University of South Florida

Digital Commons @ University of South Florida

USF Tampa Graduate Theses and Dissertations

USF Graduate Theses and Dissertations

June 2022

Video Anomaly Detection: Practical Challenges for Learning Algorithms

Keval Doshi
University of South Florida

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Computer Sciences Commons, and the Statistics and Probability Commons

Scholar Commons Citation

Doshi, Keval, "Video Anomaly Detection: Practical Challenges for Learning Algorithms" (2022). *USF Tampa Graduate Theses and Dissertations.*
https://digitalcommons.usf.edu/etd/9346

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact scholarcommons@usf.edu.

Video Anomaly Detection: Practical Challenges for Learning Algorithms

by

Keval Doshi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Yasin Yilmaz, Ph.D.
Kwang-Cheng Chen, Ph.D.
Ismail Uysal, Ph.D.
Sudeep Sarkar, Ph.D.
Michael Jones, Ph.D.

Date of Approval:
May 13, 2022

Keywords: Video Surveillance, Video Understanding, Continual Learning, Deep Learning, Interpretability

Copyright © 2022, Keval Doshi

Dedication

To my dear parents, who have supported and guided me throughout my education. Thank you for always encouraging me and giving valuable advice whenever I needed it the most.

Acknowledgments

First and foremost, I would like to sincerely thank my advisor Dr. Yasin Yilmaz, whose vast knowledge and wealth of experience have served as a constant source of inspiration during my academic journey. I especially appreciated his high degree of professionalism, humbleness and the effort he put in while carefully reviewing my papers. He provided me with an opportunity when I approached him as a masters student with nothing to offer other than enthusiasm, for which I will be eternally grateful. This dissertation would not have been possible without him.

I would like to offer my special thanks to the rest of my committee members - Dr. Kwang-Cheng Chen, Dr. Ismayil Uysal, Dr. Michael Jones, Dr. Sudeep Sarkar and the chairperson of my Ph.D. dissertation defense Dr. John Murray-Bruce, for all of their support and invaluable feedback.

I would also like to thank my lab mates Ammar Haydari, Dr. Almuthanna Nassar, Salman Shuvo and Shatha Abudalou at the Secure and Intelligent Systems Lab for our many research discussions and fruitful collaborations. I must thank my close friends Aman Chawda, Anmol Khare and Brij Naik for their companionship and encouragement through the years.

Last but not the least, I would like to thank my parents Chetan and Minal Doshi, and my brother Premal. Without their unfaltering support and guidance during the entire journey, it would have been impossible to complete this dissertation.

Table of Contents

List of Tables . . . iii

List of Figures . . . iv

Abstract . . . vi

Chapter 1: Online Anomaly Detection in Surveillance Videos with Asymptotic Bounds on False Alarm Rate . . . 1
- 1.1 Introduction . . . 1
- 1.2 Related Works . . . 3
- 1.3 Proposed Method . . . 5
  - 1.3.1 Motivation . . . 5
  - 1.3.2 Feature Selection . . . 6
  - 1.3.3 Frame Prediction . . . 6
  - 1.3.4 Object Detection . . . 8
  - 1.3.5 Feature Vector . . . 9
  - 1.3.6 Anomaly Detection . . . 10
  - 1.3.7 Threshold Selection . . . 14
- 1.4 Experiments . . . 15
  - 1.4.1 Datasets . . . 15
  - 1.4.2 Comparison with Existing Methods . . . 16
  - 1.4.3 Implementation Details . . . 16
  - 1.4.4 Impact of Sequential Anomaly Detection . . . 17
  - 1.4.5 Results . . . 17
  - 1.4.6 Computational Complexity . . . 19
- 1.5 Conclusion . . . 20

Chapter 2: A Modular and Unified Framework for Detecting and Localizing Video Anomalies . . . 21
- 2.1 Introduction . . . 21
- 2.2 Related Works . . . 23
- 2.3 Proposed Method . . . 24
  - 2.3.1 Motivation . . . 24
  - 2.3.2 Transfer Learning-Based Feature Extraction . . . 26
  - 2.3.3 Statistical Anomaly Detection . . . 27
- 2.4 Experiments . . . 32
  - 2.4.1 Datasets . . . 32
  - 2.4.2 Results . . . 33

2.4.3 Ablation Study 38
2.5 Conclusion and Discussions 40

Chapter 3: Rethinking Video Anomaly Detection 41
3.1 Introduction 41
3.2 Related Work 45
3.3 Continual Video Anomaly Detection 46
3.3.1 Problem Formulation 46
3.3.2 Continual VAD Algorithm 50
3.4 Dataset for Continual VAD 53
3.4.1 Existing Datasets 53
3.4.2 New Dataset: NOLA 54
3.5 Experiments 55
3.5.1 Discussion 58
3.6 Conclusion 58

Chapter 4: Multi-Task Learning for Video Surveillance with Limited Data 59
4.1 Introduction 59
4.2 Related Work 61
4.3 Multi-Task Video Surveillance 62
4.3.1 Problem Setup 62
4.3.2 Proposed Approach 65
4.3.3 Deep Learning-Based Feature Extraction 65
4.3.4 Anomaly Detection 70
4.4 Experiments 71
4.4.1 Datasets 71
4.4.2 Results 72
4.5 Conclusion 77

References 79

Appendix A: Proof of Theorem 1 92

Appendix B: Copyright Permissions 95

List of Tables

Table 1.1 AuROC result comparison on three datasets. 19

Table 2.1 Online detection comparison in terms of the proposed APD metric on the ShanghaiTech dataset. 34

Table 2.2 Offline anomaly localization comparison in terms of frame-level AUC on three datasets. 36

Table 2.3 Few-shot scene adaptation comparison of the proposed and the state-of-the-art [55] algorithms in terms of frame-level AUC. 37

Table 2.4 Performance of each module in terms of the frame-level AUC on the ShanghaiTech dataset. 38

Table 3.1 Comparison of existing and proposed VAD datasets. 53

Table 3.2 Performance of the proposed detector and recent state-of-the-art approaches across different continual learning splits in terms of the proposed APD metric. 55

Table 4.1 Online detection comparison in terms of the proposed APD metric on the ShanghaiTech dataset. 73

Table 4.2 Offline anomaly localization comparison in terms of frame-level AUC on three datasets. 74

Table 4.3 Performance of the proposed detector in terms of frame-level AUC for cross-domain adaptability on different cameras from the ShanghaiTech Dataset. 75

Table 4.4 Performance of the proposed detector in terms of event-level APD for cross-domain adaptability on different cameras from the ShanghaiTech Dataset. 75

Table 4.5 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. 75

List of Figures

Figure 1.1 Proposed MONAD framework. 5

Figure 1.2 Example video frames from the UCSD Ped2 dataset showing the extraction of bounding box center (location) feature in nominal training data (top row) and test data (bottom row). 8

Figure 1.3 The ROC curves of the proposed MONAD algorithm and the online version of Liu et al. [52] for a practical range of false alarm rate in the UCSD Ped 2 (left) and ShanghaiTech (right) data sets. 11

Figure 1.4 Actual false alarm periods vs. derived lower bounds for the UCSD Ped.2 (top left), ShanghaiTech (top right), and Avenue (bottom) data sets. 12

Figure 1.5 The advantage of sequential anomaly detection over single-shot detection in terms of controlling false alarms. 18

Figure 2.1 Proposed MOVAD framework. 25

Figure 2.2 Comparison of the proposed and the state-of-the-art algorithms Liu et al. [52] and Morais et al. [65] in terms of online detection capability. 34

Figure 2.3 Threshold selected according to Eq. (2.2) satisfies the desired lower bound on false alarm period (i.e., upper bound on false alarm rate) even in the non-asymptotic regime with the finite sample size of the CUHK Avenue dataset. 35

Figure 2.4 The proposed model is able to interpret the cause of the anomaly correctly. 38

Figure 2.5 The advantage of sequential anomaly detection over a single-shot detector. 39

Figure 3.1 The proposed continual learning framework. 42

Figure 3.2 Simple optical flow method performs close to the state-of-the-art methods [55, 71, 52, 17, 65, 96] on the three popular benchmark datasets in terms of the frame-level AUC metric. 43

Figure 3.3 Definitions of true alarm and false alarm. 48

Figure 3.4 Proposed VAD algorithm. 49

Figure 3.5 Comparison of the proposed and state-of-the-art algorithms Liu et al. [52] and Park et al. [70] in terms of learning from few samples on the ShanghaiTech (top) and UCSD (bottom) datasets. 56

Figure 4.1 Proposed video anomaly detection framework. 64

Figure 4.2 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. 68

Figure 4.3 Comparison of the proposed and state-of-the-art algorithms Liu et al. [52], Luo et al. [57] and Doshi et al. [16] in terms of few-shot learning. 76

Figure 4.4 Comparison of the proposed and the state-of-the-art algorithms Liu et al. [52], Luo et al. [57] and Doshi et al. [16] in terms of continual learning capability. 77

Abstract

Anomaly detection in surveillance videos is attracting an increasing amount of attention. Despite the competitive performance of several existing methods, they lack theoretical performance analysis, particularly due to the complex deep neural network architectures used in decision making. Additionally, real-time decision making is an important but mostly neglected factor in this domain. Much of the existing methods that claim to be online, depend on batch or offline processing in practice. Furthermore, several critical tasks such as continual learning, model interpretability and cross-domain adaptability are completely neglected in existing works. Motivated by these research gaps, in this dissertation we discuss our work on real-time video anomaly detection, specifically addressing challenges encountered in a practical implementation. We begin by proposing a multi-objective deep learning module along with a statistical anomaly detection module, and demonstrate its effectiveness on several publicly available data sets. Furthermore, 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. Finally, 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. 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. Particularly, we study these three problems (few-shot learning, continual learning, cross-domain adaptability) in a multi-task learning setting.

Chapter 1: Online Anomaly Detection in Surveillance Videos with Asymptotic Bounds on False Alarm Rate

1.1 Introduction

[1]The rapid advancements in the technology of closed-circuit television (CCTV) cameras and their underlying infrastructural components such as network, storage, and processing hardware have led to a sheer number of surveillance cameras implemented all over the world, and estimated to go beyond 1 billion globally, by the end of the year 2021 [49]. Video surveillance is an essential tool used in law enforcement, transportation, environmental monitoring, etc. mainly for improving security and public safety. For example, it has become an inseparable part of crime deterrence and investigation, traffic violation detection, and traffic management. However, considering the massive amounts of videos generated in real-time, manual video analysis by human operator becomes inefficient, expensive, and nearly impossible, which in turn makes a great demand for automated and intelligent methods for analyzing and retrieving important information from videos, in order to maximize the benefits of CCTV.

One of the most important, challenging and time-critical tasks in automated video surveillance is the detection of abnormal events such as traffic accidents and violations, crimes, and natural disasters. Hence, video anomaly detection has become an important research problem in the recent years. Anomaly detection in general is a vast, crucial, and challenging research topic, which deals with the identification of data instances deviating from nominal

[1] Portions of this chapter were published in Elsevier Pattern Recognition [18]. Copyright permissions from the publishers are included in Appendix B.

patterns. It has a wide range of applications, e.g., in medical health care[98], cyber-security [94], hardware security [21], aviation [63], and spacecraft monitoring [34].

Given the important role that video anomaly detection can play in ensuring safety, security and sometimes prevention of potential catastrophes, one of the main outcomes of a video anomaly detection system is the real-time decision making capability. Events such as traffic accidents, robbery, and fire in remote places require immediate counteractions to be taken in a timely manner, which can be facilitated by the real-time detection of anomalous events. Despite its importance, a very limited body of research has focused on online and real-time detection methods. Moreover, some of the methods that claim to be online heavily depend on batch processing of long video segments. For example, [52] performs a normalization step which requires the entire video.

A vast majority of the recent state-of-the-art video anomaly detection methods depend on complex neural network architectures [88]. Although deep neural networks provide superior performance on various machine learning and computer vision tasks, such as object detection [12], image classification [43], playing games [85], image synthesis[77], etc., where sufficiently large and inclusive data sets are available to train on, there is also a significant debate on their shortcomings in terms of interpretability, analyzability, and reliability of their decisions [38]. For example, [69, 86] propose using a nearest neighbor-based approach together with deep neural network structures to achieve robustness, interpretability for the decisions made by the model, and as defense against adversarial attack. Additionally, to the best of the our knowledge, none of the neural network-based video anomaly detection methods has been analyzed in terms of performance guarantees. On the other hand, statistical and nearest neighbor-based methods remain popular due to their appealing characteristics such as being amenable to performance analysis, computational efficiency, and robustness [6, 28].

Motivated by the aforementioned domain challenges and research gaps, we propose a hybrid use of neural networks and statistical k nearest neighbor (kNN) decision approach for finding anomalies in video in an online fashion. In summary, our contributions in this paper are as follows:

- We propose a novel framework composed of deep learning-based feature extraction from video frames, and a statistical sequential anomaly detection algorithm.
- We derive an asymptotic bound on the false alarm rate of our detection algorithm, and propose a technique for selecting a proper threshold which satisfies the desired false alarm rate.
- We extensively evaluate our proposed framework on publicly available video anomaly detection data sets.

The remainder of the paper is organized as: Related Work (Section 1.2), Proposed Method (Section 1.3), Experiments (Section 1.4), and Conclusion (Section 1.5).

1.2 Related Works

Semi-supervised detection of anomalies in videos, also known as outlier detection, is a commonly adopted learning technique due to the inherent limitations in availability of annotated and anomalous instances. This category of learning methods deals with learning a notion of normality from nominal training videos, and attempts to detect deviations from the learned normality notion. [7, 36]. There are also several supervised detection methods, which train on both nominal and anomalous videos. The main drawback of such methods is the difficulty in finding frame-level labeled, representative, and inclusive anomaly instances. To this end, [88] proposes using a deep multiple instance learning (MIL) approach to train on video-level annotated videos, in a weakly supervised manner. Although training on anomalous videos would enhance the detection capability on similar anomaly events, supervised methods typically suffer from unknown and novel anomaly types.

One of the key components of the video anomaly detection algorithms is the extraction of meaningful features, which can capture the difference between the nominal and anomalous events within the video. The selection of feature types has a significant impact on the identifiability of types of anomalous events in the video sequences. Many early video anomaly detection techniques and some recent ones focused on the trajectory features [2], which limits their applicability to the detection of the anomalies related to the trajectory patterns, and moving objects. For instance, [24] studied detection of abnormal vehicle trajectories such as illegal U-turn. [65] extracts human skeleton trajectory patterns, and hence is limited to only the detection of abnormalities in human behavior.

Motion and appearance features are another class of widely used features in this domain. [83] extracts motion direction and magnitudes, to detect spatio-temporal anomalies. Histogram of optical flow [5, 10], and histogram of oriented gradients [13] are some other commonly used hand-crafted feature extraction techniques used in the literature. Sparse coding based methods [99] are also applied in detection of video anomalies. They learn a dictionary of normal sparse events, and attempt to detect anomalies based on the reconstructability of video from the dictionary atoms. [64] uses sparse reconstruction to learn joint trajectory representations of multiple objects.

In contrary to the hand-crafted feature extraction, are the neural network based feature learning methods. [95] learns the appearance and motion features by deep neural networks. [56] utilizes Convolutional Neural Networks (CNN), and Convolutional Long Short Term Memory (CLSTM) to learn appearance and motion features, respectively. Neural network based approaches have been recently dominating the literature. For example, [74] trains Generative Adversarial Network (GAN) on normal video frames, to generate internal scene representations (appearance and motion), based on a given frame and its optical flow, and detects deviation of the GAN output from the normal data, by AlexNet [43]. [82] trains a GAN-like adversarial network, in which a reconstruction component learns to reconstruct the normal test frames, and attempts to train a discriminator by gradually injecting anomalies to it, while concurrently the discriminator (detector) learns to detect the anomalies injected by the reconstructor. In [17, 16], a transfer learning based approach is used for continual learning for anomaly detection in surveillance videos from a few samples.

