Multi-Task Learning for Video Surveillance with Limited Data

Keval Doshi and Yasin Yilmaz
University of South Florida
4202 E Fowler Ave, Tampa, FL 33620
{kevaldoshi, yasiny}@usf.edu

Abstract

Learning from limited data in video surveillance is important for sustainable performance while adapting to new information in a scene over time or adapting to a different scene. In a real-world scene, for an anomaly detection algorithm, all possible nominal patterns and behaviors are not typically available immediately for a single training session. In contrast, labeled nominal data patterns may become available irregularly over a long time horizon, and the anomaly detection algorithm needs to quickly learn such new patterns from limited samples for acceptable performance. Otherwise, it would suffer from frequent false alarms. Additionally, the anomaly detection algorithm needs to continually learn new nominal patterns in multiple training sessions without forgetting the previous knowledge and losing performance. Cross-domain adaptability (i.e., transfer learning to another surveillance scene) is another task where the anomaly detection algorithm has to quickly learn from limited nominal training data to achieve acceptable performance. To overcome these challenges, we design a modular framework and use it to extract semantic embeddings, which we then train on by using deep metric learning. Particularly, we study these three problems (few-shot learning, continual learning, cross-domain adaptability) in a multi-task learning setting. We also compare our proposed framework to existing state-of-the-art approaches using various evaluation metrics. The empirical results indicate that the proposed approach is able to outperform the existing approaches on all three tasks for three benchmark datasets.

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. Particularly, the ability to detect events in real-time is critical for prevention of potential catastrophes. Hence, video anomaly detection has been attracting an increasing amount of research interest, with most of the recent approaches heavily dependent on end-to-end trained complex deep learning based approaches [11, 32, 43].

In the literature, the video anomaly detection problem is formulated as detecting activities or events that diverge from those typically seen in the training data. This is particularly challenging due to the fact that most anomalies are contextual in nature, making it almost impossible to obtain a fairly representative set of anomalies. Hence, conventional supervised learning approaches are not feasible in video anomaly detection. For example, in the popular UCSD [30] and ShanghaiTech [32] video anomaly detection benchmark datasets, a person riding a bike is considered as anomalous; however, in the recently released Street Scene [44] dataset, it is considered as nominal. Hence, most of the existing approaches [2, 4, 8–10, 17, 23, 32, 40, 42, 44, 45, 55] focus on learning an all-encompassing notion of normality, and detect events that deviate from it.

A crucial task which is neglected by almost all existing algorithms is cross-domain adaptability, where a trained model is able to perform reasonably well on a completely new surveillance scene without requiring any additional training. While a similar task was discussed in [34], the proposed approach still required some training data from the new scene to fine-tune its model using meta-learning. This approach might not always be feasible since it requires a human operator to manually collect a representative set of nominal frames which also includes new activities pertaining to the surveillance scene, which is not ideal. Furthermore, this cannot be automated by using pretrained activity recognition models since each video sequence consists of several different activities occurring at once, which even the current state-of-the-art approaches cannot detect accurately.

Moreover, in the traditional formulation with a single training session, the inherent assumption that the training data includes all possible nominal activities is unrealistic. Even while considering a single scene (e.g., a static camera monitoring a particular street) setup, it is not possible to capture all possible nominal activities in a single training session. Rather, it would be more realistic to treat the nominal class as an "open set", as in continual learning [33]. 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. However, unlike humans, deep learning based approaches are unable to learn *incrementally* from new incoming data without suffering from catastrophic forgetting [25], or learn a new pattern from only a few samples. Furthermore, current approaches require training a model from scratch for each scene, even when the objective and most data patterns remain the same (e.g., for a different camera view monitoring a similar street).

For practical implementations, it is also unreasonable to assume the availability of sufficient training data for all nominal events/behaviors. This presents a novel challenge to the current approaches discussed in Section 2 as their decision functions heavily depend on Deep Neural Networks (DNNs) [10]. In the existing benchmark datasets, several frames are available for all nominal activities, which makes it relatively straightforward for recent methods to learn them. However, almost all recent approaches neglect analyzing the performance of their models in absence of sufficient training data for a certain activity.

To summarize, our contributions in this paper are as follows:

- We propose the first multi-task learning framework capable of cross adaptability, few-shot learning, and continual learning for video anomaly detection with limited data.
- We propose the first semantic embedding based approach for video anomaly detection using deep metric learning, which significantly reduces the memory and computational requirements.
- We extensively evaluate our proposed approach on each task using publicly available datasets and show that we can transition effectively between them.

