Video Anomaly Detection via Visual Cloze Tests

Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, *Senior Member, IEEE*, En Zhu, and Jianping Yin

***Abstract*— Although great progress has been sparked in video anomaly detection (VAD) by deep neural networks (DNNs), existing solutions still fall short in two aspects: 1) The extraction of video events cannot be both precise and comprehensive. 2) The semantics and temporal context are under-explored. To tackle above issues, we are inspired by cloze tests in language education and propose a novel approach named *Visual Cloze Completion* (VCC), which conducts VAD by completing *visual cloze tests* (VCTs). Specifically, VCC first localizes each video event and encloses it into a spatio-temporal cube (STC). To realize both precise and comprehensive event extraction, appearance and motion are used as complementary cues to mark the object region associated with each event. For each marked region, a normalized patch sequence is extracted from several neighboring frames and stacked into a STC. With each patch and the patch sequence of a STC regarded as a visual "word" and "sentence" respectively, we deliberately erase a certain "word" (patch) to yield a VCT. Then, the VCT is completed by training DNNs to infer the erased patch and its optical flow via video semantics. Meanwhile, VCC fully exploits temporal context by alternatively erasing each patch in temporal context and creating multiple VCTs. Furthermore, we propose localization-level, event-level, model-level and decision-level solutions to enhance VCC, which can further exploit VCC's potential and produce significant VAD performance improvement. Extensive experiments demonstrate that VCC achieves highly competitive VAD performance.**

***Index Terms*— Video anomaly detection, visual cloze tests.**

I. Introduction

VIDEO anomaly detection (VAD) [1], which aims to automatically detect abnormal events in surveillance videos, enjoys enormous potential to various security-critical realms like municipal management, traffic monitoring and emergency reaction. Formally, VAD refers to detecting suspicious video events that deviate from the normal routine. With many attempts made [1], VAD remains a challenging task. This can be ascribed to the *scarcity*, *ambiguity* and *unpredictability* of anomalies [2], which renders the direct modeling of abnormal events unrealistic. As a result, VAD usually follows the *one-class classification* setup [3]: At the training stage, videos with only normal events are collected as they are highly accessible. A normality model is then built with those normal videos. For inference, all events that do not comply with this normality model are viewed as abnormal. As the labels for anomalies and normal sub-classes are both absent, VAD is usually addressed by unsupervised or self-supervised approaches. In the literature, VAD solutions can be categorized into the classic VAD methods and recent DNN based VAD methods (reviewed in Sec. II). Classic VAD relies on hand-crafted descriptors to extract features from videos, while features are then fed into classic anomaly detection models for VAD. By contrast, DNN based VAD is inspired by DNN's success in numerous vision tasks [4]. It not only avoids complex feature engineering, but also achieves superior performance to classic VAD.

Despite the remarkable success and dominant role of DNN based VAD, we notice its two prominent issues: **(1)** *Existing methods for DNN based VAD cannot achieve a both precise and comprehensive extraction of video events in the first place.* As discussed in Sec. III-A.1.a, "precise" refers to localizing a video event with a compact bounding box, while "comprehensive" means extracting all video events without omission. Early VAD methods usually extract video events by a multi-scale sliding window [5], [6], but it often splits one object into multiple windows, which leads to imprecise extraction. Meanwhile, some VAD methods like [7], [8], and [9] simply overlook the event extraction by learning on a per-frame basis. However, such a way is vulnerable to several problems, *e.g.* scale variations due to foreground depth as well as foreground-background imbalance [10], [11]. Recently, few works [12], [13], [14] achieve more precise extraction by a pre-trained generic object detector, but another fatal "*closed world*" problem arises: The pre-trained detector is unable to recognize novel foreground, thus leading to non-comprehensive event extraction. More importantly, the subjects of many abnormal events are intrinsically novel due to VAD's nature.

(2) *Existing methods for DNN based VAD usually cannot fully exploit the video semantics and temporal context to discriminate anomalies.* As illustrated by Fig. 1, DNN based VAD typically follows two learning paradigms (*reconstruction* or *frame prediction*), but they are both unsatisfactory: Reconstruction based methods learn to reconstruct normal events and view poorly reconstructed events as abnormal. However, simple reconstruction drives DNNs to memorize low-level details rather than meaningful semantics [15], while abnormal events are also reconstructed well in many cases [16]. By contrast, frame prediction based methods aim to predict

Manuscript received 16 January 2022; revised 4 June 2023 and 23 July 2023; accepted 26 July 2023. Date of publication 31 July 2023; date of current version 15 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072465 and in part by the Science and Technology Innovation Program of Hunan Province under Grant 2022RC3061. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alptekin Küpçü. *(Guang Yu and Siqi Wang contributed equally to this work.) (Corresponding author: Zhiping Cai.)*

Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, and En Zhu are with the College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: guangyu@nudt.edu.cn; wangsiqi10c@nudt.edu.cn; zpcai@nudt.edu.cn; xinwangliu@nudt.edu.cn; enzhu@nudt.edu.cn).

Jianping Yin is with the School of Cyberspace Science, Dongguan University of Technology, Dongguan 523808, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2023.3300094, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2023.3300094

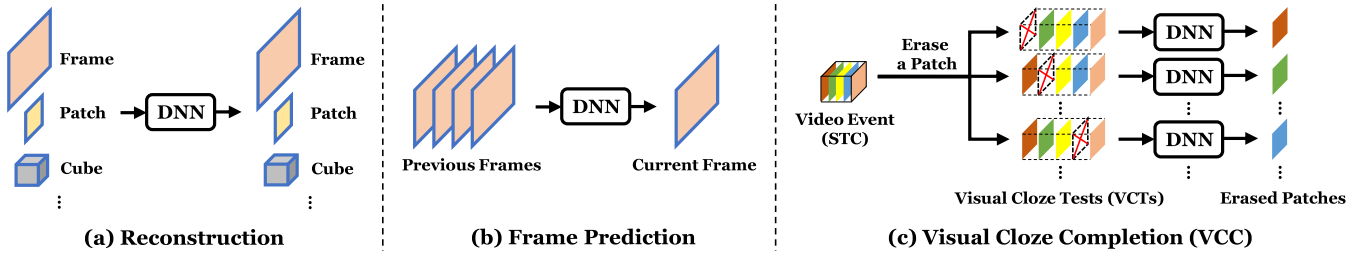

Fig. 1. Learning paradigm comparison for DNN based VAD. **(a)** Reconstruction based methods train DNN to reconstruct data collected from normal training videos. **(b)** Frame prediction based methods take previous frames as inputs of DNN to predict current frame. **(c)** VCC first encloses video events with STCs in a both precise and comprehensive manner. Different types of VCTs are then created by erasing the patch at different temporal positions. Afterwards, a separated DNN is trained to complete each type of VCTs with the generated patch or auxiliary information. Note that cubes used for reconstruction are different from STCs in VCC, as they are yielded by a relatively coarse strategy and cannot enclose video events both precisely and comprehensively.

a normal video frame from previous frames, and poorly predicted frames are believed to contain anomalies. Prediction avoids reducing training loss by simply memorizing low-level details. Nevertheless, it typically scores each video frame only by the prediction errors of a single frame, whilst the temporal context with valuable clues is not fully exploited.

Unlike many recent efforts that focus on searching better DNN architectures for reconstruction or frame prediction, we are inspired by the popular cloze test in language study, and propose a new paradigm named *Visual Cloze Completion (VCC)*. As Fig. 1 shows, the core idea of VCC is to train DNNs to complete a series of *visual cloze tests (VCTs)*, which comprises of two major steps: **(1)** Extracting video events to construct VCTs. To harvest video events in a both precise and comprehensive way, we leverage appearance and motion as mutually complementary cues to localize the foreground object region associated with each event. Based on localization results, a normalized patch sequence is extracted and stacked into one spatio-temporal cube (STC), which serves as the enclosure of a video event. With each patch in STC compared to a "word", we can view the whole patch sequence of the STC as a "sentence". In this way, a VCT can be constructed by erasing a certain "word" (patch) in the "sentence" (STC). **(2)** Learning to complete VCTs. Specifically, DNNs are trained to "answer" the VCT by inferring the erased patch, which requires DNNs to attend to the video semantics (*e.g.* high-level body parts) rather than only low-level details. Meanwhile, VCC is equipped with two ensemble strategies, *VCT type ensemble* and *modality ensemble*: VCT type ensemble creates multiple types of VCTs by alternatively erasing each patch in STC, which allows each patch in the temporal context to be considered. Modality ensemble requires DNNs to infer the erased patch's optical flow, which incorporates richer motion semantics like appearance-motion correspondence. In this way, the proposed VCC paradigm is able to handle the above two issues effectively for better VAD performance.

A preliminary version of this paper is presented in [17]. Compared with [17], we mainly extend the original work in terms of five aspects: **(1)** We improve the motion cue to achieve more robust foreground localization. This enables the localization results to be more noise-resistant, and produces less artifacts and misinterpreted video events. **(2)** We design a spatially-localized strategy to alleviate the scale variation problem during video event extraction. The strategy enables video events extracted from one local spatial region are modeled separately, which ensures video events with the comparable scale to be processed by DNNs. **(3)** We design a new DNN architecture named spatio-temporal UNet (ST-UNet), which makes it possible to build a stronger normality model for VAD. When compared with the standard UNet used in [17], ST-UNet synthesizes a recurrent network structure to accumulate temporal context information in STCs and produce high-level feature maps, which facilitates the proposed VCC paradigm to learn richer video semantics. **(4)** We design a mixed score metric and a score rectification strategy, which prove to be simple but highly effective strategies for the anomaly decision stage. **(5)** We carry out more extensive experiments on various benchmark datasets to justify the effectiveness of VCC, and more in-depth analysis and discussion are also provided. In this paper, our main contributions are summarized below:

- We propose a novel video event extraction pipeline by leveraging both appearance and motion as mutually complementary cues. To our best knowledge, this is the first VAD work to explicitly clarify the necessity of a both precise and comprehensive video foreground localization, which overcomes the "closed-world" problem and lays a firm foundation for VAD in the first place.
- We propose to conduct VAD by building and completing VCTs, which offers a promising alternative to frequently-used reconstruction or frame prediction paradigm.
- We propose VCT type ensemble and modality ensemble strategy respectively, so as to fully exploit the temporal context and motion information in video events.
- We further propose localization-level, event-level, model-level and decision-level solutions respectively to further enhance VCC, which enables us to better develop VCC's potential and obtain remarkable performance gain.

Extensive experiments demonstrate that VCC can achieve highly competitive VAD performance, and it is detailed below.