1.3 Proposed Method

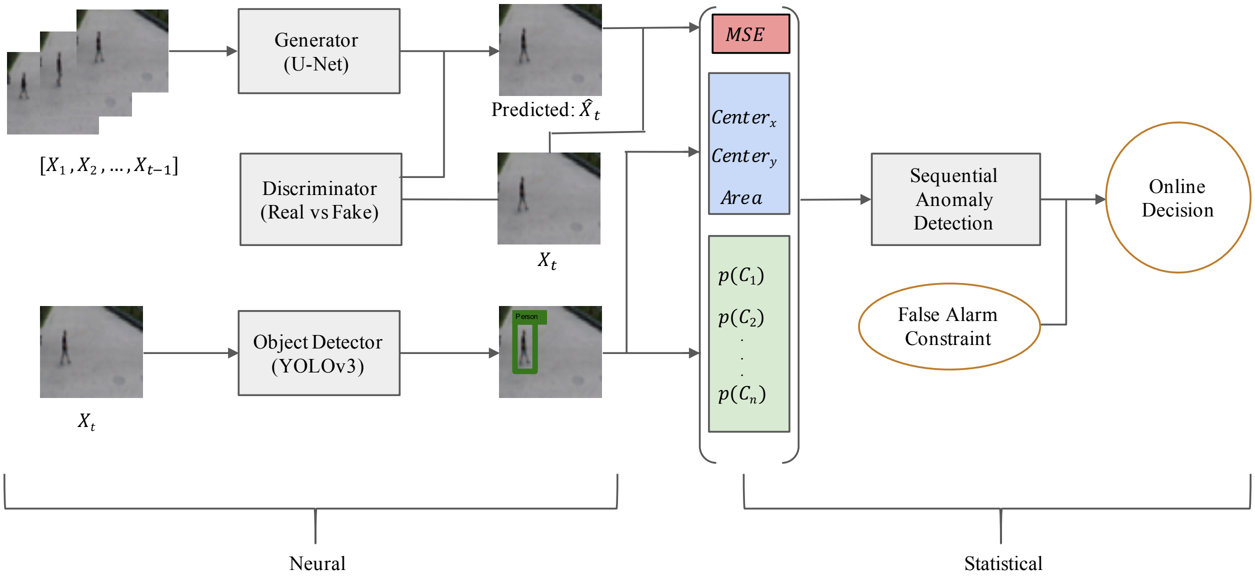


Figure 1.1: Proposed MONAD framework. At each time t, neural network-based feature extraction module provides motion (MSE), location (center coordinates and area of bounding box), and appearance (class probabilities) features to the statistical anomaly detection module, which automatically sets its decision threshold to satisfy a false alarm constraint and makes online decisions.

1.3.1 Motivation

Anomaly detection in surveillance videos is defined as the identification of unusual events which do not conform to the expectation. We base our study on two important requirements that a successful video anomaly detector should satisfy: (i) extract meaningful features which can be utilized to distinguish nominal and anomalous data; and (ii) provide a decision making strategy which can be easily tuned to satisfy a given false alarm rate. While existing works partially fulfills the first requirement by defining various constraints on spatial and temporal video features, they typically neglect providing an analytical and amenable decision strategy. Motivated by this shortcoming, we propose a unified framework called Multi-Objective Neural Anomaly Detector (MONAD[2]). Like *monads* provide a unified functional model for programming, our proposed MONAD unifies deep learning-based feature extraction and analytical anomaly detection by incorporating two modules, as shown in Fig. 1.1. The first module consists of a Generative Adversarial Network (GAN) based future frame predictor and a lightweight object detector (YOLOv3) to extract meaningful features. The second module consists of a nonparametric statistical algorithm which uses the extracted features for online anomaly detection. To the best of our knowledge, this is the first work to present theoretical performance analysis for a deep learning-based video anomaly detection method. Our MONAD framework is described in detail in the following sections.

1.3.2 Feature Selection

Most existing works focus on a certain aspect of the video such as optical flow, gradient loss or intensity loss. This in turn restrains the existing algorithms to a certain form of anomalous event which is manifested in the considered video aspect. However, in general, the type of anomaly is broad and unknown while training the algorithm. For example, an anomalous event can be justified on the basis of appearance (a person carrying a gun), motion (two people fighting) or location (a person walking on the roadway). To account for all such cases, we create a feature vector F_t^i for each object i in frame X_t at time t, where F_t^i is given by $[w_1 F_{motion}, w_2 F_{location}, w_3 F_{appearance}]$. The weights w_1, w_2, w_3 are used to adjust the relative importance of each feature category.

1.3.3 Frame Prediction

A heuristic approach for detecting anomalies in videos is by predicting the future video frame $\widehat{X}_t$ using previous video frames $\{X_1, X_2, \ldots, X_{t-1}\}$, and then comparing it to X_t through mean squared error (MSE). Instead of deciding directly on MSE, we use MSE of video frame

[2] *Monad* is a philosophical term for infinitesimal unit, and also a functional programming term for minimal structure.

prediction to obtain motion features (Section 1.3.5). GANs are known to be successful in generating realistic images and videos. However, regular GANs might face the vanishing gradient problem during learning as they hypothesize the discriminator as a classifier with the sigmoid cross entropy loss function. To overcome this problem, we use a modified version of GAN called Least Square GAN (LS-GAN) [61]. The GAN architecture comprises of a generator network G and a discriminator network D, where the function of G is to generate frames that would be difficult-to-classify by D. Ideally, once G is well trained, D cannot predict better than chance. Similar to [52], we employ a U-Net [80] based network for G and a patch discriminator for D.

For training the generator G, we follow [52], and combine the constraints on intensity, gradient difference, optical flow, and adversarial training to get the following objective function

$$\begin{aligned} L_G = \gamma_{int} L_{int}(\widehat{X}, X) + \gamma_{gd} L_{gd}(\widehat{X}, X) + \\ \gamma_{of} L_{of}(\widehat{X}, X) + \gamma_{adv} L_{adv}(\widehat{X}, X) \end{aligned} \tag{1.1}$$

where $\gamma_{int}, \gamma_{gd}, \gamma_{of}, \gamma_{adv} \geq 0$ are the corresponding weights for the losses.

- Intensity loss is the l_1 or l_2 distance between the predicted frame $\widehat{X}$ and the actual frame X, which is used to maintain similarity between pixels in the RGB space, and given by

$$L_{int}(\widehat{X}, X) = \left\| \widehat{X} - X \right\|^2 . \tag{1.2}$$

- Gradient difference loss is used to sharpen the image prediction and is given by

$$\begin{aligned} L_{gd}(\widehat{X}, X) = \sum_{i,j} \left\| |\widehat{X}_{i,j} - \widehat{X}_{i-1,j}| - |X_{i,j} - X_{i-1,j}| \right\|_1 \\ + \left\| |\widehat{X}_{i,j} - \widehat{X}_{i,j-1}| - |X_{i,j} - X_{i,j-1}| \right\|_1 \end{aligned} \tag{1.3}$$

where (i, j) denotes the spatial index of a video frame.

- Optical flow loss is used to improve the coherence of motion in the predicted frame, and is given by

$$L_{of}(\widehat{X}_{t+1}, X_{t+1}, X_t) = \left\| f(\widehat{X}_{t+1}, X_t) - f(X_{t+1}, X_t) \right\|_1 \tag{1.4}$$

  where f is a pretrained CNN-based function called Flownet, and is used to estimate the optical flow.

- Adversarial generator loss is minimized to confuse D as much as possible such that it cannot discriminate the generated predictions, and is given by

$$L_{adv}(\widehat{X}) = \sum_{i,j} \frac{1}{2} L_{MSE}(D(\widehat{X}_{i,j}), 1) \tag{1.5}$$

  where $D(\widehat{X}_{i,j}) = 1$ denotes "real" decision by D for patch (i,j), $D(\widehat{X}_{i,j}) = 0$ denotes "fake" decision, and L_{MSE} is the mean squared error function.

1.3.4 Object Detection

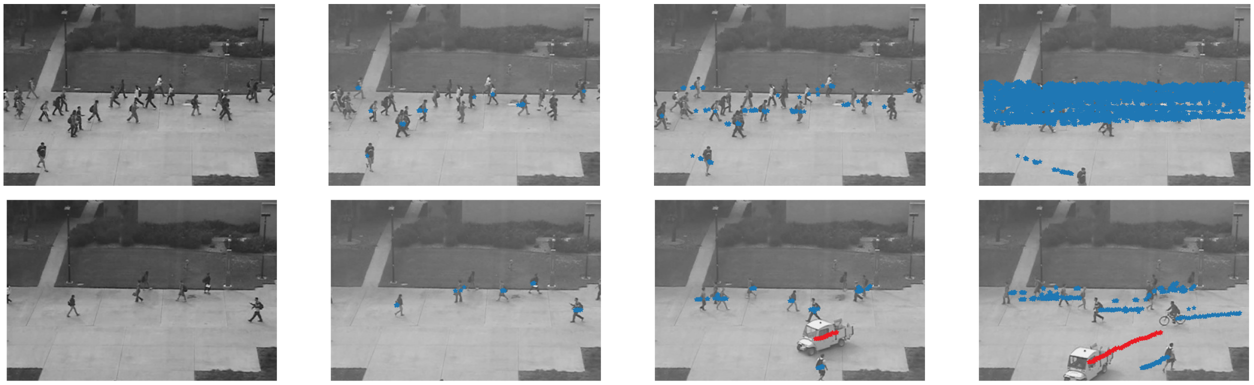


Figure 1.2: Example video frames from the UCSD Ped2 dataset showing the extraction of bounding box center (location) feature in nominal training data (top row) and test data (bottom row). Columns from left to right correspond to the first, 30th, 150th, and the last frame in all training videos (top row), and in a test video (bottom row). In the test video, the unusual path of golf cart, shown with red dots, together with the class probability and high prediction error (MSE) due to unusual speed of cart statistically contribute to the anomaly decision. Best viewed in color.

We propose to detect objects using a real-time object detection system such as You Only Look Once (YOLO) [76] to obtain location and appearance features (Section 1.3.5). The advantage of YOLO is that it is capable of processing higher frames per second on a GPU while providing the same or even better accuracy as compared to the other state-of-the-art models such as SSD and ResNet. Speed is a critical factor for online anomaly detection, so we currently prefer YOLOv3 in our implementations. For each detected object in image X_t, we get a bounding box (location) along with the class probabilities (appearance). As shown in Fig. 1.2, we monitor the center of the bounding boxes to track paths different objects might take in the training videos. Instead of simply using the entire bounding box, we monitor the center of the box and its area to obtain location features. This not only reduces the complexity, but also effectively avoids false positives in case the bounding box is not tight. In a testing video, objects diverging from the nominal paths and class probabilities will help us detect anomalies, as explained in Section 1.3.6.

1.3.5 Feature Vector

Finally, for each object i detected in a frame, we construct a feature vector as:

$$F_t^i = \begin{bmatrix} w_1 MSE(X_t, \widehat{X}_t) \\ w_2 Center_x \\ w_2 Center_y \\ w_2 Area \\ w_3 p(C_1) \\ w_3 p(C_2) \\ \vdots \\ w_3 p(C_n) \end{bmatrix}, \tag{1.6}$$

where $MSE(X_t, \widehat{X}_t)$ is the prediction error from the GAN-based frame predictor (Section 1.3.3); $Center_x, Center_y, Area$ denote the coordinates of the center of the bounding box and the area of the bounding box (Section 1.3.4); and $p(C_1), \dots, p(C_n)$ are the class probabilities for the detected object (Section 1.3.4). Hence, at any given time t, with n denoting the number of possible classes, the dimensionality of F_t^i is given by $m = n + 4$.

1.3.6 Anomaly Detection

Our goal here is to detect anomalies in streaming videos with minimal detection delays while satisfying a desired false alarm rate. We can safely hypothesize that any anomalous event would persist for an unknown period of time. This makes the problem suitable for a sequential anomaly detection framework [4]. However, since we have no prior knowledge about the anomalous event that might occur in a video, parametric algorithms which require probabilistic model and data for both nominal and anomaly cannot be used directly. Next, we explain the training and testing of our proposed nonparametric sequential anomaly detection algorithm.

The training procedure is given as follows. First, given a set of N training videos $\mathcal{V} \triangleq \{v_i : i = 1, 2, \dots, N\}$ consisting of P frames in total, we leverage the deep learning module of our proposed detector to extract M feature vectors $\mathcal{F}^M = \{F^i\}$ for M detected objects in total such that $M \geq P$. We assume that the training data does not include any anomalies. These M vectors correspond to M points in the nominal data space, distributed according to an unknown complex probability distribution. Following a data-driven approach we would like to learn a nonparametric description of the nominal data distribution. Due to its attractive traits, such as analyzability, interpretability, and computational efficiency [6, 28], we use k nearest neighbor (kNN) distance, which captures the local interactions between nominal data points, to figure out a nominal data pattern. Given the informativeness of extracted motion, location, and appearance features, anomalous instances are expected to lie further away from the nominal manifold defined by $\mathcal{F}^M$. Consequently, the kNN distance

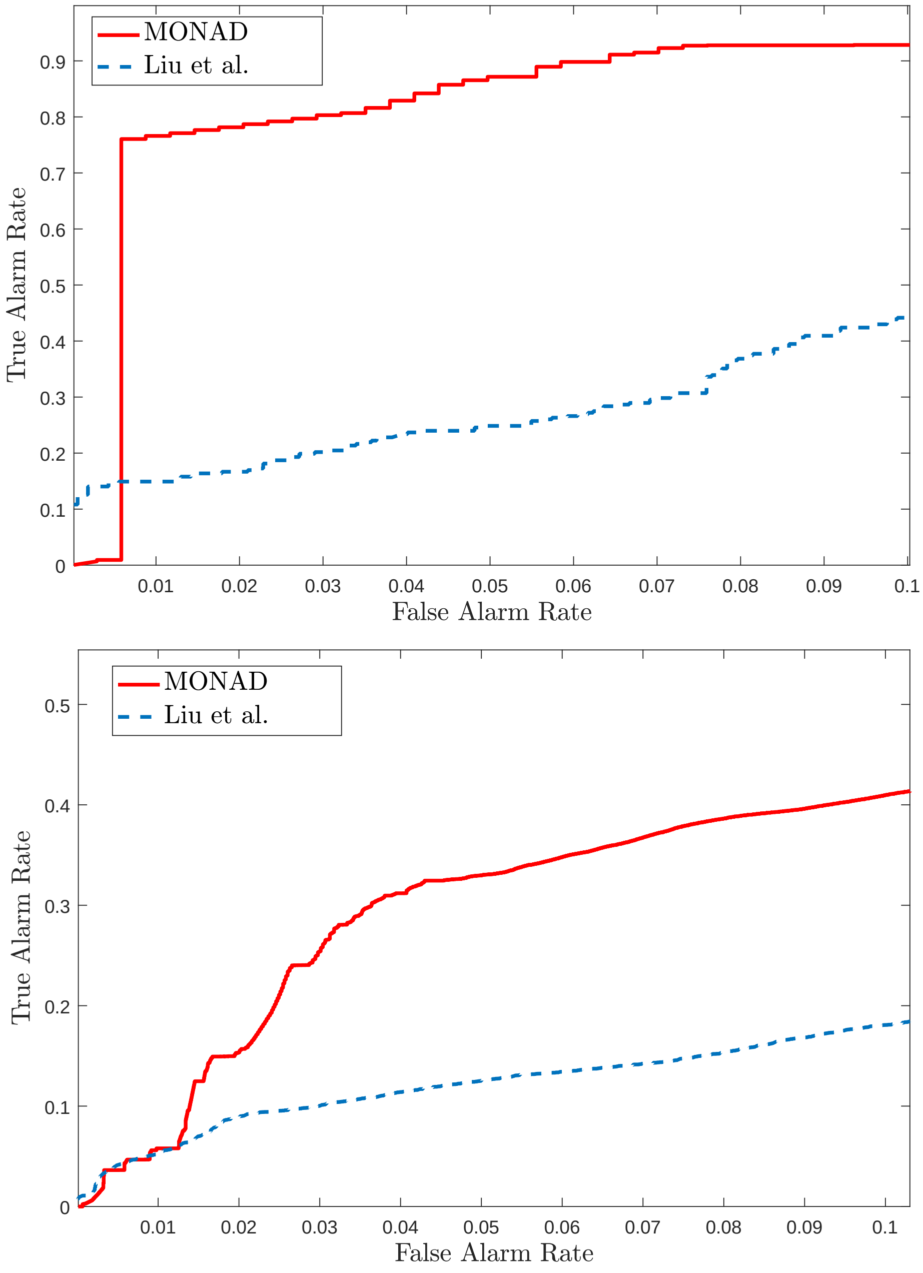


Figure 1.3: The ROC curves of the proposed MONAD algorithm and the online version of Liu et al. [52] for a practical range of false alarm rate in the UCSD Ped 2 (left) and ShanghaiTech (right) data sets.

