

Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events

Guang Yu*

Siqi Wang*

National University of Defense
Technology

{yuguangnudt,wangsiqi10c}@gmail.com

Zhiping Cai

National University of Defense
Technology
zpcai@nudt.edu.cn

En Zhu

National University of Defense
Technology
enzhu@nudt.edu.cn

Chuanfu Xu

National University of Defense
Technology

xuchuanfu@nudt.edu.cn

Jianping Yin

Dongguan University of Technology
jpyin@dgut.edu.cn

Marius Kloft

TU Kaiserslautern
kloft@cs.uni-kl.de

ABSTRACT

As a vital topic in media content interpretation, video anomaly detection (VAD) has made fruitful progress via deep neural network (DNN). However, existing methods usually follow a reconstruction or frame prediction routine. They suffer from two gaps: (1) They cannot localize video activities in a both precise and comprehensive manner. (2) They lack sufficient abilities to utilize high-level semantics and temporal context information. Inspired by frequently-used *cloze test* in language study, we propose a brand-new VAD solution named *Video Event Completion* (VEC) to bridge gaps above: First, we propose a novel pipeline to achieve both precise and comprehensive enclosure of video activities. Appearance and motion are exploited as mutually complimentary cues to localize regions of interest (RoIs). A normalized spatio-temporal cube (STC) is built from each RoI as a *video event*, which lays the foundation of VEC and serves as a basic processing unit. Second, we encourage DNN to capture high-level semantics by solving a *visual cloze test*. To build such a visual cloze test, a certain patch of STC is erased to yield an incomplete event (IE). The DNN learns to restore the original video event from the IE by inferring the missing patch. Third, to incorporate richer motion dynamics, another DNN is trained to infer erased patches' optical flow. Finally, two ensemble strategies using different types of IE and modalities are proposed to boost VAD performance, so as to fully exploit the temporal context and modality information for VAD. VEC can consistently outperform state-of-the-art methods by a notable margin (typically 1.5%–5% AUROC) on commonly-used VAD benchmarks. Our codes and results can be verified at github.com/yuguangnudt/VEC_VAD.

*Authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

MM '20, October 12–16, 2020, Seattle, WA, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7988-5/20/10...\$15.00

<https://doi.org/10.1145/3394171.3413973>

CCS CONCEPTS

• **Computing methodologies** → **Scene anomaly detection; Un-supervised learning.**

KEYWORDS

Video anomaly detection, video event completion

ACM Reference Format:

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 (MM '20)*, October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3394171.3413973>

1 INTRODUCTION

Videos play a key role in multimedia. Video anomaly detection (VAD), which performs anomaly detection by interpreting the video content automatically, has been appealing to both academia and industry, since it is valuable to various safety-critical scenarios like municipal and traffic management. Formally, VAD refers to detecting video activities that divert significantly from the observed normal routine. Despite many efforts, VAD remains challenging for two features of anomaly [5]: (1) *Scarcity*. As anomalies are usually rare, collecting real anomalies for training is often hard or even impossible. (2) *Ambiguity*. The anomaly does not possess fixed semantics and may refer to different activities based on different context, so it can be highly variable and unpredictable. Such features render modeling anomalies directly unrealistic. Thus, VAD usually adopts an *one-class classification* setup [15]. This setup collects training videos with only normal activities, which are much more accessible than anomalies, to build a normality model. Activities that do not conform to this model are then viewed as anomalies. As all training data are normal, discriminative supervised learning is usually not applicable. Instead, the unsupervised/self-supervised learning has been the commonly-used scheme in VAD. Following such a scheme, existing VAD solutions fall into two categories: (1) *Classic* VAD, which requires domain knowledge to design hand-crafted descriptors to depict high-level features (e.g. trajectory, speed) or low-level features (e.g. gradient, texture) of video activities. Extracted features are fed into classic anomaly detection methods

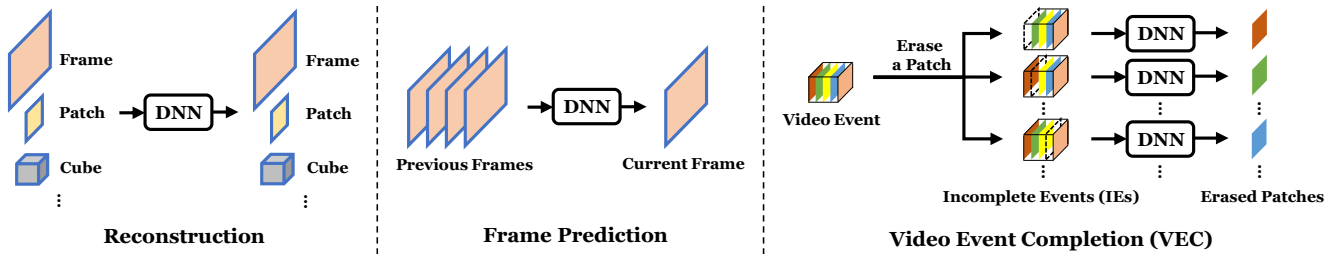


Figure 1: Typical solutions for DNN based VAD. *Left*: Reconstruction based methods train DNN to reconstruct inputs, e.g. video frames/patches/cubes. *Middle*: Frame prediction based methods take several previous frames as inputs of DNN to predict current frame. *Right*: VEC first encloses video events with spatio-temporal cubes (STCs) based on a novel pipeline that synthesizes appearance and motion cues. Then, erasing the patch at STC’s different positions produces different types of incomplete events (IEs), which serves as different “visual cloze tests”. Each DNN is trained to solve a visual cloze test, i.e. learning to complete the missing patch of a certain type of IE. Note that cubes for reconstruction differs from STCs in VEC, as they are yielded by a relatively coarse strategy (e.g. sliding windows) and cannot enclose video events both precisely and comprehensively.

like one-class support vector machine (OCSVM) to spot anomalies. Feature engineering of classic VAD can be labor-intensive and sub-optimal [40], and designed descriptors are often hard to transfer among different scenes. (2) *Deep neural network (DNN)* based VAD, which is inspired by DNN’s success in traditional vision tasks [19]. Due to DNNs’ strong capabilities in feature learning and activity modeling [11], DNN based VAD achieves superior performance and enjoys surging popularity when compared with classic VAD.