II. Related Work

A. Classic VAD

Classic VAD methods [18], [19], [20], [21], [22], [23], [24], [25] are usually comprised of two stages: The feature extraction stage based on carefully designed hand-crafted feature descriptors, as well as a separated VAD stage based on

classic anomaly detection methods. In the feature extraction stage, feature descriptors have been thoroughly explored, *e.g.* dynamic texture [21], optical flow [26], [27], spatio-temporal gradients [28], [29]. In the subsequent VAD stage, features of video events are fed into a classic anomaly detection method to model normality and discern anomalies. Various methods have been explored for this purpose, *e.g.* sparse coding and its variants [30], [31], one-class classifier [32], [33], sociology or nature inspired models [34], [35]. As feature engineering can be tedious and complex, many recent researches have turned to DNN based VAD.

B. DNN Based VAD

Instead of extracting features from video events by manually designed descriptors, DNN based VAD aims to learn proper features automatically from video events via DNNs. Learned features can be either fed into a classic anomaly detection method, or directly used for end-to-end VAD. With only roughly labeled normal videos for training, most DNN based VAD methods follow a *reconstruction* or *frame prediction* paradigm: **(1)** Reconstruction based methods learn to reconstruct normal video events in training, and assume that a poor reconstruction indicates the emergence of an abnormal event. Deep autoencoder (AE) and its variants are the most frequently-used model for reconstruction: The pioneer work from Xu et al. [36] introduces fully-connected stacked denoising AE to address VAD, and its improved version is reported in [5]; Hasan et al. [37] leverage convolutional AE (CAE) as an alternative to AE, since CAE is more suitable for modeling images and videos. Then, numerous CAE variants are explored in recent research, such as Winner-take-all CAE [38], Long Short Term Memory based CAE [39], variational AE (VAE) [40] and memory-augmented AE [16]; Wang et al. [41] propose to combine VAE and UNet to achieve more accurate pixel-wise reconstruction; Abati et al. [42] propose a combination of AE and a parametric density estimator. Besides, the cross-modality reconstruction [8], [43] is shown to be promising. Apart from AE, other types of DNNs like sparse coding based recurrent neural network (RNN) [44], [45], [46] and generative adversarial network [47], [48], [49] are also explored for reconstruction based VAD. Recently, other techniques are utilized to incorporate AE to achieve reconstruction based VAD, such as fast sparse coding [50], deep embedded clustering [51], deep k-means [52], deep support vector domain description [53]. **(2)** *Frame prediction* based methods learn to predict current frames by previous frames, while a poorly predicted frame is assumed to contain anomalies. Liu et al. [7] for the first time validate frame prediction as a useful baseline for DNN based VAD, and they also impose appearance and motion constraints to guarantee the quality of predicting normal events. Afterwards, Lu et al. [54] improve prediction by a convolutional variational RNN model. Since prediction on a per-frame basis leads to the bias towards background [10], Zhou et al. [11] introduce the attention mechanism in prediction. Other methods [55], [56], [57], [58] are also proposed to enhance prediction, such as bidirectional prediction [59], multi-timescale prediction [60], multi-space prediction [61], multi-path prediction [62]. Another natural instinct is to combine prediction with reconstruction into a hybrid paradigm [9], [63], [64], [65], [66]. In addition to reconstruction and frame prediction, other DNN based methods [67], [68], [69], [70], [71], [72], [73] are also explored. Hinami et al. [12] propose to detect and recount abnormal events by integrating a generic model and environment-dependent anomaly detectors. Wang et al. [74] introduce contrastive learning to VAD.

III. The Proposed Method

A. Basic Visual Cloze Completion (VCC)

1) Video Event Extraction: An appropriate representation of video events is the foundation for good VAD performance. To this end, we simply assume that a video event is supposed to enclose a subject (*i.e.*, a foreground object) and its activity during a temporal interval. Therefore, a natural solution is to enclose a video event by a spatio-temporal cube (STC) denoted by V. To build a STC, the spatial region of the subject on the video frame, which is viewed as the region of interest (RoI) here, should be marked by a bounding box. With the location of this RoI, a patch sequence $(p_1, \cdots, p_D)$ with D patches is extracted from the current and $(D-1)$ temporally adjacent frames to describe the activity of this subject. Since DNNs usually take fixed-size inputs, we resize those patches into $h \times w$ new patches $(p'_1, \cdots, p'_D)$ and stack them into a $h \times w \times D$ STC: $V = [p'_1; \cdots; p'_D]$. Note that we use a small D because it facilitates us to safely assume that the subject of video event stays in the RoI during the short temporal interval.

a) Motivation: To extract high-quality STCs to represent video events, the key is to localize RoIs of foreground objects. In this paper, we argue that the localization should be both *precise* and *comprehensive*. Specifically, *precise* localization expects the whole region of a foreground object to be covered by a compact bounding box, while the box contains minimal background; *comprehensive* localization requires all foreground objects to be extracted without omission. However, as discussed by the first issue in Sec. I, existing VAD methods fail to realize precise and comprehensive localization simultaneously, and we intuitively illustrate this in Fig. 2: The classic sliding window strategy often splits one foreground object by several windows (Fig. 2 (a)); Motion based localization cannot discriminate different objects and extract excessive irrelevant background (Fig. 2 (b)); The object detector that only uses appearance cues tends to omit novel or blurring objects (Fig. 2 (c)). Thus, such localization hinders DNNs from building a good normal event model and undermines VAD performance.

To this end, we recall that a video event is defined to be a foreground object and its activity. Thus, both *appearance cues* from objects and *motion cues* from their activities need to be considered for extracting RoIs. As to appearance cues, the impressive success of modern object detection [13], [75] naturally motivates us to leverage a generic object detector, which can exploit appearance cues efficiently for localization. With generic knowledge from large-scale real-world datasets like Microsoft COCO [76], the pre-trained detector is able

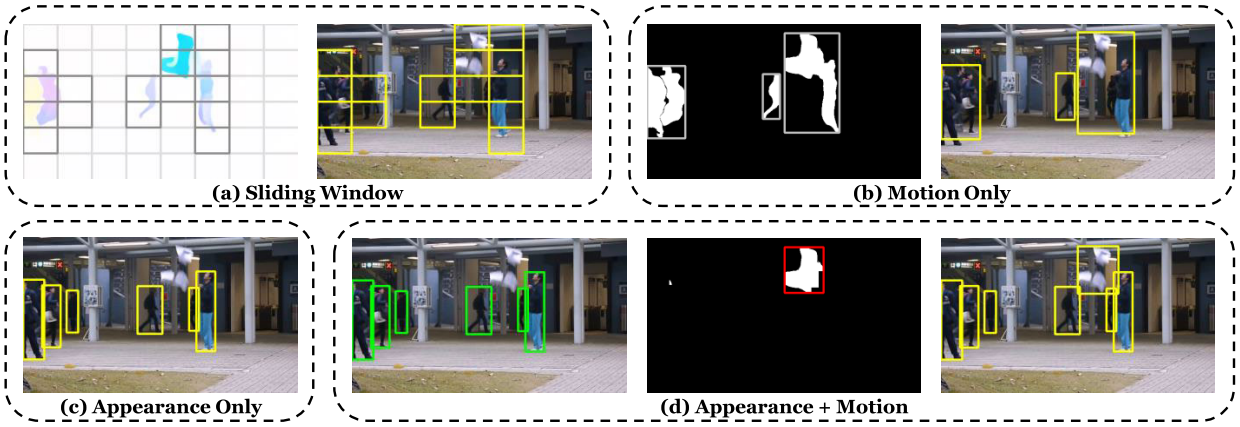

Fig. 2. Comparison of RoI localization: Imprecise localization will be produced by sliding window (a) or motion only (b), while non-comprehensive localization will be yielded by appearance only (c). Both precise and comprehensive localization can be achieved by the proposed pipeline (d).

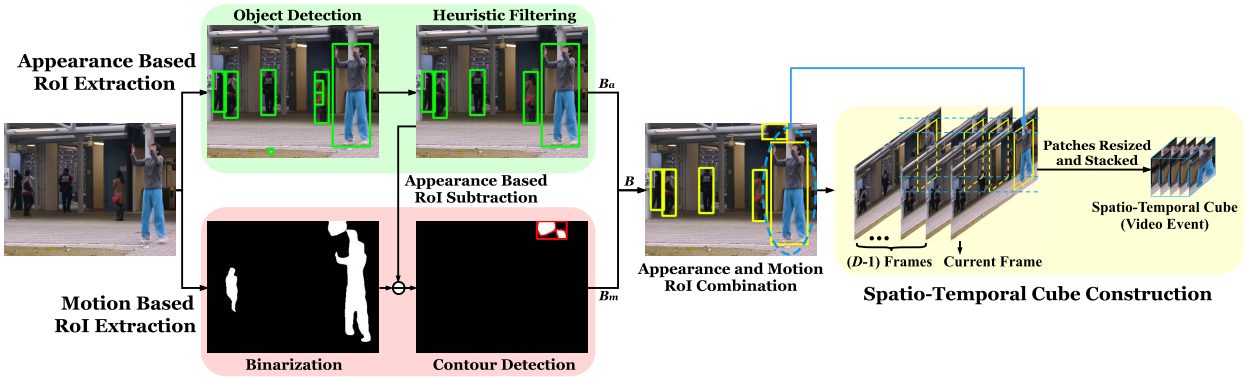

Fig. 3. Video event extraction pipeline: **(1)** Appearance based RoI extraction (green): It utilizes a pre-trained object detector and some efficient filtering rules to extract appearance based RoIs. **(2)** Motion based RoI extraction (red): First, the motion map of the current frame is binarized by magnitude into a binary map. Then, the highlighted pixels in appearance based RoIs are subtracted from the binary map. Finally, contour detection and thresholding are applied to the binary map to obtain motion based RoIs. **(3)** Spatio-temporal cube (STC) construction (yellow): For each RoI, patches from the current frame and $(D-1)$ previous frames are extracted. D patches are then resized and stacked into a STC, which encloses a video event.

to extract the majority of daily objects (*e.g.* humans and vehicles) in a highly precise manner. However, as illustrated in Sec. I, RoI extraction with only appearance cues is non-comprehensive due to the fatal "closed-world" problem. To this end, motion cues provide valuable complementary information to localize omitted foreground objects, which enables us to overcome the "closed-world" problem and accomplish more comprehensive RoI extraction. Inspired by those ideas, we propose a new pipeline with both appearance based and motion based RoI extraction, which are presented below.