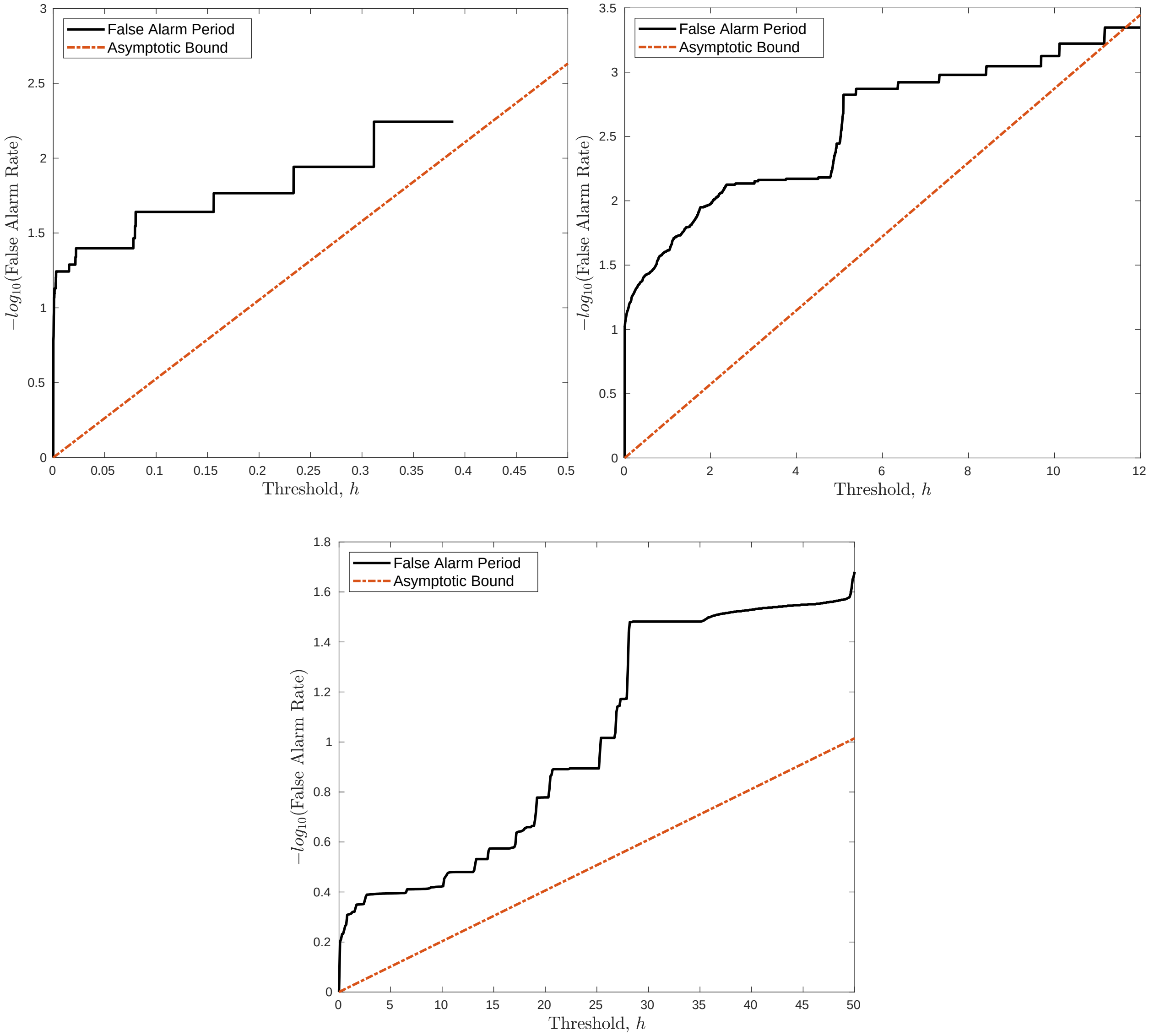


Figure 1.4: Actual false alarm periods vs. derived lower bounds for the UCSD Ped.2 (top left), ShanghaiTech (top right), and Avenue (bottom) data sets.

of anomalous instances with respect to the nominal data points in $\mathcal{F}^M$ will be statistically higher as compared to the nominal data points. The training procedure of our detector is given as follows:

1. Randomly partition the nominal dataset $\mathcal{F}^M$ into two sets $\mathcal{F}^{M_1}$ and $\mathcal{F}^{M_2}$ such that $M = M_1 + M_2$.

2. Then for each point F_i in $\mathcal{F}^{M_1}$, we compute the kNN distance d_i with respect to the points in set $\mathcal{F}^{M_2}$.

3. For a significance level α, e.g., 0.05, the $(1 - \alpha)$th percentile d_α of kNN distances $\{d_1, \dots, d_{M_1}\}$ is used as a baseline statistic for computing the anomaly evidence of test instances.

4. The maximum value of kNN distances $\{d_1, \dots, d_{M_1}\}$ is used as an upper bound (ϕ) for δ_t, given by Eq. (1.7), which is then used for selecting a threshold h, as explained in Section 1.3.7.

During the testing phase, for each object i detected at time t, the deep learning module constructs the feature vector F_t^i and computes the kNN (Euclidean) distance d_t^i with respect to the training instances in $\mathcal{F}^{M_2}$. The proposed sequential anomaly detection system then computes the instantaneous frame-level anomaly evidence δ_t:

$$\delta_t = (\max_i \{d_t^i\})^m - d_\alpha^m, \tag{1.7}$$

where m is the dimensionality of feature vector F_t^i. Finally, following a CUSUM-like procedure [4] we update the running decision statistic s_t as

$$s_t = \max\{s_{t-1} + \delta_t, 0\}, s_0 = 0. \tag{1.8}$$

For nominal data, δ_t typically gets negative values, hence the decision statistic s_t hovers around zero; whereas for anomalous data δ_t is expected to take positive values, and successive positive values of δ_t will make s_t grow. We decide that a video frame is anomalous if the decision statistic s_t exceeds the threshold h. After s_t exceeds h, we perform some fine tuning to better label video frames as nominal or anomalous. Specifically, we find the frame s_t started to grow, i.e., the last time $s_t = 0$ before detection, say τ_{start}. Then, we also determine the frame s_t stops increasing and keeps decreasing for n, e.g., 5, consecutive frames, say τ_{end}. Finally, we label the frames between τ_{start} and τ_{end} as anomalous, and continue testing for new anomalies with frame $\tau_{end} + 1$ by resetting $s_{\tau_{end}} = 0$.

1.3.7 Threshold Selection

If the test statistic crosses the threshold when there is no anomaly, this event is called a false alarm. Existing works consider the decision threshold as a design parameter, and do not provide any analytical procedure for choosing its value. For an anomaly detection algorithm to be implemented in a practical setting, a clear procedure is necessary for selecting the decision threshold such that it satisfies a desired false alarm rate. The reliability of an algorithm in terms of false alarm rate is crucial for minimizing human involvement. To provide such a performance guarantee for the false alarm rate, we derive an asymptotic upper bound on the average false alarm rate of the proposed algorithm.

Theorem 1. *The false alarm rate of the proposed algorithm is asymptotically (as $M_2 \to \infty$) upper bounded by*

$$FAR \leq e^{-\omega_0 h}, \tag{1.9}$$

where h is the decision threshold, and $\omega_0 > 0$ is given by

$$\begin{aligned} \omega_0 &= v_m - \theta - \frac{1}{\phi}\mathcal{W}\left(-\phi\theta e^{-\phi\theta}\right), \\ \theta &= \frac{v_m}{e^{v_m d_\alpha^m}}. \end{aligned} \tag{1.10}$$

In (1.10)*,* $\mathcal{W}(\cdot)$ *is the Lambert-W function,* $v_m = \frac{\pi^{m/2}}{\Gamma(m/2+1)}$ *is the constant for the* m *dimensional Lebesgue measure (i.e.,* $v_m d_\alpha^m$ *is the* m *dimensional volume of the hyperball with radius* d_α*), and* ϕ *is the upper bound for* δ_t*.*

Proof. See Appendix.

Although the expression for ω_0 looks complicated, all the terms in (1.10) can be easily computed. Particularly, v_m is directly given by the dimensionality m, d_α comes from the training phase, ϕ is also found in training, and finally there is a built-in Lambert-W function in popular programming languages such as Python and Matlab. Hence, given the training data, ω_0 can be easily computed, and based on Theorem 1, the threshold h can be chosen to asymptotically achieve the desired false alarm period as follows

$$h = \frac{-\log(FAR)}{\omega_0}. \tag{1.11}$$

1.4 Experiments

1.4.1 Datasets

We evaluate our proposed method on three publicly available video anomaly data sets, namely the CUHK avenue dataset [54], the UCSD pedestrian dataset [59], and the ShanghaiTech [57] campus dataset. Each data set presents its own set of challenges and unique characteristics such as types of anomaly, video quality, background location, etc. Hence, we treat each dataset independently and present individual results for each of them. Here, we briefly introduce each dataset that are used in our experiments.

- UCSD: The UCSD pedestrian data set is composed of two parts, namely Ped1 and Ped2. Following the work of [36, 30], we exclude Ped1 from our experiments due to its significantly lower resolution of 158 x 238 and a lack of consistency in the reported results as some recent works reported their performance only on a subset of the entire data set. Hence, we present our results on the UCSD Ped2 dataset which consists of

16 training and 12 test videos, each with a resolution of 240 x 360. All the anomalous events are caused due to vehicles such as bicycles, skateboarders and wheelchairs crossing pedestrian areas.

- Avenue: The CUHK avenue dataset consists of 16 training and 21 test videos with a frame resolution of 360 x 640. The anomalous behaviour is represented by people throwing objects, loitering and running.
- ShanghaiTech: The ShanghaiTech Campus 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 x 856.

1.4.2 Comparison with Existing Methods

We compare our proposed algorithm in Table 1.1 with state-of-the-art deep learning-based methods, as well as methods based on hand-crafted features: MPPCA [39], MPPC + SFA [59], Del et al. [14], Conv-AE [29], ConvLSTM-AE [56], Growing Gas [89], Stacked RNN [57], Deep Generic [30], GANs [73], Liu et al. [52]. A popular metric used for comparison in anomaly detection literature is the Area under the Receiver Operating Characteristic (AuROC) curve. Higher AuROC values indicate better performance for an anomaly detection system. For performance evaluation, following the existing works [11, 36, 52], we consider frame level AuROC.

1.4.3 Implementation Details

In the prediction pipeline, the U-NET based generator and the patch discriminator are implemented in Tensorflow. Each frame is resized to 256 x 256 and normalized to [-1,1]. The window size t is set to 4. Similar to [52], we use the Adam optimizer for training and set the learning rate to 0.0001 and 0.00001 for the generator and discriminator, respectively. The object detector used is YOLOv3 which is based on the Darknet architecture and is pretrained on the MS-COCO dataset. During training, we extract the bounds which have a confidence level greater than 0.6, and for testing we consider confidence levels greater than or equal to 0.4. The weights w_1, w_2 and w_3 are set to 1, 0.4 and 0.9 respectively. The sequential anomaly detection algorithm is implemented in Python.

1.4.4 Impact of Sequential Anomaly Detection

To demonstrate the importance of sequential anomaly detection in videos, we implement a nonsequential version of our algorithm by applying a threshold to the instantaneous anomaly evidence δ_t, given in (1.7), which is similar to the approach employed by many recent works [52, 88, 36]. As Figure 1.5 shows, instantaneous anomaly evidence is more prone to false alarms than the sequential MONAD statistic since it only considers the noisy evidence available at the current time to decide. Whereas, the proposed sequential statistic handles noisy evidence by integrating recent evidence over time.

1.4.5 Results

We compare our results to a wide range of methods in Table 1.1. Recently, [36] showed significant gains over the rest of the methods. However, their methodology of computing the AuROC gives them an unfair advantage as they calculate the AuROC for each video in a dataset, and then average them as the AuROC of the dataset, as opposed to the other works which concatenate all the videos first and then determine the AuROC as the dataset's score.

As shown in Table 1.1 we are able to outperform the existing results in the avenue and UCSD dataset, and achieve competitive performance in the ShanghaiTech dataset. We should note here that our reported result in the ShanghaiTech dataset is based on online decision making without seeing future video frames. A common technique used by several recent works [52, 36] is to normalize the computed statistic for each test video independently, including the ShanghaiTech dataset. However, this methodology cannot be implemented in

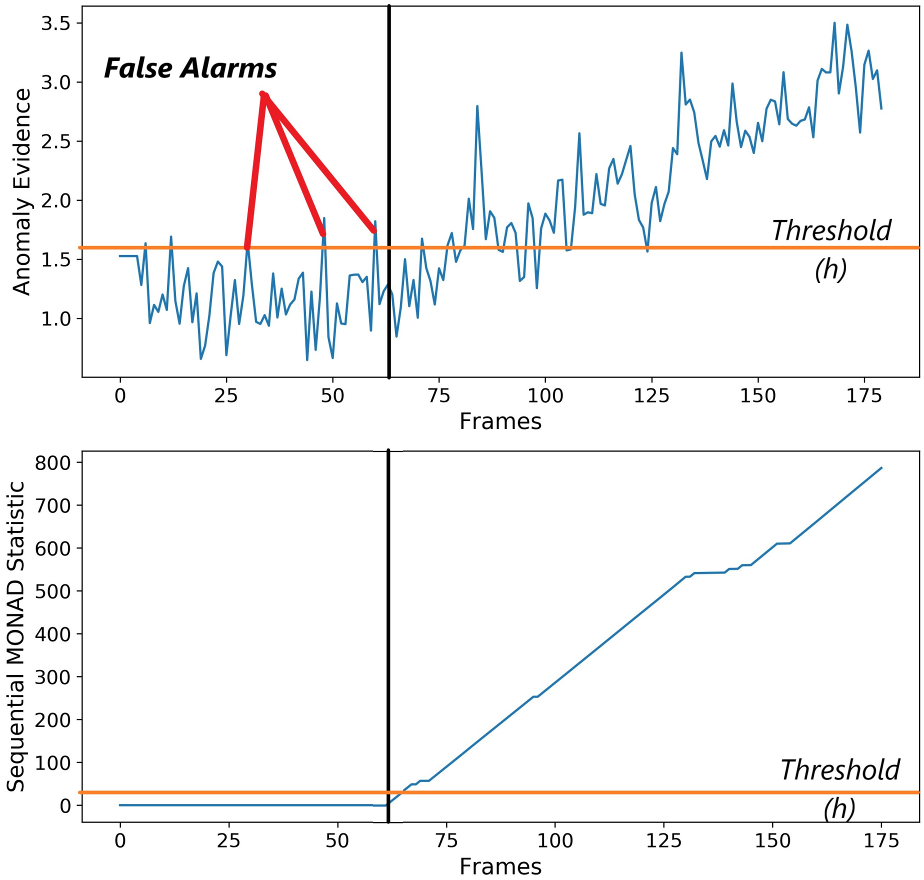


Figure 1.5: The advantage of sequential anomaly detection over single-shot detection in terms of controlling false alarms.

an online (real-time) system as it requires prior knowledge about the minimum and maximum values the statistic might take.

Hence, we also compare our online method with the online version of state-of-the-art method [52]. In that version, the minimum and maximum values of decision statistic is obtained from the training data and used for all videos in the test data to normalize the decision statistic, instead of the minimum and maximum values in each test video separately. AuROC value, which is the most common performance metric in the literature, considers the entire range $(0, 1)$ of false alarm rates. However, in practice, false alarm rate must satisfy an acceptable level (e.g., up to 10%). In Figure 1.3, on the UCSD and ShanghaiTech data sets, we compare our algorithm with the online version of [52] within a practical range of false alarm in terms of the ROC curve (true positive rate vs. false positive rate). As clearly

Table 1.1: AuROC result comparison on three datasets.

Methodology	CUHK Avenue	UCSD Ped 2	ShanghaiTech
MPPCA [39]	-	69.3	-
MPPC + SFA [59]	-	61.3	-
Del et al. [14]	78.3	-	-
Conv-AE [29]	80.0	85.0	60.9
ConvLSTM-AE [56]	77.0	88.1	-
Growing Gas [89]	-	93.5	-
Stacked RNN [57]	81.7	92.2	68.0
Deep Generic [30]	-	92.2	-
GANs [74]	-	88.4	-
Liu et al. [52]	85.1	95.4	**72.8**
Ours	**86.4**	**97.2**	70.9

seen in the figures, the proposed MONAD algorithm achieves much higher true alarm rates than [52] in both datasets while satisfying practical false alarm rates.

Finally, in Figure 1.4, we analyze the bound for false alarm rate derived in Theorem 1. For the clarity of visualization, the figure shows the logarithm of false alarm period, which is the inverse of the false alarm rate. In this case, the upper bound on false alarm rate becomes a lower bound on the false alarm period. The experimental results corroborate the theoretical bound and the procedure presented in Section 1.3.7 for obtaining the decision threshold h.

1.4.6 Computational Complexity

In this section we analyze the computational complexity of the sequential anomaly detection module, as well as the average running time of the deep learning module.

- Sequential Anomaly Detection: The training phase of the proposed anomaly detection algorithm requires computation of kNN distances for each point in $\mathcal{F}^{M_1}$ to each point in $\mathcal{F}^{M_2}$. Therefore, the time complexity of training phase is given by $\mathcal{O}(M_1 M_2 m)$. The space complexity of the training phase is $\mathcal{O}(M_2 m)$ since M_2 data instances need to be saved for the testing phase. In the testing phase, since we compute the kNN distances of a single point to all data points in $\mathcal{F}^{M_2}$, the time complexity is $\mathcal{O}(M_2 m)$.

- Deep Learning Module: The average running time for the GAN-based video frame prediction is 22 frames per second. The YOLO object detector requires about 12 milliseconds to process a single image. This translates to about 83.33 frames per second. The running time can be further improved by using a faster object detector such as YOLOv3-Tiny or a better GPU system. All tests are performed on NVIDIA GeForce RTX 2070 with 8 GB RAM and Intel i7-8700k CPU.