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) [3, 5, 30], Hidden Markov Models (HMMs) [20, 27], sparse coding [39, 54], and appearance features [6, 30], recent approaches have been completely dominated by deep learning based algorithms. Recent algorithms can be broadly classified into reconstruction based approaches [15, 17, 36, 41, 43], which try to classify frames based on the reconstruction error, and prediction based approaches [8, 11, 29, 32], which attempt to predict a future frame, primarily by using generative adversarial networks (GANs) [16]. More recently, skeletal trajectory based approaches [40, 49] have been proposed since a large proportion of anomalies in the benchmark datasets involve anomalous human poses. In such algorithms, an RNN architecture is typically used to learn nominal poses, and estimation error is used during testing to detect the level of abnormality. Apart from these approaches, [45] 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.

Several recent works propose using a GAN for detecting anomalies in videos. For example, [32] proposes a future frame prediction network which attempts to predict the future frame based on a sequence of input frames, and computes the prediction error in terms of the peak signal to noise ratio. However, such an approach cannot be practically implemented since GANs are notoriously difficult to train on few samples. Moreover, retraining a GAN from scratch to offset catastrophic forgetting is computationally infeasible.

Hence, continual learning has been recently gaining increasing research interest [26, 33, 50, 52, 53]. However, not a lot of progress has been made yet in continual learning for video anomaly detection. In [10], 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 [18, 22], which treat each object independently, and fails to capture the intricate relationship between different objects.

3. Multi-Task Video Surveillance

In the existing video anomaly detection literature, the singular goal is to detect frames/behaviors which are not previously seen in the training data. However, for a detector to be applicable in a real-world scenario, a single model needs to be able to perform multiple tasks such as knowledge sharing among different scenes, and continually learning new behaviors from a few samples. Thus, we next carefully define a multi-task problem setup for video anomaly detection, which we believe should guide future research towards more comprehensive approaches.

3.1. Problem Setup

In the recent literature, most detectors train a reconstruction or prediction based deep learning model on a batch of video frames, typically in an end-to-end fashion to learn nominal appearance or motion features. However, we argue that for general video surveillance, such a setup is not optimal since learned visual embeddings are exceedingly dependant on conditions such as illumination, view point variation, occlusion, etc. Also, the standard framework implicitly assumes that sufficient training data is available for each activity from the target scene where the detector will be deployed [34]. Such an assumption requires a human to manually annotate hours of videos from each scene to generate an anomaly-free training dataset, which is far from ideal. Motivated by these observations, we propose a general video surveillance framework which consists of the following tasks in addition to anomaly detection.

T1: Cross-Domain Adaptability: Given videos from different scenes but a similar environment, it is fair to assume that the type of nominal activities remains consistent. Then, a model trained on one scene should be able to adapt to other scenes without needing any additional training. For example, in the benchmark video surveillance datasets discussed in Section 4, the same type of nominal activities are shared.

T2: Few-Shot Learning: For a more realistic setup, we assume that the training set consists of limited samples for some nominal activities, and thus a single model should also be capable of learning patterns from those few samples. This task is particularly essential since most recent methods are deep learning based, which are notoriously difficult to train on a few samples.

T3: Continual Learning: Finally, it is crucial for a detector to learn new nominal activities without suffering from catastrophic forgetting. Specifically, the detector should not lose performance while training on new nominal data.

Recently, [34] proposed a few-shot scene adaptation framework using meta-learning. However, it collects images during testing to calibrate the model to the new scene, which we argue is not ideal since it again requires human supervision to make sure that it does not include any anomalous activity in training. Furthermore, the approach in [34] is based on the future frame prediction model [32], which uses a GAN and thus is unable to perform tasks T1 and T3. Another recent work, [10] considered T3 as a necessary objective for a practical video anomaly detector. A kNN based approach was proposed, which requires the detector to store all the extracted visual embeddings from the training data in memory, and during testing find the Euclidean distance to it. While this allows for a rehearsal-free approach, it quickly becomes infeasible since the size of the memory required grows exponentially with the number of objects detected. Also, due to the high dimensionality of the visual embeddings, using clustering based approaches to limit the memory is computationally too expensive. To the best of our knowledge, this paper is the first to propose a multi-task framework which addresses all three tasks simultaneously and to consider zero-shot cross-domain adaptability. We next present our proposed approach.

3.2. Proposed Approach

Since humans perceive a visual environment in terms of activities, we believe that it is more natural and efficient to learn video activities *semantically* rather than storing entire frames in buffer or learning high-dimensional visual embeddings. Motivated by the human visual cortex system [1] which consists of six regions of cortical hierarchy (V1-V6), we first extract the spatial information (as in V1 & V2) from the scene by using a using a semantic segmentation model. This is followed by the global motion (as in V3), i.e., the direction and speed with which different objects travel, which is extracted using an optical flow model. Finally, we recognize basic objects (as in V4) using an object detector, and form relationships between the different modalities (as in V5 & V6).