Despite the fruitful progress, DNN based VAD still suffers from two gaps. **Gap #1**: Existing DNN based VAD methods cannot localize video activities in a both precise and comprehensive manner. A standard practice in VAD is to use a sliding window with motion filtering [34, 40], but such localization is obviously imprecise. Recent DNN based VAD methods like [21, 29, 42] simply ignore this issue by learning on the whole frame, but this suffers from scale variations incurred by image depth and foreground-background imbalance problem [20, 48]. Few studies [12, 14] improve localization precision by a pre-trained object detector, but it causes a “closed world” problem—The detector only recognize objects in training data and tends to omit video novelties, which leads to non-comprehensive localization results. Such a gap degrades later video activity modeling; **Gap #2**: Existing DNN based methods lack sufficient abilities to exploit high-level semantics and temporal context in video activities. As illustrated by Fig. 1, two paradigms (*reconstruction* and *frame prediction*) dominate DNN based VAD in the literature: Reconstruction based methods learn to reconstruct inputs and detect poorly reconstructed data as anomalies. However, in this case DNNs tend to memorize low-level details rather than learning high-level semantics [18], and they even reconstruct anomalies well due to overly strong modeling power [10]; Frame prediction based methods learn to predict a normal video frame from previous video frames, and detect poorly predicted frames as anomalies. Prediction makes it hard to simply memorize details for reducing training loss, but it scores anomalies by the prediction error of a single frame, which overlooks the temporal context. Thus, neither reconstruction nor frame prediction provides a perfect solution. Unlike recent research that focuses on exploring better network architectures to improve reconstruction or frame prediction, we are

inspired by *cloze test* in language study and mitigate gaps above by proposing *Video Event Completion (VEC)* as a new DNN based VAD solution (see Fig. 1). Our contributions are summarized below:

- VEC for the first time combines both appearance and motion cues to localize video activities and extract video events. It overcomes the “closed world” problem and enables both precise and comprehensive video activity enclosure, and it lays a firm foundation for video event modeling in VEC.
- VEC for the first time designs visual cloze tests as a new learning paradigm, which trains DNNs to complete the erased patches of incomplete video events, to substitute frequently-used reconstruction or frame prediction based methods.
- VEC also learns to complete the erased patches’ optical flow, so as to integrate richer information of motion dynamics.
- VEC utilizes two ensemble strategies to fuse detection results yielded by different types of incomplete events and data modalities, which can further boost VAD performance.

2 RELATED WORK

Classic VAD. Classic VAD usually consists of two stages: Feature extraction by hand-crafted descriptors and anomaly detection by classic machine learning methods. As to feature extraction, early VAD methods usually adopt tracking [17] to extract high-level features like motion trajectory [31, 44] and destination [3]. However, they are hardly applicable to crowded scenes [26]. To this end, low-level features are extensively studied for VAD, such as dynamic texture [26], histogram of optical flow [7], spatio-temporal gradients [16, 22], 3D SIFT [6], etc. Afterwards, various classic machine learning methods are explored to perform anomaly detection, such as probabilistic models [2, 6, 26], sparse coding and its variants [7, 22, 45], one-class classifier [43], sociology inspired models [27]. However, feature extraction is the major bottleneck for classic VAD: Manual feature engineering is complicated and labor-intensive, while the designed descriptors often suffer from limited discriminative power and poor transferability among different scenes.

DNN Based VAD. DNN based VAD differs from classic VAD by learning features automatically from raw inputs with DNNs. The learned features are fed into a classic model or embedded into

DNNs for end-to-end VAD. With only normal videos for training, existing DNN based VAD basically falls into a reconstruction or frame prediction routine. They are reviewed respectively below: (1) *Reconstruction* based methods learn to reconstruct inputs from normal training videos, and assume that a large reconstruction error signifies the anomaly. Autoencoder (AE) and its variants are the most popular DNNs to perform reconstruction. For example, [39] pioneers DNN based VAD by introducing stacked denoising AE (SDAE) and propose its improvement [40]; [11] adopts convolutional AE (CAE) that are more suitable for modeling videos, while recent works explore numerous CAE variants such as Winner-take-all CAE (WTA-CAE) [37] and Long Short Term Memory based CAE (ConvLSTM-AE) [24]; [41] integrates variational AE into a two-stream recurrent framework (R-VAE) to realize VAD; [1] equips AE with a parametric density estimator (PDE-AE) for anomaly detection; [10] propose a memory-augmented AE (Mem-AE) to make AE’s reconstruction error more discriminative. In addition to AE, other types of DNNs are also used for the reconstruction purpose, e.g. sparse coding based recurrent neural network (SRNN) [25] and generative adversarial network (GAN) [32, 35]. Cross-modality reconstruction is also shown to produce good VAD performance by learning the appearance-motion correspondence (AM-CORR) [29]. (2) *Frame prediction* based methods learn to predict current frames by several previous frames, while a poor prediction is viewed as abnormal. [21] for the first time formulates frame prediction as an independent VAD method, and imposes appearance and motion constraints for prediction quality. [23] improves frame prediction by using a convolutional variational RNN (Conv-VRNN). However, prediction by a per-frame basis leads to a bias to background [20], and [48] proposes attention mechanism to ease the issue. Another natural idea is to combine prediction with reconstruction as a hybrid VAD solution: [46] design a spatio-temporal CAE (ST-CAE), in which an encoder is followed by two decoders for reconstruction and prediction purpose respectively; [28] reconstructs and predicts human skeletons by a message-passing encoder-decoder RNN (MPED-RNN); [42] integrates reconstruction into prediction by a predictive coding network based framework (AnoPCN); [36] conducts prediction and reconstruction in a sequential manner.