1.5 Conclusion

For video anomaly detection, we presented an online algorithm, called MONAD, which consists of a deep learning-based feature extraction module and a statistical decision making module. The first module is a novel feature extraction technique that combines GAN-based frame prediction and a lightweight object detector. The second module is a sequential anomaly detector, which enables performance analysis. The asymptotic false alarm rate of MONAD is analyzed, and a practical procedure is provided for selecting its detection threshold to satisfy a desired false alarm rate. Through real data experiments, MONAD is shown to outperform the state-of-the-art methods, and yield false alarm rates consistent with the derived asymptotic bounds. For future work, we plan to focus on the importance of timely detection in video [60] by proposing a new metric based on the average delay and precision.

Chapter 2: A Modular and Unified Framework for Detecting and Localizing Video Anomalies

2.1 Introduction

[3]With the increasing demand for security, increasing storage and processing capabilities, and decreasing cost of electronics, surveillance cameras have been widely deployed [100]. Due to the exponential increase in the number of CCTV cameras, the amount of video generated far surpasses our ability to manually analyze it. Automated detection of anomalies in video is challenging since the definition of “anomaly” is ambiguous – any event that does not conform to “normal” behaviors can be considered as an anomaly. For example, a person riding a bike is usually a nominal behavior, however, it may be considered as anomalous if it occurs in a restricted space.

Specifically, due to the important role video anomaly detection plays in ensuring safety, security, and sometimes prevention of potential catastrophes, a major functionality of a video anomaly detection system is the real-time decision making capability. While there is a lot of prior work on anomaly detection in surveillance videos, they mainly *focus on offline localization of anomaly in video frames* following an instance-based binary hypothesis testing approach and *ignoring the online (i.e., real-time) detection of anomalous events.* For example, most of the existing works, e.g. [36, 52, 100], employ a video normalization technique that requires an entire video segment for computation. They also typically depend on the assumption that there is an anomaly in the video segment. In practice, this assumption either will not hold for short video segments (on the order of minutes) or will cause long

[3] Portions of this chapter were published in IEEE/CVF Winter Conference on Applications of Computer Vision [19]. Copyright permissions from the publishers are included in Appendix B.

delays in detecting anomalous events for sufficiently long video segments (on the order of days).

The automated video surveillance literature lacks a clear distinction between online anomalous event detection and offline anomalous frame localization [52, 36, 71, 68, 62]. While the commonly used frame-level AUC (area under the ROC curve), which is borrowed from the instance-based binary hypothesis testing, might be a suitable metric for localizing the anomaly in video frames, it ignores the temporal nature of videos and fails to capture the dynamics of detection results, e.g., a detector that detects a late portion of an anomalous event and alarms the user after a long delay can achieve the same frame-level AUC as the detector that quickly detects the anomalous event and timely alarms the user but misses some anomalous frames afterwards. While minimizing the delay in detecting an anomalous event is critical [60], it is also necessary to control the false alarm rate. Hence, a video anomaly detector should aim to judiciously raise alarms in a timely manner.

For practical implementations, it is unrealistic to assume the availability of sufficient training data such that it encompasses all possible nominal events/behaviors. Thus, a practical framework should also be able to perform *few-shot* adaptation to new nominal scenarios over time. This presents a novel challenge to the current approaches discussed in Section 3.2 as their decision functions heavily depend on Deep Neural Networks (DNNs) [17]. DNNs typically require a large amount of training data to learn a new nominal pattern or exhibit the risk of catastrophic forgetting with incremental updates [40].

Another limitation of existing methods is the lack of interpretability due to the inclination towards end-to-end deep learning based models, leading to a semantic gap between the visual features and the real interpretation of events [65]. While such models perform well on some benchmark datasets, i.e., they are easily able to detect a certain category of anomalies, they cannot adequately generalize to other types of anomalies. For example, [65, 79, 62] propose a pose estimation based framework, and hence are only able to detect human-related anomalies.

Moreover, there is no straightforward way to modify such methods to target a different class of anomaly since they are based on intricately designed neural networks.

Our goal in this paper is to present a more systematic framework for video anomaly detection and localization, and tackle practical challenges such as few-shot adaptation, which is largely unexplored in the existing literature. In summary, our contributions in this paper are as follows:

- We present a systematic unified framework for online event detection and offline frame localization for video anomalies, and propose a new performance metric for online event detection.
- We propose a modular transfer learning based anomaly detection architecture which can be easily modified to target specific anomaly categories and can easily adapt to new scenarios using a few samples (cross-domain adaptivity).
- We introduce a statistical technique for the selection of detection threshold to satisfy a desired false alarm rate.

2.2 Related Works

There is a fast-growing body of research investigating anomaly detection in videos. A key component of computer vision problems is the extraction of meaningful features. In video surveillance, the extracted features should be capable of capturing the difference between nominal and anomalous events within a video [17]. While some methods use supervised learning to train on both nominal and anomalous events [51, 44], the majority of existing research is concentrated on semi-supervised learning due to the limitations in the availability of annotated anomalous instances. Early anomaly detection methods used handcrafted approaches which extract different types of motion information in the form of histogram of oriented gradients (HOGs) [5, 10] and optical flow. Another category is sparse coding-based methods [99], which were used to learn a dictionary of normal sparse events, and attempt to detect anomalies based on the reconstructability of video from the dictionary atoms. For example, [64] uses sparse reconstruction to learn joint trajectory representations of multiple objects. These approaches, while computationally inexpensive, often fail to capture complex anomalous patterns. The recent literature however has been dominated by Convolutional Neural Network (CNN) based methods [29, 30, 57, 75, 82, 95, 62, 71, 36] due to their significantly superior detection performance. Recently, transfer learning based object detection methods have also been frequently used [16, 17, 36, 25] to learn appearance features. The neural network-based methods can be broadly segregated into reconstruction-based methods [29, 74, 9, 36] and prediction-based methods [52, 55, 81]. However, these CNNs require a significant amount of training to adapt to a new scenario. Hence, recently few-shot learning has been gaining attention in the computer vision literature [41, 90, 87, 92, 53, 55]. However, no significant progress has been made yet in few-shot scene adaptation for video surveillance. Hence, in this work, we primarily compare our few-shot adaptation performance with [55], which proposes a meta-learning algorithm for cross-domain adaptivity.

2.3 Proposed Method

2.3.1 Motivation

In the recent anomaly detection literature, most of the proposed methods consist of training a deep neural network on available nominal samples. However, such an approach has several shortcomings. First, the applicability of such a method is limited to a few scenarios where there is a drastic change in the appearance or motion of an object. In [16], it is shown that modifying the benchmark datasets results in a significant drop in the performance of state-of-the-art algorithms. Second, to the best of our knowledge, there is no existing method that can be easily modified or extended to a new category of anomalies. For example, even recent algorithms such as [97, 52, 70] cannot detect (or be modified to detect) anomalies pertaining to changes in human poses. Third, because of the extensive use of end-to-end learning in recent algorithms, the models lack interpretability. While there are

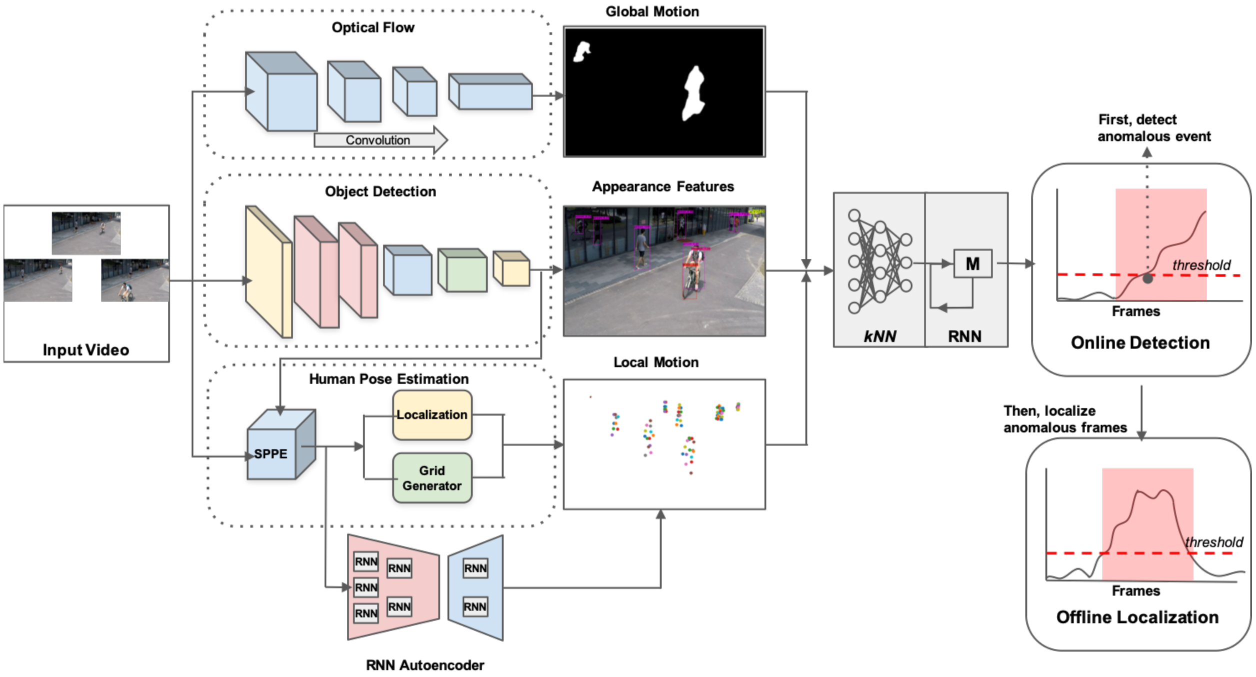


Figure 2.1: Proposed MOVAD framework. At each time t, neural network-based feature extraction module provides location (center coordinates and area of bounding box), appearance (class probabilities), global motion (optical flow), and local motion (pose estimation) features to the statistical anomaly detection module, which computes kNN distance for anomaly evidence using a fully connected neural network, and sequentially decides for anomalous events using an RNN. In human pose estimation, the single person pose estimation (SPPE) is converted to multi-person pose features.

certain supervised methods, e.g., [88], which are capable of recognizing the type of anomaly, they depend on the availability of anomalous data. Finally, existing methods also lack a clear procedure for incorporating new knowledge, and would likely necessitate significant changes to the existing architecture.

Motivated by these shortcomings, we propose a modular framework, called *Modular Online Video Anomaly Detector (MOVAD)*, consisting of deep learning-based feature extraction and statistical anomaly detection, as shown in Fig. 2.1. In particular, transfer learning based convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used to extract informative features, followed by a novel kNN-based neural network and RNN-based sequential anomaly detector.

The choice of separating feature extraction module and decision module also enables theoretical performance analysis and a closed-form expression for the detection threshold. In the following sections, we discuss our framework in detail.

2.3.2 Transfer 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 easily 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. 2.1.

- Object Appearance: A pre-trained object detection system is used to detect objects and extract appearance and spatial features. Since we do not assume any prior knowledge about the type of anomalies, and hence by extension the object classes, we use a model trained on the MS-COCO dataset. For online anomaly detection, the real-time operation is a critical factor, and hence, we currently prefer the You Only Look Once (YOLO) [76] algorithm, specifically YOLOv4, in our implementations. It should be noted that the choice of the object detector is not critical for the proposed framework, and can be adjusted according to the application. 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. Hence, to monitor the contextual motion of different objects, we propose using a pre-trained optical flow model such as Flownet 2 [35]. We hypothesize that objects with an unusually high/low optical flow intensity would exhibit an anomalous behavior. Thus, the mean and variance are for each detected object are used as our global motion features.

- 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 [22] is proposed to extract skeletal trajectories.

2.3.3 Statistical Anomaly Detection

- Anomaly Evidence: Given the various extracted features, the next step in the proposed framework is to compute an anomaly evidence score for each video frame in an online fashion. Due its favorable characteristics, such as interpretability and theoretical tractability, we use k-nearest-neighbor (kNN) distance as an anomaly evidence. For a feature vector $X_{t,i} \in \mathbb{R}^m$ representing each object i in frame t, our objective is to compute its Euclidean distance $D_{t,i}$ to the kth nearest feature vector in the nominal training set. Since kNN distance computation becomes expensive with increasing training size, for scalability, we propose training a fully connected neural network with parameters θ, which takes $X_{t,i}$ as the input and gives an accurate approximation $\tilde{D}_{t,i}(\theta)$ to $D_{t,i}$. The objective function for training the kNN neural network is given by

$$\min_{\theta} \frac{1}{N} \sum_{j=1}^{N} (D_j - \tilde{D}_j(\theta))^2 + \lambda f(\theta), \tag{2.1}$$

where N is the number of feature vectors in the training set, $\lambda f(\theta)$ is the regularization term. The number of neighbors k determines a trade-off between sensitivity to anomalies and robustness to nominal outliers. While smaller k values makes the system more sensitive to real anomalies, it may also make the system more vulnerable to nominal outliers. However, the choice of k is not critical for the detection performance since the proposed sequential detection module does not directly decide on the anomaly evidences. As shown next, through the internal memory of the RNN structure, it gathers the evidences to detect anomalous events, hence does not typically raise an alarm due to a single evidence due to an outlying frame.

- Online Anomaly Detection: To accommodate the temporal continuity of video data and detect anomalous events in an online fashion, a sequential statistical decision making method based on RNN is proposed. The anomaly evidence scores (i.e., kNN distances) from streaming video frames provide an informative time series data which typically takes large values when the anomalous event starts. However, to avoid false alarms due to outlying large evidences from nominal frames, the proposed framework does not decide using individual evidences, but instead utilizes the temporal information inherent in the evidence time series (i.e., an anomalous event consists of a number of successive anomalous frames). Specifically, it takes the streaming kNN distances $\{\tilde{D}_t\}$ as input and updates an internal state, which is then passed through ReLU activation function to yield the decision statistic s_t. The time series $\{\tilde{D}_t\}$ is obtained by taking the largest kNN distance among objects in each frame, i.e., $\tilde{D}_t = \max_i \tilde{D}_{t,i}$. The output neuron in RNN compares s_t with a threshold h to raise an alarm if $s_t \geq h$ or continue with the next frame otherwise. Note that the RNN structure can be expanded to accept multiple time series (in addition to kNN distances) and to have deeper layers if desired. While kNN distances are available for the nominal class, there is no such scores for the anomaly class to train RNN in the considered semi-supervised setup. Synthetic kNN distances are generated uniformly in the interval $(D_\alpha, 2D_{max})$ where α is a statistical significance level (e.g., $\alpha = 0.05$), D_α is the $(1-\alpha)$ percentile of nominal distances in the training set, and D_{max} is the maximum nominal distance in the training.

To circumvent the training with synthetic data, and obtain a closed-form expression for the threshold h, we also propose a simplified decision rule. Motivated by the resemblance of the memory (internal state) and ReLU operations of RNN with the minimax optimum sequential change detection algorithm CUSUM [4], we consider fixing the RNN weights to obtain the simplified decision statistic $\tilde{s}_t = \max\{\tilde{s}_{t-1} + \delta_t, 0\}$. In this update rule, the weights of internal state and input are set to one, where the input $\delta_t = \tilde{D}_t^m - D_\alpha^m$ is the normalized kNN distance, where m is the dimensionality of feature vectors $X_{t,i}$. In our experiments, the simplified detector gave very similar results to the general RNN detector. With the weights set to one, there is no need to train the RNN, and the simplified decision statistic $\tilde{s}_t$ lends itself to theoretical analysis to derive a closed-form expression for the threshold h, as explained next.

Corollary 1. *As the training size grows ($N \to \infty$), the false alarm rate of the proposed simplified detector based on $\tilde{s}_t$ is upper bounded by $FAR \leq e^{-\omega_0 h}$ and the threshold h can be set as*

$$h = \frac{-\log \beta}{\omega_0} \tag{2.2}$$

to asymptotically satisfy a desired false alarm constraint $FAR \leq \beta$. The constant ω_0 is computed from the training data and given by

$$\omega_0 = v_m - \theta - \frac{1}{\phi}\mathcal{W}\left(-\phi\theta e^{-\phi\theta}\right), \tag{2.3}$$
$$\theta = \frac{v_m}{e^{v_m D_\alpha^m}},$$

where $\mathcal{W}(\cdot)$ is the Lambert-W function, $v_m = \frac{\pi^{m/2}}{\Gamma(m/2+1)}$ is the constant for the m-dimensional Lebesgue measure (i.e., $v_m d_\alpha^m$ is the m-dimensional volume of the hyperball with radius d_α), and ϕ is the upper bound for δ_t.

Proof. During the testing phase when there is no anomaly, the feature vector with the maximum kNN distance is independently distributed due to the appearance vector of a randomly detected object. Hence, the i.i.d assumption from Theorem 1 still holds. The rest follows the proof of Theorem 1.