b) Appearance based RoI extraction: With a pre-trained generic object detector M and a raw frame I_a from videos as inputs, appearance based RoI extraction intends to yield a RoI set B_a via appearance cues of foreground objects, where $B_a \subseteq \mathbb{R}^4$ and each item $b_{ap} \in B_a$ stands for a RoI marked by a bounding box. Note that a bounding box is represented by a 4-D vector that contains the coordinates of its top-left and bottom-right vertex. As the green part in Fig. 3 shows, I_a is sent into M to produce a preliminary RoI set B_{ap} by selecting those output bounding boxes with confidence scores larger than a threshold T_s. The output class labels from M are discarded, *i.e.* M is only used to provide localization and no fine-grained class information is exploited. Afterwards, two efficient rules are designed to remove RoIs that are evidently undesirable: **(1)** RoI area threshold T_a, which eliminates overly small RoIs. **(2)** Overlapping ratio T_o, which deletes nested or significantly overlapped RoIs in B_{ap}. Thus, we can assure that extracted RoIs based on appearance cues provide precise localization of most foreground objects in daily life.

c) Motion based RoI extraction: To localize foreground objects beyond the coverage of the pre-trained detector, motion based RoI extraction exploits motion cues to provide a supplementary bounding box set B_m. Specifically, as can be seen in the red part of Fig. 3, we first introduce a motion map I_m, which contains the motion magnitude of each pixel on the current frame, as our motion cues. To obtain I_m, the most straightforward way is to compute temporal gradients with two consecutive frames [17], and other sophisticated means can also be explored for better computation of I_m (discussed in Sec. III-B.1). With such a motion map, we can simply binarize the motion map by a threshold T_b and yield a binary map that manifests RoI regions with drastic motion. Rather than a direct localization on this binary map, we first substract appearance based RoIs in B_a from the map, which facilitates more precise motion based RoI extraction for two reasons: First, such

Algorithm 1 The Proposed RoI Extraction Pipeline

Input: Frame I_a and its motion map I_m, pre-trained object detector M, threshold $T_s, T_a, T_o, T_b, T_{ar}$
Output: RoIs represented by a bounding box set B
$B_{ap} \leftarrow ObjDet(I_a, M, T_s)$ # Object detection
$B_a = \{\}$ # Rule based filtering
for $b_{ap} \in B_{ap}$ **do**
  if $Area(b_{ap}) > T_a$ & $Overlap(b_{ap}, B_{ap}) < T_o$ **then**
    $B_a = B_a \cup \{b_{ap}\}$
  end if
end for
$I_m^{(b)} \leftarrow Bin(I_m, T_b)$ # Motion map binarization
$I_m^{(b)} \leftarrow RoISub(I_m^{(b)}, B_a)$ # Remove RoIs in B_a
$\mathcal{C} \leftarrow ContourDet(I_m^{(b)})$ # Contour detection
$B_m = \{\}$
for $c \in \mathcal{C}$ **do**
  $b_m = ContourBox(c)$ # Get contour bounding box
  if $Area(b_m) > T_a$ & $\frac{1}{T_{ar}} < AspectRatio(b_m) < T_{ar}$ **then**
    $B_m = B_m \cup \{b_m\}$
  end if
end for
$B = B_a \cup B_m$

subtraction encourages the localization process to focus on those omitted foreground objects and produce more precise RoIs for them, otherwise the overlap of multiple objects will generate large connected RoIs (see Fig. 2 (b)). Second, the subtraction also avoids redundant computing overhead. Finally, we conduct contour detection to obtain the RoI contour, which then gives us the corresponding bounding box b_m. Similarly, we adopt two filtering rules (RoI area threshold T_a and maximum aspect-ratio threshold T_{ar}) to refine motion based RoIs into a set B_m. The final RoI set B is the union of two complementary RoI sets $B = B_a \cup B_m$, and the entire RoI extraction pipeline is summarized by Algorithm 1 and Fig. 3. With those RoIs, we are able to extract high-quality STCs and then construct VCTs for DNNs to solve.

2) Visual Cloze Tests (VCTs): With extracted high-quality STCs, our next step is to build a normality model by some learning paradigm. However, just as we discussed by the second issue in Sec. I, frequently-used reconstruction or frame prediction paradigm cannot fully exploit video semantics and temporal context information. To remedy this problem, we are inspired by popular *cloze test*, which requires students to complete an incomplete text with certain words or phases deliberately erased, so as to test students' grasp of the semantics and their ability to exploit context information [77]. In natural language processing (NLP), this idea has been explored to build large-scale language model [78]. Since video semantics and context information are also of paramount importance to discriminating abnormal events, we are naturally inspired to design *visual cloze tests (VCTs)* as a counterpart of cloze test in computer vision. As we assume a video event to be enclosed by a STC, the patch sequence of the STC naturally corresponds to a visual "sentence" that describes the video event, while a patch p'_i can be viewed a visual "word". With such an analog, a VCT can be built by erasing any patch p'_i from a STC. To complete the VCT, DNNs are required to give an inferred patch $\tilde{p}'_i$, which is supposed to be as close to p'_i as possible.

VCTs bring two benefits: **(1)** To complete a VCT, DNNs are encouraged to capture video semantics in STC. For example, consider a video event that describes a walking person. DNNs must attend to the motion of some key high-level parts (*e.g.* the forwarding leg and swinging arm) in STC to realize a good completion. **(2)** Since any patch in a STC can be erased to create a VCT, we can readily build multiple VCTs by erasing every possible patch. In this way, the temporal context is fully exploited by considering each patch in this context for completion. In this paper, our VCC method performs VAD by two types of VCT completion, *appearance completion* and *motion completion*, and equip them with two *ensemble strategies*. We illustrate appearance and motion completion for a type-i VCT in Fig. 4 and detail them below.

a) Appearance completion: With the j-th extracted event denoted by the STC $V_j = [p'_{j,1}; \cdots; p'_{j,D}]$, a VCT $V_j^{(i)} = [p'_{j,1}; \cdots p'_{j,i-1}; p'_{j,i+1}; \cdots p'_{j,D}]$, $i \in \{1, \cdots D\}$ is built by erasing the i-th patch $p'_{j,i}$ of V_j (see the blue part in Fig. 4). It should be noted that any VCT built by erasing the i-th patch of a STC is called a *type-i VCT*, and all type-i VCTs are collected as the *type-i VCT set* $\mathcal{V}^{(i)} = \{V_1^{(i)}, \cdots V_N^{(i)}\}$, where N is the number of extracted video events (STCs). Afterwards, as shown by green part in Fig. 4, a type-i VCT $V_j^{(i)}$ in $\mathcal{V}^{(i)}$ and its corresponding erased patch $p'_{j,i}$ are viewed as the input and completion goal respectively to train a generative DNN $f_a^{(i)}$, which then generates a patch $\tilde{p}'_{j,i} = f_a^{(i)}(V_j^{(i)})$ to fill the "blank" of the VCT $V_j^{(i)}$. $f_a^{(i)}$ can be implemented by multiple network architectures, *e.g.* a standard UNet used in basic VCC, and we will explore a more sophisticated solution later (see Sec. III-B.3). To train $f_a^{(i)}$, we minimize the *appearance loss*, which computes the difference between the erased patches and inferred patches for type-i VCT set $\mathcal{V}^{(i)}$:

$$\mathcal{L}_a^{(i)} = \frac{1}{N} \sum_{j=1}^{N} \| \tilde{p}'_{j,i} - p'_{j,i} \|_2^2 \tag{1}$$

Note that we slightly abuse the notation by viewing $p'_{j,i}$ ($\tilde{p}'_{j,i}$) as the column vector yielded by concatenating all columns of the original 2D patch $p'_{j,i}$ ($\tilde{p}'_{j,i}$). Since the goal of appearance completion is normalized small patches, we discover that a simple appearance loss like Eq. 1 is sufficient for yielding high-quality completions for VCT. Thus, adversarial training in frame based VAD methods like [7] is unnecessary for our patch based completion. It should be noted that the DNN $f_a^{(i)}$ only handles VCTs from the type-i VCT set $\mathcal{V}_j^{(i)}$, which enables $f_a^{(i)}$ to be more specialized and easier to train.

Since DNNs are trained to complete VCTs created by only normal events, it is difficult for DNNs to complete VCTs from unseen abnormal events, which would lead to larger completion errors for abnormal events. Thus, we can measure the quality of completions to compute the anomaly score to realize VAD. To this end, we can flexibly select any score metric $\mathcal{S}_a^{(i)}(\tilde{p}'_{j,i}, p'_{j,i})$, such as mean square error (MSE) or

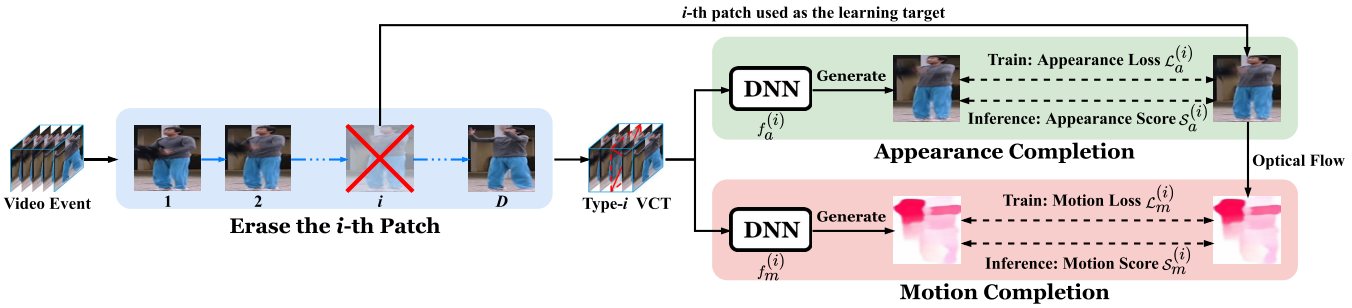

Fig. 4. The basic procedure of VCC based VAD approach: **(1)** Constructing a type-i VCT (blue): The i-th patch of a STC is erased to build a type-i VCT, while the erased patch is used as the learning target of appearance completion. **(2)** Appearance completion (green): To complete the VCT, a DNN takes the patches in VCT as input and learns to generate the erased patch. **(3)** Motion completion (red): Another DNN takes the VCT as input and learns to generate the optical flow patch of the erased patch.