3 VIDEO EVENT COMPLETION (VEC)

3.1 Video Event Extraction

In this paper, video event extraction aims to enclose video activities by numerous normalized spatio-temporal cubes (STCs). A STC is then viewed as a video event, which serves as the basic processing unit in VEC. A both *precise* and *comprehensive* localization of video activities is the key to video event extraction. Ideally, *precise* localization expects the subject of a video activity is intactly extracted with minimal irrelevant background, while *comprehensive* localization requires all subjects associated with video activities are extracted. As explained by **Gap #1** (Sec. 1) and the intuitive comparison in Fig. 2, existing DNN based VAD methods fail to realize precise and comprehensive localization simultaneously, which undermines the quality of video activity modeling and VAD performance. Hence, we exploit appearance and motion as mutually complementary cues to achieve both precise and comprehensive localization (Fig. 2 (d)). This new pipeline is detailed as follows:

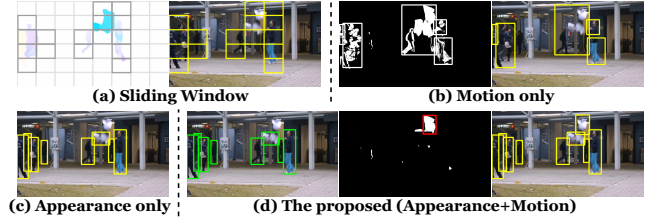


Figure 2: Comparison of localization strategies: Sliding window (a) or motion only (b) produces imprecise localization, while appearance only (c) yields non-comprehensive localization results. The proposed (d) pipeline achieves more precise and comprehensive localization simultaneously.

Motivation. As video activities are behaviors conducted by certain subjects in videos, we consider both *appearance cues* from those subjects and *motion cues* from their behaviors to localize regions of interest (RoIs) that enclose those video activities. To utilize appearance cues, a natural solution is modern object detection [4]. With pre-training on large-scale public datasets like Microsoft COCO, pre-trained detectors can precisely localize RoIs with frequently-seen objects. Therefore, we use a pre-trained object detector to realize *precise* localization of most video activities by detecting their associated subjects, e.g. humans. However, pre-trained detectors only detect objects in the “closed world” formed by known object classes in the training dataset. This leads to a fatal problem: Anomalies, which are often novel classes outside the “closed world”, will be omitted. To this end, motion cues like temporal gradients are proposed as complimentary information to accomplish more *comprehensive* RoI localization. More importantly, we argue that appearance and motion cues should not be isolated: RoIs already localized by appearance should be filtered when exploiting motion cues, which reduces computation and makes motion based RoI localization more precise (see Fig. 2 (d)). As illustrated by the overview in Fig. 3 and Algorithm 1, we elaborate each component of the new video event extraction pipeline below.

Appearance Based RoI Extraction. Given a video frame I and a pre-trained object detector model M , our goal is to obtain a RoI set B_a based on appearance cues from subjects of video activities, where $B_a \subseteq \mathbb{R}^4$ and each entry of B_a refers to a RoI enclosed by a bounding box. The bounding box is denoted by the coordinates of its top-left and bottom-right vertex, which is a 4-dimensional vector. As shown by the green module in Fig. 3, we first feed I into M , and obtain a preliminary RoI set B_{ap} with confidence scores above the threshold T_s (class labels are discarded). Then, we introduce two efficient heuristic rules to filter unreasonable RoIs: (1) RoI area threshold T_a that filters out overly small RoIs. (2) Overlapping ratio T_o that removes RoIs that are nested or significantly overlapped with larger RoIs in B_{ap} . In this way, we ensure that extracted RoIs can precisely enclose subjects of most everyday video activities.

Motion Based RoI Extraction. To enclose those activities outside the “closed world”, motion based RoI extraction aims to yield a complementary bounding box set B_m based on motion cues. We leverage the temporal gradients of frames as motion cues and complementary information. As shown by the red module of Fig. 3, we first binarize the absolute values of temporal gradients by a

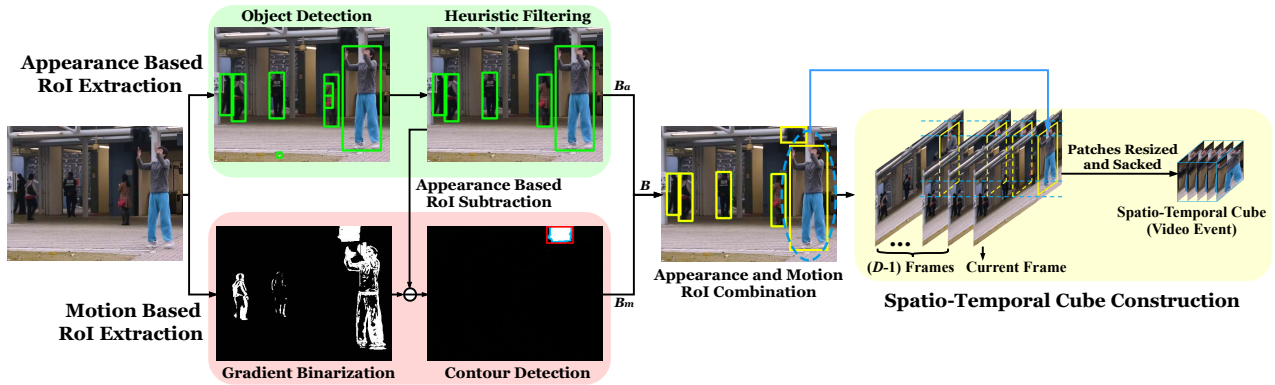


Figure 3: Pipeline of video event extraction: (1) Appearance based ROI extraction (green): Appearance based ROIs are extracted with a pre-trained object detector and filtered based on efficient heuristic rules. (2) Motion based ROI extraction (red): First, temporal gradients are binarized by magnitude into a binary map. Then, highlighted pixels in appearance based ROIs are subtracted from the binary map. Finally, contour detection and simple heuristics are applied to the binary map for final motion based ROIs. (3) Spatio-temporal cube (STC) extraction (yellow): For each ROI, corresponding patches from current frame and $(D - 1)$ previous frames are extracted. D patches are then resized and stacked into a STC, which represents a video event.