However, unlike the existing approaches, instead of using the extracted features as a visual embedding to perform the tasks (T1-T3), we propose using the labels {location, appearance, motion} to extract a semantic embedding by using a Word2Vec model, as shown in Fig. 1. Such transformation has several advantages. Firstly, it is significantly easier to cluster similar labels as compared to high-dimensional visual embeddings, thus reducing the computational complexity. Secondly, this allows us to generalize better to different nominal activities, since in the Euclidean space, two similar activities such as a "person walking on the sidewalk" and a "person walking on the road" are quite apart, however, in the semantic space they are quite close. Finally, this also allows us to transfer knowledge between different scenarios since semantic embeddings are independent of spatial information. For example, a change in the location of road would render the learned visual features useless, whereas it would not affect the learned semantic features.

3.3. Deep Learning-Based Feature Extraction

In general, the end-to-end training of DNNs for video anomaly detection necessitates focusing on a particular aspect in which anomalies may occur, such as object appearance or motion or pose, and extracting only those features. However, even in the same scene, anomalous events may be manifested in different aspects. Hence, advanced video anomaly detectors should utilize features from multiple aspects together. For instance, biological vision systems extracts different features in the visual cortex such as appearance, global motion, and local motion [1]. To this end, we propose a flexible feature extraction module that can work with various modalities, which enables a plug-and-play modular architecture. This means although appearance, global motion, and local motion features are considered in this paper, the proposed framework can be eas-

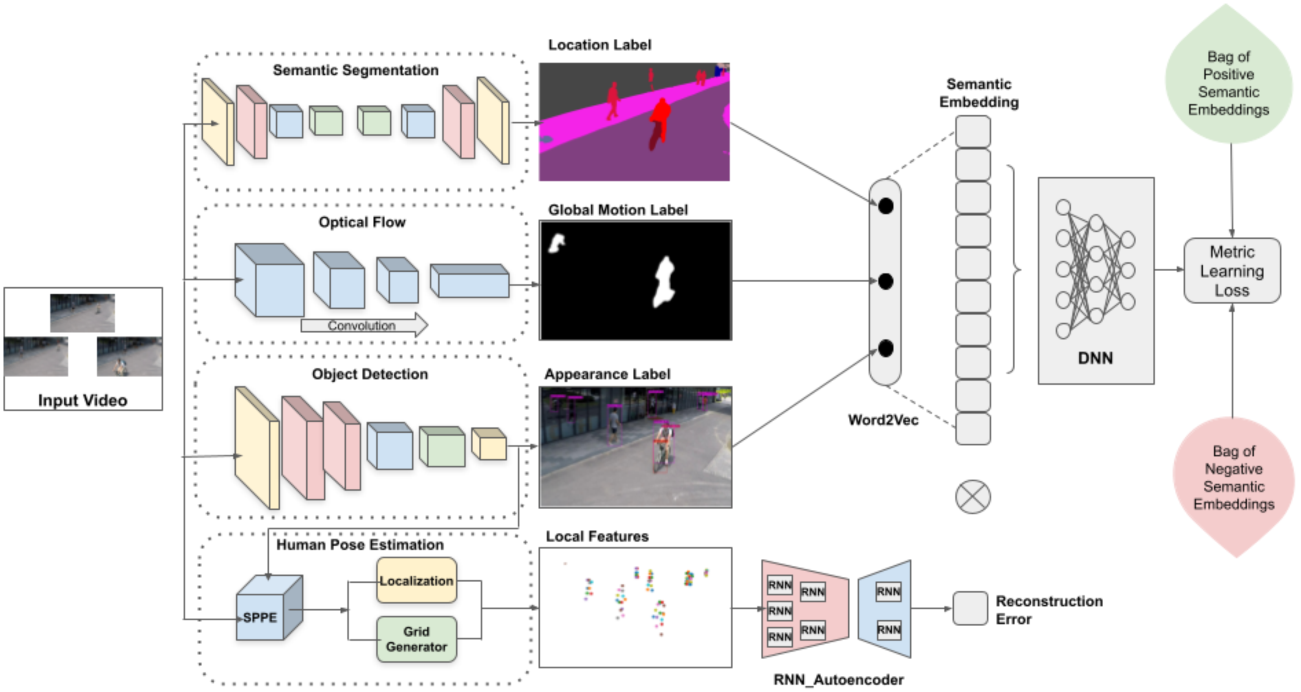


Figure 1. Proposed video anomaly detection framework. At each time t, neural network-based feature extraction module provides location, appearance and global motion labels, and local motion (pose estimation) reconstruction error, which is then used to form a semantic embedding which represents the detected activity. This is then used to train a deep neural network using metric learning, which outputs the anomaly score.

ily modified to add new feature extractors or remove existing ones. Furthermore, entirely retraining a video anomaly detector for new scene/domain is typically not necessary since most domains share the same feature types (appearance, global motion, local motion, etc.). As a result, to significantly reduce the training computational complexity, a transfer learning approach is utilized in the proposed framework. We next explain the considered feature extractors, which work in parallel as shown in Fig. 1.