Peak Signal to Noise Ratio (PSNR) [7], to compute completion errors and yield the anomaly score of patch $p'_{j,i}$. In fact, our preliminary work [17] shows that choosing $\mathcal{S}_a^{(i)}(\tilde{p}'_{j,i}, p'_{j,i})$ to be MSE proves to be very effective to score anomalies, but we will show that a mixed score that combines different metrics can boost the VAD performance (see Sec. III-B.4).

b) Motion completion: Motion is the other important attribute of videos, so we also take motion information into account when building and completing VCTs. For this purpose, dense optical flow can be leveraged as a highly accessible and effective representation of per-pixel motion in videos. Concretely, it estimates the motion displacement (dx, dy) of the pixel at the position (x, y) between two consecutive frames with time interval dt, which are assumed to satisfy $P(x, y, t) = P(x + dx, y + dy, t + dt)$, where $P(x, y, t)$ denotes the pixel intensity at the position (x, y) for time t. Optical flow can be computed by either classic methods or DNN based methods [79]. For efficiency, we estimate the dense optical flow by a pre-trained FlowNetv2 model [80]. With estimated optical flow map of each frame, we can obtain optical flow patches $(o_{j,1}, \cdots o_{j,D})$ that correspond to video patches $(p_{j,1}, \cdots p_{j,D})$ in V_j, and resize them into $h \times w$ patches $(o'_{j,1}, \cdots o'_{j,D})$. Motion completion requires a DNN $f_m^{(i)}$ to infer the optical flow patch of the erased patch $p'_{j,i}$ by $V_j^{(i)}$, *i.e.* $\tilde{o}'_{j,i} = f_m^{(i)}(V_j^{(i)})$, so as to make the inferred optical flow $\tilde{o}'_{j,i}$ to be as close to $o'_{j,i}$ as possible. Similar to appearance completion, $f_m^{(i)}$ is trained with motion loss $\mathcal{L}_m^{(i)}$:

$$\mathcal{L}_m^{(i)} = \frac{1}{N}\sum_{j=1}^{N} \|\tilde{o}'_{j,i} - o'_{j,i}\|_2^2 \tag{2}$$

Likewise, we use the same way as appearance to define the motion anomaly score $\mathcal{S}_m^{(i)}(\tilde{o}'_{j,i}, o'_{j,i})$ during inference. With motion completion, we encourage the DNN to infer the motion statistics from the temporal context provided by VCTs, which enables it to consider richer video semantics like appearance-motion correspondence of foreground objects.

c) Ensemble strategies: Ensemble technique enables one to establish a more effective model by joining several models [81]. We propose to equip VCTs with two ensemble strategies, so as to fully unleash its potential: **(1)** *VCT type ensemble*. To fully exploit the temporal context for VAD, each patch in the temporal context of a video event (STC) should be involved when computing the video event's anomaly score. To this end, we notice that one STC will produce D different VCTs, thus making it possible to consider each patch in the temporal context for completion. Therefore, we propose to compute the final appearance anomaly score for a video event by an ensemble of scores, which are obtained by completing all different types of VCTs created from this event:

$$\mathcal{S}(V_j) = \frac{1}{D}\sum_{i=1}^{D} \mathcal{S}^{(i)} \tag{3}$$

When $\mathcal{S}_a^{(i)}(\tilde{p}'_{j,i}, p'_{j,i})$ or $\mathcal{S}_m^{(i)}(\tilde{o}'_{j,i}, o'_{j,i})$ is used as $\mathcal{S}^{(i)}$, we yield the final appearance anomaly score $\mathcal{S}_a(V_j)$ or motion anomaly score $\mathcal{S}_m(V_j)$. **(2)** *Modality ensemble*. To fuse results from appearance and motion to yield the overall anomaly score, we use a weighted sum of $\mathcal{S}_a(V_j)$ and $\mathcal{S}_m(V_j)$ to compute the overall anomaly score $S(V_j)$ for a video event V_j:

$$\mathcal{S}(V_j) = w_a \frac{\mathcal{S}_a(V_j) - \bar{\mathcal{S}}_a}{\sigma_a} + w_m \frac{\mathcal{S}_m(V_j) - \bar{\mathcal{S}}_m}{\sigma_m} \tag{4}$$

where $\bar{\mathcal{S}}_a, \sigma_a, \bar{\mathcal{S}}_m, \sigma_m$ are the means and standard deviations of appearance and motion scores for all events in the training set, which are used to normalize appearance and motion anomaly scores into the same scale. To score a frame, the maximum of all events' scores on a frame is the frame score.

B. Enhanced Visual Cloze Completion

Although basic VCC already performs satisfactorily, this section will further introduce localization-level, event-level, model-level and decision-level enhancement solutions, which aggregate into the enhanced VCC for even better performance.

1) Localization-Level Enhancement: As discussed in Sec. III-A.1.c, localization of motion based RoIs requires to compute the motion map I_m, which is yielded by computing temporal gradients in our previous work [17]. Nevertheless, computing I_m by temporal gradients suffer from two major drawbacks: First, the appearance of foreground objects will impose a significant influence on the magnitude of temporal gradients, which makes them less reliable to reflect motion. For example, two pedestrians with the same speed may produce different temporal gradients when they are in clothes with

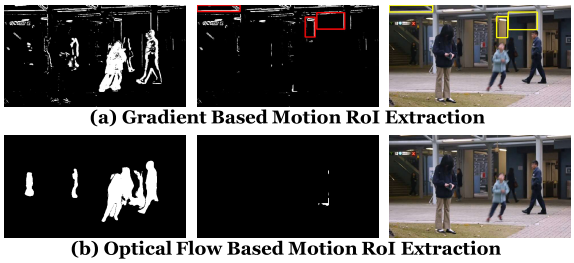

Fig. 5. Comparison of different motion cues for motion based RoI extraction.

different colors. Second, temporal gradients are susceptible to low-level noises. These noises can be pervasive in real-world videos due to various factors like illumination changes or gentle vibration of camera. Disturbed by low-level noises, temporal gradients could generate massive low-level artifacts in motion map (see Fig. 5 (a)). Even after the rule based filtering, some artifacts are still misinterpreted as RoIs.

Motivated by observations above, we propose to employ optical flow maps as more robust motion cues in this paper. When compared with temporal gradients, optical flow is blessed with several strengths: First, optical flow is insensitive to appearance, as it is based on correspondence rather than intensity changes. Second, optical flow, which is estimated by pre-trained FlowNetv2 model, is more robust to low-level noises, owing to FlowNetv2's correlation layer for high-level features and stacked architecture for noise reduction [80]. Third, optical flow is required for motion completion (see Sec. III-A.2.b), so no additional computation is actually required. As the example shows in Fig. 5 (b), we can yield a smoother binary map with less low-level artifacts to indicate motion regions when the optical flow map is used as I_m, and misinterpreted RoIs are effectively removed. As a result, improving motion cues by optical flow is able to enhance VCC by more precise motion based RoI localization, which leads to better performance. Besides, optical flow's robustness also makes it easier to determine the binarization threshold T_b, as value of T_b can be unified among different datasets regardless of the differences in objects' appearance and scenarios.

2) Event-Level Enhancement: Video events in many real-world scenarios are influenced by varied foreground depth. With different depth, the same type of video events may exhibit different sizes and scales, which pose an important challenge to modeling and inference. For example, in a typical scene from UCSDped1 dataset (see Fig. 6), pedestrians at the bottom left corner have obviously larger size and motion (optical flow) magnitude than those at the top right corner. Hence, video events from the same category (pedestrian walking) suffer from a large intra-class difference that undermines the one-class learning for VAD. Although our video event extraction scheme somewhat alleviates this problem by normalizing all RoIs into the same size, it does not address this problem from the root for two reasons: First, the normalization is performed spatially, while the motion magnitude cannot be adjusted by spatial interpolation. Second, foreground objects with different depths have different levels of clarity, which cannot be changed by normalization.

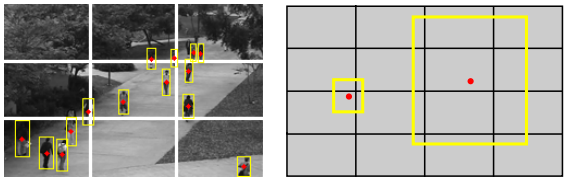

Fig. 6. The spatially-localized strategy for block-based learning and inference.

To mitigate this issue, we design a spatially-localized strategy to assist video event extraction and modeling. As shown in Fig. 6, the core idea of the spatially-localized strategy is to divide the video frame into several local spatial regions, *a.k.a.* blocks. Since each block only covers a local spatial region, we can safely assume that all foreground objects in this block share similar depth. Afterwards, video events in one block are modeled or tested in a separated manner, so as to enable DNNs to only handle video events with a comparable scale. For example, when dividing the video frame into 9 blocks according to Fig. 6 (left), each block separately collects the assigned video events to train independent DNNs for learning and inference. To assign a video event to a block, a simple and natural criterion is to assign it to the block that *enjoys maximum overlap with the video event's bounding box*. Based on this criterion, we propose to introduce a simple theorem below to efficiently determine the assignment of video events:

Theorem 1: Given an arbitrary 2-D rectangle b in a 2-D plane, the plane is uniformly partitioned into rectangular local regions $\{R_i\}_{i=1}^{\infty}$ *with any size. If the rectangle b's geometric center* c_b *and the* k_{th} *local region* R_k *satisfy* $c_b \in R_k$*, the overlap area* $\mathcal{O}(b, R_k)$ *of b and* R_k *satisfies*:

$$\mathcal{O}(b, R_k) = \sup_i \{\mathcal{O}(b, R_i)\} \tag{5}$$

The above conclusion can be generalized to any n-D hyperspace, where n is an positive integer.

The proof of Theorem 1 is given in supplementary material. Theorem 1 reveals that we can simply determine if a video event belongs to a block by checking whether the center of its bounding box (red dots in Fig. 6) lies in this block, since it guarantees a maximum overlap. Note that Theorem 1 holds only when the video frame is uniformly partitioned into rectangle blocks like Fig. 6. However, more fine-grained irregular division is also applicable: One can simply select a single frame from the training videos, and manually divide the frame into several irregular blocks that better describe the depth of different spatial regions, which actually requires minimal cost and labor. Since the surveillance videos usually share fixed background, the division can be fixed in later process. More importantly, a well-designed block division allows the block-based models to detect location-specific anomalies, which can be difficult to detect and often ignored by most VAD methods. Taking UCSDped1 dataset as an example, people walking on the sidewalk are normal, while people walking on the lawn are anomalous. When the lawn is divided into a separate block, people who walk on the grass can be easily detected as abnormal in inference, as almost no objects appear on the grass in the training videos.

