Effective Video Abnormal Event Detection by Learning a Consistency-Aware High-Level Feature Extractor

Guang Yu
National University of Defense Technology
Changsha, China

Siqi Wang
National University of Defense Technology
Changsha, China

Zhiping Cai
National University of Defense Technology
Changsha, China

Xinwang Liu
National University of Defense Technology
Changsha, China

Chengkun Wu
National University of Defense Technology
Changsha, China

ABSTRACT

With pure normal training videos, video abnormal event detection (VAD) aims to build a normality model, and then detect abnormal events that deviate from this model. Despite of some progress, existing VAD methods typically train the normality model by a *low-level* learning objective (*e.g.* pixel-wise reconstruction/prediction), which often overlooks the *high-level* semantics in videos. To better exploit high-level semantics for VAD, we propose a novel paradigm that performs VAD by learning a **C**onsistency-**A**ware high-level **F**eature **E**xtractor (CAFE). Specifically, with a pre-trained deep neural network (DNN) as teacher network, we first feed raw video events into the teacher network and extract the outputs of multiple hidden layers as their high-level features, which contain rich high-level semantics. Guided by high-level features extracted from normal training videos, we train a student network to be the high-level feature extractor of normal events, so as to explicitly consider high-level semantics in training. For inference, a video event can be viewed as normal if the student extractor produces similar high-level features to the teacher network. Second, based on the fact that consecutive video frames usually enjoy minor differences, we propose a consistency-aware scheme that requires high-level features extracted from neighboring frames to be consistent. Our consistency-aware scheme not only encourages the student extractor to ignore low-level differences and capture more high-level semantics, but also enables better anomaly scoring. Last, we also design a generic framework that can bridge high-level and low-level learning in VAD to further ameliorate VAD performance. By flexibly embedding one or more low-level learning objectives into CAFE, the framework makes it possible to combine the strengths of both high-level and low-level learning. The proposed method attains state-of-the-art results on commonly-used benchmark datasets.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547944

CCS CONCEPTS

• **Computing methodologies** → **Scene anomaly detection**; **Unsupervised learning**.

KEYWORDS

Video anomaly detection, high-level features, video semantics

ACM Reference Format:
Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, and Chengkun Wu. 2022. Effective Video Abnormal Event Detection by Learning a Consistency-Aware High-Level Feature Extractor. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547944

1 INTRODUCTION

As a vital subarea of multimedia content understanding, video abnormal event detection (VAD) [53] aims to detect unexpected abnormal events that violate the normal convention in surveillance videos. Since VAD can avoid labor-intensive and tiresome manual monitoring and checking, it has shown a great potential in diverse application scenarios, *e.g.* public safety management and information forensics [45, 79]. However, VAD is still a very challenging task. Due to the ambiguous and rare nature of anomaly [10], abnormal event occurs with a low probability and unbounded semantics. Therefore, it is often difficult or even infeasible to construct a training set that covers sufficient abnormal events for learning and detection. Accordingly, a more reasonable VAD solution is to use available and abundant normal videos to build a normality model that can outline normal events. Events that do not conform to the description of this model are considered abnormal during inference.

Enormous efforts have been made to address VAD. With the rise and surging popularity of deep neural network (DNN) [35] in various tasks, VAD solutions gradually transition from handcrafted feature based classic methods to DNN based methods. Unlike conventional VAD methods that usually suffer from complex feature engineering and sub-optimal video representations [70], DNN based methods can automatically learn high-quality features from videos to realize effective and end-to-end VAD. Such a favorable property enables DNN based methods to dominate recent VAD research. As DNN based methods prosper in VAD, it is noted that most of them are guided by a *low-level* learning objective, *e.g.* pixel-wise reconstruction [24, 50] or prediction [37, 39, 45] of pixel-level video data

like raw frames or optical flow. Since abnormal training videos are assumed to be absent in VAD, such a low-level objective offers a convenient way to create learning targets of DNN, while anomalies can be detected by simply computing pixel-wise errors.

Despite the fruitful progress of low-level learning in VAD, it is faced with a critical limitation in essence: While anomaly is a high-level concept defined by humans, the low-level objective tends to force those methods to overly focus on pixel-level details, and high-level semantics are usually not explicitly considered. Concretely, the objective of low-level learning is to minimize the differences between outputs and targets based on certain per-pixel losses. However, such pixel-wise metrics are insufficient for data like images, as they lack the ability to capture high-level semantic differences that are meaningful to human visual perception [32, 34]. Nevertheless, those high-level semantic differences have a vital influence on tasks like VAD. For example, when pedestrian walking is viewed as a normal event, two different pedestrians are supposed to be semantically similar, but per-pixel metrics could produce a very large intra-class difference that may disrupt the learning of DNN.

To remedy the deficiency of low-level learning for VAD, we notice that the modern DNN (*e.g.* convolutional neural networks) pre-trained on a public generic image dataset (*e.g.* ImageNet [15]) proves to be a surprisingly effective extractor of high-level semantic features, which are more meaningful to perception [19, 32]. Operations like convolution and pooling enable the pre-trained DNN to be invariant to small deformations or variations, but sensitive to salient high-level structures like texture and edges [16]. Motivated by such a favorable property of the pre-trained DNN, we propose a novel and effective VAD paradigm by learning a **C**onsistency-**A**ware high-level **F**eature **E**xtractor (CAFE). Specifically, we first employ a pre-trained DNN as the teacher network, and it enables us to extract high-level features that contain rich semantics from its multiple hidden layers by feeding original video events into the network. By using high-level features from pure *normal* training videos as learning targets, a smaller student network is then trained from scratch to be a high-level feature extractor, which is supposed to produce similar high-level features to the teacher network for normal video events. This pipeline, which transfers knowledge from a pre-trained teacher network to a student network based on intermediate features, is also known as feature based distillation [23, 64] (reviewed in Sec. 2.2). In this way, high-level semantics can be explicitly considered when building the normality model for VAD. Since the student network only learns to extract features of normality, it is likely to yield different features from the teacher network when faced with anomalies, which makes it possible to discriminate anomalies during inference. Second, as temporally adjacent video frames usually enjoy tiny differences that a human would barely notice, we encourage the student network to ignore the low-level differences and capture more high-level semantics by proposing a consistency-aware scheme, which requires the student network to assign consistent features for the foreground on neighboring frames. Besides, we also show that our consistency-aware scheme contributes to better anomaly scoring. Last, we design a generic framework that can synthesize the low-level learning objective into the proposed CAFE paradigm in a plug-and-play manner. In this way, the classic low-level learning can be seamlessly combined with the proposed high-level VAD paradigm CAFE, which can give rise to further VAD performance enhancement.

In summary, our contributions are listed as follows:

- To explicitly consider high-level semantics in videos, we propose a brand-new paradigm that can conduct VAD by learning a high-level feature extractor for the normal video events. To our best knowledge, this is also the first work to address VAD by a feature based distillation pipeline.
- To exploit the temporal information in videos, we devise a consistency-aware scheme. It not only encourages the feature extractor to neglect low-level differences among consecutive video frames and further focus on high-level semantics, but also enables better anomaly scoring in inference.
- To combine the strengths of both high-level and low-level learning for better VAD, we design a generic framework that can bridge them through embedding the low-level objective into the proposed high-level VAD paradigm.