Object Appearance: Object detection has received a lot of attention in recent years. Broadly, object detectors can be classified into single-stage and two-stage detectors. In single-stage object detectors such as YOLO (You Only Look Once) [47] and SSD (Single Shot Multibox Detector), the object detection task is treated as a simple regression problem, and directly outputs the bounding box coordinates. On the other hand, two-stage detectors such as Faster R-CNN [48] use a region proposal network first to generate regions of interest and then do object classification and bounding box regression. These methods are typically slower and take considerably longer, but are much better at detecting small objects. While single stage detectors are more efficient, we noticed that removing the false detections due to the lower accuracy accrues additional computational overhead, thus negating the advantage of using such detectors. To this end, following [31], we train a Faster R-CNN model which uses a Squeeze and Excitation Network (SENet) [21], since they generalize extremely well across different scenarios. SENet has a depth of 152 and uses a K-means clustering algorithm to cluster anchors, with the distance metric defined as:

$$D(box, centroid) = 1 - IoU(box, centroid),$$

where IoU denotes the intersection of union. Using the object detector, we extract the bounding box (location) as well as the class probabilities (appearance) for each object detected in a given frame. Instead of directly using the bounding box coordinates, we instead compute the center and area of the box and leverage them as our spatial features. During testing, any object belonging to a previously unseen class and/or deviating from the known nominal paths contributes to an anomalous event alarm.

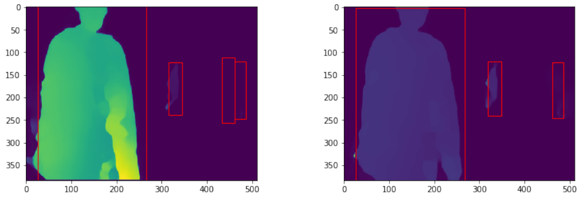
Global Motion: Apart from spatial and appearance features, capturing the motion of different objects is also critical for detecting anomalies in videos. We propose a novel modification of an optical flow model known as perspective based optical flow. Optical flow is widely used in the existing literature to extract motion features. While computing the optical flow from frames, it is a common occurrence that objects closer to the camera covers a larger portion and hence even a slight movement by such an object results in a significantly larger optical flow. Since in video surveillance large optical flow values essentially mean an anomaly, any object close to the camera could cause unnecessary false alarms. To prevent such occurrences, we propose a perspective-based optical flow approach which leverages object detection to normalize the optical flow. While perspective mapping has been widely used for detecting vehicles, crowds, and license plates, to the best of our knowledge, it has yet to be used for optical flow mapping. For obtaining the perspective-based optical flow, we assume that the difference between the actual heights of people detected in the videos is negligible. Then, the optical flow can be written as a function of the width and height of the bounding box, given by

$$O_1 = f\left(\frac{w_1 * h_1}{H_1}\right),$$

where O_1 is the optical flow intensity for a detected person, H_1 is the actual height, w_1 and h_1 are the width and height of the bounding box, respectively. Then, assuming H_1 to be constant for each detected person, we compute the normalized optical flow intensity as

$$O_{n1} = \frac{O_1}{w_1 * h_1}. \quad (1)$$

To also account for cases where the size of the detected person is too small and optical flow might not be very accurate, we set the minimum size of the person that can be detected in the image as 10. In Fig. 2, we see that in the first case there is an unusually high optical flow intensity because of a person passing near the camera, which would lead to false alarms. However, as shown in the second case, by using the perspective-based optical flow, we successfully reduce the intensity of the optical flow.



(a) Optical flow without perspective mapping.

(b) Optical flow with perspective mapping.

Figure 2. Objects closer to the camera have a significantly higher optical flow intensity even when they are moving at a nominal speed, which leads to false alarms. By using perspective-based optical flow, we successfully normalize such cases and prevent false alarms.