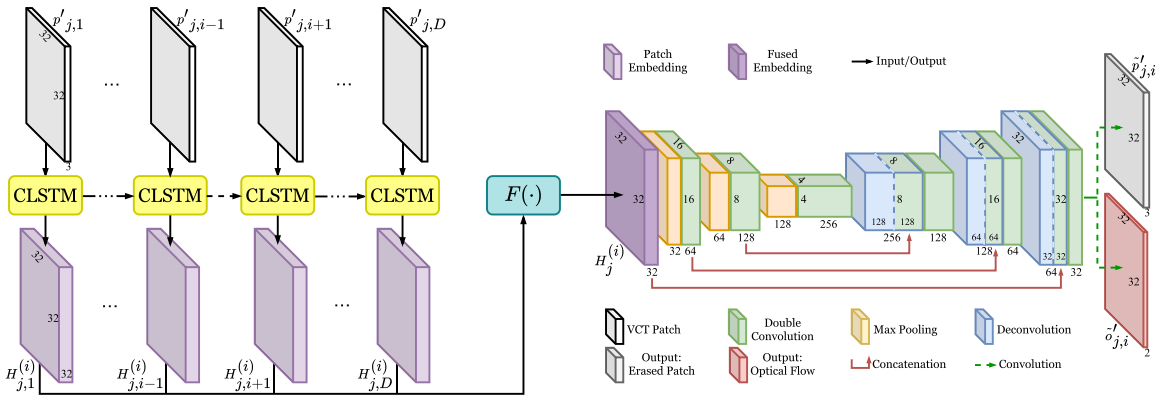

Fig. 7. ST-UNet architecture: Parameters of CLSTM is shared. Appearance completion network $f_a^{(i)}$ and motion completion network $f_m^{(i)}$ share the same ST-UNet architecture, except that $f_a^{(i)}$ has 3 output channels corresponding to erased patch $p'_{j,i}$, while $f_m^{(i)}$ has 2 output channels corresponding to optical flow $o'_{j,i}$.

3) Model-Level Enhancement: As discussed above, VCC can serve as a paradigm-level solution to fully exploit the video semantics and temporal context information for VAD. However, the standard UNet [82] used in basic VCC [17] mainly focus on the spatial information of a patch, while it does not model the temporal correlation among patches in a STC explicitly. Thus, it is natural to develop a model-level solution that is specifically tailored for this goal. To endow UNet with the ability to explicitly model temporal information, we naturally resort to convolutional long-short-term-memory (CLSTM) [83], the well-known model for handling temporal correlation. By combining CLSTM and UNet, we design a new DNN architecture named spatio-temporal UNet (ST-UNet), which is more compatiable with our VCC paradigm. Specifically, the core idea of ST-UNet is to synthesize a CLSTM module into the UNet module. For a type-i VCT $V_j^{(i)} = [p'_{j,1}; \cdots p'_{j,i-1}; p'_{j,i+1}; \cdots p'_{j,D}]$, $i \in \{1, \cdots D\}$, each time the t_{th} patch $p'_{j,t}$ is fed into the CLSTM module to compute $I_{j,t}^{(i)}$, $F_{j,t}^{(i)}$, $O_{j,t}^{(i)}$, which correspond the control signals of input gate, forget gate and output gate respectively:

$$\begin{aligned} I_{j,t}^{(i)} &= sigmoid(W_{pe} \otimes p'_{j,t} + W_{he} \otimes H_{j,t-1}^{(i)} + b_e) \\ F_{j,t}^{(i)} &= sigmoid(W_{pf} \otimes p'_{j,t} + W_{hf} \otimes H_{j,t-1}^{(i)} + b_f) \\ O_{j,t}^{(i)} &= sigmoid(W_{po} \otimes p'_{j,t} + W_{ho} \otimes H_{j,t-1}^{(i)} + b_o) \end{aligned} \tag{6}$$

where $W_{pe}, W_{pf}, W_{po}, W_{he}, W_{hf}, W_{ho}$ are learnable convolutional kernels, and b_e, b_f, b_o denote the associated biases. $H_{j,t-1}^{(i)}$ represents a high-level embedding of the previous patch $p'_{j,t-1}$, which aims to involve the influence of temporal history, while $\otimes$ denotes the convolution operation. With $I_{j,t}^{(i)}, F_{j,t}^{(i)}$ to control the influx and outflux of the past and present information (recorded in $C_{t-1}^{(i)}$ and $\tilde{C}_{j,t}^{(i)}$), the current cell state $C_{j,t}^{(i)}$ in CLSTM module can be calculated as:

$$\begin{aligned} \tilde{C}_{j,t}^{(i)} &= tanh(W_{pc} \otimes p'_{j,t} + W_{hc} \otimes H_{j,t-1}^{(i)} + b_c) \\ C_{j,t}^{(i)} &= F_{j,t}^{(i)} \circ C_{t-1}^{(i)} + I_{j,t}^{(i)} \circ \tilde{C}_{j,t}^{(i)} \end{aligned} \tag{7}$$

where W_{pc}, W_{hc} and b_c denotes the convolutional weights and bias, and $\circ$ is the Hadamard product. With the cell state $C_{j,t}^{(i)}$ and the control signal for output gate $O_{j,t}^{(i)}$, the high-level embedding $H_{j,t}^{(i)}$ of current patch $p'_{j,t}$ is computed by:

$$H_{j,t}^{(i)} = O_{j,t}^{(i)} \circ tanh(C_{j,t}^{(i)}) \tag{8}$$

In this way, each patch $p'_{j,t}$ in the VCT $V_j^{(i)}$ is sequentially fed into CLSTM and transformed into a high-level embedding $H_{j,t}^i$. $H_{j,t}^i$ is not only expected to abstract the current patch into a high-level embedding with richer semantics, but also involve the temporal history information of past patches. Since $H_{j,t}^i$ contains richer semantics and temporal context information, we can collect all high-level embeddings and compute an overall embedding $H_j^{(i)}$ for the type-i VCT of j_{th} video event, $V_j^{(i)}$, with a fusion function $F(\cdot)$:

$$H_j^{(i)} = F(H_{j,1}^{(i)}, \cdots H_{j,i-1}^{(i)}, H_{j,i+1}^{(i)}, \cdots H_{j,D}^{(i)}) \tag{9}$$

$F(\cdot)$ can be implemented by various means, *e.g.* an element-wise operator or a convolution layer, while we simply choose $F(\cdot)$ to be the element-wise summation in this paper. $H_j^{(i)}$ enables us to maximally record the video semantics and temporal context information from the VCT. Therefore, instead of raw patches from STCs, $H_j^{(i)}$ is then fed into the UNet module to obtain the completion results to the VCT:

$$\begin{aligned} \tilde{p}'_{j,i} &= U_a(H_j^{(i)}) \\ \tilde{o}'_{j,i} &= U_m(H_j^{(i)}) \end{aligned} \tag{10}$$

where U_a and U_m represents the case for appearance and motion completion respectively. Compared with the preliminary work that adopts standard UNet for VCC [17], the proposed ST-UNet enjoys three advantages: First, the introduction of CLSTM module enables us to explicitly model the temporal correlation of patches in STCs on the model level; Second, the model can maximally exploit the temporal context by fusing the high-level embedding of all patches in the VCT into an overall embedding; Third, by feeding the overall embedding rather than raw patches into the UNet module, we can encourage DNNs to achieve a better utilization of high-level video semantics for VCT completion. Our later evaluation suggests that using ST-UNet as DNN architecture constantly outperforms UNet in VCC.

4) Decision-Level Enhancement: At the decision stage, the anomaly score metric and post-processing also exert a huge influence on VAD performance. Although our preliminary work [17] show that MSE could be an effective score metric, it suffers from weaknesses, *e.g.* excessive emphasis on per-pixel error and negligence of high-level structure. To overcome such weaknesses, we propose to introduce Structural Similarity (SSIM) [84] as a supplementary score metric to MSE. Taking a video event V_j and its inferred STC $\tilde{V}_j$ as an example, the SSIM based anomaly score is computed as follows:

$$\mathcal{S}_{ss}(V_j, \tilde{V}_j) = 1 - \frac{(2\mu_{V_j}\mu_{\tilde{V}_j} + c_1)(2\sigma_{V_j\tilde{V}_j} + c_2)}{(\mu_{V_j}^2 + \mu_{\tilde{V}_j}^2 + c_1)(\sigma_{V_j}^2 + \sigma_{\tilde{V}_j}^2 + c_2)} \quad (11)$$

where $(\mu_{V_j}, \sigma_{V_j})$ and $(\mu_{\tilde{V}_j}, \sigma_{\tilde{V}_j})$ denote the mean and standard deviation of pixel intensity for V_j and $\tilde{V}_j$ respectively. $\sigma_{V_j\tilde{V}_j}$ is the covariance between the pixels in V_j and $\tilde{V}_j$, while c_1 and c_2 are constants. By mixing the MSE based anomaly score $\mathcal{S}_{mse}(V_j, \tilde{V}_j) = \|V_j - \tilde{V}_j\|_2^2$ and SSIM based anomaly score $\mathcal{S}_{ss}(V_j, \tilde{V}_j)$, we can yield an enhanced anomaly score:

$$\mathcal{S}(V_j, \tilde{V}_j) = \frac{\mathcal{S}_{mse}(V_j, \tilde{V}_j) - \bar{\mathcal{S}}_{mse}}{\sigma_{mse}} + w_{ss}\frac{\mathcal{S}_{ss}(V_j, \tilde{V}_j) - \bar{\mathcal{S}}_{ss}}{\sigma_{ss}} \quad (12)$$

where $\bar{\mathcal{S}}_{mse}, \sigma_{mse}, \bar{\mathcal{S}}_{ss}, \sigma_{ss}$ are the means and standard deviations of MSE and SSIM based anomaly scores for all STCs in the training set. In addition to the above process that applies SSIM to appearance completion, we also apply this process to motion completion by replacing $(V_j, \tilde{V}_j)$ with optical flow and its inferred result.

In addition to the anomaly score metric, post-processing based score rectification is another effective way to refine the obtained anomaly scores. The motivation for score rectification stems from the observation that video events are continuous, so the anomaly scores of adjacent video frames are supposed to be close. Therefore, it is natural for us to rectify the anomaly score of current video frame with those anomaly scores yielded by previous temporally adjacent frames. Suppose that the anomaly score for the l_{th} frame and its W previous frames are $\mathcal{S}_l, \mathcal{S}_{l-1}, \cdots \mathcal{S}_{l-W}$, we propose the general formulation below to calculate the rectified anomaly score $\hat{\mathcal{S}}_l$:

$$\hat{\mathcal{S}}_l = \frac{1}{Z}\sum_{m=0}^{W} \omega_m \mathcal{S}_{l-m} \quad (13)$$

where ω_m is a non-negative weight. $Z = \sum_m \omega_m$ is a normalizing factor. There are multiple ways to set the weight ω_m. For example, when $\omega_m = 1$, the post-processing is equivalent to the temporal moving average. Besides, one can also set ω_m to obtain a 1-D Gaussian or median filter. We will compare different types of rectification strategies in later experiments, and the results show that even the simplest form of score rectification can provide sound rectification.