Algorithm 1 Appearance and Motion based ROI Extraction

Input: Frame I and its gradient map G , pre-trained object detector M , threshold $T_s, T_a, T_o, T_g, T_{ar}$

Output: ROIs represented by a bounding box set B

```

1:  $B_{ap} \leftarrow ObjDet(I, M, T_s)$  # Detecting activity subjects
2:  $B_a = \{\}$  # Heuristic filtering
3: for  $b_{ap} \in B_{ap}$  do
4:   if  $Area(b_{ap}) > T_a$  and  $Overlap(b_{ap}, B_{ap}) < T_o$  then
5:      $B_a = B_a \cup \{b_{ap}\}$ 
6:   end if
7: end for
8:  $G_b \leftarrow GradBin(G, T_g)$  # Gradient binarization
9:  $G_b \leftarrow RoISub(G_b, B_a)$  # Subtract appear. based ROIs
10:  $C \leftarrow ContourDet(G_b)$  # Contour detection
11:  $B_m = \{\}$ 
12: for  $c \in C$  do
13:    $b_m = BoundingBox(c)$  # Get contour bounding box
14:   if  $Area(b_m) > T_a$  and  $\frac{1}{T_{ar}} < AspectRatio(b_m) < T_{ar}$  then
15:      $B_m = B_m \cup \{b_m\}$ 
16:   end if
17: end for
18:  $B = B_a \cup B_m$ 

```

threshold T_g , so as to yield a binary map that indicates regions with intense motion. Instead of using this map directly, we propose to subtract appearance based ROIs B_a from the map, which benefits motion based ROI extraction in two ways: First, the subtraction of appearance based ROIs enables us to better localize objects that are not detected by appearance cues, otherwise the gradient map of multiple objects may be overlapped and jointly produce large and imprecise ROIs (see Fig. 2 (b)). Second, the subtraction reduces the computation. Finally, we propose to perform contour detection to yield the contour and its corresponding bounding box b_m , while simple heuristics (ROI area threshold T_a and maximum aspect-ratio

threshold T_{ar}) are used to obtain final ROI set B_m . Based on two complementary ROI sets, the final ROI set $B = B_a \cup B_m$. The whole ROI extraction process is formally presented in Algorithm 1.

Spatio-temporal Cube Construction. Finally, we use each ROI in B to build a spatio-temporal cube (STC) as the video event, which represents the fundamental unit to enclose video activities. As shown by yellow module in Fig. 3, we not only extract the patch p_1 in the ROI from current frame, but also extract corresponding patches p_2, \dots, p_D by this ROI from previous $(D - 1)$ frames. In this way, we incorporate the temporal context into the extracted video event. To normalize video activities with different scales, we resize patches from the same ROI into $H \times W$ new patches p'_1, \dots, p'_D , which are then stacked into a $H \times W \times D$ STC: $C = [p'_1; \dots; p'_D]$.

3.2 Visual Cloze Tests

As explained by **Gap #2** in Sec. 1, previous methods typically rely on a reconstruction or frame prediction paradigm. They cannot fully exploit high-level semantics and temporal context information. As a remedy, we propose to build *visual cloze tests* as a new learning paradigm, so as to model normal video activities represented by STCs above. We present it in terms of the following aspects:

Motivation. We are inspired by *cloze test*, an extensively-used exercise in language learning and instruction. It requires completing a text with its certain words erased. Cloze test aims to test students' ability to understand the vocabulary semantics and language context [38]. Recently, learning to solve cloze tests is also shown to be a fairly effective pretraining method in natural language processing (NLP), which enables DNN to capture richer high-level semantics from text [8]. This naturally inspires us to compare the patch sequence of a STC to the word sequence in classic cloze test. Similarly, we can erase a certain patch p'_i in video event (STC) to build a visual cloze test, which is solved by completing the resultant *incomplete event* (IE) with the DNN's inferred \hat{p}'_i . Such a learning paradigm benefits DNN in two aspects: (1) In order to complete such a visual cloze test, DNN is encouraged to capture high-level semantics in

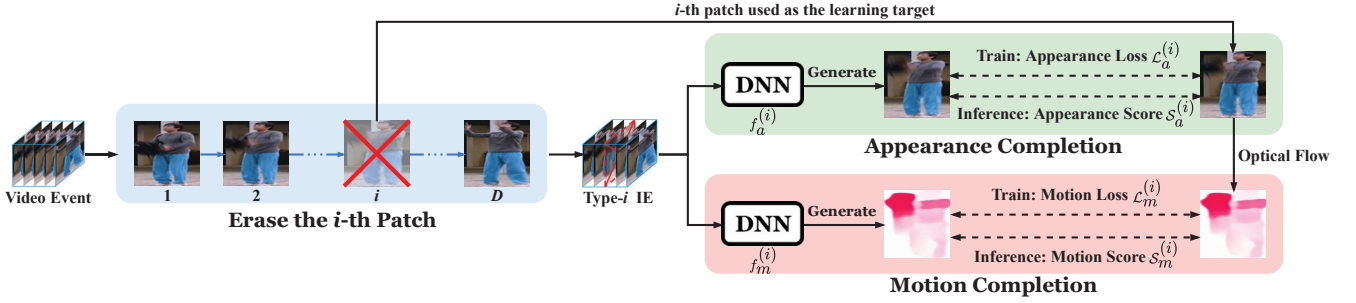


Figure 4: Visual cloze tests with a type- i incomplete event (IE): (1) Erase the i -th patch (blue): The i -th patch of a STC is erased to build a type- i IE, while the erased patch is used as the learning target of appearance completion. (2) Appearance completion (green): To complete the IE, a DNN takes the IE as input and learns to generate the erased patch. (3) Motion completion (red): A DNN takes the IE as input and learns to generate the optical flow patch that corresponds to the erased patch.

STC. For example, suppose a video event contains a walking person. DNN must attend to those key moving parts (e.g. the forwarding leg and swinging arm) in the patch to achieve a good completion. This makes visual cloze test more meaningful than frequently-used reconstruction, which tends to memorize every low-level detail of inputs to minimize training loss. (2) Since any patch in a STC can be erased to generate an IE, we can build multiple cloze tests by erasing the patch at different temporal positions. This enables us to fully exploit the temporal context by enumerating all possible IEs for cloze tests. By contrast, prediction based VAD methods only consider prediction errors of a single frame to detect anomalies, which involves poor temporal context information in video activities. As shown in Fig. 4, we detail VEC’s components below.