Although the expression for ω_0 looks complicated, all the terms in Eq. (2.3) can be easily computed. Particularly, v_m is directly given by the number of features m, D_α comes from the training phase, ϕ is also found in training, and finally there is a built-in Lambert-W function in popular programming languages such as Python and Matlab. Hence, given the training data, ω_0 can be easily computed, and the threshold h can be chosen using Eq. (2.2) to asymptotically achieve the desired false alarm rate $\beta \in (0, 1)$.

Decision threshold h is a key parameter that is common to all existing anomaly detection algorithms, and yet is often overlooked. Since an alarm is raised when the test statistic crosses the threshold, choosing an appropriate threshold is critical for controlling the number of false alarms and minimizing the need for human involvement. In a practical setting, without a clear procedure for selecting the decision threshold, an exhaustive empirical process is needed to calibrate the threshold for an acceptable false alarm rate.

- New Performance Metric for Online Detection: Low detection delay is a crucial requirement in most video-related applications such as autonomous driving [50] and automated video surveillance. However, the detection delay, which is the time required by an algorithm to detect an anomalous event, is largely unexplored in the field of video anomaly detection. The popular performance metric in the video anomaly detection literature, AUC, cannot effectively evaluate the performance of online anomaly detection algorithms [45]. Hence, we present a new performance metric called APD (Average Precision as a function of Delay), which is based on average detection delay and precision. The proposed delay metric is given by

  $$\text{APD} = \int_0^1 P(\gamma)\ \mathrm{d}\gamma, \tag{2.4}$$

  where γ denotes the normalized average detection delay, and P denotes the precision. The average detection delay is normalized by the largest possible delay either defined by a performance requirement or the length of natural cuts in the video stream such as the video segments in the benchmark datasets (See Sec. 2.4.1).

- Offline Localization: Once an anomalous event is detected, the detection instance is marked as the starting point, and the decision statistic is updated as usual to determine the end point. When the decision statistic drops consecutively for a number of frames (e.g., five frames is found to be a good number in our experiments), the beginning of the drop window is marked as the end point. Finally, the frames between the start and end points are labeled as anomalous.

- Implementation Details: In our implementation, we fix the number of neighbors as $k = 10$. However, as indicated in Section 2.3.3, the choice of k is not sensitive and does not significantly affect the performance of the detector. The detection performance is controlled by the decision threshold h, which can be mathematically set by following Eq. (2.3). For the kNN regression network, 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. The feature vector is 18-dimensional for each detected object, and consists of 15 class probabilities (appearance), mean and variance of optical flow in the bounding box (global motion), and prediction error of pose if human (local motion). Global and local motion features are normalized to [0,1] using the min and max values from the training data.

2.4 Experiments

In this section, we first briefly discuss the benchmark datasets and the evaluation metrics. Then, we provide a detailed comparison between the proposed algorithm and the state-of-the-art algorithms in terms of online detection and offline localization. We also evaluate our few-shot adaptation performance.

2.4.1 Datasets

We consider four publicly available benchmark datasets, namely the CUHK Avenue dataset, the UCSD pedestrian dataset, the ShanghaiTech campus dataset, and the UR fall 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. The anomalous events are caused due to 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.

- 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. The resolution for each video frame is 480 × 856.

- UR Fall: While the UR fall dataset is not popularly used for video anomaly detection, it has recently been proposed for testing the generalization capability of anomaly detection algorithms [55]. This dataset contains 70 depth videos collected with a Microsoft Kinect camera in a nursing home and the anomalies consist of a person falling in a closed room.

2.4.2 Results

- Online Detection: Since the proposed online detection formulation is event-based as compared to frame-based, 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 [52, 36, 65, 70] 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 [36, 55, 68] do not have their implementation details/code publicly available, while others are end-to-end [68, 71, 79] and cannot be implemented to work in an online fashion. Hence, we compare our method with the online versions of [52, 65, 58]. As shown in Fig. 2.2, our proposed algorithm achieves a better performance than the other algorithms in terms of quick detection and achieving high precision in alarms. This result is also summarized in Table 4.1 in terms of the APD values.

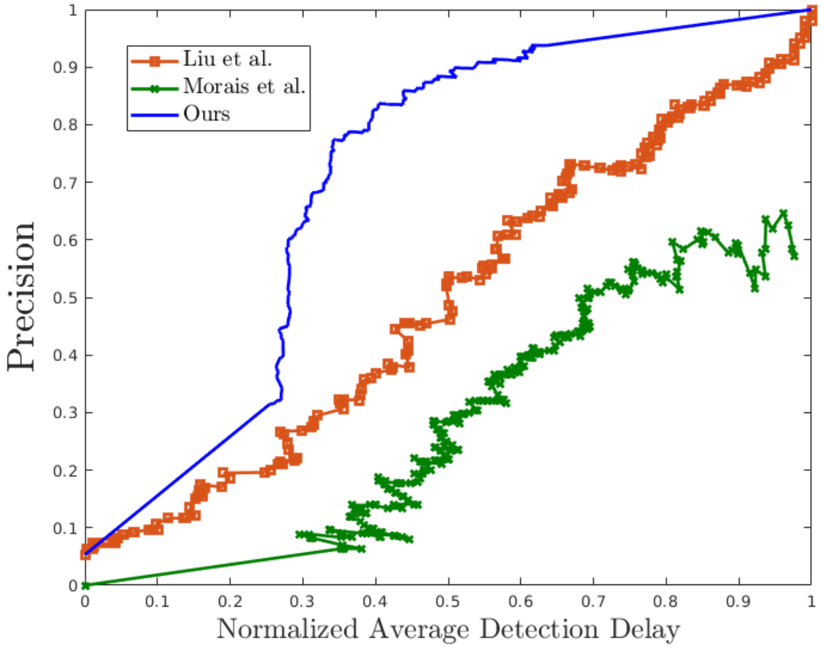


Figure 2.2: Comparison of the proposed and the state-of-the-art algorithms Liu et al. [52] and Morais et al. [65] in terms of online detection capability. The proposed algorithm has a significantly higher precision for any given detection delay.

Table 2.1: Online detection comparison in terms of the proposed APD metric on the ShanghaiTech dataset. Higher APD value represents a better online anomaly detection performance.

Online Detection	
Methodology	**APD**
Liu et al. [52]	0.504
Morais et al. [65]	0.324
Luo et al. [58]	0.447
Ours	**0.705**

- Threshold Selection: We next evaluate the non-asymptotic use of the asymptotic threshold expression given in Eq. (2.2). As shown in Fig. 2.3, even with the limited data size of the CUHK Avenue dataset, the derived expression satisfies the desired

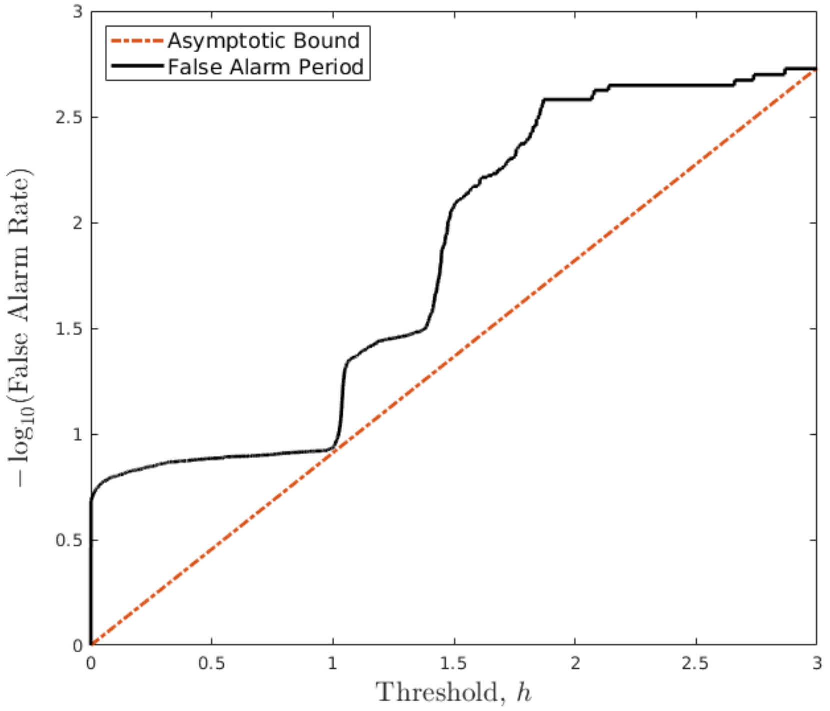


Figure 2.3: Threshold selected according to Eq. (2.2) satisfies the desired lower bound on false alarm period (i.e., upper bound on false alarm rate) even in the non-asymptotic regime with the finite sample size of the CUHK Avenue dataset.

upper bound on the false alarm rate, which corresponds to a lower bound on the false period (inverse rate) in the figure.

- Offline Localization: To show the offline localization capability of our algorithm, we also compare our algorithm to a wide range of state-of-the-art methods, as shown in Table 4.2, using the 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 [71]. Hence, for offline anomaly localization, we consider frame-level AUC criterion. While [36] recently showed significant gains over the other algorithms, their methodology of computing the average AUC over an entire dataset gave them an unfair advantage. Specifically, as opposed to determining the AUC on the concatenated videos, first the AUC for each video segment was computed and then those AUC values were averaged. As shown in Ta-

Table 2.2: Offline anomaly localization comparison in terms of frame-level AUC on three datasets.

Offline Localization			
Methodology	**CUHK Avenue**	**UCSD Ped 2**	**ShanghaiTech**
MPPCA [39]	-	69.3	-
Del et al. [14]	78.3	-	-
Conv-AE [29]	80.0	85.0	60.9
ConvLSTM-AE[56]	77.0	88.1	-
Growing Neural Gas [89]	-	93.5	-
Stacked RNN[57]	81.7	92.2	68.0
Deep Generic [30]	-	92.2	-
GANs [73]	-	88.4	-
Future Frame [52]	85.1	95.4	72.8
Skeletal Trajectory [65]	-	-	73.4
Multi-timescale Prediction [79]	82.85	-	**76.03**
Memory-guided Normality [70]	88.5	97.0	70.5
Ours	**88.7**	**97.2**	73.62

ble 4.2, 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 [79] is the only one that outperforms ours on the ShanghaiTech dataset since the anomalies are mostly caused by previously unseen human poses and [79] 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 [65], they cannot work on the UCSD dataset.

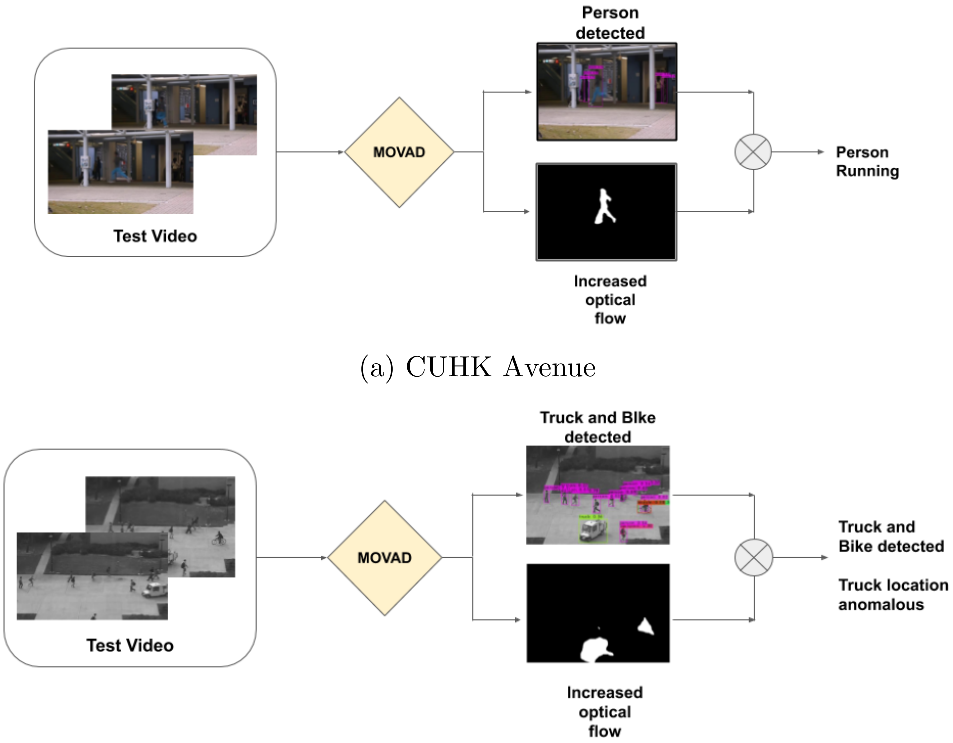
- Few-Shot Scene Adaptation: Our goal here is to compare the few-shot scene adaptation capability of the proposed algorithm and see how well it can generalize to new scenarios. In this case, we only use a few scenes from a specific scenario to adapt. However, few-shot scene adaptation is mostly unexplored and to the best of our knowledge only [55] discusses it. Hence, following the experimental setup defined in [55], we use K-shots to adapt to a new scenario, where 1-shot is a sequence of 10 frames. From [55], we consider the Pre-trained baseline that learns the model from videos available during training

Table 2.3: Few-shot scene adaptation comparison of the proposed and the state-of-the-art [55] algorithms in terms of frame-level AUC. The proposed algorithm is able to quickly adapt to new scenarios.

Target	**Methods**	**1-shot (K=1)**	**5-shot (K=5)**	**10-shot (K=10)**
UCSD Ped 2	Pre-trained (ShanghaiTech)	81.95	81.95	81.95
	Pre-trained (UCF Crime)	62.53	62.53	62.53
	r-GAN (ShanghaiTech)	91.19	91.8	92.8
	r-GAN (UCF Crime)	83.08	86.41	90.21
	Ours	**93.19**	**95.91**	**96.01**
CUHK Avenue	Pre-trained (ShanghaiTech)	71.43	71.43	71.43
	Pre-trained (UCF Crime)	71.43	71.43	71.43
	r-GAN (ShanghaiTech)	76.58	77.1	78.79
	r-GAN (UCF Crime)	72.62	74.68	79.02
	Ours	**80.18**	**80.21**	**80.68**
UR Fall	Pre-trained (ShanghaiTech)	64.08	64.08	64.08
	Pre-trained (UCF Crime)	50.87	50.87	50.87
	r-GAN (ShanghaiTech)	75.51	78.7	83.24
	r-GAN (UCF Crime)	74.59	79.08	81.85
	Ours	**86.11**	**88.7**	**91.28**

and then directly applies the model in testing without any adaptation. Moreover, we also compare with a few-shot scene-adaptive anomaly detection model using a meta-learning framework proposed in [55] called r-GAN. They use a GAN-based framework similar to [52] and MAML algorithm for meta-learning.

As compared to the pre-trained and r-GAN models, which need considerable training on either the ShanghaiTech or UCF Crime [88] dataset, our transfer learning based algorithm (pre-trained on generic datasets such as MS-COCO) is able to leverage our optical flow model which requires minimal computation to establish a baseline and adapt the decision parameter h to a new scene. Due to the lack of available training data, we are unable to use the local motion and appearance features meaningfully, and hence our features are only dependant on the optical flow statistics. However, as shown in Table 2.3, we are still able to outperform the compared methods in terms of the frame-level AUC.



(a) CUHK Avenue

(b) UCSD

Figure 2.4: The proposed model is able to interpret the cause of the anomaly correctly.

2.4.3 Ablation Study

In Table 2.4, we present the results for each module of the proposed MOVAD framework on the ShanghaiTech dataset. While it is clear that optical flow is the major contributor among all the modules in this dataset, each module serves a specific purpose. In this dataset, although several recent works perform closely to the proposed framework, a distinguishing advantage of MOVAD is its interpretability. By leveraging the statistical nature of our

Table 2.4: Performance of each module in terms of the frame-level AUC on the ShanghaiTech dataset.

