Rethinking Video Anomaly Detection - A Continual Learning Approach

Keval Doshi
University of South FLorida
4202 E Fowler Ave, Tampa, FL 33620
kevaldoshi@usf.edu

Yasin Yilmaz
University of South FLorida
4202 E Fowler Ave, Tampa, FL 33620
yasiny@usf.edu

Abstract

While video anomaly detection has been an active area of research for several years, recent progress is limited to improving the state-of-the-art results on small datasets using an inadequate evaluation criterion. In this work, we take a new comprehensive look at the video anomaly detection problem from a more realistic perspective. Specifically, we consider practical challenges such as continual learning and few-shot learning, which humans can easily do but remains to be a significant challenge for machines. A novel algorithm designed for such practical challenges is also proposed. For performance evaluation in this new framework, we introduce a new dataset which is significantly more comprehensive than the existing benchmark datasets, and a new performance metric which takes into account the fundamental temporal aspect of video anomaly detection. The experimental results show that the existing state-of-the-art methods are not suitable for the considered practical challenges, and the proposed algorithm outperforms them with a large margin in continual learning and few-shot learning tasks.

1. Introduction

With an ever-increasing number of closed-circuit television (CCTV) cameras and the subsequent amount of video data generated continuously in real-time, it has now become inefficient and nearly impossible for human operators to manually analyze the collected data. Even though automated video surveillance has attracted much research interest in recent years, learning *continually* from new data remains largely unexplored. While the vast majority of recent anomaly detection methods perform competitively on the three popular benchmark datasets (UCSD Pedestrian [16], CUHK Avenue [19], and ShanghaiTech Campus [17]), we believe that progress in this domain has become stagnant. This can be attributed to several factors, such as a flawed problem formulation, lack of a comprehensive dataset, and an inadequate evaluation criterion.

Traditionally, the video anomaly detection (VAD) problem is formulated as detecting behaviors or patterns that are previously unseen in the training data. However, such a formulation has an underlying assumption that the training data includes all possible nominal patterns, which is impractical. The main challenge in VAD is the “open set” nature of the nominal class for behaviors and patterns. Since the data domain of VAD is the real-world behaviors and patterns, it is not possible to confine the nominal class to a static (i.e., fixed) training set even for a specific scene (e.g., a static camera monitoring a particular street). A more realistic problem formulation can be provided by the Continual Learning framework [18]. A practical VAD algorithm must continually train [1] on new nominal video data arriving irregularly over time. As opposed to the standard classification setup, where training on a fixed dataset is followed by testing, in the continual learning setup, training and testing episodes are interleaved, resulting in an ever-growing training dataset, as shown in Fig. 1. The main challenge in this setup is to incrementally learn new nominal patterns from sequentially arriving new training data without forgetting the past knowledge obtained from previous training data.

The current practice for performance evaluation in VAD also follows the standard binary classification setup. Considering each video frame as an independent instance to be classified as nominal or anomalous, the existing performance criterion uses the area-under-the-curve (AUC) metric, which computes the area under the ROC curve (true positive rate (TPR) vs. false positive rate (FPR)). This commonly used frame-level AUC metric is not adequate to evaluate the overall VAD performance. In real-world scenes, usually the main objective is to detect anomalous activities rather than anomalous frames. Even though both tasks might seem similar, they each serve a different purpose. While anomalous activity detection is crucial for raising an alarm in a timely manner, and hence must be online, anomalous frame localization on the other hand is used to capture anomalous activities for future analysis, and thus can be

[1] Not continuously. In CL, it is natural to have gaps between training episodes. The key point is the ability to incrementally train on sequential data arriving over a long time horizon without forgetting the past.

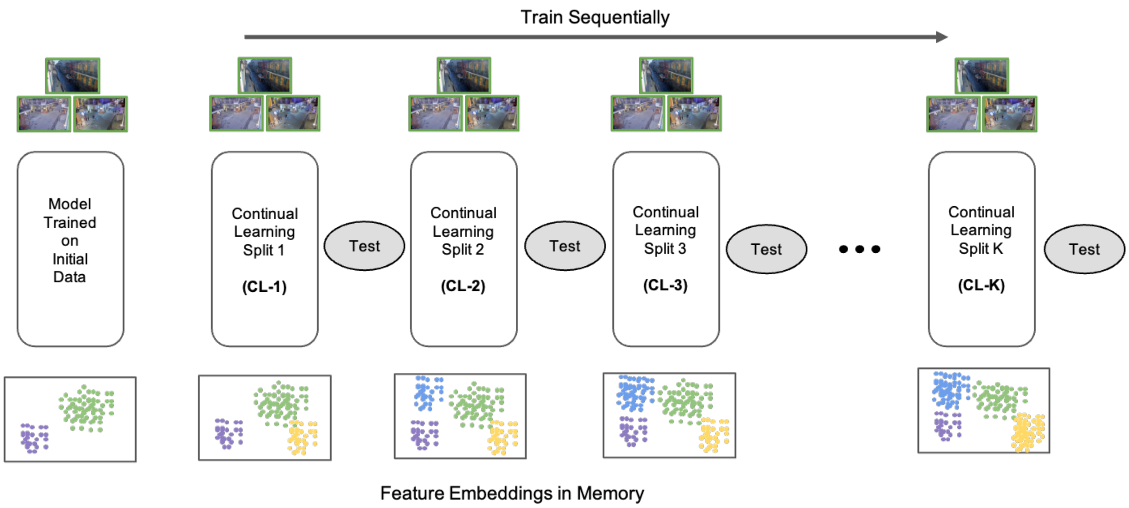
Train Sequentially
Model Trained on Initial Data
Continual Learning Split 1 (CL-1)
Test
Continual Learning Split 2 (CL-2)
Test
Continual Learning Split 3 (CL-3)
Test
• • •
Continual Learning Split K (CL-K)
Test
Feature Embeddings in Memory