Experiments on commonly-used datasets show that the proposed method is able to achieve state-of-the-art VAD performance. In particular, our method attains 92.6% frame-level AUC on Avenue, which is a highly competitive result among VAD literature. In the rest of paper, Sec. 2 reviews representative VAD methods and knowledge distillation; Sec. 3 introduces our method in detail; Sec. 4 presents results of evaluation and discussion; Sec. 5 concludes the paper.

2 RELATED WORK

2.1 Video Anomaly Detection

Conventional VAD methods [4, 13, 14, 40, 47, 77] first extract video representations by handcrafted feature extractors, and then model the representations with classic anomaly detection methods. However, the design of handcrafted feature extractors requires considerable domain expertise [35], and the video representations extracted by handcrafted extractors can be sub-optimal for VAD [70]. Since DNNs can perform powerful automatic representation learning, recent methods usually utilize DNNs to achieve more effective VAD. Existing DNN based VAD methods typically follow a low-level learning objective, *e.g.* pixel-wise reconstruction/prediction, which has shown favorable effectiveness in DNN based VAD. The learned video representations are either input into classic anomaly detection methods to identify anomalies [30, 61, 70], or embedded into DNNs for end-to-end VAD [9, 71, 72, 78]. Recent methods usually perform end-to-end VAD due to its simplicity and efficiency, while large reconstruction/prediction errors are often used to indicate the occurrence of abnormal events during inference. To improve the quality and discriminative power of low-level learning, various networks or techniques have been explored such as convolutional auto-encoder [24], U-Net [37], adversarial learning [21, 57, 74], memory module [22, 39, 51], foreground localization [73, 80] and transformer [18]. In addition to ordinary pixel-wise reconstruction/prediction that uses data from the same modality as inputs and outputs (*e.g.* video frame prediction), some works also carry out cross-modality reconstruction/prediction [11, 50]. A common practice is to utilize raw images to generate their corresponding optical flow or gradients for learning appearance-motion correspondence. As analyzed in Sec. 1, low-level learning based methods

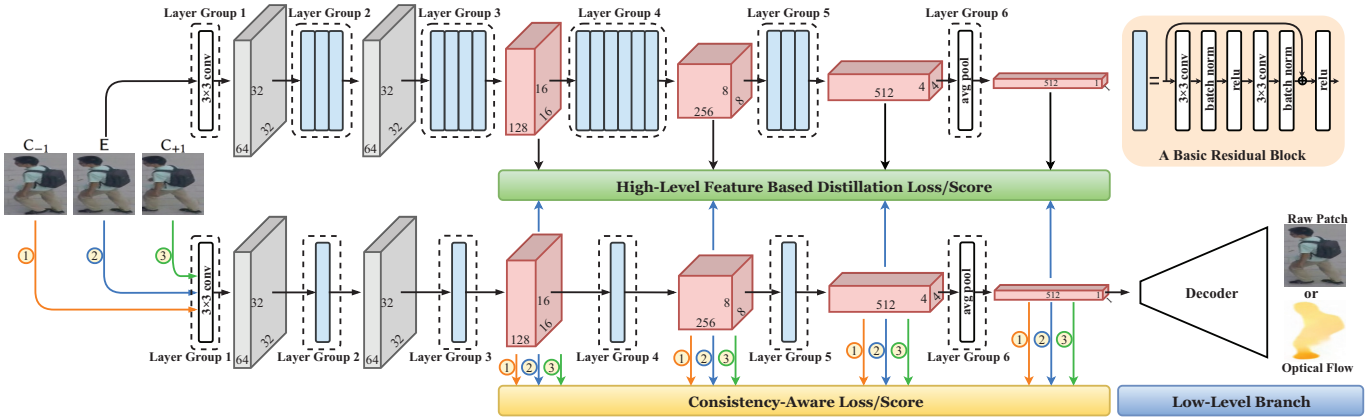

Figure 1: The overview of the proposed VAD method, which consists of three components: (1) Learning a high-level feature extractor: A smaller student network (*e.g.* ResNet-9) is trained to mimic a pre-trained teacher network (*e.g.* ResNet-34 that discards the fully connected layer) by maximizing the similarity of high-level features they produce, so as to be a high-level feature extractor of normal video events. (2) Consistency-aware scheme: The student network is trained to endow temporally adjacent foreground patches with consistent high-level features, so as to capture more high-level semantics and enable better anomaly scoring. (3) Bridging high/low-level learning: An additional low-level learning objective (*e.g.* pixel-wise reconstruction) is embedded into the proposed high-level VAD paradigm (CAFE) in a plug-and-play manner to further boost VAD performance.

tend to overly emphasize pixel-level details and ignore high-level semantics in videos. By contrast, the proposed CAFE can explicitly take high-level semantics of videos into account during VAD by learning a consistency-aware high-level feature extractor.

2.2 Knowledge Distillation

Knowledge distillation (KD) refers to the method that transfers information from a network to another network [64]. Basically, KD follows a teacher-student framework. In this framework, the network that provides knowledge is called teacher, while the network that receives the knowledge is called student. Buciluǎ *et al.* [7] pioneer KD for model compression. Then, Hinton *et al.* [28] explore the extension of KD, which distills a large teacher network into a small student network by minimizing the differences between the logits (*i.e.* the inputs of the last Softmax layer) produced by the two networks. Later, Romero *et al.* [56] extend the logit based KD to feature based KD by introducing features (*i.e.* the outputs of hidden layers) for distillation. Generally, feature based KD aims to minimize the differences between the features extracted from multiple hidden layers of the teacher and student network, which enables richer and more flexible information transferring. As an effective technique, KD has been widely applied to various fields, such as semantic segmentation [26, 38], speech emotion recognition [2], object detection [12, 66], domain adaptation [5, 8] and optical flow estimation [3, 60].

Despite its vast applications, KD is hardly explored in the realm of VAD. In the literature, we notice that a few works follow or integrate a KD pipeline to perform other anomaly detection tasks. However, most of them [6, 46, 58, 63, 65, 76] are not designed for video data, and they are unable to consider the temporal information, which actually plays a vital role in video analysis. By contrast, the proposed VAD method enables us to preserve the temporal correlation of continuous video frames and exploit motion cues on the temporal dimension during distillation. As far as we know, only one work explores a logit based KD pipeline for VAD by using the logits or class probabilities for distillation [20]. However, it claims that KD is insufficient for VAD, since it attains unsatisfactory performance (*e.g.* 73.7% AUC on Avenue dataset). As a comparison, we for the first time demonstrate that VAD can be carried out in a highly effective manner by customizing a feature based KD pipeline. To our best knowledge, this is actually the first work that propose to perform VAD by feature based KD.