ShanghaiTech	
Module	**AUC**
Object Detection	0.594
Optical Flow	0.703
Pose Estimation	0.652

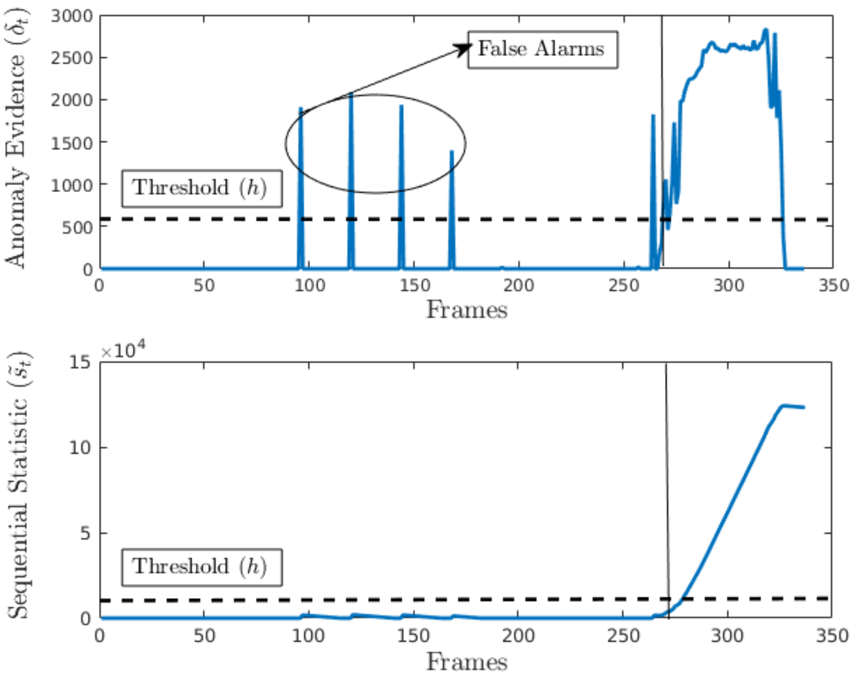


Figure 2.5: The advantage of sequential anomaly detection over a single-shot detector. It is seen that a sequential detector can significantly reduce the number of false alarms.

decision making module, it is possible to determine the cause of increase in the decision statistic. In Fig. 2.4, we present a sample scenario from the CUHK Avenue and UCSD datasets, in which the proposed detector is able to evaluate the statistics from each module and justify the cause of the anomaly. However, since there is no ground truth available in terms of the description of the anomaly, we were unable to quantitatively evaluate the interpretability performance of MOVAD.

- Impact of Sequential Detection: To emphasize the significance of the proposed sequential detection method, we compare a nonsequential version of our algorithm by applying a threshold to the instantaneous anomaly evidence δ_t (Sec. 2.3.3), which is similar to the approach employed by many recent works [52, 88, 36]. As shown in Fig. 2.5, the proposed sequential statistic handles noisy evidence by integrating recent evidence over time. On the other hand, the instantaneous anomaly evidence is more prone to false alarms since it only considers the noisy evidence available at the current time to decide. Specifically, without sequential detection, the APD presented in Table 2.1 for the proposed framework reduces to 0.673.

2.5 Conclusion and Discussions

For video anomaly detection, we presented a modular framework called MOVAD, which consists of an interpretable transfer learning based feature extractor, and a novel kNN-RNN based sequential anomaly detector. Mathematical analysis was provided for false alarm rate and threshold selection. Following the timely detection requirement in practical settings, MOVAD first detects anomalous events in an online fashion, and then deals with localizing the anomalous video frames. Online detection of anomalous events is largely overlooked in the video anomaly detection literature, thus a new performance metric was also introduced to compare algorithms in terms of online anomaly detection in videos. Through extensive testing on the benchmark datasets, we show that MOVAD significantly outperforms the state-of-the-art methods for online detection while performing competitively for offline localization.

While being able to capture anomalies in various video aspects, such as object appearance and motion, the proposed method currently is not optimized for specific anomaly types. For instance, it is not able to detect unexpected human poses as the optical flow does not change significantly (see Appendix A). For future work, we plan to focus on continual and self-supervised learning for MOVAD.

Chapter 3: Rethinking Video Anomaly Detection

3.1 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 [47], CUHK Avenue [54], and ShanghaiTech Campus [52]), 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.

[4]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

[4] Portions of this chapter were published in IEEE/CVF Winter Conference on Applications of Computer Vision [20]. Copyright permissions from the publishers are included in Appendix B.

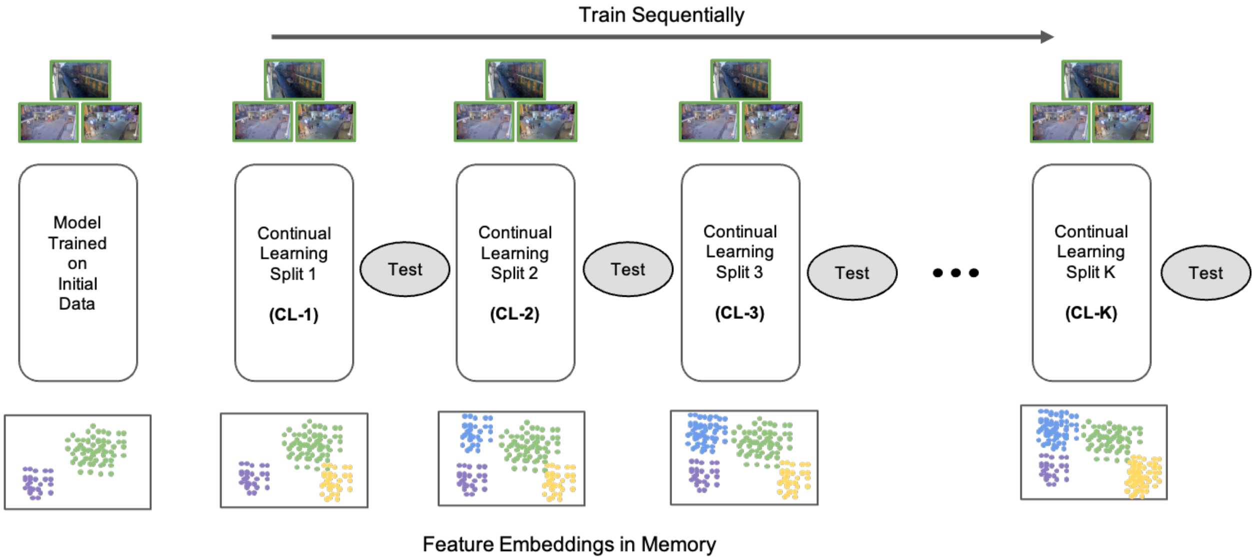

Train Sequentially
Model Trained on Initial Data
Continual Learning Split 1 (CL-1)
Test
Continual Learning Split 2 (CL-2)
Test
Continual Learning Split 3 (CL-3)
Test
• • •
Continual Learning Split K (CL-K)
Test
Feature Embeddings in Memory

Figure 3.1: 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.

[53]. A practical VAD algorithm must continually train [5] 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. 3.1. 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

[5] 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.

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 offline. The existing VAD literature lacks a clear distinction between the anomalous activity detection and anomalous frame localization tasks [52, 36, 79, 65]. 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.

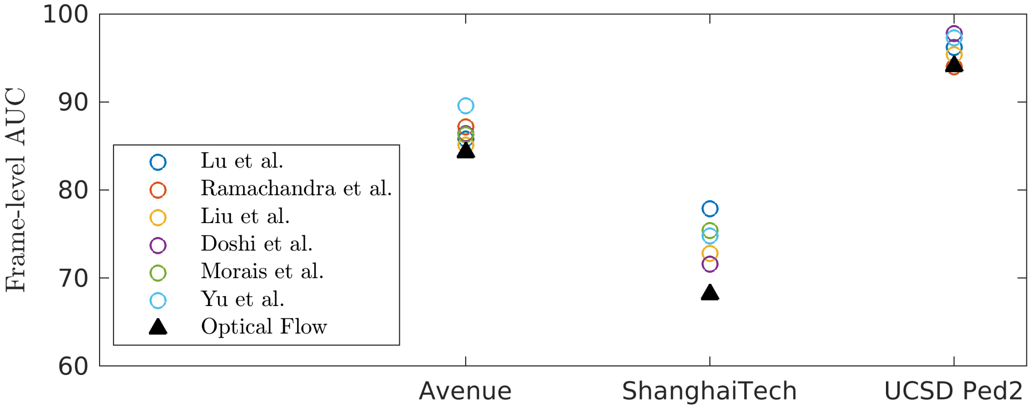


Figure 3.2: Simple optical flow method performs close to the state-of-the-art methods [55, 71, 52, 17, 65, 96] on the three popular benchmark datasets in terms of the frame-level AUC metric.

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 [47], Avenue [54], and ShanghaiTech [52] 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. 3.2, we compare the performance of the recent state-of-the-art methods [55, 71, 52, 17, 65, 96] 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 [71], 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.

3.2 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) [5, 10, 47], Hidden Markov Models [42, 32], sparse coding [99, 64], and appearance features [11, 47], recent approaches have been completely dominated by deep learning based algorithms. Recent algorithms can be broadly classified into reconstruction based approaches [26, 29, 57, 67, 70], which try to classify frames based on the reconstruction error, and prediction based approaches [52, 46, 15, 18], which attempt to predict a future frame, primarily by using generative adversarial networks (GANs) [27]. More recently, skeletal trajectory based approaches [65, 79] 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, [72] 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 [41, 90, 87, 92, 53]. However, not a lot of progress has been made yet in continual learning for VAD. In [17], 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 [36, 30], 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.

3.3 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 [53]. 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. 3.4, will help guide the future VAD research towards practical and reproducible solutions.

3.3.1 Problem Formulation

Although a stream of video frames $F = \{f_1, f_2, ...\}$ 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 $X = \{x_1, x_2, ...\}$.

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 $\min \mathbb{E}_1[T_i - \tau_i]$, where $\mathbb{E}_1$ 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

  $$\mathrm{ADD} = \frac{1}{N}\sum_{i=1}^{N}(T_i - \tau_i), \tag{3.1}$$

  with N denoting the number of anomalous activities. Considering a longest tolerable delay $\delta_{\max}$, if there is no alarm within the duration $[\tau_i, \tau_i + \delta_{\max}]$ after anomalous activity i happens, the delay is set to be the maximum value, i.e., $T_i - \tau_i = \delta_{\max}$. 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), $\frac{1}{N}\sum_{j=1}^{N} 1\{T_i \geq \tau_i\}$. The indicator function $1\{A\}$ 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., $1\{T_i \geq \tau_i\}$, ADD assigns the specific cost of detection delay $\delta_i = T_i - \tau_i$.

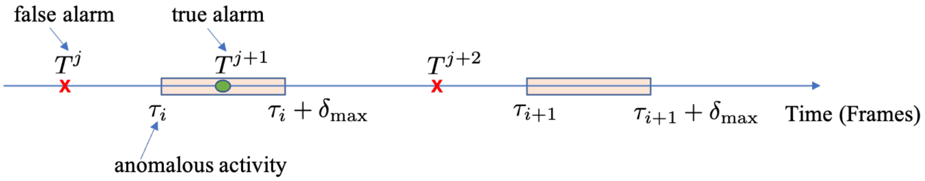


Figure 3.3: Definitions of true alarm and false alarm. The anomalous activity i is successfully detected with alarm time $T_i = T^{j+1}$, whereas the anomalous activity $i+1$ is missed.

- 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. 3.3, an alarm j is a true alarm if it is raised within the relevant duration of an anomalous activity, i.e., $T^j \in \cup[\tau_i, \tau_i + \delta_{\max}]$, 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., $[\tau_i, \tau_i + \delta_{\max}] \cap [\tau_{i+1}, \tau_{i+1} + \delta_{\max}] = \emptyset, \forall i$. 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 $(T^j \in \cup[\tau_i, \tau_i + \delta_{\max}])$, which gives the alarm precision. Empirically, the alarm precision is computed as

$$P = \frac{1}{M} \sum_{j=1}^{M} 1\{T^j \in \cup[\tau_i, \tau_i + \delta_{\max}]\}, \tag{3.2}$$

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$/\delta_{\max}$. Mathematically, APD is given by

$$\text{APD} = \int_0^1 P(\alpha)\, \mathrm{d}\alpha, \tag{3.3}$$

  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.

  As compared to the APD metric defined in Eq. (2.4) for instance based detections, here we consider event based detections.

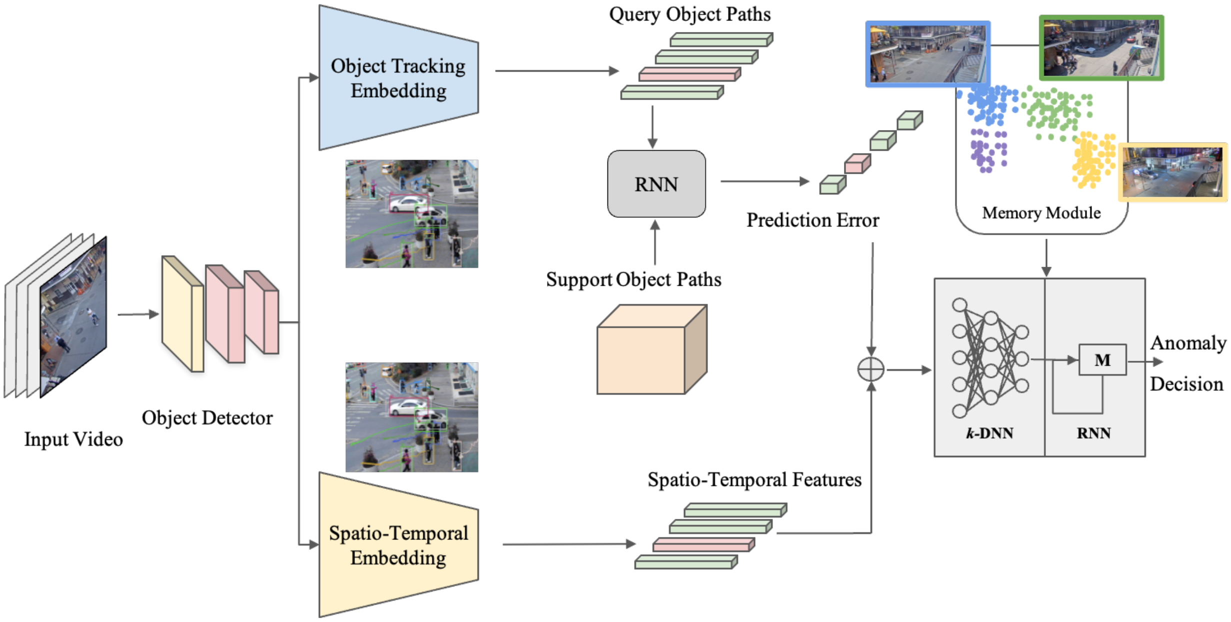


Figure 3.4: Proposed VAD algorithm. Object tracking features and spatio-temporal object features form the feature vector, whose kNN distance with respect to the nominal vectors is used to make an anomaly decision within an RNN structure. The use of kNN distances facilitates effective continual learning.

- 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. 3.1). 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.,

$$\text{APD}_k \geq \text{APD}_{k-1}, \forall k. \tag{3.4}$$

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.

3.3.2 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 [91]. 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. 3.4, we first detect objects in each frame by using a pretrained object detector, such as YOLO-v4 [76]. 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 [93], 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.

Table 3.1: Comparison of existing and proposed VAD datasets. Ground truth refers to the type of anomaly labeling.

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**	**990,000**	**450,000**	**Spatial, Temporal**	**1280 x 720**	Audio also available

3.4 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 [88], 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 [71]. 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).

3.4.1 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.

- 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.

3.4.2 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[6] 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 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 3.1. 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 3.3.1. One split is used for initial

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

Table 3.2: Performance of the proposed detector and recent state-of-the-art approaches across different continual learning splits in terms of the proposed APD metric.

Method	**CL-1**	**CL-2**	**CL-3**	**CL-4**	**CL-5**	**CL-6**	**CL-7**	**CL-8**	**CL-9**	**CL-10**
Future Frame Prediction [52]	0.137	0.149	0.173	0.205	0.211	0.232	0.202	0.22	0.245	0.271
MNAD [70]	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

training, and the rest 10 splits are used to evaluate the continual learning performance (Fig. 3.1).

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. 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 [17, 66] 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 [52] and the Memory guided Normality (MNAD) method [70]. 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 [36] 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.

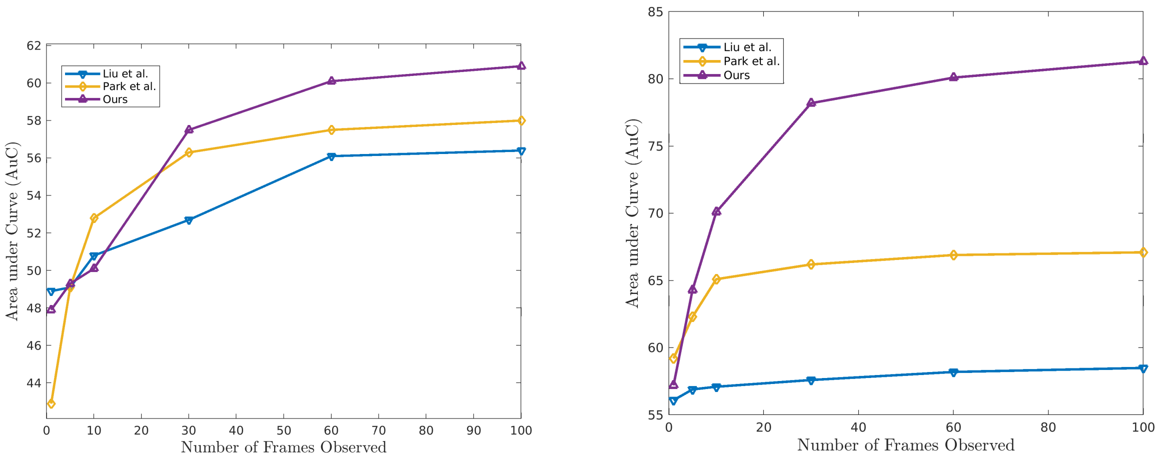


Figure 3.5: Comparison of the proposed and state-of-the-art algorithms Liu et al. [52] and Park et al. [70] in terms of learning from few samples on the ShanghaiTech (top) and UCSD (bottom) datasets.