Figure 1. The proposed continual learning framework. Training data consists of a number of splits, used to update the algorithm and knowledge base. After each update, the model is evaluated on the entire test set.

offline. The existing VAD literature lacks a clear distinction between the anomalous activity detection and anomalous frame localization tasks [17, 12, 30, 23]. The standard frame-level AUC metric is only suitable for anomalous frame localization. For online activity detection, it is imperative to evaluate the performance in terms of activities and also consider the detection delay in performance evaluation. An ideal VAD algorithm should minimize the average delay in detecting anomalous activities and avoid false alarms as much as possible.

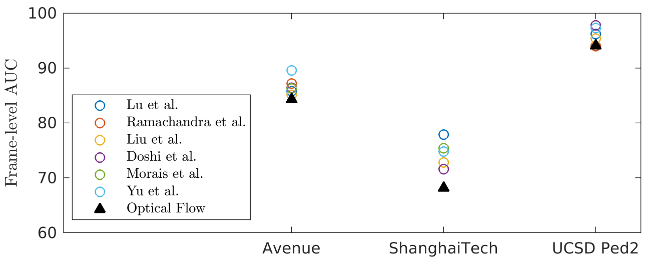

Figure 2. Simple optical flow method performs close to the state-of-the-art methods [20, 27, 17, 5, 23, 37] on the three popular benchmark datasets in terms of the frame-level AUC metric.

The popular benchmark datasets in VAD are prepared for the traditional classification setup based on static training, whose shortcomings are explained above. In these datasets, anything not seen in the training data is labeled as anomalous, which causes a very limited nominal class and a superficial definition for anomaly. For example, in the UCSD [16], Avenue [19], and ShanghaiTech [17] datasets, the nominal behaviors mainly consists of walking people. Such a limited nominal class enables optical flow based approaches to perform increasingly well on these datasets. In Fig. 2, we compare the performance of the recent state-of-the-art methods [20, 27, 17, 5, 23, 37] on these benchmark datasets with respect to a simple optical flow based algorithm, which only computes the average optical flow in a frame. Even such a rudimentary approach is able to perform competitively with respect to the state-of-the-art models, demonstrating the skewness in the benchmark datasets. Furthermore, in these datasets, a person using a bike or skateboard is always considered as anomalous. Even in the more recent Street Scene dataset [27], certain activities like loitering and dog walking on the sidewalk are considered anomalous irrespective of their context. However, in real life, such activities are fairly common and would be considered anomalous only under certain circumstances, such as riding a bike against the flow of traffic or loitering after midnight. Finally, none of the existing datasets/algorithms take into consideration practical challenges such as different weather and lighting conditions, shifts in the activity levels based on the day and time, and adapting to different views due to a moving camera. Hence, for the advancement of VAD, a significantly more comprehensive dataset that can shift the focus to evaluating the continual learning performance of VAD algorithms is required.

Another important limitation of the current state-of-the-art methods is the inherent assumption that each test video segment includes an anomalous activity. In practice, for this assumption to hold, the length of video segments may need to be extremely long since in real-world scenes anomalous activities typically occur infrequently. On the contrary, the video segments in the existing benchmark datasets are a few minutes long and always labeled by some anomlaous frames, which do not necessarily correspond to real-world anomalies. Thus, most of the existing methods are designed to find anomalous frames in each video segment, which will result in many false alarms in a real-world scenario.

Motivated by the above research gaps in VAD, in this paper, we

- design a framework for continual learning and propose a new performance metric based on detection delay and alarm precision;
- introduce a new comprehensive dataset for continual learning in VAD;
- propose a novel algorithm that significantly outperforms the state-of-the-art methods in online activity detection and continual learning, and provide guidance for future algorithm design.