Appearance Completion. Given j -th video event represented by the STC $C_j = [p'_{j,1}; \dots; p'_{j,D}]$, we first erase the i -th patch $p'_{j,i}$ of C_j to build an IE $C_j^{(i)} = [p'_{j,1}; \dots; p'_{j,i-1}; p'_{j,i+1}; \dots; p'_{j,D}]$ as a cloze test, $i \in \{1, \dots, D\}$ (blue module in Fig. 4). All IEs built by erasing the i -th patch of the STC are collected as the type- i IE set $C^{(i)} = \{C_1^{(i)}, \dots, C_N^{(i)}\}$, where N is the number of extracted video events (STCs). Afterwards, as shown by red module in Fig. 4, a type- i IE $C_j^{(i)}$ in $C^{(i)}$ and its corresponding erased patch $p'_{j,i}$ are used as the input and learning target respectively to train a generative DNN $f_a^{(i)}$ (e.g. autoencoder, generative adversarial networks, U-Net, etc.), which aims to generate a patch $\tilde{p}'_{j,i} = f_a^{(i)}(C_j^{(i)})$ to complete the IE $C_j^{(i)}$ into the original event C_j . To train $f_a^{(i)}$, here we can simply minimize the pixel-wise appearance loss $\mathcal{L}_a^{(i)}$ for type- i IE set $C^{(i)}$:

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

where $\|\cdot\|_p$ denotes p -norm. In our experiments, we found that choosing $p = 2$ already works well. To further improve the fidelity of generated patch, other values of p or techniques like adversarial training can also be explored to design $\mathcal{L}_a^{(i)}$. For inference, any error measure $S_a^{(i)}(\tilde{p}'_{j,i}, p'_{j,i})$ can be used to yield the appearance anomaly score of patch $p'_{j,i}$, such as mean square error (MSE) or Peak Signal to Noise Ratio (PSNR) [21]. Empirically, choosing $S_a^{(i)}(\tilde{p}'_{j,i}, p'_{j,i})$

to be simple MSE has been effective enough to score anomalies, and poorly completed STCs with higher MSE are more likely to be anomalies. Since appearance completion is actually a challenging learning task for DNN, an independent DNN $f_a^{(i)}$ is trained to handle one IE type $C^{(i)}$. Otherwise, using one DNN for all IE types will degrade the performance of appearance completion.

Motion Completion. Motion is another type of valuable prior in videos. Optical flow, which estimates the pixel-wise motion velocity and direction between two consecutive frames, is a popular low-level motion representation in videos. We feed two consecutive frames into a pre-trained FlowNet model [13], and a forward pass can yield the optical flow efficiently. For each STC C_j , we extract optical flow patches $o_{j,1}, \dots, o_{j,D}$ that correspond to video patches $p_{j,1}, \dots, p_{j,D}$, and also resize them into $H \times W$ patches $o'_{j,1}, \dots, o'_{j,D}$. Motion completion requires a DNN $f_m^{(i)}$ to infer the optical flow patch $o'_{j,i}$ of the erased patch $p'_{j,i}$ by the type- i IE $C_j^{(i)}$, i.e. $\tilde{o}'_{j,i} = f_m^{(i)}(C_j^{(i)})$. $f_m^{(i)}$ can be trained by the motion loss $\mathcal{L}_m^{(i)}$:

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

Likewise, we also adopt $p = 2$ for $\mathcal{L}_m^{(i)}$ and simple MSE to compute the motion anomaly score $S_m^{(i)}(\tilde{o}'_{j,i}, o'_{j,i})$ during inference. With motion completion, we encourage DNN to infer the motion statistics from the temporal context provided by IEs, which enables VEC to consider richer motion dynamics. The process of both appearance and motion completion for a type- i IE are shown in Fig. 4.

Ensemble Strategies. Ensemble is a powerful tool that combines multiple models into a stronger one [9]. During inference, we propose two ensemble strategies to improve the VEC performance: (1) *IE type ensemble*. Erasing a different patch in STC produces a different IE, which contains a different patch combination as temporal context. To fully exploit all possible temporal context for VAD, we compute the final appearance anomaly score by an ensemble of scores, which are yielded by multiple DNNs for different IE types:

$$S_a(C_j) = \frac{1}{D} \sum_{i=1}^D S_a^{(i)}(\tilde{p}'_{j,i}, p'_{j,i}) \quad (3)$$

IE type ensemble is also applicable to the final motion score $\mathcal{S}_m(C_j)$:

$$\mathcal{S}_m(C_j) = \frac{1}{D} \sum_{i=1}^D \mathcal{S}_m^{(i)}(o'_{j,i}, o'_{j,i}) \quad (4)$$

(2) *Modality ensemble*. Since two different modalities, raw pixels and optical flow, are considered to perform completion in VEC, we must fuse their results to yield the overall anomaly score. For simplicity, we use a weighted sum of $\mathcal{S}_a(C_j)$ and $\mathcal{S}_m(C_j)$ to compute the overall anomaly score $\mathcal{S}(C_j)$ for a video event C_j :

$$\mathcal{S}(C_j) = w_a \cdot \frac{\mathcal{S}_a(C_j) - \bar{\mathcal{S}}_a}{\sigma_a} + w_m \cdot \frac{\mathcal{S}_m(C_j) - \bar{\mathcal{S}}_m}{\sigma_m} \quad (5)$$

where $\bar{\mathcal{S}}_a, \sigma_a, \bar{\mathcal{S}}_m, \sigma_m$ denote the means and standard deviations of appearance and motion scores for all normal events in training, which are used to normalize appearance and motion scores into the same scale. In addition to this straightforward weighting strategy, other more sophisticated strategies like late fusion [40] are also applicable to achieve better modality ensemble performance.