3 THE PROPOSED METHOD

3.1 Preprocessing

For learning and inference, we need to extract video events from videos as basic processing units in the first place. Video events can be represented by the entire video frames [37] or foreground patches on frames [30]. To yield reasonable video events, we notice that recent VAD works have demonstrated the importance of foreground localization [21], which can avoid the learning bias towards redundant and meaningless background [73, 80]. Similar to the process of video event construction in [73], we first localize each foreground object by a bounding box on video frames. For each localized object on frames, we use its bounding box to crop a foreground patch. Finally, we resize the foreground patch to a fixed size $H \times W$, and regard the resized patch as a video event E, which serves as the basic processing unit in the proposed method.

3.2 Learning a High-Level Feature Extractor

3.2.1 Motivation. Abnormal events refer to those events that violate the conventional daily cognition of humans, which is a high-level concept based on human visual perception. For instance, the sudden appearance of a car on the sidewalk can be regarded as an abnormal event by humans. In this case, the abnormality of the car

is reflected by the high-level semantic differences from the walking or running pedestrians on the road, *e.g.* the differences of shape and texture. However, as discussed in Sec. 1, existing DNN based VAD methods are typically guided by a low-level learning objective like pixel-wise reconstruction/prediction, which often excessively memorizes pixel-level details and lacks explicit consideration of high-level semantics in videos. Consequently, they often capture insufficient high-level semantic features, which hinders them from discriminating abnormal events [22, 51, 68]. Motivated by the fact that modern DNNs pre-trained on a generic image dataset can serve as an effective extractor of high-level semantic features [19, 32], we make a student DNN learn to be a similar high-level feature extractor by imitating a pre-trained teacher DNN. Such an idea benefits VAD in two aspects: (1) With high-level features extracted by the pre-trained teacher DNN as learning targets, we expect the student network to produce similar features, which encourages the student network to attend to high-level semantics in videos. (2) The hierarchical representation learning ability of DNN [35] makes it possible to extract features on multiple different layers, which enables us to introduce multi-level semantic information into the learning process of VAD. Our pipeline is detailed below.

3.2.2 Pipeline. With a pre-trained teacher DNN $\mathcal{T}$, we aim to train a student DNN $\mathcal{S}$ to be a high-level feature extractor of normal video events by training $\mathcal{S}$ to output similar high-level features to $\mathcal{T}$. Formally, given a video event E, we first feed it into the teacher network $\mathcal{T}$ and student network $\mathcal{S}$ respectively, so as to compute outputs of DNNs' multiple hidden layers as high-level features. We denote the i-th layer's outputs of $\mathcal{T}$ and $\mathcal{S}$ as $\mathcal{T}(\mathsf{E}, i)$ and $\mathcal{S}(\mathsf{E}, i)$ respectively. Then, we can select certain high-level layers from $\mathcal{T}$ and $\mathcal{S}$ and build two ordered layer index set $K_{\mathcal{T}}$ and $K_{\mathcal{S}}$. To enable $\mathcal{S}$ to mimic $\mathcal{T}$, we assume that $K_{\mathcal{T}}$ shares the same cardinality with $K_{\mathcal{S}}$, *i.e.* $C = |K_{\mathcal{T}}| = |K_{\mathcal{S}}|$. For convenience, we simply require that the extracted feature $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ has the same shape with $\mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})$, where $K_{\mathcal{S}}^{(j)}/K_{\mathcal{T}}^{(j)}$ denotes the j-th layer index in $K_{\mathcal{S}}/K_{\mathcal{T}}$ and $1 \le j \le C$. Without losing the generality, when $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ does not share the shape of $\mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})$, one can use a transformer layer/function to unify their shape [64]. With $K_{\mathcal{T}}$ and $K_{\mathcal{S}}$, we can build two feature sets $\mathsf{F}_{\mathcal{T}}(\mathsf{E}) = \{\mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)}) | 1 \le j \le C\}$ and $\mathsf{F}_{\mathcal{S}}(\mathsf{E}) = \{\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}) | 1 \le j \le C\}$. To encourage $\mathcal{S}$ to be a high-level feature extractor, we maximize the similarity of extracted features produced by $\mathcal{T}$ and $\mathcal{S}$, which can be formulated as the following objective:

$$\max_{\mathcal{S}} \sum_{j=1}^{C} Sim(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}), \mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})) \tag{1}$$

where $Sim(\cdot)$ is a pre-defined similarity measure.

As stated in Sec. 3.1, we extract the patch of each localized foreground as a video event E. As for the teacher $\mathcal{T}$ and student network $\mathcal{S}$, any popular DNN backbone can be used as the network architecture of $\mathcal{T}$ and $\mathcal{S}$. We choose ResNet [25] as our network architecture, which has been widely-used due to its effectiveness. To be more specific, we use a ResNet-34 and a smaller ResNet-9 (where the number 34 and 9 represent the number of weighted layers) as the teacher network $\mathcal{T}$ and the student network $\mathcal{S}$ respectively, and

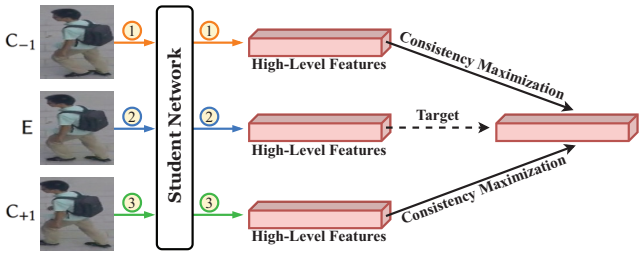

Figure 2: Illustration of the proposed consistency-aware scheme, which aims to maximize the consistency between high-level features of consecutive foreground patches.

their detailed structures are shown in Fig. 1. A smaller $\mathcal{S}$ not only reduces the computational cost, but also improves its discriminative ability by limiting its generalization to anomalies [58]. Given $\mathcal{T}$ (*i.e.* ResNet-34) pre-trained on a generic image dataset, we discard its last fully connected layer, which is a task-specific layer for classification. It should be noted that $\mathcal{S}$ (*i.e.* ResNet-9) does not have the fully connected layer. Then, we extract the outputs of the last four layer groups as $C = 4$ high-level features (see Fig. 1). This is because the features from the higher layer groups can better capture high-level semantics and omit meaningless detailed pixel values [19], which is also validated by experiments in Sec. 4.3.2. To calculate the similarity between the high-level features $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ and $\mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})$ for training, any similarity measure can be explored. Empirically, we find that the simple mean square error (MSE) can already be a highly effective similarity measure. Thus, the objective in (1) can be converted to minimize the following loss $\mathcal{L}_{kd}$:

$$\mathcal{L}_{kd} = \sum_{j=1}^{C} \|\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}) - \mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})\|_2^2 \tag{2}$$

For simplicity, we slightly abuse the notation in Eq. (2) by viewing tensors like $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ as a vector, which is also followed by other expressions in this paper. At the training stage, only video events from pure normal training videos are fed into $\mathcal{T}$ and $\mathcal{S}$ for learning. In other words, we expect $\mathcal{S}$ to be an effective high-level feature extractor for normality only. During inference, as $\mathcal{S}$ has limited capacity and it has not been trained to extract features from unseen anomalies, it is unlikely to produce good high-level features for anomalies like the pre-trained $\mathcal{T}$. Hence, we can calculate the feature discrepancies between the high-level features $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ and $\mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})$ to yield a distillation anomaly score $\mathcal{A}_{kd}(\mathsf{E})$ for the video event E. A higher score $\mathcal{A}_{kd}(\mathsf{E})$ indicates that the event E is more likely to be anomalous. By learning a high-level feature extractor, we provide a novel and elegant way to explicitly consider the high-level semantics within videos.