2. Related Work

Anomaly detection in videos has been extensively studied for several years. While early approaches focused on using handcrafted motion features such as histogram of oriented gradients (HOGs) [1, 2, 16], Hidden Markov Models [14, 11], sparse coding [38, 22], and appearance features [3, 16], recent approaches have been completely dominated by deep learning based algorithms. Recent algorithms can be broadly classified into reconstruction based approaches [7, 9, 21, 25, 26], which try to classify frames based on the reconstruction error, and prediction based approaches [17, 15, 4, 6], which attempt to predict a future frame, primarily by using generative adversarial networks (GANs) [8]. More recently, skeletal trajectory based approaches [23, 30] have been proposed since a large proportion of anomalies in the benchmark datasets involve anomalous poses. In such algorithms, an RNN architecture is typically used to learn nominal human poses, and estimation error is used during testing to detect the level of abnormality. Apart from these approaches, [28] proposed a Siamese network to learn spatio-temporal patches and detect an anomaly using the dissimilarity between patches. While these methods perform competitively on the benchmark datasets, they are completely dependant on complex neural networks and mostly end-to-end trained. This makes them notoriously difficult to train on new data, which is crucial in complex temporal applications such as VAD. Furthermore, there is no clear procedure for these methods to adapt to different nominal baselines.

Continual learning has been recently gaining increased research interest [13, 33, 31, 35, 18]. However, not a lot of progress has been made yet in continual learning for VAD. In [5], a modular transfer learning based architecture is proposed to extract appearance and motion features, and a CUSUM based approach is used to continually learn nominal patterns. However, it is only briefly discussed and the algorithm is evaluated only in terms of the false alarm rate on a single YouTube video. Furthermore, the algorithm uses an object-centric framework similar to [12, 10], which treat each object independently, and fails to capture the intricate relationship between different objects. Whereas, our proposed method tracks each object while also capturing spatial information relative to other objects in the frame.

3. Continual Video Anomaly Detection

Ideally, when a video anomaly detection system acquires new information, it should be capable of updating its definition of nominal patterns/behaviors to avoid false alarms. However, this is not straightforward with the existing algorithms since they are extensively dependant on end-to-end trained deep neural networks that are prone to *catastrophic forgetting* when trained incrementally, i.e., they tend to forget previously learned information when trained sequentially on a new task [18]. Hence, we first carefully define a framework for continual learning in the context of video anomaly detection. Then, we propose a new metric for assessing the online activity detection performance that, and an effective algorithm for continual VAD. We believe the new problem formulation and the new dataset, introduced in Sec. 4, will help guide the future VAD research towards practical and reproducible solutions.

3.1. Problem Formulation

Although a stream of video frames $F = \{f_1, f_2, \dots\}$ is a standard data structure for general video processing, for anomaly detection, a video frame is not a natural data unit due to two main reasons: lack of temporal continuity and interpretability. Firstly, the task of classifying video frames as nominal or anomalous ignores the temporal continuity in video frames, which is the main characteristic that differentiates video from a sequence of images. Activities happening in a video are the cause of temporal continuity, e.g., running person, falling object, etc. Also, since humans perceive a visual environment in terms of activities, the results of classifying video activities are much more interpretable than frame classification results. Therefore, we consider a data structure of streaming video activities $X = \{x_1, x_2, \dots\}$.

An activity x_i can be typically defined in terms of action, e.g., playing basketball, or object(s)-action pair, e.g., car crashing. An activity may involve multiple objects, e.g., people walking. The index i denotes the order of activity x_i in terms of starting time. If multiple activities start at the same frame, they can be ordered randomly. In a given frame, there can be multiple activities or no activity.

While we use activity as a data unit, it should be noted that for the anomaly detection task there is no need to explicitly recognize the activities in a video, setting it apart from the activity recognition task. Two competing objectives make VAD a meaningful and challenging problem:

raise an alarm as soon as possible when an anomalous activity takes place, and raise an alarm only when it is relevant.

Detection Delay: The first objective of quickly detecting anomalous activities can be mathematically written as $\min_1[T_i - \tau_i]$, where $_1$ denotes the expectation with respect to the probability distribution of anomalous activities, τ_i is the starting time of anomalous activity i, and $T_i \geq \tau_i$ is the alarm time. Empirically, the average detection delay can be computed as

$$\text{ADD} = \frac{1}{N}\sum_{i=1}^{N}(T_i - \tau_i), \tag{1}$$

with N denoting the number of anomalous activities. Considering a longest tolerable delay δ_{max}, if there is no alarm within the duration $[\tau_i, \tau_i + \delta_{\text{max}}]$ after anomalous activity i happens, the delay is set to be the maximum value, i.e., $T_i - \tau_i = \delta_{\text{max}}$. Note that the considered objective of minimizing the average detection delay covers as a special case the traditional classification objective of minimizing false negative rate (a.k.a. misdetection rate), $\frac{1}{N}\sum_{j=1}^{N} T_i \geq \tau_i$. The indicator function A takes the value 1 when the condition A holds, otherwise 0. Minimizing the false negative rate (FNR) is the same as its more popular version, maximizing the true positive rate (TPR), as FNR $= 1 -$ TPR. Instead of using the generic cost of 1 for each missed anomalous activity, i.e., $T_i \geq \tau_i$, ADD assigns the specific cost of detection delay $\delta_i = T_i - \tau_i$.