- 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. 3.3.1. In this experiment, we use the k-DNN and experience replay based version of our algorithm. From Table 3.2, 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 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. 3.3.1. 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 datasets that occurs several times. Fig. 3.5 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.

3.5.1 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, 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.

3.6 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. 3.3) and the new dataset (Sec. 3.4) 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.

Chapter 4: Multi-Task Learning for Video Surveillance with Limited Data

4.1 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 [52, 18, 70].

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 [47] and ShanghaiTech [52] video anomaly detection benchmark datasets, a person riding a bike is considered as anomalous; however, in the recently released Street Scene [71] dataset, it is considered as nominal. Hence, most of the existing approaches [72, 71, 3, 7, 15, 29, 37, 16, 17, 52, 68, 100, 65] 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 [55], 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 [53]. 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 [40], 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 4.2 as their decision functions heavily depend on Deep Neural Networks (DNNs) [17]. 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.

4.2 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) [5, 10, 47], Hidden Markov Models (HMMs) [42, 32], sparse coding [99, 64], and appearance features [11, 47], recent approaches have been completely dominated by deep learning based algorithms. Recent algorithms can be broadly classified into reconstruction based approaches [26, 29, 57, 67, 70], which try to classify frames based on the reconstruction error, and prediction based approaches [52, 46, 15, 18], which attempt to predict a future frame, primarily by using generative adversarial networks (GANs) [27]. More recently, skeletal trajectory based approaches [65, 79] 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, [72] 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 dependent on complex neural networks and mostly end-to-end trained.

Several recent works propose using a GAN for detecting anomalies in videos. For example, [52] 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 [41, 90, 87, 92, 53]. However, not a lot of progress has been made yet in continual learning for video anomaly detection. In [17], 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 [36, 30], which treat each object independently, and fails to capture the intricate relationship between different objects.

4.3 Multi-Task Video Surveillance

In the existing video anomaly detection literature, the singular goal is to detect frames or 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.

4.3.1 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 [55]. 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 4.4, 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, [55] 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 [55] is based on the future frame prediction model [52], which uses a GAN and thus is unable to perform tasks T1 and T3. Another recent work, [17] 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.

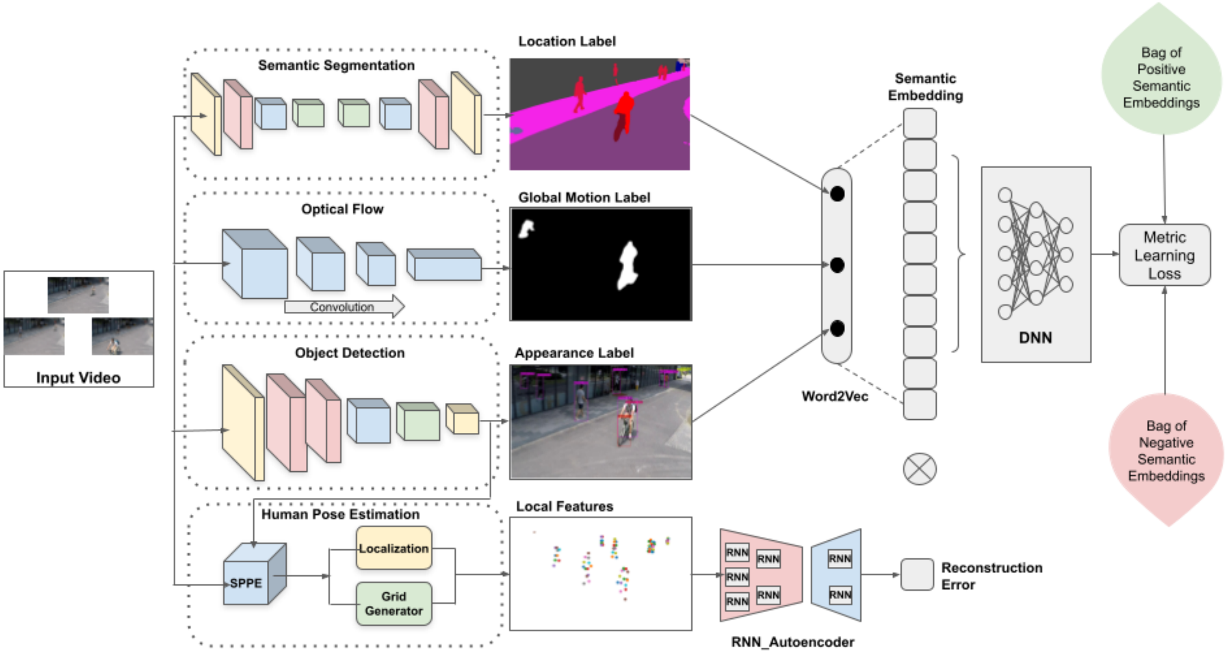


Figure 4.1: Proposed video anomaly detection framework. At each time t, neural network-based feature extraction module provides location, appearance and global motion labels, and local motion (pose estimation) reconstruction error, which is then used to form a semantic embedding which represents the detected activity. This is then used to train a deep neural network using metric learning, which outputs the anomaly score.

4.3.2 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. 4.1. 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.

4.3.3 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 easily 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. 4.1.

- 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) [76] 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 [78] 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 [48], we train a Faster R-CNN model which uses a Squeeze and Excitation Network (SENet) [33], 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.

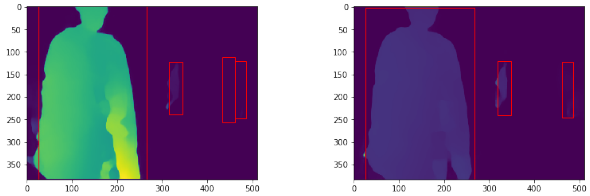
- Global Motion: Apart from spatial and appearance features, capturing the motion of different objects is also critical for detecting anomalies in videos. We propose a novel modification of an optical flow model known as perspective based optical flow. Optical flow is widely used in the existing literature 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

$$O_{n1} = \frac{O_1}{w_1 * h_1}. \tag{4.1}$$

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. 4.2, 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 4.2: 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 [22] 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 [7]. 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 [17] 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.

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

4.3.4 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. 4.1, in our proposed approach, we pose anomaly detection as a regression problem. We want the anomalous semantic video embeddings to 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 [31] and is defined as:

  $$\mathcal{L} = \max(0, m + \|f(a) - f(p)\| - \|f(a) - f(n)\|), \tag{4.2}$$

  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 [52, 36, 71, 68, 62], 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 [68, 71, 79, 55, 52]. Some recent works [79, 56] 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 [45], 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.

4.4 Experiments

In this section, we present the performance of the proposed approach on the tasks defined in Section 4.3.1. 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.

4.4.1 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 [19], 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).

4.4.2 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 [52, 36, 65, 70] 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 [36, 55, 68] do not have their implementation details/code publicly available, while others are end-to-end [68, 71, 79] and cannot be implemented to work in an online fashion. Hence, we compare our method with the online versions of [52, 65, 58]. 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 4.1 in terms of the APD value.

Table 4.1: Online detection comparison in terms of the proposed APD metric on the ShanghaiTech dataset. Higher APD value represents a better online anomaly detection performance.

Online Detection	
Method	**APD**
Liu et al. [52]	0.504
Morais et al. [65]	0.324
Luo et al. [58]	0.447
Ours	**0.675**

- 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 4.2, 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 [71]. Hence, for anomaly localization, we consider the frame-level AUC criterion. As shown in Table 4.2, 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 [79] is the only one that outperforms ours on the ShanghaiTech dataset since the anomalies are mostly caused by previously unseen human poses and [79] 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 [65], they cannot work on the UCSD dataset.

Table 4.2: Offline anomaly localization comparison in terms of frame-level AUC on three datasets.

Anomaly Localization (AUC)			
Method	**CUHK Avenue**	**UCSD Ped 2**	**ShanghaiTech**
MPPCA [39]	-	69.3	-
Del et al. [14]	78.3	-	-
Conv-AE [29]	80.0	85.0	60.9
ConvLSTM-AE[56]	77.0	88.1	-
Growing Neural Gas [89]	-	93.5	-
Stacked RNN[57]	81.7	92.2	68.0
Deep Generic [30]	-	92.2	-
GANs [73]	-	88.4	-
Future Frame [52]	85.1	95.4	72.8
Skeletal Trajectory [65]	-	-	73.4
Multi-timescale Prediction [79]	82.85	-	**76.03**
Memory-guided Normality [70]	88.5	97.0	70.5
Ours	**86.4**	**95.6**	70.12

- 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 [55] discusses a similar few-shot adaptation concept. However, the proposed approach discussed in [55] requires several anomaly-free video frames for adapting their model to the new scene, which might not always be feasible. Particularly, in [55], a GAN-based framework is used in [55] similar to [52], and MAML algorithm [23] is used for meta-learning. As shown in Tables 4.3–4.5, 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.

Table 4.3: 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 [57]	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 [52]	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 4.4: Performance of the proposed detector in terms of event-level APD 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 [57]	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 [52]	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

- 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. 4.3, it is seen that the proposed algorithm clearly outperforms the state-of-the-art algorithms [16, 52, 57] 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.

Table 4.5: 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.

Frame-level AUC		
Approach	**ShanghaiTech**	**Avenue**
Stacked RNN [57]	0.643	0.724
Future Frame Prediction [52]	0.652	0.749
Skeletal Trajectory [65]	0.683	-
Ours	0.689	0.79

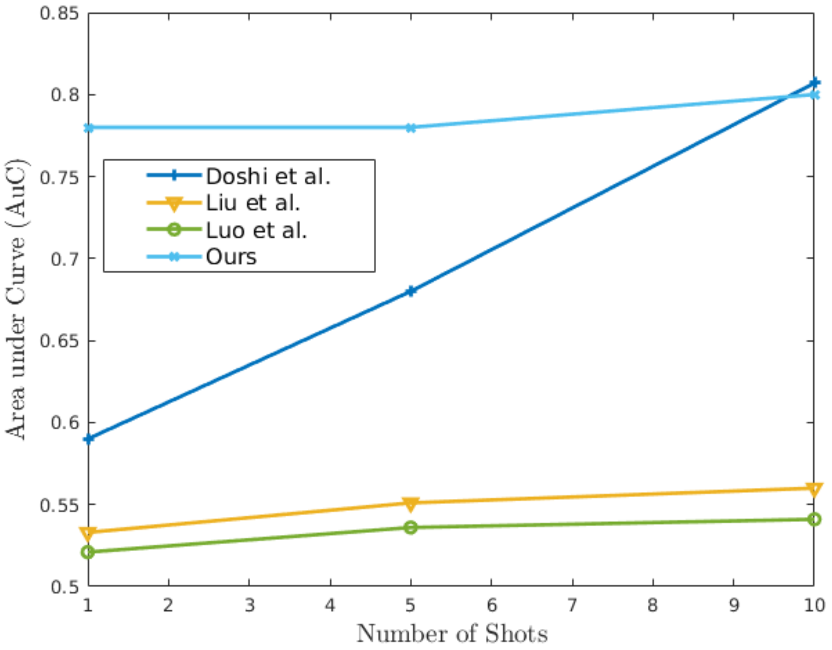


Figure 4.3: Comparison of the proposed and state-of-the-art algorithms Liu et al. [52], Luo et al. [57] and Doshi et al. [16] in terms of few-shot learning. Together with the original training data, some frames for bike riding are used to train the algorithms. The proposed algorithm achieves high performance even with one shot.

- 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 4.4, it is seen that the proposed algorithm clearly outperforms the state-of-the-art algorithms [17, 52, 57] 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.

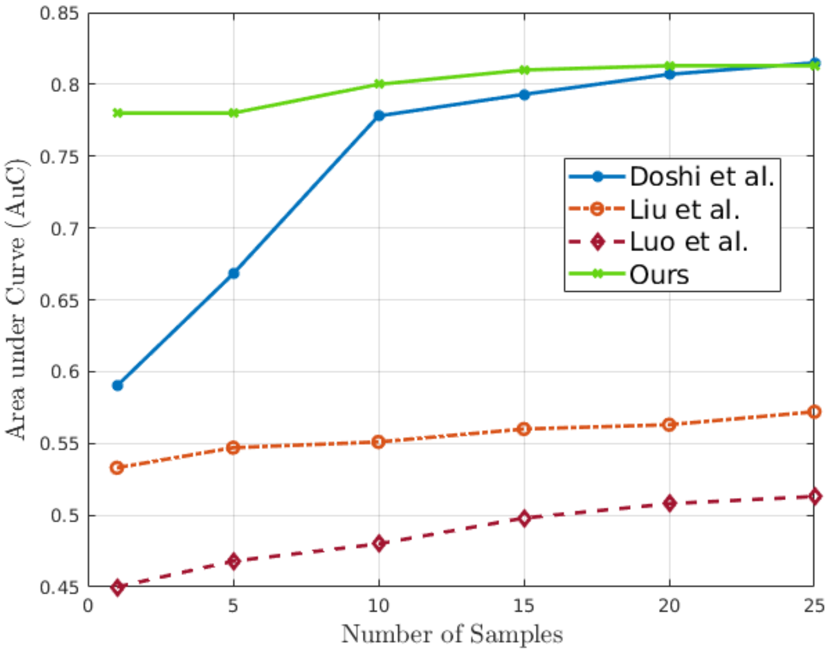


Figure 4.4: Comparison of the proposed and the state-of-the-art algorithms Liu et al. [52], Luo et al. [57] and Doshi et al. [16] in terms of continual learning capability. Different than few-shot learning, here the new data (bike frames) are used to incrementally update the algorithms after the initial training on the training data. While training with new samples, the proposed algorithm maintains superior performance compared to the state-of-the-art methods.

4.5 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*.

[2] Nadeem Anjum and Andrea Cavallaro. Multifeature object trajectory clustering for video analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1555–1564, 2008.

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

[4] Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993.

[5] 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.

[6] George H Chen, Devavrat Shah, et al. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018.

[7] 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.

[8] Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[9] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017.

[10] 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.

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

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] Keval Doshi and Yasin Yilmaz. Rethinking video anomaly detection-a continual learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3961–3970, 2022.

[21] Rana Elnaggar, Krishnendu Chakrabarty, and Mehdi B Tahoori. Hardware trojan detection using changepoint-based anomaly detection techniques. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 27(12):2706–2719, 2019.

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

[23] 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.

[24] Zhouyu Fu, Weiming Hu, and Tieniu Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–602. IEEE, 2005.

[25] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A scene-agnostic framework with adversarial training for abnormal event detection in video. *arXiv preprint arXiv:2008.12328*, 2020.

[26] 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.

[27] 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.

[28] Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. In *Advances in Neural Information Processing Systems*, pages 10921–10931, 2019.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 387–395, 2018.

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

[36] 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.

[37] 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.

[38] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in neural information processing systems*, pages 5541–5552, 2018.

[39] 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.

[40] 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.

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

[42] 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.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[44] Federico Landi, Cees GM Snoek, and Rita Cucchiara. Anomaly locality in video surveillance. *arXiv preprint arXiv:1901.10364*, 2019.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] L. Lin and N. Purnell. A world with a billion cameras watching you is just around the corner. *The Wall Street Journal, https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402*, 2019.

[50] Shih-Chieh Lin, Yunqi Zhang, Chang-Hong Hsu, Matt Skach, Md E Haque, Lingjia Tang, and Jason Mars. The architectural implications of autonomous driving: Constraints and acceleration. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 751–766, 2018.

[51] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1490–1499, 2019.

[52] 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.

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

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

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

[60] Huizi Mao, Xiaodong Yang, and William J Dally. A delay metric for video object detection: What average precision fails to tell. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 573–582, 2019.

[61] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[62] 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.

[63] Bryan Matthews. Automatic Anomaly Detection with Machine Learning. https://ntrs.nasa.gov/citations/20190030491, 2019.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[75] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1896–1904. IEEE, 2019.

[76] 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.

[77] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[78] 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.

[79] 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.

[80] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[81] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402, 2018.

[82] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018.

[83] Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2112–2119. IEEE, 2012.

[84] Tony C Scott, Greg Fee, and Johannes Grotendorst. Asymptotic series of generalized lambert w function. *ACM Communications in Computer Algebra*, 47(3/4):75–83, 2014.

[85] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[86] Chawin Sitawarin and David Wagner. Defending against adversarial examples with k-nearest neighbor. *arXiv preprint arXiv:1906.09525*, 2019.

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

[88] 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.

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

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] Yang Xiang, Ke Li, and Wanlei Zhou. Low-rate ddos attacks detection and traceback by using new information metrics. *IEEE transactions on information forensics and security*, 6(2):426–437, 2011.

[95] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[96] 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.

[97] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020.