3.3 Consistency-Aware Scheme

3.3.1 Motivation. As introduced in Sec. 3.1 and Sec. 3.2, the video event is represented by a single foreground patch, while the valuable temporal information in videos has not been considered so far. This is because the teacher network $\mathcal{T}$ is pre-trained with image data, and it can only accept a single image as input for feature

extraction. To take temporal information into account, we notice the fact that the differences between two video frames in a very short time interval are usually quite small, and human visual perception is usually insensitive to such tiny differences. To endow the student network $\mathcal{S}$ with such insensitivity to those detail differences, we propose a consistency-aware scheme that requires $\mathcal{S}$ to obtain consistent high-level features from foreground patches of temporally adjacent frames in training (see Fig. 2). In this way, our scheme not only considers temporal context of videos, but also further encourages $\mathcal{S}$ to focus on high-level semantic features.

3.3.2 Consistency-aware Scheme. Given a video event E represented by a foreground patch, we crop two patches by the same location of E from two neighboring frames. Similarly, we resize two patches into $H \times W$ to yield the context patches C_{-1} and C_{+1}, which serve as the temporal context of E. We assume $\mathsf{E}, \mathsf{C}_{-1}, \mathsf{C}_{+1}$ to be semantically consistent as they are in the same short period. Similar to the process in Sec. 3.2, we feed C_{-1} and C_{+1} into the student network $\mathcal{S}$ to yield two feature sets $\mathsf{F}_{\mathcal{S}}(\mathsf{C}_{-1}) = \{\mathcal{S}(\mathsf{C}_{-1}, K_{\mathcal{S}}^{(j)}) | 1 \le j \le C\}$ and $\mathsf{F}_{\mathcal{S}}(\mathsf{C}_{+1}) = \{\mathcal{S}(\mathsf{C}_{+1}, K_{\mathcal{S}}^{(j)}) | 1 \le j \le C\}$ respectively. It should be noted that two additional forward passes through the student network $\mathcal{S}$ will not significantly increase the computational cost, as $\mathcal{S}$ is a lightweight DNN like ResNet-9. To encourage the student network $\mathcal{S}$ to assign consistent high-level features to the patches C_{-1}, E and C_{+1}, we maximize the consistency of their high-level features $\mathcal{S}(\mathsf{C}_{-1}, K_{\mathcal{S}}^{(j)})$, $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ and $\mathcal{S}(\mathsf{C}_{+1}, K_{\mathcal{S}}^{(j)})$, which can be formulated as the following objective:

$$\max_{\mathcal{S}} \sum_{j=1}^{C} Con(\mathcal{S}(\mathsf{C}_{-1}, K_{\mathcal{S}}^{(j)}), \mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}), \mathcal{S}(\mathsf{C}_{+1}, K_{\mathcal{S}}^{(j)})) \tag{3}$$

where $Con(\cdot)$ is a pre-defined consistency measure of three inputs. One can use plenty of ways to define $Con(\cdot)$, and here we simply define it based on a similarity measure $Sim(\cdot)$ by the following form $Con(x, y, z) = Sim(y, x) + Sim(y, z)$. When we also adopt MSE as the similarity measure, the objective (3) can be transformed to minimize the following loss $\mathcal{L}_{ca}$:

$$\mathcal{L}_{ca} = \frac{1}{2} \sum_{j=1}^{C} (\|\mathcal{S}(\mathsf{C}_{-1}, K_{\mathcal{S}}^{(j)}) - \mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})\|_2^2 + \|\mathcal{S}(\mathsf{C}_{+1}, K_{\mathcal{S}}^{(j)}) - \mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})\|_2^2) \tag{4}$$

As to inference, since $\mathcal{S}$ is merely trained to preserve the temporal consistency of normal video events, the consistency of abnormal events cannot be guaranteed in inference. Meanwhile, as anomalies are often viewed to be more unpredictable than normality [37], we also expect abnormal events to be less consistent than normal events. Therefore, we can also leverage the consistency of the consecutive patches to be a standard to discriminate anomalies. In this way, we can calculate a consistency based anomaly score $\mathcal{A}_{ca}(\mathsf{E})$.

It is noted that learning a high-level feature extractor and the consistency-aware scheme are then assembled into the proposed CAFE paradigm. As a new solution that differs from previous VAD routine, CAFE alone has already been able to achieve very competitive VAD performance in our later empirical evaluations, while we also develop a framework to further exploit its potential below.

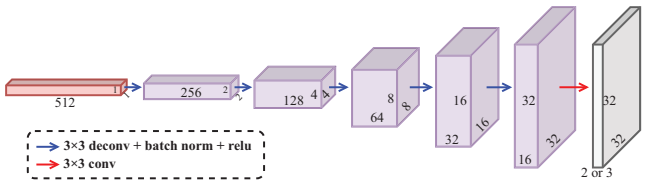

Figure 3: The architecture of the low-level learning branch $\mathcal{D}$ for pixel-wise reconstruction of the raw video event E or its optical flow $\mathsf{O}(\mathsf{E})$. The red square represents the last high-level feature $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)})$ produced by the student network $\mathcal{S}$.

3.4 Bridging High/Low-Level Learning

3.4.1 Motivation. Despite the aforementioned limitation, we notice that classic low-level learning still has some important strengths in VAD. For example, low-level learning provides a convenient way to incorporate motion cues, which are usually represented by pixel-level temporal gradients or optical flow in VAD, into the learning process. To combine the strengths of both low-level and high-level learning for VAD, we design the generic framework below to bridge them and further enhance VAD performance.

3.4.2 Framework. To incorporate a low-level learning objective, *e.g.* pixel-wise reconstruction, we propose to introduce an additional low-level learning branch $\mathcal{D}$. Given the event E and its feature set $\mathsf{F}_{\mathcal{S}}(\mathsf{E}) = \{\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}) | 1 \le j \le C\}$ produced by the student network $\mathcal{S}$, we train the branch $\mathcal{D}$ to learn a pixel-level target G by taking the high-level feature set $\mathsf{F}_{\mathcal{S}}(\mathsf{E})$ as the input. In other words, it is equivalent to the following objective:

$$\max_{\mathcal{D}} Sim(\mathcal{D}(\mathsf{F}_{\mathcal{S}}(\mathsf{E})), \mathsf{G}) \tag{5}$$