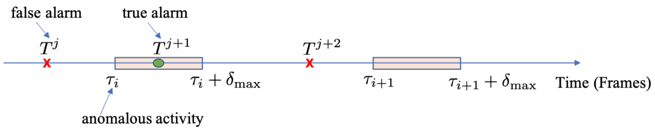

Figure 3. Definitions of true alarm and false alarm. The anomalous activity i is successfully detected with alarm time $T_i = T^{j+1}$, whereas the anomalous activity $i+1$ is missed.

Alarm Precision: The second objective of alarming only when necessary is equivalent to the well-known precision metric of binary classification. Maximizing the alarm precision means maximizing the ratio of Number of true alarms/Number of all alarms. As illustrated in Fig. 3, an alarm j is a true alarm if it is raised within the relevant duration of an anomalous activity, i.e., $T^j \in \cup[\tau_i, \tau_i + \delta_{\text{max}}]$, otherwise it is a false alarm. We combine close anomalous activities into a single one, e.g., car crashing and people are running, such that the anomalous activity intervals do not overlap, i.e., $[\tau_i, \tau_i + \delta_{\text{max}}] \cap [\tau_{i+1}, \tau_{i+1} + \delta_{\text{max}}] = \emptyset, \forall i$. If multiple alarms are raised within an anomalous activity interval, only the first one is considered as true alarm, and the rest is ignored. Mathematically, we want to maximize the probability $(T^j \in \cup[\tau_i, \tau_i + \delta_{\text{max}}])$, which gives the alarm precision. Empirically, the alarm precision is computed as

$$P = \frac{1}{M}\sum_{j=1}^{M} T^j \in \cup[\tau_i, \tau_i + \delta_{\text{max}}], \tag{2}$$

where $M = |\{T^j\}|$ is the number of all alarms and $|\cdot|$ denotes the cardinality of a set. Note that the alarm precision is much easier to calculate than false alarm/positive rate (FPR), another commonly used metric in binary classification. While the normalization term M in precision, i.e., number of all alarms, is easy to know, false alarm rate requires the number of all nominal activities, which is not easy to find.

Average Precision Delay: In order to obtain a single metric for conveniently comparing VAD algorithms, we propose a new metric called *Average Precision Delay (APD)*, which combines average detection delay and alarm precision. Similar to the way the popular AUC metric summarizes TPR and FPR, APD measures the area under the Precision vs. normalized ADD (NADD) curve. To map ADD into $[0, 1]$, we normalize it by the maximum delay, i.e., NADD $=$ ADD$/\delta_{\text{max}}$. Mathematically, APD is given by

$$\text{APD} = \int_0^1 P(\alpha)\,\mathrm{d}\alpha, \tag{3}$$

where α denotes NADD, and P denotes the precision. A highly successful algorithm with an APD value close to 1 must have high precision and low delay in its alarms.

Continual Learning: In the proposed continual learning framework, the VAD algorithm is trained in multiple sessions over time using several batches of nominal data, called splits (Fig. 1). In practice, training splits may arrive irregularly with varying sizes. Following the common practice in VAD, no labels are provided with the training splits. Although training data is assumed to be nominal, some level of contamination with anomalous activities may be tolerated depending on the robustness of the VAD algorithm to outliers. The objective in the continual learning setup is to improve the APD performance consistently with each training split k, i.e.,

$$\text{APD}_k \geq \text{APD}_{k-1}, \forall k. \tag{4}$$

The APD value is measured after each training split using all the available test data. Assessing the performance on a comprehensive test dataset is important to see if the algorithm suffers from catastrophic forgetting. If the algorithm is not suitable for continual learning, it may start to lose performance although more training data and accordingly more knowledge becomes available. On the contrary, a successful continual VAD algorithm will consistently improve its APD performance with more training splits.