4 EVALUATION

4.1 Experimental Settings

Evaluation is performed on three most commonly-used benchmark datasets for VAD: UCSDped2 [26], Avenue [22] and ShanghaiTech [25]. For video event extraction, cascade R-CNN [4] pre-trained on COCO dataset is used as object detector as it achieves a good trade-off between performance and speed. Other parameters are set as follows for UCSDped2, Avenue and ShanghaiTech respectively: Confidence score threshold T_s : (0.5, 0.25, 0.5); RoI area threshold T_a : ($10 \times 10, 40 \times 40, 8 \times 8$); Overlapping ratio T_o : (0.6, 0.6, 0.65); Gradient binarization threshold T_g : (18, 18, 15); Maximum aspect-ratio threshold $T_{ar} = 10$. For cube construction, we set $H = W = 32$ and $D = 5$. As to VEC, we adopt U-Net [33] as the basic network architecture of generative DNNs (see Fig. 5), which are optimized by the default Adam optimizer in PyTorch [30]. Considering the dataset scale, DNNs are trained by 5, 20, 30 epochs with a batch size 128 on UCSDped2, Avenue and ShanghaiTech respectively. For anomaly scoring, we set (w_a, w_m) to be (0.5, 1), (1, 1) and (1, 0.5) for UCSDped2, Avenue and ShanghaiTech respectively. For quantitative evaluation, we adopt the most frequently-used metric: Area Under the Receiver Operating Characteristic curves (AUROC) that are computed with frame-level detection criteria [26]. Frame-level Equal Error Rate (EER) [26] is also reported in the supplementary material. We run experiments on a PC with 64 GiB RAM, Nvidia Titan Xp GPUs and a 3.6GHz Intel i7-9700k CPU.

4.2 Comparison with State-of-the-Art methods

Within our best knowledge, we extensively compare VEC’s performance with 18 state-of-the-art DNN based VAD methods reviewed in Sec. 2. Note that we exclude [14] from comparison as it actually uses a different evaluation metric from commonly-used frame-level AUROC, which leads to an unfair comparison. As discussed in Sec. 2, existing methods can be categorized into reconstruction based, frame prediction based and hybrid methods in Table 1. As to VEC, we design two configurations for IE type ensemble: (1) VEC-A: IE type ensemble is applied to appearance completion, while it is

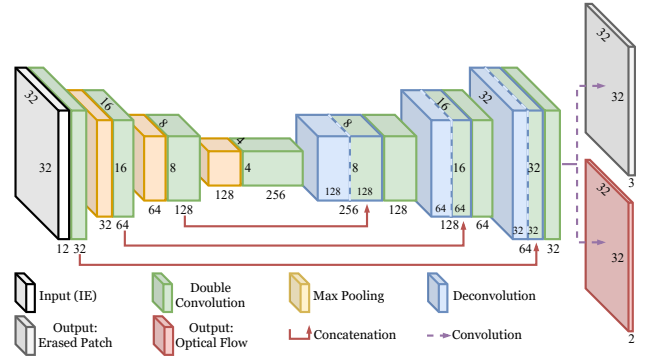


Figure 5: DNN architecture used in our experiments. Appearance completion network $f_a^{(i)}$ and motion completion network $f_m^{(i)}$ share the same U-Net architecture, except that $f_a^{(i)}$ has 3 output channels (images) while $f_m^{(i)}$ has 2 (optical flow).

not applied to motion completion (i.e. only type- D IEs are used to train the DNN for motion completion). (2) VEC-AM: IE type ensemble is applied to both appearance completion and motion completion. Besides, modality ensemble is applied to both VEC-A and VEC-AM. The results are reported in Table 1, and we visualize the yielded frame-level ROC curves in Fig. 6. We draw the following observations: First, both VEC-A and VEC-AM consistently outperform existing state-of-the-art DNN based VAD methods on three benchmarks. In particular, we note that VEC achieves notable performance gain on recent challenging benchmarks (Avenue and ShanghaiTech) with constantly $\geq 3\%$ and $\geq 1\%$ AUROC improvement respectively against all state-of-the-art methods. Meanwhile, we note that VEC-A even achieves over 90% frame-level AUROC on Avenue, which is the best performance ever achieved on Avenue dataset to our knowledge. In terms of the comparison between VEC-A and VEC-AM, two configurations yield fairly close performance, despite of slight differences on different datasets. As a consequence, in addition to those thoroughly-studied reconstruction or prediction based methods, the proposed VEC provides a highly promising alternative for DNN based VAD with state-of-the-art performance.

4.3 Detailed Analysis

Ablation Studies. To show the role of the proposed video event extraction and ensemble strategies, we perform corresponding ablation studies and display the results in Table 2: (1) As to video event extraction, we compare four practices for localizing video activities: Frame (FR, i.e. no localization at all), multi-scale sliding windows with motion filtering (SDW), appearance based RoI extraction only (APR) and the proposed appearance and motion based RoI extraction (APR+MT). Note that we did not report SDW’s results on ShanghaiTech, since it produces excessive STCs that are beyond the limit of our hardware, which is actually an important downside of SDW. There are several observations: First, with the same ensemble strategies, the proposed APR+MT constantly outperforms other methods by a sensible margin. Specifically, APR+MT has an obvious advantage (2.7%-4.6% AUROC gain) over FR and SDW, which are commonly-used strategies of recent VAD methods. Interestingly, we

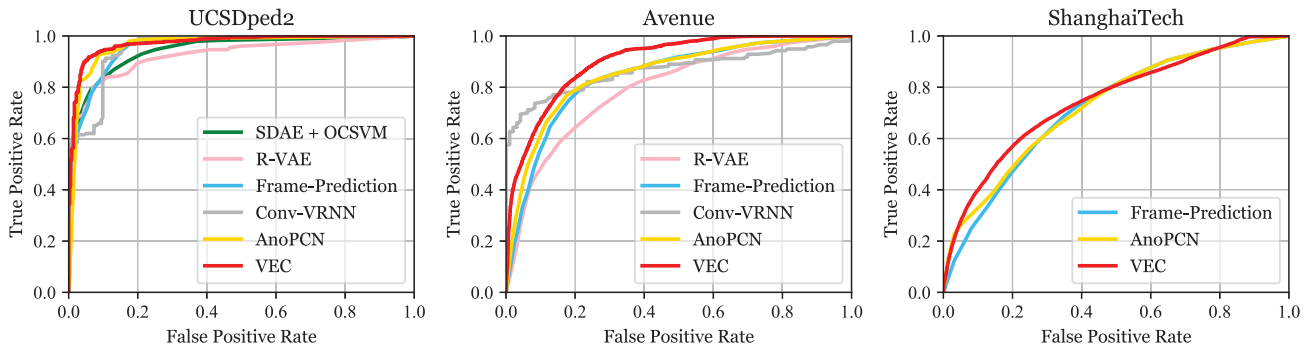


Figure 6: Comparison of frame-level ROC curves on different VAD benchmark datasets.