Concretely, we explores two types of learning targets here: First, to explicitly integrate motion cues in videos, we can set the learning target to be E's corresponding optical flow patch by $\mathsf{G} = \mathsf{O}(\mathsf{E})$. To yield $\mathsf{O}(\mathsf{E})$, we utilize a pre-trained FlowNet2 model [29] to efficiently calculate the optical flow map of the video frame that contains E. Based on E's location on the original frame, we crop an optical flow patch from the optical flow map and resize it into $\mathsf{O}(\mathsf{E})$ with the size $H \times W$. Second, we simply follow many pixel-level reconstruction based VAD methods and choose the raw video event to be the learning target $\mathsf{G} = \mathsf{E}$. For efficiency, we simply select the last high-level feature $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)})$ from the entire feature set $\mathsf{F}_{\mathcal{S}}(\mathsf{E})$ to be the input of $\mathcal{D}$. Then, we implement the low-learning branch $\mathcal{D}$ by a fully convolutional generative network (shown in Fig. 3). When MSE is used as the similarity measure as before, the objective in (5) is equivalent to minimizing the loss $\mathcal{L}_{lo}$ below:

$$\mathcal{L}_{lo} = \|\mathcal{D}(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)})) - \mathsf{G}\|_2^2 \tag{6}$$

It should be noted that one can establish two or more low-learning branches to simultaneously learn two or more low-level learning targets. For example, we can use a motion branch $\mathcal{D}^{(m)}$ and an appearance branch $\mathcal{D}^{(a)}$ to learn the target $\mathsf{G}^{(m)} = \mathsf{O}(\mathsf{E})$ and $\mathsf{G}^{(a)} = \mathsf{E}$ respectively. In this case, the loss term $\mathcal{L}_{lo}$ is defined by:

$$\mathcal{L}_{lo} = \|\mathcal{D}^{(a)}(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)})) - \mathsf{E}\|_2^2 + \|\mathcal{D}^{(m)}(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)})) - \mathrm{O}(\mathsf{E})\|_2^2 \tag{7}$$

Just like the common practice in existing low-level learning based VAD methods, we can also calculate the differences between the low-level branch output $\mathcal{D}(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)}))$ and $\mathrm{O}(\mathsf{E})$ or E to yield a motion anomaly score $\mathcal{A}_m(\mathsf{E})$ or appearance anomaly score $\mathcal{A}_a(\mathsf{E})$, which can be added to existing anomaly scores to assist VAD.

3.5 Training Procedure

To train the proposed model, we can combine the loss terms defined above to yield the overall training loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{kd} + \lambda_{ca}\mathcal{L}_{ca} + \lambda_{lo}\mathcal{L}_{lo} \tag{8}$$

where λ_{ca} and λ_{lo} are hyper-parameters to balance each loss item. Note that we view the high-level VAD paradigm as our fundamental solution, for which $\lambda_{lo} = 0$ and the overall loss becomes $\mathcal{L} = \mathcal{L}_{kd} + \lambda_{ca}\mathcal{L}_{ca}$. In the training process, we only update the parameters of the student network $\mathcal{S}$ and fix the the parameters of the pre-trained teacher network $\mathcal{T}$. In addition, we introduce the consistency-aware loss $\mathcal{L}_{ca}$ after a number of training epochs, so as to enable the student network $\mathcal{S}$ to produce high-quality high-level features first.

3.6 Anomaly Inference

For inference, we design three types of anomaly scores to evaluate a video event E: (1) Distillation anomaly score $\mathcal{A}_{kd}(\mathsf{E})$, which measures the discrepancies between the high-level features $\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)})$ and $\mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})$ $(1 \le j \le C)$ produced by $\mathcal{S}$ and $\mathcal{T}$ respectively:

$$\mathcal{A}_{kd}(\mathsf{E}) = -\sum_{j=1}^{C} Sim(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}), \mathcal{T}(\mathsf{E}, K_{\mathcal{T}}^{(j)})) \tag{9}$$

(2) Consistency based anomaly score $\mathcal{A}_{ca}(\mathsf{E})$, which computes the differences between the high-level features (produced by $\mathcal{S}$) of three adjacent foreground patches $\mathsf{C}_{-1}, \mathsf{E}, \mathsf{C}_{+1}$:

$$\mathcal{A}_{ca}(\mathsf{E}) = -\sum_{j=1}^{C} Con(\mathcal{S}(\mathsf{C}_{-1}, K_{\mathcal{S}}^{(j)}), \mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(j)}), \mathcal{S}(\mathsf{C}_{+1}, K_{\mathcal{S}}^{(j)})) \tag{10}$$

(3) Low-level learning based anomaly score $\mathcal{A}_{lo}(\mathsf{E})$, which measures the differences between the learning target G and $\mathcal{D}$'s output:

$$\mathcal{A}_{lo}(\mathsf{E}) = -Sim(\mathcal{D}(\mathcal{S}(\mathsf{E}, K_{\mathcal{S}}^{(C)})), \mathrm{G}) \tag{11}$$

Finally, we can yield the overall anomaly score $\mathcal{A}(\mathsf{E})$ by a weighted sum of the above anomaly scores:

$$\mathcal{A}(\mathsf{E}) = \omega_{kd}\frac{\mathcal{A}_{kd}(\mathsf{E}) - \mu_{kd}}{\sigma_{kd}} + \omega_{ca}\frac{\mathcal{A}_{ca}(\mathsf{E}) - \mu_{ca}}{\sigma_{ca}} + \omega_{lo}\frac{\mathcal{A}_{lo}(\mathsf{E}) - \mu_{lo}}{\sigma_{lo}} \tag{12}$$

where $\mu_{kd}, \sigma_{kd}, \mu_{ca}, \sigma_{ca}, \mu_{lo}, \sigma_{lo}$ denote the means and standard deviations of distillation anomaly score, consistency based score and low-level learning based score of all normal video events in training, which are utilized to standardize the three scores to the same scale. In our implementations, we simply use MSE as the similarity measure in the above defined anomaly scores. For the consistency based anomaly score $\mathcal{A}_{ca}(\mathsf{E})$, we adopt the form of $Con(\cdot)$ defined in Sec. 3.3, *i.e.* $Con(x, y, z) = Sim(y, x) + Sim(y, z)$.

To obtain the anomaly score of each video frame for frame-level evaluation, we take the maximum of all video events' anomaly scores on a frame as the score of this frame. Following previous works, we apply a sliding window to smooth the scores of frames.