IV. Empirical Evaluations

A. Experimental Setup

To evaluate the proposed VCC approach, we mainly conduct empirical evaluations on the following VAD benchmark datasets: UCSDped2 [21], Avenue [28] and ShanghaiTech [7], which are three most commonly-used datasets for DNN based VAD. To further validate the effectiveness of VCC, we additionally test our approach on three other VAD datasets: UCSDped1 [21], UMN [92] and Subway Exit [92], which are less reported for DNN based VAD but frequently used for evaluating earlier classic VAD methods. The quantitative evaluation of VAD is usually exercised under either the *frame-level criteria* or the *pixel-level criteria* [21]. Details of benchmark datasets and evaluation criteria are given in supplementary material. With either criteria, we can compute area under the curve (AUC) of Receiver Operation Characteristic Curve (ROC) and equal error rate (EER) as quantitative performance measure. We perform both frame-level and pixel-level evaluation for UCSDped1, UCSDped2, Avenue and ShanghaiTech dataset, while only frame-level evaluation is performed on other datasets. Implementation details of VCC are also presented in supplementary material.

B. Performance Comparison

1) Commonly-Used Datasets: To facilitate the comparison with vast VAD methods, our empirical evaluation is mainly carried on three most commonly-used benchmark datasets in DNN based VAD: UCSDped2, Avenue and ShanghaiTech. On those datasets, we have conducted an extensive comparison with VAD methods in the literature, including both representative classic VAD methods and state-of-the-art DNN based VAD methods. Note that we do not compare [13] as it is evaluated by a different way from common practice. As to the proposed VCC, we focus on the VAD performance of the enhanced VCC, while basic VCC's performance [17] is also listed for a reference. The AUC results under frequently-used criteria are shown in Table I, while EER results and pixel-level AUC on Avenue and ShanghaiTech are given in supplementary material. Besides, we also visualize typical frame-level ROC curves in Fig. 8 for an intuitive comparison. With those results, we are able to draw the following observations: **(1)** First, enhanced VCC has attained highly competitive VAD performance on those commonly-used datasets, and it outperforms 43 out of 45 methods by a notable margin. Specifically, enhanced VCC almost conquers UCSDped2 dataset with 99.0% frame-level AUC, which typically leads compared VAD methods by about 2% AUC. Meanwhile, under the more strict pixel-level criteria, VCC exhibits even more obvious advantage by 96.8% pixel-level AUC. When it comes to more challenging Avenue and ShanghaiTech, enhanced VCC also achieves satisfactory performance: While the frame-level AUC of most VAD methods are below 90% and 80% on Avenue and ShanghaiTech respectively, our VCC yields 92.2% AUC on Avenue and 80.2% AUC on ShanghaiTech, which are close or comparable to more recent state-of-the-art works [71], [72]. **(2)** Compared with basic VCC counterpart [17],

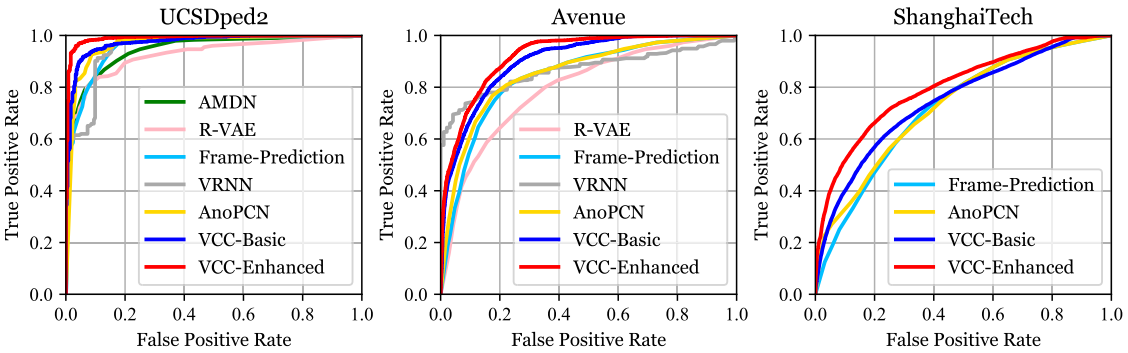

Fig. 8. Comparison of frame-level ROC curves on UCSDped2, Avenue and ShanghaiTech dataset.

TABLE I
AUC COMPARISON ON UCSDPED2, AVENUE AND SHANGHAITECH

Type	Method	UCSDped2 Frame	UCSDped2 Pixel	Avenue Frame	ShanghaiTech Frame
Classic Methods	SF [34]	55.6%	-	-	-
	MDT [21]	82.9%	-	-	-
	SR [26]	82.9%	-	-	-
	MAC [85]	90.0%	73.7%	-	-
	GAS [35]	94.1%	-	-	-
	Unmasking [86]	82.2%	-	80.6%	-
Reconstruction Based Methods	CAE [37]	90.0%	-	70.2%	-
	AMDN [5]	90.8%	-	-	-
	SRNN [44]	92.2%	-	81.7%	68.0%
	WTA-CAE [38]	96.6%	89.3%	82.1%	-
	AM-GAN [47]	93.5%	-	-	-
	LSTM [39]	88.1%	-	77.0%	-
	R-VAE [40]	92.4%	-	79.6%	-
	PDE-AE [42]	95.4%	-	-	72.5%
	Mem-AE [16]	94.1%	-	83.3%	71.2%
	AM-Corr [8]	96.2%	-	86.9%	-
	AnomalyNet [45]	94.9%	52.8%	86.1%	-
	SRNN-AE [46]	92.2%	-	83.5%	69.6%
	GEPC [51]	-	-	-	76.1%
	OGNet [49]	98.1%	-	-	-
	Clustering-AE [52]	96.5%	-	86.0%	73.3%
	FSCN [50]	92.8%	-	85.5%	-
	DeepOC [53]	96.9%	95.0%	86.6%	-
Frame Prediction Based Methods	Frame-Prediction [7]	95.4%	-	85.1%	72.8%
	VRNN [54]	96.1%	-	85.8%	-
	Att-prediction [11]	96.0%	-	86.0%	-
	Multi-Prediction [60]	-	-	82.9%	76.0%
	SIGNet [59]	96.2%	48.4%	86.8%	-
	Online [87]	97.2%	-	86.4%	70.9%
	Multispace [61]	95.4%	-	86.8%	73.6%
	DMMNet [88]	94.5%	-	-	-
	Bi-Prediction [89]	96.6%	-	87.8%	-
	Multipath-Pred. [62]	96.3%	-	88.3%	76.6%
Hybrid of Recon. and Pred.	ST-CAE [63]	91.2%	-	80.9%	-
	MPED-RNN [64]	-	-	-	73.4%
	Mem-Guided [65]	97.0%	-	88.5%	70.5%
	AnoPCN [9]	96.8%	-	86.2%	73.6%
	Predictive-AE [90]	95.8%	-	87.4%	-
Other DNN Based Methods	Recounting [12]	92.2%	89.1%	-	-
	TCP [67]	88.4%	-	-	-
	NNC [68]	-	-	88.9%	-
	Cluster-Att [74]	-	-	87.0%	79.3%
	Scene-Aware [91]	-	-	89.6%	74.7%
	SSMTL [71]	97.5%	-	91.5%	82.4%
	SSPCAB [72]	-	-	92.9%	83.6%
Proposed VCC	Basic	97.3%	93.0%	89.6%	74.8%
	Enhanced	99.0%	96.8%	92.2%	80.2%

TABLE II
PERFORMANCE COMPARISON ON UCSDPED1

Method	Frame-level AUC	Frame-level EER	Pixel-level AUC	Pixel-level EER
SF [34]	67.5%	-	19.7%	-
MDT [21]	81.4%	25.0%	44.1%	58.0%
SR [26]	89.5%	19.0%	50.2%	53.0%
LSA [93]	92.7%	16.0%	-	-
SCL [28]	91.8%	15.0%	63.8%	41.0%
GPR [23]	83.8%	23.7%	63.3%	37.3%
MAC [85]	85.0%	-	65.0%	-
GAS [35]	93.8%	-	65.1%	-
Unmasking [86]	68.4%	-	52.4%	-
CAE [37]	81.0%	27.9%	-	-
WTA-CAE [38]	91.9%	14.8%	68.7%	35.7%
AMDN [5]	92.1%	16.0%	67.2%	40.1%
AM-GAN [47]	97.4%	8.0%	70.3%	35.0%
LSTM [39]	75.5%	-	-	-
ST-CAE [63]	92.3%	15.3%	-	-
R-VAE [40]	75.0%	32.4%	-	-
Frame-Prediction [7]	83.1%	-	-	-
TCP [67]	95.7%	8.0%	64.5%	40.8%
AnomalyNet [45]	83.5%	25.2%	45.2%	-
VRNN [54]	86.3%	-	-	-
Att-prediction [11]	83.9%	-	-	-
FSCN [50]	82.4%	25.2%	-	-
DeepOC [53]	83.5%	23.4%	63.1%	-
SIGNet [59]	86.0%	-	51.6%	-
DMMNet [88]	86.0%	-	-	-
Bi-Prediction [89]	89.0%	-	-	-
Multipath-Pred. [62]	83.4%	-	-	-
VCC	87.7%	20.5%	76.3%	27.9%

enhanced VCC enables a significant performance advancement (1.7%-5.4% AUC gain). Such progress has justified our solutions to enhance VCC on multiple levels.

2) Other Datasets: We additionally conduct experiments on three other VAD datasets: UCSDped1, UMN and Subway exit. They are less used in recent DNN based VAD research due to some inherent limitations (detailed in supplementary material). Nevertheless, we still test the enhanced VCC on them to offer a comprehensive evaluation. With fewer DNN based methods, we also include representative classic VAD methods as a reference. The comparison is presented in Table II-IV. From those results, we note that VCC also achieves satisfactory VAD performance, even though those datasets are less suitable for DNN based VAD: On UCSDped1 dataset, we note that the proposed VCC obtains fairly competitive performance under frame-level criteria (87.7% AUC and 20.5% EER). Meanwhile, under the more strict pixel-level criteria, VCC is the best performer (76.3% AUC and 27.9% EER) with an evident advantage (typically 6%-10% AUC gain) over the compared methods. As for UMN dataset, we follow the frequently-used practice to report the average AUC by calculating frame-level AUC on three individual scenes [45]. Like most VAD methods, VCC achieves near-perfect performance on the relatively simple UMN dataset, and produces an average frame-level AUC above 99%; Although the sparse Subway Exit provides less video event data for training DNNs than other datasets, VCC still yields a decent performance (over 91% frame-level AUC), which is readily comparable to most existing methods.