Local Motion: To study the social behavior in a video, it is an important factor to study the human motion closely. For inanimate objects like cars, trucks, bikes, etc., monitoring the optical flow is sufficient to judge whether they portray some sort of anomalous behavior. However, with regard to humans, we also need to monitor their poses to determine whether an action is anomalous or not. Hence, using a pre-trained multi-person pose estimator such as AlphaPose [13] is proposed to extract skeletal trajectories.

Location: Generalizing to different locations is a crucial step for seamless cross-domain adaptability. Specifically, activities occurring at similar locations need to be grouped together, or else it can lead to false alarms. For example, if the training data considers a person walking on the road as nominal, a similar activity such as walking on the sidewalk should not be considered as anomalous during inference. Hence, to recognize different background locations, we use a hierarchical multi-scale semantic segmentation model trained on the Cityscapes dataset [1]. The model uses HRNet-OCR as backbone and is more memory efficient than other approaches. It uses an attention based approach to combine multi-scale predictions.

Semantic Embedding: Finally, we define each detected activity for every frame in the form of its output label from each model. This simple transformation allows for several advantages. First, due to its small dimensionality, clustering similar semantic labels is significantly easier compared to clustering visual embeddings. This reduces the computational and memory cost, which is one of the issues [10] suffers from. Furthermore, it also allows for easy interpretability of the detected activity, which is a missing aspect in almost all recent works. Finally, practical challenges such as few-shot learning and continual learning can be easily implemented in the proposed approach. Hence, given semantic labels for each detected activity, we generate its corresponding semantic embedding by using a Word2Vec model and then average across them to form a 300-dimensional semantic feature vector. The reconstruction error is then concatenated with the semantic feature vector, to form the semantic label embedding for each detected activity. We also generate semantic embeddings for pseudo-abnormal activities, which are then used to determine if an activity is nominal or anomalous, by learning a new distance metric.

3.4. Anomaly Detection

Deep Metric Learning: Annotating anomalous frames in videos is a particularly challenging task. On the other hand, describing nominal and anomalous behaviors using semantic labels is relatively straightforward. Hence, we propose to learn a distance metric using a fully connected deep neural network. As shown in Fig. 1, in our proposed approach, we pose anomaly detection as a regression problem. We want the anomalous semantic video embeddings to

[1] https://github.com/NVIDIA/semantic-segmentation

have higher anomaly scores than the normal embeddings. To this end, we propose training a fully connected neural network with a custom loss function to learn a distance metric. The loss function is based on the triplet loss [19] and is defined as:

$$\mathcal{L} = \max(0, m + \|f(a) - f(p)\| - \|f(a) - f(n)\|), \quad (2)$$

where $f(\cdot)$ represents the semantic embedding function, a, p, n are the anchor, positive and negative semantic labels respectively. The margin m is used to determine the boundary after which the negative samples contribute to the loss.

On the other hand, localizing the anomalies temporally or spatially is not a time sensitive task and hence can be performed in an offline fashion. However, in previous works [22, 32, 38, 42, 44], there is a lack of distinction between *online detection* and *offline localization*. The majority of existing works are not suitable for online detection as they perform batch processing [32, 34, 42, 44, 49]. Some recent works [35, 49] use online methods like LSTM networks, but also require a normalization of decision statistic over a video segment, which prevents online detection. Moreover, as discussed in [28], traditional metrics such as precision and recall cannot effectively evaluate the performance of online anomaly detection algorithms. Hence, a new performance metric is needed for online anomalous event detection in videos.

Implementation Details: In our implementations, we use SENet for object detection, Flownet 2 for optical flow, AlphaPose for pose estimation and HRNet for semantic segmentation. The semantic embeddings are extracted using a Word2Vec model, and then input to a deep neural network with 3 layers consisting of 10 neurons each. The DNN is trained using a triplet loss. Global and local motion features are normalized to [0,1] using the min and max values from the training data.

4. Experiments

In this section, we present the performance of the proposed approach on the tasks defined in Section 3.1. We first present the performance of our model in terms of the online anomaly detection and anomalous frame localization on the three benchmark datasets. Then, we evaluate the cross-domain adaptability performance on the ShanghaiTech Campus dataset, which is the largest publicly available dataset for video anomaly detection, and consists of videos captured from 13 different cameras, and the CUHK Avenue dataset. To evaluate the first task of cross-domain adaptability, we use the learned model from the first camera in the ShanghaiTech dataset and test it on the data from all the other cameras from ShanghaiTech, as well as the CUHK Avenue dataset. For the second task of few-shot learning, we analyze the performance of the proposed approach with respect to the number of frames required to learn the new patterns. Finally, for the third task of continually learning new patterns, we check whether the performance of the proposed algorithm consistently improves with each training session. We also present the performance of our combined model on a real-world dataset.