4 EVALUATION

4.1 Experimental Settings

To evaluate the proposed method, we conduct experiments on the three commonly-used VAD datasets: UCSDped2 [47], Avenue [40] and ShanghaiTech [37]. Following the most frequently-used evaluation metric in VAD, we adopt frame-level Area Under Curve (AUC) [47] of the Receiver Operating Characteristic (ROC) curve for quantitative evaluation. A higher AUC indicates a better VAD performance. As for video event extraction, we follow the pipeline in [73] to localize video foreground and extract video events, and the patch size $H \times W$ is set to 32×32. The teacher network $\mathcal{T}$ is implemented by a ResNet-34 architecture in Fig. 1. Due to the size of video events in our approach, we adopt the publicly available CIFAR-100 dataset [33] for the pre-training of $\mathcal{T}$, as the dataset provides diverse 32×32 generic images. Specifically, we train $\mathcal{T}$ to perform classification on CIFAR-100 for 200 epochs, while the batch size is set to 128. $\mathcal{T}$ is pre-trained by SGD optimizer in PyTorch [52] with an initial learning rate=0.1, momentum factor=0.9 and weight decay=5e-4. As shown in Fig. 1 and Fig. 3, we adopt a ResNet-9 network and a fully convolutional decoder to implement the student network $\mathcal{S}$ and the low-level learning branch $\mathcal{D}$ respectively. $\mathcal{S}$ and $\mathcal{D}$ are optimized by the default Adam optimizer in PyTorch [52] in an end-to-end manner. Other parameters for training $\mathcal{S}$ and $\mathcal{D}$ are set as follows: We set $\lambda_{ca} = 0.1$ and $\lambda_{lo} = 1$ for all experiments. The batch size is set to be 256. The number of training epochs is set to 50, while the consistency-aware scheme is introduced into training after 30 training epochs as mentioned in Sec. 3.5. Due to the evident differences in characteristics of data and anomalies on each benchmark dataset, the weight of each anomaly score, $(\omega_{kd}, \omega_{ca}, \omega_{lo})$, are typically set to be $(0.5, 1.0, 0.1)$, $(1, 0.5, 0.1)$ and $(0.5, 1, 0.5)$ for UCSDped2, Avenue and Shanghaitech respectively. The sliding window size for the smoothing of frame anomaly scores is set to 10, 20 and 20 for UCSDped2, Avenue and ShanghaiTech respectively. More details are provided in supplementary material.

4.2 Comparison with State-of-the-art Methods

VAD has been explored by a huge number of works. To verify the effectiveness of the proposed method, we have extensively compared it with 36 state-of-the-art VAD methods. It should be noted that we do not compare works like [20, 30] as they calculate frame-level AUC in a different way. For our method, we test four configurations in total. The basic configuration only deploys the proposed high-level VAD paradigm (*i.e.* CAFE) that learns a consistency-aware high-level feature extractor, while no low-level learning (LL) is performed. Meanwhile, three additional configurations are devised to integrate one or two low-level learning objectives into the high-level CAFE paradigm: Learning the optical flow patch $\mathrm{O}(\mathsf{E})$ (CAFE w/ LL-O), learning the raw video event (CAFE w/ LL-E), as well as learning both optical flow and raw video event (CAFE w/ LL-O & LL-E). The comparison is displayed in Table 1, from which we can draw the following conclusions: (1) Without the aid of any

Table 1: AUC comparison with state-of-the-art VAD methods.

Method	UCSDped2	Avenue	ShanghaiTech
CAE [24]	90.0%	70.2%	-
AMDN [70]	90.8%	70.2%	-
ST-CAE [78]	91.2%	80.9%	-
sRNN [43]	92.2%	81.7%	68.0%
WTA-CAE [61]	96.6%	82.1%	-
LSTM-AE [42]	88.1%	77.0%	-
AM-GAN [55]	93.5%	-	-
Recounting [27]	92.2%	-	-
TCP [54]	88.4%	-	-
Frame-Prediction [37]	95.4%	85.1%	72.8%
AnomalyNet [79]	94.9%	86.1%	-
AnoPCN [71]	96.8%	86.2%	73.6%
Attention-Prediction [80]	96.0%	86.0%	-
PDE-AE [1]	95.4%	-	72.5%
Mem-AE [22]	94.1%	83.3%	71.2%
AM-Correspondence [50]	96.2%	86.9%	-
NNC [31]	-	88.9%	-
MPED-RNN [49]	-	-	73.4%
MLAD [62]	99.2%	71.5%	-
Multispace [75]	95.4%	86.8%	73.6%
DeepOC [69]	96.9%	86.6%	-
GEPC [48]	-	-	76.1%
OGNet [74]	98.1%	-	-
BMAN [36]	96.6%	90.0%	76.2%
Clustering-AE [11]	96.5%	86.0%	73.3%
r-GAN [41]	96.2%	85.8%	77.9%
SIGNet [17]	96.2%	86.8%	-
Multipath-Prediction [67]	96.3%	88.3%	76.6%
Mem-Guided [51]	97.0%	88.5%	70.5%
Scene-Aware [59]	-	89.6%	74.7%
VEC [73]	97.3%	90.2%	74.8%
SRNN-AE [44]	92.2%	83.5%	69.6%
AMMCN [9]	96.6%	86.6%	73.7%
MPN [45]	96.9%	89.5%	73.8%
HF^2 [39]	99.3%	91.1%	76.2%
CT-D2GAN [18]	97.2%	85.9%	77.7%
CAFE	98.3%	90.7%	75.8%
CAFE w/ LL-O	98.3%	91.3%	77.1%
CAFE w/ LL-E	98.3%	90.4%	75.3%
CAFE w/ LL-O & LL-E	98.4%	92.6%	77.0%

low-level learning objective or motion cues like optical flow, our brand-new VAD paradigm (*i.e.* CAFE) that learns a consistency-aware high-level feature extractor has already been able to achieve pretty promising VAD performance by itself. As indicated by the results in Table 1, CAFE alone achieves a top-3 and top-2 frame-level AUC on UCSDped2 and Avenue respectively among the compared 36 methods. On the most challenging ShanghaiTech dataset, it also defeats the majority of state-of-the-art methods by a fairly satisfactory 75.8% frame-level AUC, even though it has not used optical flow to boost performance like most of recent VAD methods do. Such observations demonstrate the effectiveness to consider high-level

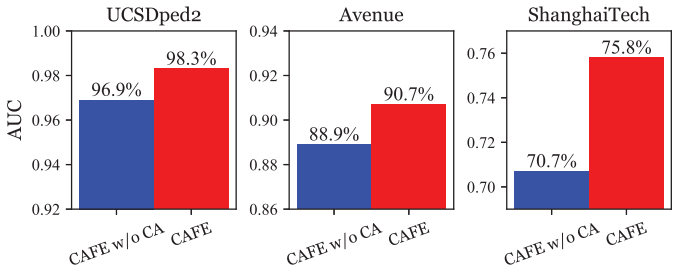

Figure 4: Performance comparison between CAFE and CAFE without consistency-aware scheme (CAFE w/o CA).

semantics for detecting abnormal events in videos. (2) The potential of the proposed VAD paradigm (CAFE) can be further unleashed by combining it with low-level learning objectives, which allows it to produce state-of-the-art VAD performance. Specifically, it is observed that CAFE equipped with two low-level learning branches (CAFE w/ LL-O & LL-E) achieves the best overall performance among four configurations, and the improvement is especially evident on Avenue and ShanghaiTech dataset (1.9% and 1.2% AUC gain when compared with CAFE alone). When compared with state-of-the-art counterparts, it achieves 92.6% frame-level AUC on Avenue dataset, which is a highly competitive result among VAD literature. On both UCSDped2 and ShanghaiTech dataset, it ranks the third place among 36 methods, and the performance gap between the best performer is less than 1% AUC in both cases. Besides, it is worth noting that no VAD method consistently achieves the best performance on all datasets in our comparison. Such results validate our method as a highly competitive VAD solution among vast VAD methods. (3) Adding a different low-level learning objective exerts a different influence on the performance. As shown in Table 1, using optical flow O(E) as the low-level learning target constantly strengthens the performance of CAFE, which justifies the necessity to introduce motion information into VAD. By contrast, using raw video events as a single low-level learning target (CAFE w/ LL-E) does not bring improvement, while the performance is even slightly degraded (within 0.5% AUC) on Avenue and ShanghaiTech dataset. A possible reason can be that CAFE has fully exploited the information within raw video frame pixels, while an additional frame pixel based low-level target does not contribute to VAD performance anymore. Interestingly, we discover that using both E and O(E) (CAFE w/ LL-O & LL-E) tends to outperform the case where only O(E) is used as learning target (CAFE w/ LL-O) on UCSDped2 and Avenue dataset, which implies that appearance and motion can be mutually-complementary clues for VAD.