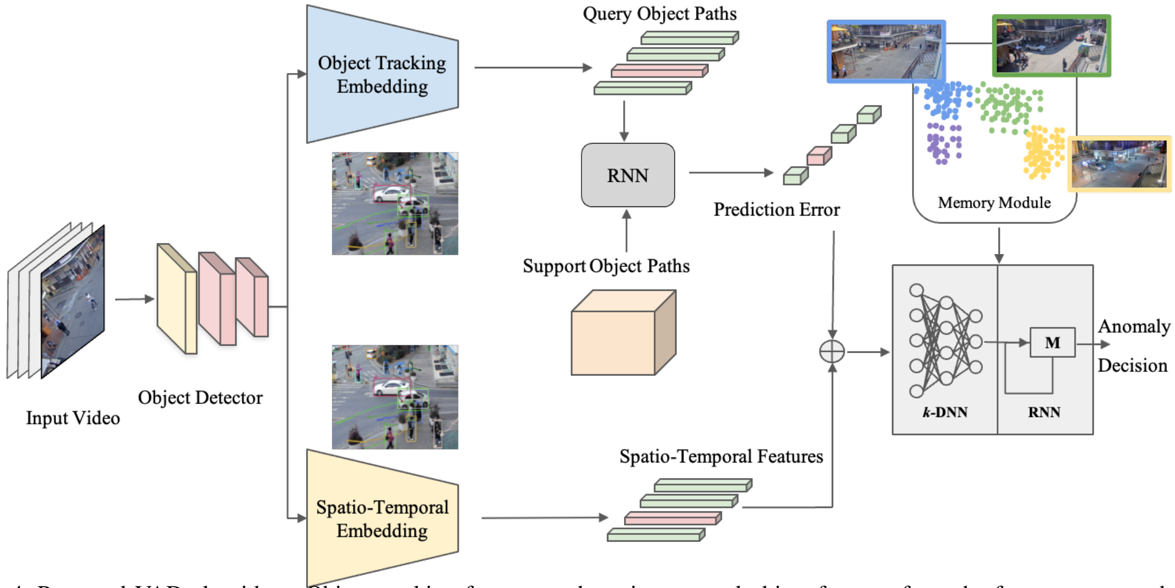

Figure 4. Proposed VAD algorithm. Object tracking features and spatio-temporal object features form the feature vector, whose kNN distance with respect to the nominal vectors is used to make an anomaly decision within an RNN structure. The use of kNN distances facilitates effective continual learning.

3.2. Continual VAD Algorithm

Due to the the tendency of deep neural networks to forget previously learned information when the network is trained sequentially on multiple tasks, end-to-end trained VAD models are not suitable when it comes to continual learning. Even though experience replay has shown promising results on toy examples recently, it still cannot be scaled up to problems with complex tasks since constantly retraining on all previously learned tasks is highly inefficient, and the amount of data that would have to be stored quickly becomes unmanageable [34]. However, in this work, we show that this challenge can be addressed by treating continual learning with a two-stage approach: by first extracting a low dimensional feature embedding for each frame using end-to-end deep learning models and then employing k-Nearest-Neighbors (kNN) based RNN model to prevent catastrophic forgetting.

As shown in Fig. 4, we first detect objects in each frame by using a pretrained object detector, such as YOLO-v4 [29]. Then, we use the extracted bounding boxes to construct a feature embedding to represent the spatio-temporal activities observed in the frame. Particularly, we monitor the number of objects detected per object class, the number of object classes observed, the day of the week and the time of the day the video frame belongs to. To limit the computational complexity, we discretize the day dimension into two categories as weekday and weekend. Similarly, we discretize the time of the day into four categories as active and inactive times of day and night. In addition, to extract more intricate features from each detected object, we also employ a re-identification and tracking algorithm called DeepSORT [36], which performs real-time path tracking of each detected object. The extracted object paths are provided to an RNN to make predictions about the future path. The prediction errors for all object paths are then stacked into a feature vector together with the spatio-temporal features.

Next, the kNN distance of the feature vector is computed with respect to the set of nominal feature vectors stored in the memory module. As explained next, we consider two different ways of computing the kNN distance for continual learning purposes. The single-dimensional time series of kNN distances provides evidence for anomalies since the frames from anomalous activities typically lie farther away in the feature space from the nominal frames. However, to leverage the temporal continuity among frames, we do not directly decide for each frame using its kNN distance; we use an RNN structure to capture the temporal dependency in kNN distances and decide using that sequential information. To train the RNN with anomalous frames, we use synthetic kNN distances generated uniformly between the 95th percentile of nominal kNN distances and its double.

Continual Learning: We propose two approaches for continual learning, which are based on two different ways of computing the kNN distance. The first one is based on exact kNN distance computation and is particularly useful for continually learning nominal behaviors when the amount of training data is still tractable. In this approach, we incrementally update the memory module with the kNN distance of new features from each training split. However, with many training splits over a long time horizon, the exact computation of kNN distance may be prohibitive as the nominal training set grows. For long-term scalability, we propose a second approach which estimates the kNN distance using a fully-connected deep neural network (k-DNN). To continually update k-DNN, we use experience replay, i.e., in addition to the most recent feature vector and its kNN value, previous feature vectors and kNN values are also used to update k-DNN. The second approach has the advantage of being computationally efficient during testing, especially when the training set is large.

Implementation Details: For the kNN regression network (k-DNN), we use a fully-connected deep neural network with 3 hidden layers consisting of 20 neurons each. We empirically chose the simplest network that gave a sufficiently low prediction error. A single hidden layer LSTM with a two input time steps is used for the decision RNN. The YOLO object detector is trained on the MS-COCO dataset with 80 classes, and the DeepSORT object tracker is trained on the MOT16 dataset. For path prediction, an LSTM with three hidden layers with 20 input time steps is used. We remove trajectories which last for less than 50 frames. All the features are normalized to $[0, 1]$ using the maximum and minimum values from training. The entire pipeline is able to run at approximately 18 fps on a RTX 2070 GPU, which can be significantly improved by using a better GPU or more lightweight models. Moreover, to maintain real-time performance, the videos can also be analyzed at lower fps. For the maximum detection delay, we set a limit of 5 minutes, which we believe is sufficient for detecting any type of anomaly.