| Method | | UCSDped2 | Avenue | ShanghaiTech |
|--------------------------------|---------------------------------------|--------------|--------------|--------------|
| Reconstruction Based Methods | CAE [11] | 85.0% | 80.0% | 60.9% |
| | SDAE + OCSVM [40] | 90.8% | - | - |
| | SRNN [25] | 92.2% | 81.7% | 68.0% |
| | GAN [32] | 93.5% | - | - |
| | ConvLSTM-AE [24] | 88.1% | 77.0% | - |
| | WTA-CAE + OCSVM [37] | 96.6% | 82.1% | - |
| | R-VAE [41] | 92.4% | 79.6% | - |
| | PDE-AE [1] | 95.4% | - | 72.5% |
| | Mem-AE [10] | 94.1% | 83.3% | 71.2% |
| | AM-CORR [29] | 96.2% | 86.9% | - |
| AnomalyNet [47] | 94.9% | 86.1% | - | |
| Frame Prediction Based Methods | Frame-Prediction [21] | 95.4% | 84.9% | 72.8% |
| | Conv-VRNN [23] | 96.1% | 85.8% | - |
| | Attention-Prediction [48] | 96.0% | 86.0% | - |
| Hybrid Methods | ST-CAE [46] | 91.2% | 80.9% | - |
| | Skeleton-Trajectories + MPED-RNN [28] | - | - | 73.4% |
| | AnoPCN [42] | 96.8% | 86.2% | 73.6% |
| Prediction&Reconstruction [36] | 96.3% | 85.1% | 73.0% | |
| Proposed | VEC-A | 96.9% | 90.2% | 74.7% |
| | VEC-AM | 97.3% | 89.6% | 74.8% |

Table 1: AUROC comparison between the proposed VEC and state-of-the-art VAD methods.

note that SDW performs worse than FR on UCSDped2 and Avenue. This indicates that an imprecise localization of video activities even degrades VAD performance, and it also justifies the importance of a precise localization. Meanwhile, when compared with APR, the proposed APR+MT brings evident improvement by 1.8%, 2.5% and 1.2% AUROC gain on UCSDped2, Avenue and ShanghaiTech respectively. Such observations demonstrate that a more comprehensive localization of video activities will contribute to VAD performance. (2) As to ensemble strategies, we compare three cases: Not using IE type ensemble (for both appearance and motion completion), not using modality ensemble ($w_m = 0$), and both IE type and modality ensemble are used (VEC-AM). We yield the following observations: First, IE type ensemble contributes to VAD performance by 1.3%, 2.1% and 0.4% AUROC on UCSDped2, Avenue and ShanghaiTech respectively, which justifies the importance to fully exploit temporal context. Second, modality ensemble enables a remarkable 8%

AUROC gain on UCSDped2. This is because UCSDped2 contains low-resolution gray-scale frames, and motion clues are more important for detecting anomalies. For Avenue and ShanghaiTech with high-resolution colored frames, modality ensemble also enables over 1% AUROC improvement, although using the modality of raw pixel only already leads to satisfactory performance.

Visualization. To show how visual cloze tests in VEC helps discriminating anomalies in a more intuitive way, we visualize generated patches and optical flow of representative normal/abnormal video events in Fig. 7. Heat maps are used for a better visualization of the pixel-wise completion errors. By Fig. 7, it is worth noting several phenomena: First of all, VEC can effectively complete normal events and their optical flow. For normal events, minor completion errors are observed to be distributed around foreground contour in a relatively uniform manner, and their optical flow can also be

| Dataset | Video Event Extraction | | | | Ensemble | | AUROC |
|--------------|------------------------|-----|-----|--------|----------|----------|--------------|
| | FR | SDW | APR | APR+MT | IE Type | Modality | |
| 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 2: Ablation Studies for VEC.

soundly recovered. By contrast, abnormal events produces prominent completion errors in terms of both raw pixel and optical flow completion. Next, it is noted that the distribution of anomalies' completion errors is highly non-uniform. As shown by heat maps, large completion errors are often observed at those regions that have clear high-level semantics, e.g. the bicycle that the man was riding with (UCSDped2), falling paper with its shape and position wrongly inferred (Avenue), the backpack that was thrown (ShanghaiTech). By contrast, other regions are endowed with relatively smaller errors. Such observations imply that VEC indeed attends to those parts with high-level semantics in abnormal events.

Other Remarks. (1) VEC adopts a similar U-Net architecture to previous works, but it achieves significantly better performance, which exactly verifies visual cloze tests' effectiveness as a new learning paradigm. Thus, better network architecture can be explored, while techniques like adversarial training and attention mechanism are also applicable. In fact, VEC with only type- D IEs can be viewed as predicting the last patch of STCs (as shown in Table 2, it is also better than frame prediction [21]). Besides, when visual cloze tests in VEC are replaced by plain reconstruction, experiments report a 3% to 7% AUROC loss on benchmarks, which demonstrates that our video event extraction and visual cloze tests are both indispensable. (2) Details of VEC's computation cost and parameter sensitivity are also discussed in the supplementary material.