4.3 Discussion

4.3.1 The Role of Consistency-Aware Scheme. As presented in Sec. 3.3, consistency-aware scheme is an inevitable component of our CAFE solution. To demonstrate the effectiveness of consistency-aware scheme, we conduct an ablation study by evaluating CAFE without consistency-aware scheme (CAFE w/o CA). To be more specific, we neither employ the consistency-aware loss $\mathcal{L}_{ca}$ for training nor use the consistency based score $\mathcal{A}_{ca}(\mathsf{E})$ in inference, *i.e.* we set $\lambda_{ca} = \omega_{ca} = 0$ in the experiments. We compare the

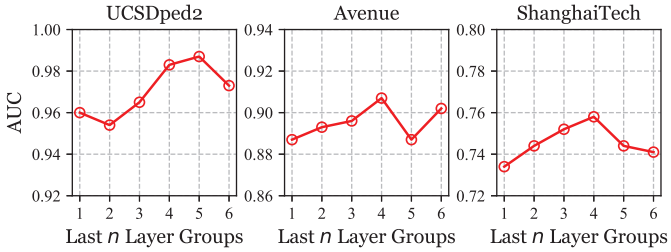

Figure 5: The influence of used layer group numbers. The number n on the x-axis indicates that the last n layer groups (shown in Fig. 1) are utilized for feature based distillation.

performance between CAFE w/o CA and the original CAFE. As shown in Fig. 4, CAFE consistently outperforms CAFE w/o CA by a notable margin on the three datasets. Specifically, consistency-aware scheme brings 1.4%, 1.8% and 5.1% AUC improvement on UCSDped2, Avenue and ShanghaiTech dataset respectively. In particular, we notice that the most significant performance gain is achieved on the most challenging ShanghaiTech dataset, while it is such an evident improvement that allows CAFE to be comparable to state-of-the-art VAD methods in this case. As a consequence, the proposed consistency-aware scheme plays a key role in utilizing the temporal information and high-level semantics for VAD.

4.3.2 The Influence of Layer Group Numbers. As detailed in Sec. 3.2, we extract multi-level features from the last $n = 4$ layer groups of the teacher and student DNN. Thus, it is natural to investigate how the variation of n influences the VAD performance. As the results in Table 5 suggest, we obtain two observations: First, when n is increased from 1 to 4, the VAD performance generally enjoys an ascending trend with approximately 2% AUC gain (the only exception is $n = 2$ on UCSDped2 dataset). Such an improvement justifies the necessity to exploit richer semantics by incorporating high-level features from multiple layers of DNNs. Second, when n is larger than 4, the VAD performance tends to be degraded in most cases. The reason can be ascribed to that some extracted features are not from high-level layers when n is set to a large value like $n = 6$, *e.g.* the first and second layer group that are very close to the input layer (see Fig. 1). Such a choice of layer group also verifies the necessity of using high-level features for VAD.

4.3.3 Using Low-level Learning Only. Another natural question is how our method will behave when only the low-level learning objective is used. To answer this question, we explore an example by using the reconstruction of optical flow as the only learning target in our method, *i.e.* we remove the loss term $\mathcal{L}_{kd}$ and $\mathcal{L}_{ca}$ in Eq. (8) and using $\mathcal{A}_{lo}$ as the only anomaly score. In this case, using low-level learning only (LL-O) yields evidently worse performance than CAFE with LL-O (CAFE w/ LL-O) on all datasets (95.3%, 79.8% and 76.3% AUC), especially on Avenue dataset. Meanwhile, it is also outperformed by CAFE on UCSDped2 and Avenue dataset and only performs comparably on ShanghaiTech. Such a comparison reveals again that using low-level learning only is insufficient for VAD.

4.3.4 The Architecture of Student Network. As we explained in Sec. 3.2.2, the student network $\mathcal{S}$ is supposed to be a smaller DNN with less capacity than the teacher network $\mathcal{T}$, so as to focus on normality and avoid generalization to the anomalies. To verify this motivation, we additionally test a larger ResNet-17 as the student network $\mathcal{S}$. Similar to the structure in Fig. 1, ResNet-17 is derived from the standard ResNet-18 [25] by discarding the last fully connected layer. As shown in Table 2, the frame-level AUC witnesses an obvious fall on all datasets when $\mathcal{S}$ is set to the larger ResNet-17, while the performance on UCSDped2 dataset even suffers from a 6% AUC loss. Thus, such observations validate our design of CAFE.

Table 2: The influence of student network on AUC.

Architecture	UCSDped2	Avenue	ShanghaiTech
ResNet-9	98.3%	90.7%	75.8%
ResNet-17	92.3%	89.0%	74.4%

4.3.5 Computational Cost. We use Python to implement CAFE on a PC, which is equipped with an Intel i9-10900X CPU and two NVIDIA 2080Ti GPUs. In this environment, CAFE takes about 0.06s, 0.07s and 0.11s to extract video events and infer score for each frame on UCSDped2, Avenue and ShanghaiTech respectively.

4.3.6 The Influence of Different Anomaly Scores. As we illustrated in Sec. 3.6, the final anomaly score of each video event is obtained by a weighted sum of three different anomaly scores, so we discuss the sensitivity of VAD performance to their weights in this section. We put the results and analysis in supplementary material due to page limit. According to the results, the maximum AUC fluctuation caused by weight variation is 0.8%, 1.5% and 0.5% on UCSDped2, Avenue and ShanghaiTech dataset respectively.

5 CONCLUSION

In this paper, we develop a novel VAD paradigm named CAFE that can accomplish highly effective abnormal event detection from videos. Unlike previous DNN based VAD solutions that typically rely on a low-level learning objective to train the network, we for the first time propose to discriminate anomalies by learning a high-level feature extractor with a feature based distillation pipeline, which facilitates us to explicitly consider valuable high-level semantics. To preserve the temporal correlation in videos, we expand the distillation pipeline by a novel consistency-aware scheme. The proposed scheme encourages adjacent video frames to be assigned with similar high-level features. In addition, we devise a generic framework that can embed one or more low-level learning objectives, so as to incorporate the strengths of both high-level and low-level learning for VAD. Empirical evaluations on commonly-used datasets substantiate the effectiveness of the proposed method.