[98] Haibin Zhang, Jiajia Liu, and Nei Kato. Threshold tuning-based wearable sensor fault detection for reliable medical monitoring using bayesian network model. *IEEE Systems Journal*, 12(2):1886–1896, 2018.

[99] 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.

[100] 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.

Appendix A: Proof of Theorem 1

In [4][page 177], for CUSUM-like algorithms with independent increments, a lower bound on the average false alarm period is given as follows

$$E_\infty[T] \geq e^{\omega_0 h},$$

where h is the detection threshold, and $\omega_0 \geq 0$ is the solution to $E[e^{\omega_0 \delta_t}] = 1$.

To analyze the false alarm period, we need to consider the nominal case. In that case, since there is no anomalous object at each time t, the selection of object with maximum kNN distance in $\delta_t = (\max_i\{d_t^i\})^m - d_\alpha^m$ does not necessarily depend on the previous selections due to lack of an anomaly which could correlate the selections. Hence, in the nominal case, it is safe to assume that δ_t is independent over time.

We firstly derive the asymptotic distribution of the frame-level anomaly evidence δ_t in the absence of anomalies. Its cumulative distribution function is given by

$$P(\delta_t \leq y) = P((\max_i\{d_t^i\})^m \leq d_\alpha^m + y).$$

It is sufficient to find the probability distribution of $(\max_i\{d_t^i\})^m$, the mth power of the maximum kNN distance among objects detected at time t. As discussed above, choosing the object with maximum distance in the absence of anomaly yields independent m-dimensional instances $\{F_t\}$ over time, which form a Poisson point process. The nearest neighbor ($k = 1$) distribution for a Poisson point process is given by

$$P(\max_i\{d_t^i\} \leq r) = 1 - \exp(-\Lambda(b(F_t, r)))$$

where $\Lambda(b(F_t, r))$ is the arrival intensity (i.e., Poisson rate measure) in the m-dimensional hypersphere $b(F_t, r)$ centered at F_t with radius r [8]. Asymptotically, for a large number of training instances as $M_2 \to \infty$, under the null (nominal) hypothesis, the nearest neighbor distance $\max_i\{d_t^i\}$ of F_t takes small values, defining an infinitesimal hyperball with homogeneous intensity $\lambda = 1$ around F_t. Since for a homogeneous Poisson process the intensity is written as $\Lambda(b(F_t, r)) = \lambda|b(F_t, r)|$ [8], where $|b(F_t, r)| = \frac{\pi^{m/2}}{\Gamma(m/2+1)} r^m = v_m r^m$ is the Lebesgue measure (i.e., m-dimensional volume) of the hyperball $b(F_t, r)$, we rewrite the nearest neighbor distribution as

$$P(\max_i\{d_t^i\} \leq r) = 1 - \exp\left(-v_m r^m\right),$$

where $v_m = \frac{\pi^{m/2}}{\Gamma(m/2+1)}$ is the constant for the m-dimensional Lebesgue measure.

Now, applying a change of variables we can write the probability density of $(\max_i\{d_t^i\})^m$ and δ_t as

$$f_{(\max_i\{d_t^i\})^m}(y) = \frac{\partial}{\partial y}\left[1 - \exp\left(-v_m y\right)\right], \tag{A.1}$$

$$= v_m \exp(-v_m y), \tag{A.2}$$

$$f_{\delta_t}(y) = v_m \exp(-v_m d_\alpha^m) \exp(-v_m y) \tag{A.3}$$

Using the probability density derived in (A.3), $E[e^{\omega_0 \delta_t}] = 1$ can be written as

$$1 = \int_{-d_\alpha^m}^{\phi} e^{\omega_0 y} v_m e^{-v_m d_\alpha^m} e^{-v_m y} dy, \tag{A.4}$$

$$\frac{e^{v_m d_\alpha^m}}{v_m} = \int_{-d_\alpha^m}^{\phi} e^{(\omega_0 - v_m) y} dy, \tag{A.5}$$

$$= \left.\frac{e^{(\omega_0 - v_m) y}}{\omega_0 - v_m}\right|_{-d_\alpha^m}^{\phi}, \tag{A.6}$$

$$= \frac{e^{(\omega_0 - v_m)\phi} - e^{(\omega_0 - v_m)(-d_\alpha^m)}}{\omega_0 - v_m}, \tag{A.7}$$

where $-d_\alpha^m$ and ϕ are the lower and upper bounds for $\delta_t = (\max_i\{d_t^i\})^m - d_\alpha^m$. The upper bound ϕ is obtained from the training set.

As $M_2 \to \infty$, since the mth power of $(1-\alpha)$th percentile of nearest neighbor distances in training set goes to zero, i.e., $d_\alpha^m \to 0$, we have

$$e^{(\omega_0 - v_m)\phi} = \frac{e^{v_m d_\alpha^m}}{v_m}(\omega_0 - v_m) + 1. \tag{A.8}$$

We next rearrange the terms to obtain the form of $e^{\phi x} = a_0(x+\theta)$ where $x = \omega_0 - v_m$, $a_0 = \frac{e^{v_m d_\alpha^m}}{v_m}$, and $\theta = \frac{v_m}{e^{v_m d_\alpha^m}}$. The solution for x is given by the Lambert-W function [84] as $x = -\theta - \frac{1}{\phi}\mathcal{W}(-\phi e^{-\phi\theta}/a_0)$, hence

$$\omega_0 = v_m - \theta - \frac{1}{\phi}\mathcal{W}\left(-\phi\theta e^{-\phi\theta}\right). \tag{A.9}$$

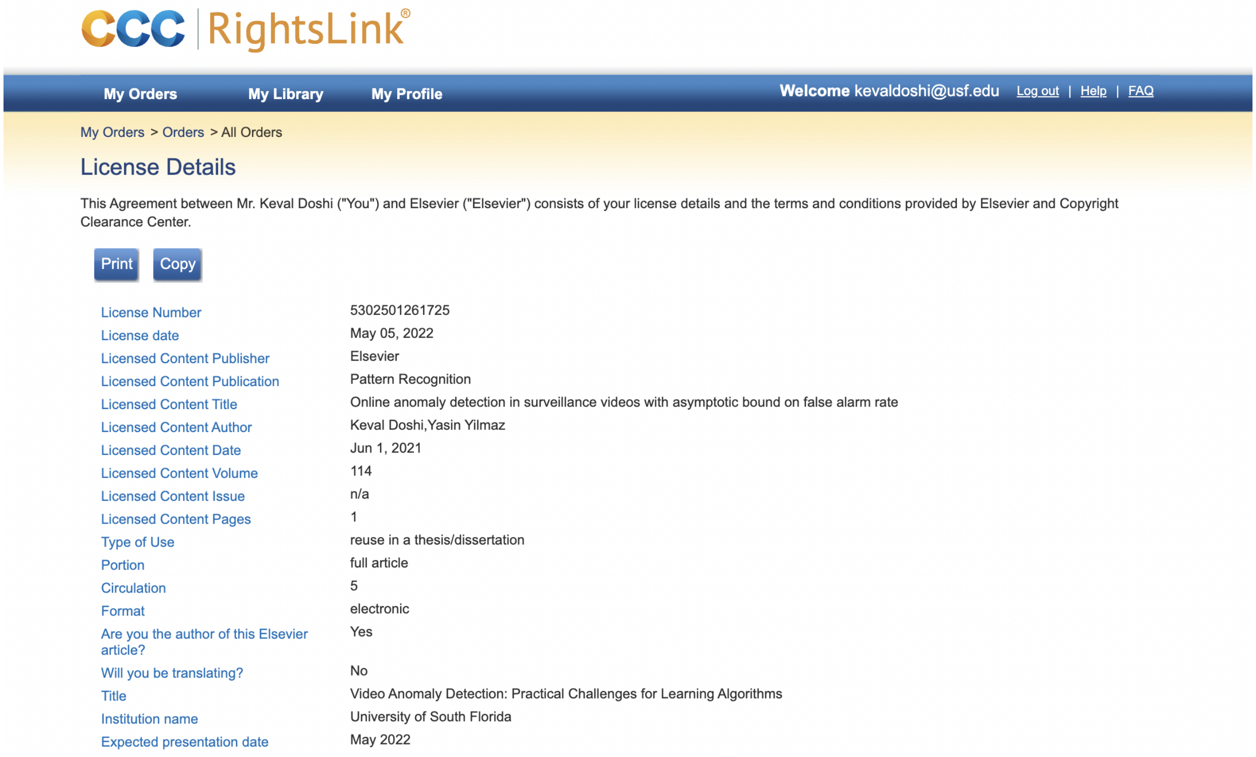
Finally, since the false alarm rate (i.e., frequency) is the inverse of false alarm period $E_\infty[T]$, we have

$$FAR \leq e^{-\omega_0 h},$$

where h is the detection threshold, and ω_0 is given above.

Appendix B: Copyright Permissions

The permission below is for the reproduction of material in Chapter 1.

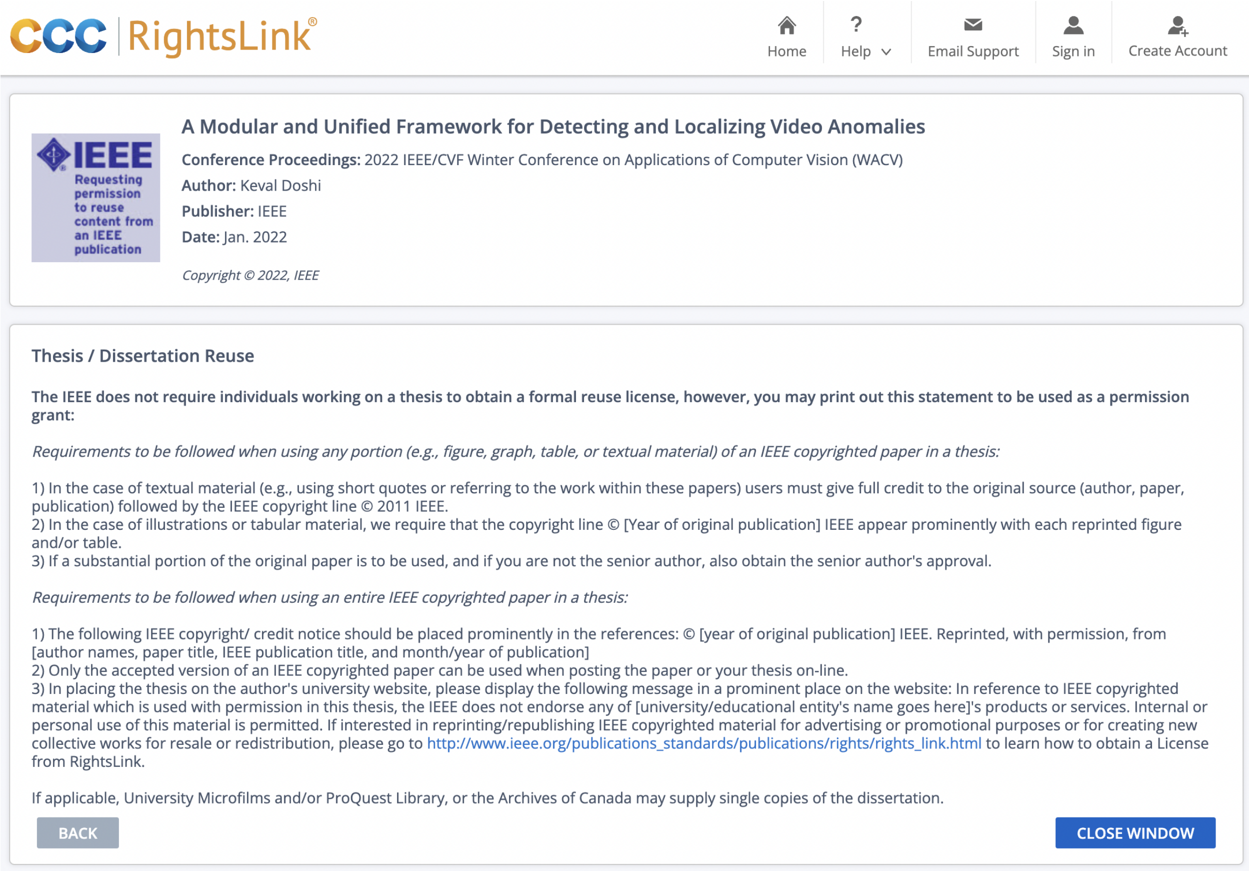


License Details

This Agreement between Mr. Keval Doshi ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5302501261725
License date	May 05, 2022
Licensed Content Publisher	Elsevier
Licensed Content Publication	Pattern Recognition
Licensed Content Title	Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate
Licensed Content Author	Keval Doshi,Yasin Yilmaz
Licensed Content Date	Jun 1, 2021
Licensed Content Volume	114
Licensed Content Issue	n/a
Licensed Content Pages	1
Type of Use	reuse in a thesis/dissertation
Portion	full article
Circulation	5
Format	electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Title	Video Anomaly Detection: Practical Challenges for Learning Algorithms
Institution name	University of South Florida
Expected presentation date	May 2022

The permission below is for the reproduction of material in Chapter 2.



A Modular and Unified Framework for Detecting and Localizing Video Anomalies

Conference Proceedings: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Author: Keval Doshi

Publisher: IEEE

Date: Jan. 2022

Copyright © 2022, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

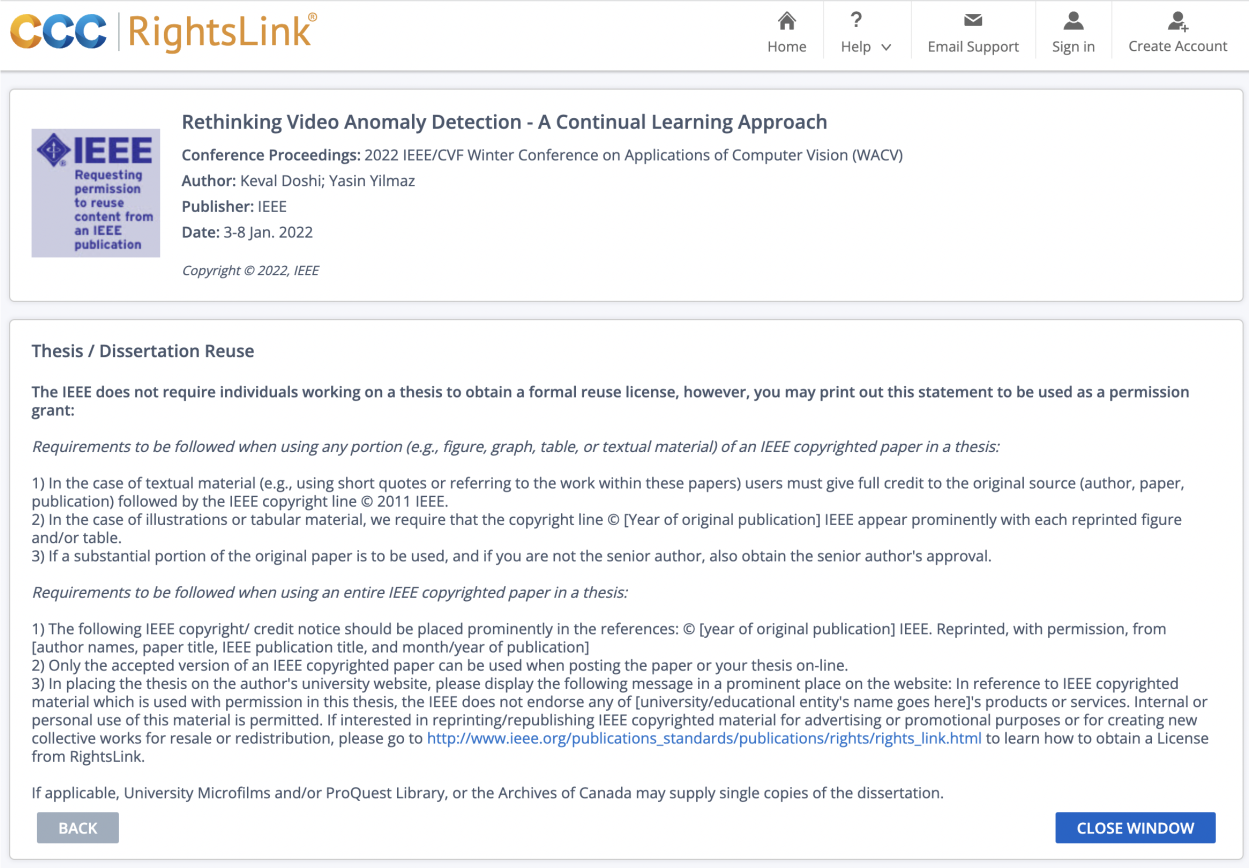
1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

The permission below is for the reproduction of material in Chapter 3.


CCC RightsLink®
Home
Help
Email Support
Sign in
Create Account
IEEE
Requesting permission to reuse content from an IEEE publication

Rethinking Video Anomaly Detection - A Continual Learning Approach

Conference Proceedings: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Author: Keval Doshi; Yasin Yilmaz

Publisher: IEEE

Date: 3-8 Jan. 2022

Copyright © 2022, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.