrates that the proposed noise augmentation strategy is able to aid the dynamic variation module for better capturing the real anomaly and achieving better understanding of the intrinsics of anomalies. When we manually synthetic some anomalies to aid the training process, our method achieves 0.51% and 1.60% performance gain over the training strategy where no noise augmentations are adopted. The performance promotions demonstrate that our synthetic anomaly data is able to provide extra useful supervision, indicating that larger abnormal detection dataset is needed for sufficient training of abnormal detection methods. If both augmentation strategies are combined, the proposed method is able to achieve much better performance than the two separate augmentation strategies. It indicates that the proposed two augmentation strategies are beneficial for the understanding of the anomaly concept by suppressing the interference coming from the environment as well as hardware failures, and generating pseudo signals that simulating the occurrence of anomalies.

TABLE VI
ABLATION STUDIES OF VARIOUS FEATURE SOURCES ON UCF-CRIM E BENCHMARK. + SYMBOL INDICATES THE BASE MODEL IS EQUIPPED WITH HIGH-ORDER CONTEXT ENCODING MODULE

Feature Source	Overall AUC(%)	Anomaly Subset AUC(%)
I3D [40]	72.43	51.23
I3D +	74.18	55.04
R(2+1)D [41]	72.47	47.96
R(2+1)D +	75.59	51.25
TSN [38]	82.95	63.65
TSN +	84.44	64.95

4) Speed Analysis: The whole model can run at 44 FPS on a single RTX 2080Ti GPU. Among them, the feature extraction model—-TSN with BN-Inception backbone [38] runs at 45 FPS, with the input frame resolution set as 224×224. Then, based on the extracted feature, it only consumes 1.81 ms (runs at 550 FPS) to predict the anomaly scores. In summary, our model is applicable in the on-line applications.

V. CONCLUSION

In this work, we focused on anomaly localization in surveillance videos and proposed a weakly supervised anomaly localization network that deeply exploring the temporal context in consecutive segments. Our model encoded temporal dynamic variations as well as high-level semantic information, and leveraged both of them for anomaly detection and localization. Furthermore, we devised a weak supervision enhancement strategy. The accuracy of anomaly localization was greatly improved under the introduced supervision of video noise augmentation and pseudo-location data. We also collected a new traffic anomaly detection dataset for evaluating methods under realistic scenarios on roads. SOTA methods were verified on UCF-Crime dataset and our TAD dataset. The experimental results showed that the proposed anomaly detector has performed significantly better than previous methods.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[3] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2458–2465.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR)*, Jun. 2018, pp. 6479–6488.

[5] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug- and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[6] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–12.

[7] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6536–6545.

[8] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

[9] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 189–196.

[10] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[11] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2720–2727.

[12] B. Ramachandra and M. J. Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2569–2578.

[13] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3313–3320.

[14] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1446–1453.

[15] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2054–2060.

[16] B. Antic and B. Ommer, "Video parsing for abnormality detection," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2415–2422.

[17] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[18] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11996–12004.

[19] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 334–339.

[20] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.

[21] R. Bensch, N. Scherf, J. Huisken, T. Brox, and O. Ronneberger, "Spatiotemporal deformable prototypes for motion anomaly detection," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 502–523, May 2017.

[22] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[23] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3449–3456.

[24] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2112–2119.

[25] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1975–1981.

[26] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2921–2928.

[27] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017.

[28] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 56–62.

[29] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[30] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[31] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[32] K. P. Adhiya, S. R. Kolhe, and S. S. Patil, "Tracking and identification of suspicious and abnormal behaviors using supervised machine learning technique," in *Proc. Int. Conf. Adv. Comput., Commun. Control (ICAC3)*, 2009, pp. 96–99.

[33] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29573–29588, Nov. 2018.

[34] D. Li, S. Yan, M. Zhao, and T. W. S. Chow, "Spatiotemporal tree filtering for enhancing image change detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8805–8820, 2020.

[35] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust neighborhood preserving projection by nuclear/L2, 1-norm regularization for image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607–1622, Apr. 2017.

[36] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 273–280.

[37] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 136–153.

[38] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36.

[39] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 1–39, 2011.

[40] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[41] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

Hui Lv received the B.S. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology (NUST), Nanjing, China, in 2017. He is currently pursuing the Ph.D. degree in computer science and technology with NUST. His research interests include computer vision, pattern recognition, data mining, and deep learning.

Chuanwei Zhou received the B.S. degree from the College of Elite Education, Nanjing University of Science and Technology (NUST), Nanjing, China, in 2017. He is currently pursuing the Ph.D. degree in computer science and technology with NUST. His research interests include computer vision, pattern recognition, data mining, and deep learning.

Zhen Cui (Member, IEEE) received the B.S. degree from Shandong Normal University in 2004, the M.S. degree from Sun Yat-sen University in 2006, and the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences, in 2014. He also spent half a year as a Research Assistant on Nanyang Technological University (NTU) from June 2012 to December 2012. He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), from 2014 to 2015. He is currently a Professor with the Nanjing University of Science and Technology, China. His research interests include deep learning, computer vision, and pattern recognition.

Chunyan Xu received the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, in 2015. From 2013 to 2015, she was a Visiting Scholar with the Department of Electrical and Computer Engineering, National University of Singapore. She is currently a Lecturer with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Her research interests include computer vision, manifold learning, and deep learning.

Yong Li is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include machine learning, affective computing, and computer vision. He especially focuses on face and facial expression analysis, and self-supervised learning.

Jian Yang received the Ph.D. degree from the Nanjing University of Science and Technology (NUST) on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a Postdoctoral Researcher with the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University. From 2006 to 2007, he was also a Postdoctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering, NUST. He is the author of more than 100 scientific articles in pattern recognition and computer vision. His journal articles have been cited more than 4000 times in the ISI Web of Science, and 9000 times in the Web of Scholar Google. His research interests include pattern recognition, computer vision, and machine learning. He is also a Fellow of IAPR. He is/was an Associate Editor of *Pattern Recognition Letters*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*.