4. Dataset for Continual VAD

The popular benchmark datasets (UCSD, Avenue, ShanghaiTech) in VAD are not sufficiently comprehensive for the continual learning framework. There is a recent multi-scene dataset, UCF Crime [32], which is significantly larger and more complex than the popular benchmarks. However, having been collected from various YouTube videos this multi-scene dataset is also not suitable for continual learning since the sheer heterogeneity in the dataset causes incompatibility issues [27]. For instance, an obvious anomalous activity in one scene cannot be detected since a very similar activity has appeared as nominal in a quite different scene. Hence, instead of a multi-scene setup with spatial richness (i.e., comprehensive data over various scenes), we focus on a single-scene setup with a new dataset that provides temporal richness (i.e., comprehensive data over time).

4.1. Existing Datasets

The three popular benchmark datasets for VAD are discussed below.

UCSD Ped 2: The UCSD Pedestrian dataset is one of the most widely used VAD datasets. Due to the small resolution of the UCSD Ped 1 videos, most recent works only consider the UCSD Ped 2 dataset. The Ped 2 dataset consists of 16 training videos and 12 test videos. The anomalous activities are caused by vehicles such as bicycles, skateboards and wheelchairs. Despite being widely used as a benchmark dataset, most anomalies are obvious and can be easily detected from a single frame.

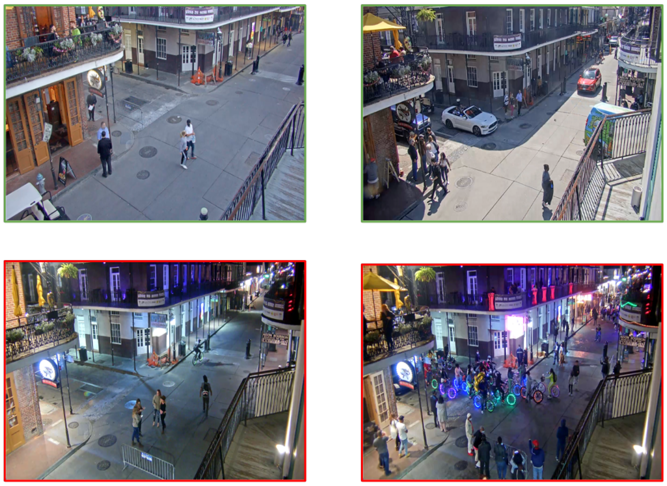

Figure 5. Sample frames from nominal (top row) and anomalous (bottom row) activities in the proposed NOLA dataset.

CUHK Avenue: Another popular dataset is the CUHK Avenue dataset, which consists of short video clips taken from a single outdoor surveillance camera. It contains 16 training and 21 test videos with a frame resolution of 360 × 640. While it is more challenging than the UCSD dataset, the anomalies are staged and the labeling of the anomalous instances is not consistent.

ShanghaiTech: The ShanghaiTech dataset is one of the largest and most challenging datasets available for anomaly detection in videos. It consists of 330 training and 107 test videos from 13 different scenes, which sets it apart from the other available datasets. The resolution for each video frame is 480 × 856. However, the videos are captured from 13 different cameras, which makes it a multi-scene formulation. On the other hand, treating it as 13 different datasets severely limits the number of available training frames for each scene.

4.2. New Dataset: NOLA

We introduce a new dataset which consists of 110 training video segments in 11 splits and 50 test segments captured over an entire week from a single moving camera[2] from a famous street in New Orleans, Louisiana, USA. To maintain consistency and avoid unrealistic normalization assumptions, all the training and testing video segments are

[2] https://www.earthcam.com/usa/louisiana/neworleans/bourbonstreet/?cam=catsmeow2

Dataset	**Total Frames**	**Training Frames**	**Testing Frames**	**Ground Truth**	**Resolution**	**Note**
UCSD Ped1	14,000	6800	7200	Spatial, Temporal	238 x 158	–
UCSD Ped2	4560	2550	2010	Spatial, Temporal	360 x 240	–
Subway	125,475	22,500	102,975	Temporal	512 x 384	2 scenes
CUHK Avenue	30,652	15,328	15,324	Spatial, Temporal	640 x 360	–
UMN	3,855	N/A	N/A	Temporal	320 x 240	Frames not directly available
ShanghaiTech	317,398	274,515	42,883	Spatial, Temporal	856 x 480	13 scenes
Street Scene	203,257	56,847	146,410	Spatial, Temporal	1280 x 720	–
NOLA (proposed)	**1,440,000**	**450,000**	**990,000**	**Spatial, Temporal**	**1280 x 720**	Audio also available

Table 1. Comparison of existing and proposed VAD datasets. Ground truth refers to the type of anomaly labeling.

clipped at 9000 frames, extracted at 30 frames per second. Overall, the dataset consists of 990,000 training frames and 450,000 testing frames, making it significantly larger than any other available dataset, as shown in Table 1. The dataset was manually collected, cleaned and annotated by the authors. The training set is split into 11 smaller batches to evaluate the performance in terms of continual learning, as described in Section 3.1. One split is used for initial training, and the rest 10 splits are used to evaluate the continual learning performance (Fig. 1).