4.1. Datasets

We consider three publicly available benchmark datasets, namely the CUHK Avenue dataset, the UCSD pedestrian dataset, and the ShanghaiTech campus dataset.

UCSD Ped 2: The UCSD pedestrian dataset is one of the most widely used video anomaly detection datasets. Due to the low resolution of the UCSD Ped 1 videos, we only consider the UCSD Ped 2 dataset. The Ped 2 dataset consists of 16 training videos and 12 test videos.

CUHK Avenue: Another popular dataset is the CUHK Avenue dataset, which consists of short video clips taken from a single outdoor surveillance camera looking at the side of a building with a pedestrian walkway in front of it. It contains 16 training and 21 test videos with a frame resolution of 360 × 640.

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.

For online detection, we evaluate existing approaches using the Average Precision Delay (APD) metric proposed in [12], which computes the area under the precision and average detection delay curve. For offline localization, we leverage the traditional Area under the ROC Curve metric (AUC).

4.2. Results

Online Detection: Since the proposed online detection formulation is event-based as compared to the classical frame-based formulation, it only considers an anomaly as a single event irrespective of the duration over which it occurs. In this setup, we present our results only on the ShanghaiTech dataset as the UCSD and CUHK Avenue datasets have fewer than 50 anomalous events, which is not enough for a reliable average performance comparison. A common technique used by several recent works [22, 32, 40, 43] is to normalize the computed statistic for each test video independently, including the ShanghaiTech dataset. However, this methodology cannot be implemented in an online (real-time) system as it requires the prior knowledge of the minimum and maximum values the statistic might take. Moreover, many recent methods [22, 34, 42] do not have their implementation details/code publicly available, while others are end-to-end [42, 44, 49] and cannot be implemented to work in an online fashion. Hence, we compare our method

Online Detection	
Method	**APD**
Liu et al. [32]	0.504
Morais et al. [40]	0.324
Luo et al. [37]	0.447
Ours	**0.675**

Table 1. Online detection comparison in terms of the proposed APD metric on the ShanghaiTech dataset. Higher APD value represents a better online anomaly detection performance.

Anomaly Localization (AUC)			
Method	**CUHK Avenue**	**UCSD Ped 2**	**ShanghaiTech**
MPPCA [24]	-	69.3	-
Del et al. [7]	78.3	-	-
Conv-AE [17]	80.0	85.0	60.9
ConvLSTM-AE [35]	77.0	88.1	-
Growing Neural Gas [51]	-	93.5	-
Stacked RNN [36]	81.7	92.2	68.0
Deep Generic [18]	-	92.2	-
GANs [46]	-	88.4	-
Future Frame [32]	85.1	95.4	72.8
Skeletal Trajectory [40]	-	-	73.4
Multi-timescale Prediction [49]	82.85	-	**76.03**
Memory-guided Normality [43]	88.5	97.0	70.5
Ours	**86.4**	**95.6**	70.12

Table 2. Offline anomaly localization comparison in terms of frame-level AUC on three datasets.

with the online versions of [32, 37, 40]. Our proposed algorithm achieves a better performance than the other algorithms in terms of quick detection and achieving high precision in alarms, as indicated by Table 1 in terms of the APD value.

Anomalous Frame Localization: To show the anomaly localization capability of our algorithm, we also compare our algorithm to a wide range of state-of-the-art methods, as shown in Table 2, using the commonly used frame-level AUC criterion. The pixel-level criterion, which focuses on the spatial localization of anomalies, can be made equivalent to the frame-level criterion through simple post-processing techniques [44]. Hence, for anomaly localization, we consider the frame-level AUC criterion. As shown in Table 2, our proposed algorithm outperforms the existing algorithms on the UCSD Ped 2 and CUHK Avenue datasets, and performs competitively on the ShanghaiTech dataset. The multi-timescale framework [49] is the only one that outperforms ours on the ShanghaiTech dataset since the anomalies are mostly caused by previously unseen human poses and [49] extensively monitors them using a past-future trajectory prediction based framework. However, this causes their performance to severely degrade on the CUHK Avenue dataset, and similar to [40], they cannot work on the UCSD dataset.

Cross Domain Adaptability: In this case, we only train our model on the training videos from a single camera in the ShanghaiTech dataset and evaluate its performance on the test videos from the rest of the cameras, and also on the Avenue dataset. Cross-domain scene adaptation is mostly unexplored and to the best of our knowledge only [34] discusses a similar few-shot adaptation concept. However, the proposed approach discussed in [34] requires several anomaly-free video frames for adapting their model to the new scene, which might not always be feasible. Particularly, in [34], a GAN-based framework is used in [34] similar to [32], and MAML algorithm [14] is used for meta-learning. As shown in Tables 3–5, considering zero-shot adaptability the proposed approach is able to outperform the state-of-the-art methods in terms of the frame-level AUC, as well as the proposed APD metric. In both of the considered datasets, behaviors that are considered anomalous are the same, which satisfies our inherent assumption.