TABLE III
AUC COMPARISON ON UMN

SF [34]	SR [26]	LSA [93]	MAC [85]	SCD [94]	GAS [35]	Unmasking [86]	AM-GAN [47]	TCP [67]	AnomalyNet [45]	NNC [68]	VCC
96.0%	97.8%	98.5%	98.3%	91.0%	99.7%	95.1%	99.0%	98.8%	99.6%	99.3%	99.4%

TABLE IV
AUC COMPARISON ON SUBWAY EXIT

MDT [21]	SR [26]	LSA [93]	AMDN [5]	CAE [37]	SCD [94]	Unmasking [86]	LSTM [39]	FCN [6]	NNC [68]	SRNN-AE [46]	SIGNet [59]	DeepOC [53]	VCC
89.7%	80.2%	88.4%	87.9%	80.7%	82.4%	85.7%	87.7%	90.2%	95.1%	89.7%	95.7%	89.5%	91.4%

C. Detailed Analysis

1) Ablation Studies: For our VCC framework, the proposed video event extraction scheme and ensemble strategies in VCTs are important components. To justify their necessity, we conduct ablation studies with the basic VCC framework under the frame-level criteria. We conduct the following studies and show results in Table V: **(1)** We compare four different event extraction schemes (FR: frame based learning, SDW: sliding windows with motion filtering, APR: appearance based extraction only, APR+MT: the proposed joint extraction based on appearance and motion). Note that SDW is not used on ShanghaiTech, because it produces excessive RoIs that are beyond our hardware capacity. We can draw several conclusions: First, the proposed APR+MT constantly possesses a noticeable advantage over other methods. As shown by row 1, 2, 3, 6 of each dataset in Table V, the performance of APR+MT leads FR and SDW by 2.7%-4.6% AUC, while both FR and SDW are widely-used strategies in literature. In particular, it is found that SDW even underperforms FR on both UCSDped2 and Avenue. This observation suggests that an imprecise localization of RoIs can be counterproductive to VAD performance. As to the comparison of APR and APR+MT, the latter strategy prevails by 1.8%, 2.5% and 1.2% AUC on UCSDped2, Avenue and ShanghaiTech respectively, which validates the importance of comprehensive localization. **(2)** We compare three configurations of ensemble strategies in VCC: Without any VCT type ensemble, without modality ensemble ($w_m = 0$), with both VCT type (for both appearance and motion completion) and modality ensemble. By comparing row 4, 5, 6 of each dataset in Table V, we come to two conclusions: First, through fully exploiting temporal context with the proposed VCT type ensemble, a 1.3%, 2.1% and 0.4% AUC improvement is achieved on UCSDped2, Avenue and ShanghaiTech respectively. Second, introducing motion information by modality ensemble constantly leads to better performance. Specifically, modality ensemble produces a significant AUC elevation (up to 8%) on UCSDped2, while more than 1% AUC improvement is also achieved on Avenue and ShanghaiTech. The great performance leap on UCSDped2 can be attributed to the fact that its gray-scale frames contain limited appearance information, while motion plays a more important role in discriminating anomalies.

2) Effectiveness of Enhancement Solutions: As discussed in Sec. IV-B, enhanced VCC outperforms basic VCC by a large margin. This section will provide a more detailed comparison that shows how enhanced VCC improves performance by each enhancement solution. As can be seen in Table VI, the comparison is made on four levels: **(1)** Localization-level. As shown by row 1-2 of each dataset in Table VI, using optical flow (OF) constantly outperforms gradients (GD) by 0.3% to 1.4% AUC on all datasets, which validates OF as better motion cues for RoI localization. **(2)** Event-level. As shown by row 2-3 of each dataset in Table VI, the spatially-localized strategy (Block) enables tangible performance gain on those datasets that are evidently influenced by varied foreground depth and scales. For example, it produces 7.9% and 2.4% AUC improvement on UCSDped1 and ShanghaiTech respectively. **(3)** Model-level. From row 3-4 of each dataset in Table VI, we can observe a consistent improvement up to 1.7% AUC brought by ST-UNet across all datasets, which validates the effectiveness to exploit semantics and temporal context information by the model-level solution. **(4)** Decision-level. First, based on results of row 4-5 of each dataset, the mixed score metric achieves comparable or superior VAD performance to the original MSE. In particular, an approximately 1% AUC improvement is observed on ShanghaiTech, which is the most challenging benchmark with abundant high-level foreground structure. Second, by row 5-6 of each dataset in Table VI, our score rectification (SR) based on temporal averaging enables us to obtain an unanimous and notable AUC growth (0.5%-4.4% AUC gain). Besides, we compare several different ways to set the weight ω_m in Eq. 13, discussion of which is provided in supplementary material due to page limit.

TABLE V
ABLATION STUDIES OF VCC FRAMEWORK

Datase	FR	SDW	APR	APR+MT	VCT Type	Modality	AUC
UCSDped2	✔				✔	✔	94.6%
		✔			✔	✔	93.3%
			✔		✔	✔	95.5%
				✔		✔	96.0%
				✔	✔		89.6%
				✔	✔	✔	**97.3%**
Avenue	✔				✔	✔	86.8%
		✔			✔	✔	85.2%
			✔		✔	✔	87.1%
				✔		✔	87.5%
				✔	✔		88.2%
				✔	✔	✔	**89.6%**
ShanghaiTech	✔				✔	✔	70.2%
		✔			✔	✔	-
			✔		✔	✔	73.6%
				✔		✔	74.4%
				✔	✔		73.5%
				✔	✔	✔	**74.8%**

TABLE VI
COMPARISON BETWEEN ENHANCED VCC AND BASIC VCC

Dataset	GD	OF	Block	UNet	ST-UNet	MSE	Mixed	SR	AUC
UCSDped2	✔			✔		✔			97.3%
		✔		✔		✔			97.6%
		✔	✔	✔		✔			97.6%
		✔	✔		✔	✔			98.1%
		✔	✔		✔		✔		98.2%
		✔	✔		✔		✔	✔	99.0%
Avenue	✔			✔		✔			89.6%
		✔		✔		✔			90.0%
		✔	✔	✔		✔			90.0%
		✔	✔		✔	✔			91.7%
		✔	✔		✔		✔		91.7%
		✔	✔		✔		✔	✔	92.2%
ShanghaiTech	✔			✔		✔			74.8%
		✔		✔		✔			74.8%
		✔	✔	✔		✔			77.2%
		✔	✔		✔	✔			77.6%
		✔	✔		✔		✔		78.5%
		✔	✔		✔		✔	✔	80.2%
UCSDped1	✔			✔		✔			78.8%
		✔		✔		✔			79.3%
		✔	✔	✔		✔			87.2%
		✔	✔		✔	✔			87.3%
		✔	✔		✔		✔		87.3%
		✔	✔		✔		✔	✔	87.7%
UMN	✔			✔		✔			94.4%
		✔		✔		✔			95.1%
		✔	✔	✔		✔			97.3%
		✔	✔		✔	✔			98.1%
		✔	✔		✔		✔		98.1%
		✔	✔		✔		✔	✔	99.4%
Subway Exit	✔			✔		✔			84.9%
		✔		✔		✔			86.3%
		✔	✔	✔		✔			86.3%
		✔	✔		✔	✔			86.8%
		✔	✔		✔		✔		87.0%
		✔	✔		✔		✔	✔	91.4%

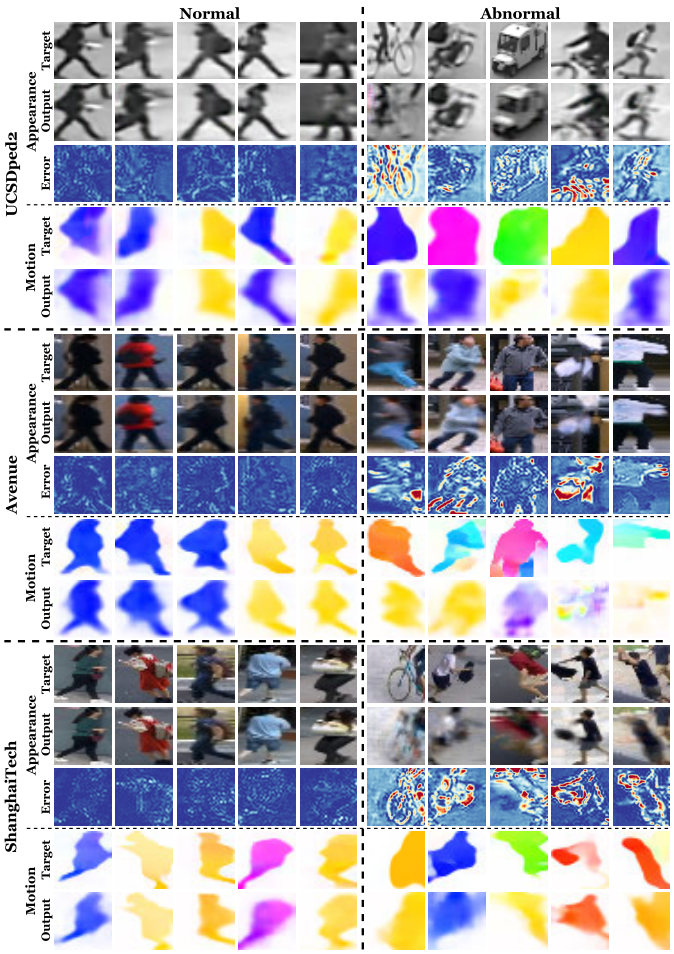

Fig. 9. Visualization of erased patches and their optical flow (Target), corresponding completed patches (Output) by VCC and appearance completion error maps (Error). In the error map warmer color indicates larger error.

3) Visualization: In Fig. 9, we choose some representative normal and abnormal video events from datasets and plot VCC's generated patches and optical flow maps, so as to demonstrate how VCC discerns anomalies intuitively. The magnitude of pixel-level completion errors are visualized by heat maps. Given the visualization in Fig. 9, we discover several interesting facts: **(1)** For normal video events, VCC is able to infer the erased patches and their optical flow satisfactorily. In case of normality, most completion errors are moderate. By contrast, abnormal video events incurs sharp and prominent errors in completion, while their motion completion also suffers from obvious errors in both magnitude and direction of optical flow. **(2)** The completion errors of normal events are spread around the object contour in a relatively uniform manner. Unlike normality, the distribution of anomalies' completion errors is evidently non-uniform, while most of them are semantically meaningful. As heat maps show, intense completion errors of anomalies are often concentrated on some high-level parts of anomalous foreground objects, *e.g.* the riding bicycle, thrown paper in the sky, the waving backpack or fast-moving body parts like legs and hands.

4) Additional Remarks: Apart from previous discussion, we make some additional remarks on our VCC approach: **(1)** Connections to frame prediction and reconstruction. In fact, when the whole frame is considered as one video event and only type-*D* VCTs are completed, VCC is equivalent to frame prediction. In other words, frame prediction can be viewed a special case of VCC, but VCC enjoys superior VAD performance to frame prediction based methods (see Table I) and the case using only type-*D* VCTs (see Table V). Besides, when the core component of VCC, *i.e.* VCTs, are replaced by plain reconstruction of STCs, our experiments usually report a 3% to 7% AUC loss, which validates the necessity of VCTs. **(2)** We report and analyze the computational cost of VCC in supplementary material. Besides, possible acceleration are also discussed in supplementary material.