In contrast to existing datasets, the proposed dataset consists of videos captured during day and night, as well as on various days of the week. This information is also provided in the form of metadata, which we believe is especially crucial since the expected amount of activity is directly related to the day and time. The proposed dataset is especially challenging because the anomalies are contextual in nature and require a deeper understanding of the videos. For example, loitering is considered as nominal during daytime, but anomalous during night. Other examples of anomalous events include a person carrying a snake, a vehicle moving in the wrong direction, sudden appearance of several bikes, etc. as anomalous. Sample frames from nominal and anomalous activities are given in Fig. 5. To detect such an anomaly, an algorithm will need to understand the behaviors with respect to the day and time. Also, since the camera alternates between two different views of the same street, each with an independent nominal baseline, it is challenging to adapt to such contextual changes. There is also audio data available in the NOLA dataset, which is not used in this work but may be helpful in future studies by providing extra information [3].

5. Experiments

In this section, we compare the continual learning capability of the proposed algorithm and state-of-the-art VAD methods. While there are a few approaches [5, 24] which attempt to continuously learn nominal behaviors from a toy dataset, their objective is to minimize the false alarm rate by updating their baseline model without considering the detection delay or TPR performance. However, to the best of our knowledge, since there is no existing approach that is designed for continual VAD, we modify two existing state-of-the-art approaches, namely the Future Frame Prediction method [17] and the Memory guided Normality (MNAD) method [26]. The future frame prediction method proposes a GAN architecture to learn appearance and motion features and aims to predict the future frames. Its detection is based on the assumption that a previously unseen activity causes a higher prediction error. On the other hand, the MNAD approach proposes a reconstruction based approach using autoencoders. We chose these two algorithms since their codes were readily available, and they could be tweaked to learn both incrementally and in batches. We also attempted to implement a more recent algorithm proposed in [12] since they also propose an object-centric approach more akin to our proposed algorithm; however, our version was unable to achieve a score close to their reported results.

Results on the Proposed NOLA Dataset: We first study the continual learning performance of the proposed and benchmark algorithms on the new NOLA dataset using the setup introduced in Sec. 3.1. In this experiment, we use the k-DNN and experience replay based version of our algorithm. From Table 2, we can see that the proposed algorithm clearly outperforms the two benchmark algorithms across all splits. Particularly, the proposed algorithm performs well at detecting anomalous activities such as a vehicle moving in the wrong direction and a person loitering after midnight. Since the initial training data consists mainly of videos captured during a weekday, we first see several false alarms caused due to test videos from weekend, which exhibits a significantly higher activity level. These false alarms gradually decrease after each split as we continually learn new baselines. In contrast, we see performance decrease for the benchmark algorithms on several splits, indicating that they suffer from catastrophic forgetting. For instance, although the future frame prediction algorithm has shown competitive performance on the existing benchmark datasets, we see that it is not capable of predicting more complex scenarios. Specifically, even after training on several thousand frames of people using a bicycle, the algorithm gives a high prediction error whenever it sees a similar

[3] The entire dataset will be publicly available

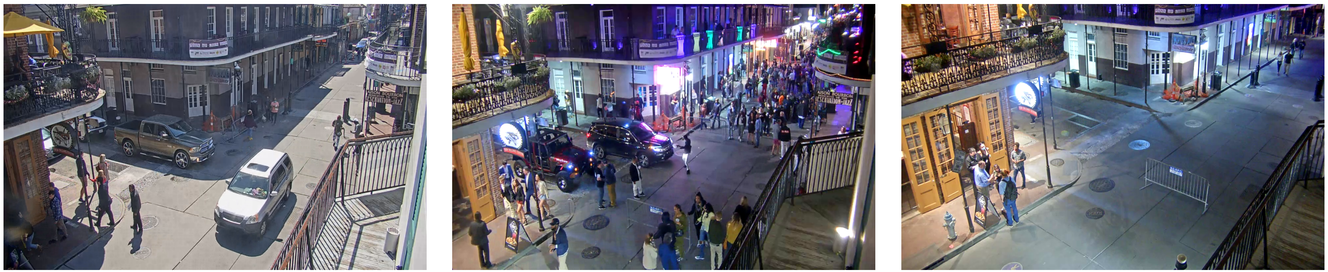

Figure 6. Fail cases.