5 CONCLUSION

In this paper, we propose VEC as a new solution to DNN based VAD. VEC first extracts STCs by exploiting both appearance and motion cues, which enables both precise and comprehensive video event extraction. Subsequently, motivated by the widely-used cloze test, VEC learns to solve visual cloze tests, i.e. training DNNs to infer deliberately erased patches from incomplete video events/STCs, so

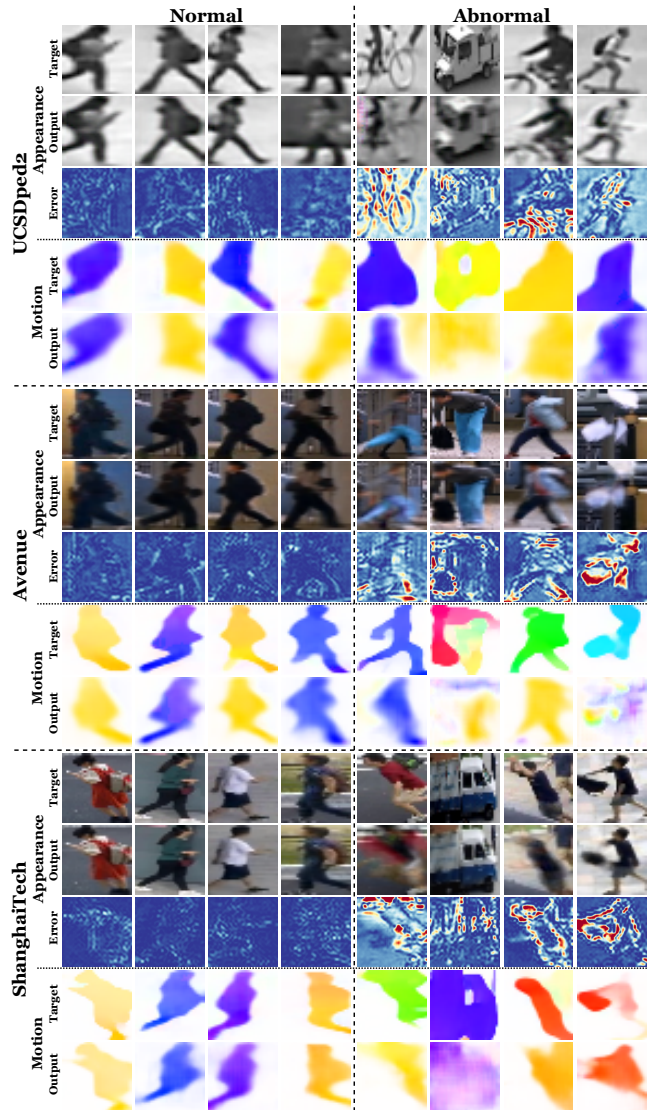


Figure 7: Visualization of erased patches and their optical flow (Target), completed patches (Output) by VEC and completion errors (Error). Brighter color indicates larger errors.

as to learn better high-level semantics. Motion modality is also involved by using DNNs to infer the erased patches' optical flow. Two ensemble strategies are further adopted to fully exploit temporal context and motion dynamics, so as to enhance VAD performance.

ACKNOWLEDGMENTS

The work is supported by National Natural Science Foundation of China under Grant No. 61702539, Hunan Provincial Natural Science Foundation of China under Grant No. 2018JJ3611, No. 2020JJ5673, NUDT Research Project under Grant No. ZK-18-03-47, ZK20-10, and The National Key Research and Development Program of China (2018YFB0204301, 2018YFB1800202, SQ2019ZD090149). Siqi Wang, Zhiping Cai and Jianping Yin are corresponding authors.

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] Borislav Antić and Björn Ommer. 2011. Video parsing for abnormality detection. In *2011 International Conference on Computer Vision*. IEEE, 2415–2422.
- [3] Arslan Basharat, Alexei Gritai, and Mubarak Shah. 2008. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
- [4] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.
- [5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.
- [6] Kai-Wen Cheng, Yie-Tarnng 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.
- [7] 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.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
- [9] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.
- [10] 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)*.
- [11] 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.
- [12] 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.
- [13] Eddy Ilg, Nikolaus Mayer, Tomoy 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.
- [14] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Juliana 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.
- [15] Shehroz S Khan and Michael G Madden. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29, 3 (2014), 345–374.
- [16] Louis Kratz and Ko Nishino. 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1446–1453.
- [17] Long Lan, Xinchao Wang, Gang Hua, Thomas S Huang, and Dacheng Tao. 2020. Semi-online Multi-people Tracking by Re-identification. *International Journal of Computer Vision* (2020), 1–19.
- [18] 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.
- [19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
- [20] Kun Liu and Huadong Ma. 2019. Exploring Background-bias for Anomaly Detection in Surveillance Videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1490–1499.
- [21] 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.
- [22] Cewu Lu, Jianping Shi, and Jiayia Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*. 2720–2727.
- [23] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. 2019. Future Frame Prediction Using Convolutional VRNN for Anomaly Detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–8.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] Ramin Mehran, Alexis Oyama, and Mubarak Shah. 2009. Abnormal crowd behavior detection using social force model. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] Claudio Picciarelli, Christian Micheloni, and Gian Luca Foresti. 2008. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology* 18, 11 (2008), 1544–1554.
- [32] 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.
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [34] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. 2018. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding* 172 (2018), 88–97.
- [35] 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.
- [36] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. 2020. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters* 129 (2020), 123–130.
- [37] 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.
- [38] Wikipedia. 2019. Cloze test. https://en.wikipedia.org/wiki/Cloze_test.
- [39] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. In *Proceedings of the British Machine Vision Conference*. 8.1–8.8.
- [40] 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.
- [41] Shiyang Yan, Jeremy S Smith, Wenjin Lu, and Bailing Zhang. 2018. Abnormal Event Detection from Videos using a Two-stream Recurrent Variational Autoencoder. *IEEE Transactions on Cognitive and Developmental Systems* (2018).
- [42] 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.
- [43] Jie Yin, Qiang Yang, and Jeffrey Junfeng Pan. 2008. Sensor-based abnormal human-activity detection. *IEEE Transactions on Knowledge and Data Engineering* 20, 8 (2008), 1082–1090.
- [44] Tianzhu Zhang, Hanqing Lu, and Stan Z Li. 2009. Learning semantic scene models by object classification and trajectory clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1940–1947.
- [45] 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.
- [46] 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.
- [47] 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).
- [48] 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).