Few-Shot Learning: Unlike the original UCSD and ShanghaiTech datasets, where an individual riding a bike is considered abnormal, we presume that this is a nominal activity with few training samples in this case. However, the remaining anomalous events in the UCSD dataset, such as a skateboarder or a cart passing by, are still considered anomalous. Our goal here is to compare the few-shot learning capability of the proposed and state-of-the-art algorithms and see how well they adapt to new patterns. In this case, together with the available training data, we also train on a few samples of a person riding a bike. In Fig. 3, it is seen that the proposed algorithm clearly outperforms the state-of-the-art algorithms [9, 32, 36] in terms of few-shot learning performance. It is important to note that for video applications, 10 shots (i.e., frames) correspond to less than a second in real time.

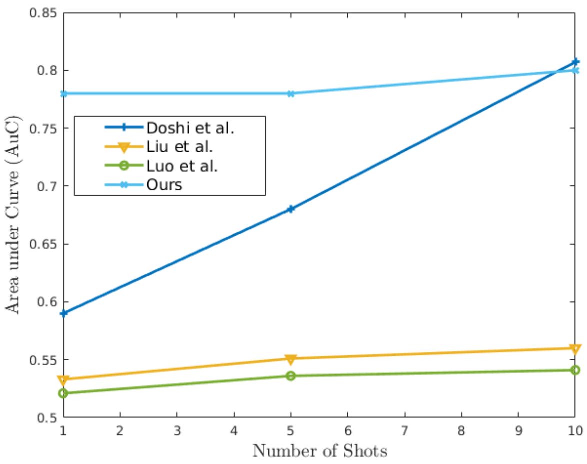


Figure 3. Comparison of the proposed and state-of-the-art algorithms Liu et al. [32], Luo et al. [36] and Doshi et al. [9] in terms of few-shot learning. Together with the original training data, some frames for bike riding are used to train the algorithms. The proposed algorithm achieves high performance even with one shot.

Method	Cam-1	Cam-2	Cam-3	Cam-4	Cam-5	Cam-6	Cam-7	Cam-8	Cam-9	Cam-10	Cam-11	Cam-12
Stacked RNN [36]	0.6412	0.6083	0.6116	0.6231	0.6834	0.6951	0.6482	0.6294	0.6867	0.6789	0.6924	0.6485
Future Frame Prediction [32]	0.6780	0.6178	0.6632	0.6588	0.6984	0.7351	0.6814	0.6186	0.6743	0.6789	0.6548	0.6509
Ours	0.7529	0.7065	0.7613	0.6813	0.7843	0.8137	0.7888	0.6258	0.7064	0.663	0.7531	0.7193

Table 3. Performance of the proposed detector in terms of frame-level AUC for cross-domain adaptability on different cameras from the ShanghaiTech Dataset.

Method	Cam-1	Cam-2	Cam-3	Cam-4	Cam-5	Cam-6	Cam-7	Cam-8	Cam-9	Cam-10	Cam-11	Cam-12
Stacked RNN [36]	0.401	0.442	0.4874	0.5012	0.4378	0.4275	0.487	0.5031	0.4145	0.4612	0.4365	0.4457
Future Frame Prediction [32]	0.4730	0.4356	0.4647	0.4537	0.512	0.5832	0.5534	0.5203	0.5043	0.4989	0.4762	0.4831
Ours	0.6482	0.5671	0.6743	0.6980	0.6944	0.5963	0.6175	0.5958	0.5734	0.61	0.6482	0.6725

Table 4. Performance of the proposed detector in terms of event-level APD for cross-domain adaptability on different cameras from the ShanghaiTech Dataset.

Frame-level AUC		
Approach	**ShanghaiTech**	**Avenue**
Stacked RNN [36]	0.643	0.724
Future Frame Prediction [32]	0.652	0.749
Skeletal Trajectory [40]	0.683	-
Ours	0.689	0.79

Table 5. Overall performance of each model in terms of frame-level AUC for cross-domain adaptability when trained on camera 1 from the ShanghaiTech dataset and tested on the entire ShanghaiTech and Avenue datasets.