Method	CL-1	CL-2	CL-3	CL-4	CL-5	CL-6	CL-7	CL-8	CL-9	CL-10
Future Frame Prediction [17]	0.137	0.149	0.173	0.205	0.211	0.232	0.202	0.22	0.245	0.271
MNAD [26]	0.162	0.21	0.219	0.262	0.251	0.289	0.311	0.271	0.295	0.28
Ours	0.235	0.239	0.243	0.296	0.317	0.323	0.325	0.375	0.377	0.401

Table 2. Performance of the proposed detector and recent state-of-the-art approaches across different continual learning splits in terms of the proposed APD metric.

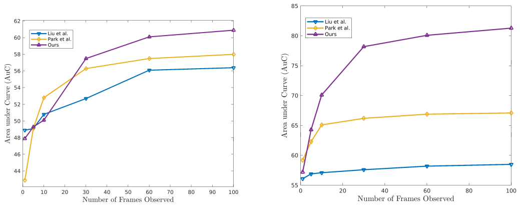

Figure 7. Comparison of the proposed and state-of-the-art algorithms Liu et al. [17] and Park et al. [26] in terms of learning from few samples on the ShanghaiTech (top) and UCSD (bottom) datasets.

activity in the test videos. This result shows why it is imperative for VAD algorithms to be evaluated on more comprehensive datasets.

Results on Existing Benchmark Datasets: To further analyze the performance of our model and to provide a fair comparison with the benchmark algorithms, we also provide performance evaluation results on the benchmark datasets using the popular frame-level AUC metric. However, since these datasets are significantly smaller, it is not possible to split them similar to the continual learning framework proposed in Sec. 3.1. Hence, we design a specific scenario in which the objective is to learn a new activity type which was unavailable in the training dataset. Specifically, we choose a person riding a bicycle as our new nominal activity, since it is the only anomalous case which is common in UCSD Ped 2 and ShanghaiTech[4] datasets that occurs several times. Fig. 7 shows that our proposed algorithm outperforms the benchmark algorithms even with the classical metric on the existing benchmark datasets. Since the datasets are relatively small here, we employ the incremental version of the proposed algorithm based on exact kNN distances.

5.1. Discussion

While the proposed detector is able to detect several kinds of anomalies, it is tuned to learn continuously and reduce the number of false alarms rather than analyze each frame intricately. Hence, in Fig. 6, we analyze a few cases in which the proposed detector is unable to raise an alarm. In the first case, the anomaly is due to a person carrying a snake in a crowded street. In the second one, we see a person deliberately stopping a car by dancing in front of it. Finally, in the third one, we see a couple arguing with the restaurant owners. To detect such anomalies, a VAD algorithm needs to have a much deeper understanding of the intricate relationships between each detected object and how it affects its surroundings. Nevertheless, this also presents the richness of the proposed NOLA dataset, and how it can help improve future VAD algorithms.

6. Conclusion

We presented a new framework and a new comprehensive dataset for continual learning in video anomaly detection. We hope the new problem formulation (Sec. 3) and the new dataset (Sec. 4) will help guide the future VAD research towards practical and reproducible solutions. We also presented a novel video anomaly detector capable of learning continuously both incrementally and through experience replay. Through extensive testing on the proposed NOLA dataset and available benchmark datasets, we show that the proposed algorithm outperforms two of the state-of-the-art approaches in continual learning, as well as in terms of the standard frame-level AUC metric. For future work, we plan on leveraging audio and video in a multi-modal setup for improved detection performance.

[4] We choose a subset of the entire dataset to test on since the videos are from several different cameras. The exact split is described in the Supplementary material.

References

[1] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1932–1939. IEEE, 2009.

[2] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):673–682, 2016.

[3] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011.

[4] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020.

[5] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 254–255, 2020.

[6] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021.

[7] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

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

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

[11] Timothy Hospedales, Shaogang Gong, and Tao Xiang. A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1165–1172. IEEE, 2009.

[12] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. 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*, pages 7842–7851, 2019.

[13] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[14] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1446–1453. IEEE, 2009.

[15] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2019.

[16] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

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

[18] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*, 2017.

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

[20] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. *arXiv preprint arXiv:2007.07843*, 2020.

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

[22] Xuan Mo, Vishal Monga, Raja Bala, and Zhigang Fan. Adaptive sparse representations for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(4):631–645, 2013.

[23] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019.

[24] Ramy Mounir, Roman Gula, Jörn Theuerkauf, and Sudeep Sarkar. Temporal event segmentation using attention-based perceptual prediction model for continual learning. *arXiv preprint arXiv:2005.02463*, 2020.

[25] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1273–1283, 2019.

[26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[27] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020.

[28] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2598–2607, 2020.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[30] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020.

[31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[32] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[33] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[34] Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14, 2020.

[35] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[36] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[37] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020.

[38] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. IEEE, 2011.