V. CONCLUSION AND LIMITATION

This paper proposes VCC as a new solution to DNN based VAD. VCC first utilizes appearance and motion as complimentary cues to extract RoIs of foreground objects, so as to accomplish both precise and comprehensive video event extraction. By erasing a certain patch, each video event is transformed into a VCT. To solve a VCT, DNNs are trained to infer the erased patch and its optical flow, which spurs DNNs to capture video semantics rather than low-level details. Subsequently, VCC is then equipped with VCT type ensemble and modality ensemble, which enable VCC to fully exploit spatio-temporal context and consider richer

motion information. To further ameliorate VCC, we develop a series of practical enhancement solutions, which lead to the enhanced VCC. Extensive empirical evaluations justify VCC as a highly effective VAD solution that achieves highly competitive performance under both frame-level and pixel-level criteria.

Limitations: Despite the effectiveness of VCC, we also observe some limitations: (1) It can be difficult for VCC to detect static anomalies like loiterers. Since we extract video events from a few temporally adjacent frames (*e.g.* 5 frames), abnormal loiterers and normal pedestrians often behave similarly in such a short time period, which makes it hard for VCC to distinguish between the two behavioral patterns. (2) As introduced in Sec. IV-C.4, the current VCC implemented by Python has not yet met the requirements of real-time video processing in spite of the acceptable inference speed. (3) To extract high-quality video events, we need to manually adjust the parameters like the thresholds in Algorithm 1 by considering the characteristics of different datasets. Although these parameters can be fixed for a typical video scene once determined, they still rely on manual empirical selection.

REFERENCES

[1] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2293–2312, May 2022.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.

[3] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowl. Eng. Rev.*, vol. 29, no. 3, pp. 345–374, Jan. 2014.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2017.

[6] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88–97, Jul. 2018.

[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.*, Jun. 2018, pp. 6536–6545.

[8] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.

[9] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1805–1813.

[10] K. Liu and H. Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1490–1499.

[11] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, "Attention-driven loss for anomaly detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4639–4647, Dec. 2020.

[12] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3639–3647.

[13] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7834–7843.

[14] M. I. Georgescu, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "A background-agnostic framework with adversarial training for abnormal event detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4505–4523, Sep. 2022.

[15] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1558–1566.

[16] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[17] G. Yu et al., "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 583–591.

[18] 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.*, Jun. 2008, pp. 1–8.

[19] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1544–1554, Nov. 2008.

[20] T. Zhang, H. Lu, and S. Z. Li, "Learning semantic scene models by object classification and trajectory clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1940–1947.

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

[22] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1590–1599, Oct. 2013.

[23] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2909–2917.

[24] M. U. K. Khan, H.-S. Park, and C.-M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 2, pp. 541–556, Feb. 2019.

[25] X. Hu, Y. Huang, X. Gao, L. Luo, and Q. Duan, "Squirrel-cage local binary pattern and its application in video anomaly detection," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 1007–1022, Apr. 2019.

[26] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.

[27] 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. 2558–2567.

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

[29] 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.*, Jun. 2009, pp. 1446–1453.

[30] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, Jun. 2011, pp. 3313–3320.

[31] G. Chen et al., "NeuroAED: Towards efficient abnormal event detection in visual surveillance with neuromorphic vision sensor," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 923–936, 2021.

[32] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 8, pp. 1082–1090, Aug. 2008.

[33] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 988–998, Jun. 2014.

[34] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.

[35] Q. Sun, H. Liu, and T. Harada, "Online growing neural gas for anomaly detection in changing surveillance scenes," *Pattern Recognit.*, vol. 64, Apr. 2016, Art. no. S0031320316302771.

[36] 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.

[37] 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.

[38] H. Tran and D. Hogg, "Anomaly detection using a convolutional winner-take-all autoencoder," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 139. 1–139. 12.

[39] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.

[40] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Abnormal event detection from videos using a two-stream recurrent variational autoencoder," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 1, pp. 30–42, Mar. 2020.

[41] T. Wang et al., "Generative neural networks for anomaly detection in crowded scenes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1390–1399, May 2019.

[42] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.

[43] H. T. Vu, T. D. Nguyen, T. Le, W. Luo, and D. Q. Phung, "Robust anomaly detection in videos using multilevel representations," in *Proc. AAAI*, 2019, pp. 5216–5223.

[44] 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.

[45] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[46] W. Luo et al., "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1070–1084, Mar. 2021.

[47] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.

[48] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[49] M. Z. Zaheer, J.-H. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14171–14181.

[50] P. Wu, J. Liu, M. Li, Y. Sun, and F. Shen, "Fast sparse coding networks for anomaly detection in videos," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107515.

[51] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10536–10544.

[52] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Proc. ECCV*, 2020, pp. 329–345.

[53] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2609–2622, Jul. 2020.

[54] Y. Lu, K. M. Kumar, S. S. Nabavi, and Y. Wang, "Future frame prediction using convolutional VRNN for anomaly detection," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[55] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, "Few-shot scene-adaptive anomaly detection," in *Proc. ECCV*, 2020, pp. 125–141.

[56] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15420–15429.

[57] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proc. AAAI*, 2021, pp. 938–946.

[58] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, "Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5546–5554.

[59] Z. Fang, J. Liang, J. T. Zhou, Y. Xiao, and F. Yang, "Anomaly detection with bidirectional consistency in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1079–1092, Mar. 2022.

[60] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 2615–2623.

[61] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3694–3706, Sep. 2021.

[62] X. Wang et al., "Robust unsupervised video anomaly detection by multipath frame prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2301–2312, Jun. 2022.

[63] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal AutoEncoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1933–1941.

[64] 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. 11988–11996.

[65] 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. 14360–14369.

[66] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13568–13577.

[67] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 1689–1698.

[68] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, "Detecting abnormal events in video using narrowed normality clusters," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2019, pp. 1951–1960.

[69] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2587–2596.

[70] G. Pang, C. Yan, C. Shen, A. V. D. Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12170–12179.

[71] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12737–12747.

[72] N.-C. Ristea et al., "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13566–13576.

[73] A. Barbalau et al., "SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection," *Comput. Vis. Image Understand.*, vol. 229, Mar. 2023, Art. no. 103656.

[74] Z. Wang, Y. Zou, and Z. Zhang, "Cluster attention contrast for video anomaly detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2463–2471.

[75] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[76] T. Y. Lin, *Microsoft COCO: Common Objects in Context*. Berlin, Germany: Springer, 2014.

[77] Wikipedia. (2019). *Cloze Test*. [Online]. Available: https://en.wikipedia.org/wiki/Cloze_test

[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[79] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: A survey," *Comput. Vis. Image Understand.*, vol. 134, pp. 1–21, May 2015.

[80] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1647–1655.

[81] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.* Cham, Switzerland: Springer, 2000, pp. 1–15.

[82] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Berlin, Germany: Springer, 2015.

[83] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. NIPS*, 2015, pp. 1–9.

[84] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[85] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, "Combining motion and appearance cues for anomaly detection," *Pattern Recognit.*, vol. 51, pp. 443–452, Mar. 2016.
[86] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2914–2922.
[87] K. Doshi and Y. Yilmaz, "Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate," *Pattern Recognit.*, vol. 114, Jun. 2021, Art. no. 107865.
[88] S. Li, J. Fang, H. Xu, and J. Xue, "Video frame prediction by deep multi-branch mask network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1283–1295, Apr. 2021.
[89] D. Chen, P. Wang, L. Yue, Y. Zhang, and T. Jia, "Anomaly detection in surveillance video based on bidirectional prediction," *Image Vis. Comput.*, vol. 98, Jun. 2020, Art. no. 103915.
[90] Y. Lai, R. Liu, and Y. Han, "Video anomaly detection via predictive autoencoder with gradient-based attention," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[91] C. Sun, Y. Jia, Y. Hu, and Y. Wu, "Scene-aware context reasoning for unsupervised abnormal event detection in videos," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 184–192.
[92] 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.
[93] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2112–2119.
[94] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016.

Guang Yu received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, National University of Defense Technology, Changsha, China. His works have been published on CVPR and ACM MM. His current research interests include anomaly detection, video abnormal event detection, and unsupervised learning. He serves as a reviewer for CVPR and ICCV.

Siqi Wang received the B.S. and Ph.D. degrees in computer science and technology from the National University of Defense Technology, China. He is currently an Assistant Research Professor with the State Key Laboratory of High-Performance Computing (HPCL), National University of Defense Technology. His works have been published on leading conferences and journals, such as NeurIPS, AAAI, ACM MM, ICPR, *Pattern Recognition*, IEEE TRANSACTIONS ON CYBERNETICS, and *Neurocomputing*. His current research interests include anomaly/outlier detection, pattern recognition, and unsupervised learning. He serves as a reviewer for several international journals, including the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, *Artificial Intelligence Review*, and *International Journal of Machine Learning and Cybernetics*.

Zhiping Cai received the bachelor's, master's, and Ph.D. degrees in computer science and technology from the National University of Defense Technology (NUDT), China, in 1996, 2002, and 2005, respectively. He is currently a Full Professor with the College of Computer Science and Technology, National University of Defense Technology, Changsha, China. His current research interests include network security and edge computing.

Xinwang Liu (Senior Member, IEEE) received the Ph.D. degree from the National University of Defense Technology (NUDT), China. He is currently a Professor with the School of Computer Science and Technology, NUDT. He has published more than 60 peer-reviewed papers, including those in highly regarded journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, NeurIPS, ICCV, CVPR, AAAI, and IJCAI. His current research interests include kernel learning and unsupervised feature learning.

En Zhu received the Ph.D. degree from the National University of Defense Technology (NUDT), China. He is currently a Professor with the School of Computer Science, NUDT, China. He has published more than 60 peer-reviewed papers, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, AAAI, and IJCAI. His current research interests include pattern recognition, image processing, machine vision, and machine learning. He was awarded the China National Excellence Doctoral Dissertation.

Jianping Yin received the Ph.D. degree from the National University of Defense Technology (NUDT), China. He is currently a Distinguished Professor with the Dongguan University of Technology. He has published more than 150 peer-reviewed papers, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, AAAI, and IJCAI. His current research interests include pattern recognition and machine learning. He was awarded the China National Excellence Doctoral Dissertation Supervisor and the National Excellence Teacher. He served on the technical program committees of more than 30 international conferences and workshops.