Continual Learning: Due to the lack of existing benchmark datasets for continual learning in surveillance videos, we follow the same modification to the original ShanghaiTech dataset as in the few-shot learning scenario, and presume that riding a bike is a nominal behavior. Our aim is to compare the proposed and state-of-the-art algorithms' continuous learning capabilities for video surveillance to see how well they respond to new trends. The algorithms are initially trained on the original training data, and then incrementally updated using the bike frames. In Figure 4, it is seen that the proposed algorithm clearly outperforms the state-of-the-art algorithms [10, 32, 36] in terms of continual learning performance. Note that the proposed method does not use the local motion reconstruction error for the UCSD dataset since pose estimation does not work well with low quality videos.

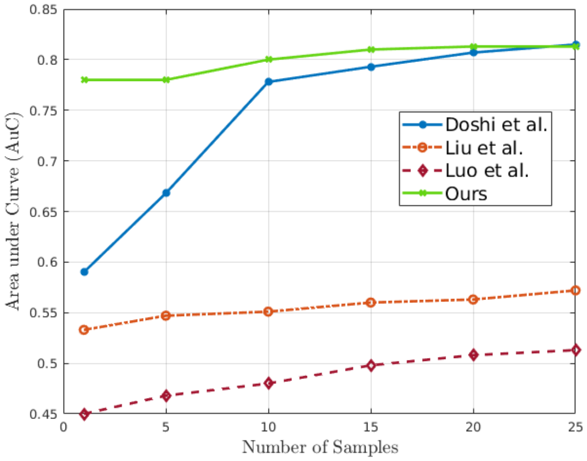


Figure 4. Comparison of the proposed and the state-of-the-art algorithms Liu et al. [32], Luo et al. [36] and Doshi et al. [9] in terms of continual learning capability. Different than few-shot learning, here the new data (bike frames) are used to incrementally update the algorithms after the initial training on the training data. While training with new samples, the proposed algorithm maintains superior performance compared to the state-of-the-art methods.

5. Conclusion

For video anomaly detection, we present a multi-task framework, which consists of cross-domain adaptability, few-shot learning, and continual learning. A modular method which consists of an interpretable transfer learning based feature extractor, and a novel anomaly detector using semantic embedding and deep metric learning was proposed. The proposed method first detects anomalous events in an online manner, and then deals with localizing the anomalous video frames, following the necessity for timely detection in realistic settings. Since online detection of anomalous events is widely ignored in the video anomaly detection literature, a new performance metric for comparing algorithms in terms of online detection was developed. Through extensive testing on the benchmark datasets, we show that the proposed approach significantly outperforms the state-of-the-art methods in cross-domain adaptability, few-shot learning, and continual learning.

References

[1] Visual system. https://en.wikipedia.org/wiki/Visual_system. 3

[2] Borislav Antić and Björn Ommer. Video parsing for abnormality detection. In *2011 International Conference on Computer Vision*, pages 2415–2422. IEEE, 2011. 1

[3] 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

[4] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. 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*, pages 2909–2917, 2015. 1

[5] 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. 2

[6] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011. 2

[7] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision*, pages 334–349. Springer, 2016. 7

[8] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020. 1, 2

[9] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 934–935, 2020. 1, 7, 8

[10] 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. 1, 2, 3, 5, 8

[11] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021. 1, 2

[12] Keval Doshi and Yasin Yilmaz. A modular and unified framework for detecting and localizing video anomalies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3982–3991, 2022. 6

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 5

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 7

[15] 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. 2

[16] 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. 2

[17] 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. 1, 2, 7

[18] 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. 2, 7

[19] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015. 6

[20] 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. 2

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[22] 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. 2, 6

[23] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using narrowed normality clusters. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1951–1960. IEEE, 2019. 1

[24] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009. 7

[25] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

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

[27] 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. 2

[28] Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 38–44. IEEE, 2015. 6

[29] 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. 2

[30] 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. 1, 2

[31] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 586–587, 2020. 4

[32] 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. 1, 2, 3, 6, 7, 8

[33] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*, 2017. 2

[34] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. *arXiv preprint arXiv:2007.07843*, 2020. 1, 3, 6, 7

[35] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017. 6, 7

[36] 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. 2, 7, 8

[37] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 7

[38] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020. 6

[39] 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. 2

[40] 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. 1, 2, 6, 7, 8

[41] 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. 2

[42] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12173–12182, 2020. 1, 6

[43] 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. 1, 2, 6, 7

[44] 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. 1, 6, 7

[45] 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. 1, 2

[46] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018. 7

[47] 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. 4

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 4

[49] 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. 2, 6, 7

[50] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2

[51] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201, 2017. 7

[52] 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. 2

[53] 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. 2

[54] 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. 2

[55] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550, 2019. 1