ACKNOWLEDGMENTS

The work is supported by National Natural Science Foundation of China (62006236, 62072465), NUDT Research Project (ZK20-10) and HPCL Autonomous Project (202101-15). Guang Yu and Siqi Wang contributed equally to this work. Siqi Wang and Zhiping Cai are corresponding authors. E-mail: guangyu@nudt.edu.cn.

REFERENCES

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

[2] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*. 292–301.

[3] Filippo Aleotti, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. 2020. Learning end-to-end scene flow by distilling single tasks knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10435–10442.

[4] Borislav Antić and Björn Ommer. 2011. Video parsing for abnormality detection. In *2011 International Conference on Computer Vision*. IEEE, 2415–2422.

[5] Shuang Ao, Xiang Li, and Charles Ling. 2017. Fast generalized distillation for semi-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4183–4192.

[7] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 535–541.

[8] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11457–11466.

[9] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. 2021. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In *AAAI*.

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.

[11] Y. Chang, Z. Tu, Wei Xie, and J. Yuan. 2020. Clustering Driven Deep Autoencoder for Video Anomaly Detection. In *ECCV*.

[12] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems* 30 (2017).

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

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

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[16] Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems* 29 (2016).

[17] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and F. Yang. 2020. Anomaly Detection With Bidirectional Consistency in Videos. *IEEE transactions on neural networks and learning systems* PP (2020).

[18] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. 2021. Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection. *Proceedings of the 29th ACM International Conference on Multimedia* (2021).

[19] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[20] Mariana-Iuliana Georgescu, Antonio Bărbălău, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. Anomaly Detection in Video via Self-Supervised and Multi-Task Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 12737–12747.

[21] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. A Background-Agnostic Framework with Adversarial Training for Abnormal Event Detection in Video. *IEEE transactions on pattern analysis and machine intelligence* PP (2021).

[22] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *The IEEE International Conference on Computer Vision (ICCV)*.

[23] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

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

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[26] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. 2019. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 578–587.

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

[28] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *ArXiv* abs/1503.02531 (2015).

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

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

[31] Radu Tudor Ionescu, Sorina Smeureanu, M. Popescu, and B. Alexe. 2019. Detecting Abnormal Events in Video Using Narrowed Normality Clusters. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), 1951–1960.

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[34] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning*. 1558–1566.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[36] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. 2020. BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection. *IEEE Transactions on Image Processing* 29 (2020), 2395–2408.

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

[38] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. 2019. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2604–2613.

[39] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13588–13597.

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

[41] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. 2020. Few-shot Scene-adaptive Anomaly Detection. In *ECCV*.

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

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

[44] Weixin Luo, W. Liu, Dongze Lian, J. Tang, Lixin Duan, Xi Peng, and Shenghua Gao. 2021. Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), 1070–1084.

[45] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. 2021. Learning Normal Dynamics in Videos With Meta Prototype Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15425–15434.

[46] Rongrong Ma, Guansong Pang, Ling Chen, and Anton van den Hengel. 2022. Deep Graph-level Anomaly Detection by Glocal Knowledge Distillation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 704–714.

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

[48] Amir Markovitz, Gilad Sharir, Itamar Friedman, L. Zelnik-Manor, and S. Avidan. 2020. Graph Embedded Pose Clustering for Anomaly Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 10536–10544.

[49] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. 2019. Learning Regularity in Skeleton Trajectories for

Anomaly Detection in Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11996–12004.
[50] Trong-Nguyen Nguyen and Jean and Meunier. 2019. Anomaly Detection in Video Sequence with Appearance-Motion Correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*. 1273–1283.
[51] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning Memory-Guided Normality for Anomaly Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 14360–14369.
[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 8024–8035.
[53] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. 2020. A Survey of Single-Scene Video Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
[54] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, E. Sangineto, and N. Sebe. 2018. Plug-and-Play CNN for Crowd Motion Analysis: An Application in Abnormal Event Detection. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018), 1689–1698.
[55] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. 2017. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1577–1581.
[56] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. *CoRR* abs/1412.6550 (2015).
[57] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. 2018. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3379–3388.
[58] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14902–14912.
[59] Che Sun, Y. Jia, Yao Hu, and Y. Wu. 2020. Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos. *Proceedings of the 28th ACM International Conference on Multimedia* (2020).
[60] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. 2020. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4654–4665.
[61] Hanh TM Tran and David Hogg. 2017. Anomaly detection using a convolutional winner-take-all autoencoder. In *Proceedings of the British Machine Vision Conference 2017*. British Machine Vision Association.
[62] Hung Thanh Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Q. Phung. 2019. Robust Anomaly Detection in Videos Using Multilevel Representations. In *AAAI*.
[63] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. 2021. Student-Teacher Feature Pyramid Matching for Anomaly Detection. *arXiv preprint arXiv:2103.04257* (2021).
[64] Lin Wang and Kuk-Jin Yoon. 2021. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE transactions on pattern analysis and machine intelligence* PP (2021).
[65] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. 2021. Glancing at the patch: Anomaly localization with global and local feature comparison. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 254–263.
[66] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4933–4942.
[67] X. Wang, Zhengping Che, Ke Yang, Bo Jiang, Jian-Bo Tang, Jieping Ye, Jingyu Wang, and Q. Qi. 2021. Robust Unsupervised Video Anomaly Detection by Multi-Path Frame Prediction. *IEEE transactions on neural networks and learning systems* PP (2021).
[68] Ziming Wang, Yuexian Zou, and Zeming Zhang. 2020. Cluster Attention Contrast for Video Anomaly Detection. *Proceedings of the 28th ACM International Conference on Multimedia* (2020).
[69] P. Wu, Jing Liu, and Fang Shen. 2020. A Deep One-Class Neural Network for Anomalous Event Detection in Complex Scenes. *IEEE Transactions on Neural Networks and Learning Systems* 31 (2020), 2609–2622.
[70] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding* 156 (2017), 117–127.
[71] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. 2019. AnoPCN: Video Anomaly Detection via Deep Predictive Coding Network. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1805–1813.
[72] Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, and Chengkun Wu. 2022. Deep Anomaly Discovery From Unlabeled Videos via Normality Advantage and Self-Paced Refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13987–13998.
[73] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. 2020. Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. In *Proceedings of the 28th ACM International Conference on Multimedia*. 583–591.
[74] M. Zaheer, Jin ha Lee, M. Astrid, and Seungik Lee. 2020. Old Is Gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 14171–14181.
[75] Y. Zhang, Xiushan Nie, Rundong He, Meng Chen, and Y. Yin. 2020. Normality Learning in Multispace for Video Anomaly Detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2020), 1–1.
[76] Zhiwei Zhang, Shifeng Chen, and Lei Sun. 2021. P-KDGAN: progressive knowledge distillation with GANs for one-class novelty detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 3237–3243.
[77] Bin Zhao, Li Fei-Fei, and Eric P Xing. 2011. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*. IEEE, 3313–3320.
[78] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1933–1941.
[79] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. 2019. AnomalyNet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security* (2019).
[80] Joey Tianyi Zhou, Le Zhang, Zhiwen Fang, Jiawei Du, Xi Peng, and Xiao Yang. 2019. Attention-Driven Loss for Anomaly Detection in Video Surveillance. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).