Intelligent Systems with Applications 18 (2023) 200236

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Review

Unveiling the performance of video anomaly detection models — A benchmark-based review

Francisco Caetano [a,b,*], Pedro Carvalho [a,c], Jaime S. Cardoso [a,b]

[a] INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal
[b] Faculty of Engineering (FEUP), University of Porto, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal
[c] School of Engineering (ISEP), Polytechnic of Porto, Rua Dr. António Bernardino de Almeida 431, 4200-072, Porto, Portugal

ARTICLE INFO

Keywords:
Anomaly detection
Deep learning
Computer vision
Benchmark

ABSTRACT

Deep learning has recently gained popularity in the field of video anomaly detection, with the development of various methods for identifying abnormal events in visual data. The growing need for automated systems to monitor video streams for anomalies, such as security breaches and violent behaviours in public areas, requires the development of robust and reliable methods. As a result, there is a need to provide tools to objectively evaluate and compare the real-world performance of different deep learning methods to identify the most effective approach for video anomaly detection. Current state-of-the-art metrics favour weakly-supervised strategies stating these as the best-performing approaches for the task. However, the area under the ROC curve, used to justify this statement, has been shown to be an unreliable metric for highly unbalanced data distributions, as is the case with anomaly detection datasets. This paper provides a new perspective and insights on the performance of video anomaly detection methods. It reports the results of a benchmark study with state-of-the-art methods using a novel proposed framework for evaluating and comparing the different models. The results of this benchmark demonstrate that using the currently employed set of reference metrics led to the misconception that weakly-supervised methods consistently outperform semi-supervised ones.

1. Introduction

Anomaly Detection has been a relevant field in Machine Learning since it began to gain traction. As far as Computer Vision is concerned, anomaly detection is an important task on the account of its prevailing applications in video surveillance, scene understanding and video summarization. Several surveys provide a comprehensive literature review while inspecting the current challenges and studying future opportunities for the application of such methods (Pang et al., 2021, Caetano et al., 2022). Video anomaly detection methods are developed under one of two main approaches to the problem: semi-supervised strategies or weakly-supervised ones.

Generally, semi-supervised methods for anomaly detection in the literature fall in the category of One-Class Classification (OCC) (Chandola et al., 2009). These models are trained on normality and assume that it is not possible to properly reconstruct an abnormal event that has never been learnt. Hence, a frame that greatly differs from the captured one is likely to represent abnormal or unexpected events. In these cases, the abnormality is detected based on the information learnt from the normal class only. This formulation comprises an open-set approach to anomaly detection, i.e., it is assumed that the types of anomalies that can occur are unbounded, whilst the normality pattern is well-defined (Acsintoae et al., 2022). One-Class Classification is present in reconstruction-based and prediction-based semi-supervised methods.

Weakly-supervised strategies differ from semi-supervised strategies by obtaining video-level labels, which allow for the training of the models using normal and abnormal snippets (Pang et al., 2019). This approach comprises a closed-set evaluation scenario, i.e., the training and test anomalies belong to the same bounded action categories. Essentially, weakly-supervised approaches can be subdivided into two classes: encoder-agnostic methods (Sultani et al., 2018, Zhang et al., 2019, Wan et al., 2020) that leverage task agnostic features of videos extracted from a vanilla feature encoder to estimate the anomaly scores of each frame and encoder-based methods; encoder-based methods (Zhu & Newsam, 2019, Zhong et al., 2019) which train both the feature encoder and classifier simultaneously.

* Corresponding author at: INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal.
E-mail addresses: francisco.t.espirito@inesctec.pt (F. Caetano), pedro.m.carvalho@inesctec.pt (P. Carvalho), jaime.cardoso@inesctec.pt (J.S. Cardoso).

https://doi.org/10.1016/j.iswa.2023.200236
Received 23 February 2023; Received in revised form 27 April 2023; Accepted 14 May 2023
Available online 23 May 2023

According to recent benchmarks in the available literature (Tian et al., 2021, Feng et al., 2021, Wu & Liu, 2021), weakly-supervised strategies comprise the best-performing approaches for deep video anomaly detection. The area under the ROC curve, the metric employed to compare and rank the performance of state-of-the-art methods, reveals that weakly-supervised strategies can achieve a higher score in its analysis, although it does not explain the reasons why that occurs. Moreover, adapting datasets originally developed for semi-supervised approaches to be used by weakly-supervised ones generates different training and testing sets. Therefore, a fair comparison between different distributions of the same set of videos cannot be drawn. A deeper analysis and benchmarking of deep video anomaly detection is severely lacking in the available literature, as a common framework must be provided first.

This work appears as the first critical review on the state-of-the-art metrics for benchmarking deep video anomaly detection models, making the following three major contributions:

- Experimental comparison of state-of-the-art video anomaly detection methods and an in-depth analysis of the context that defines normalcy and abnormality in the publicly available datasets for this task;
- Proposal of a new benchmark framework that evaluates the real-world performance of deep anomaly detection models and that automatically defines the model's optimal operating point;
- Novel insight on state-of-the-art anomaly detection model performance, as the previously employed set of metrics led to the misconception that weakly-supervised methods consistently outperformed semi-supervised ones.

This document is organised as follows: Section 2 introduces the working principles of state-of-the-art works in deep video anomaly detection. Section 3 presents a detailed analysis of each of the four methods' strategies is presented. Moreover, Section 4 introduces the proposed benchmark methodology is presented. Sections 5 and 6 depict the results achieved in the proposed benchmark. Additionally, Section 7 the attained results are discussed. Section 8 presents the main conclusions of this review. Finally, Section 9 introduces future work to expand on the main findings of this article.

2. Related work

This section intends to provide a comprehensive review of previous research in the field of video anomaly detection. The existing contributions' gaps, limitations, and strengths are discussed. In addition, a summary of the most relevant publicly available datasets for these tasks is provided.

2.1. Semi-supervised methods

Reconstruction-based methods try to reconstruct the current frame, using previous and present information. The deep autoencoder, ConvAE, proposed by Hasan et al. (2016), became the first anomaly detection approach to leverage the reconstruction error as an estimator for abnormality. This work was quickly followed by Conv3D-AE (Sabokrou et al., 2016, Zhao et al., 2017), suggesting a 3D convolutional neural network to encode the motion and content information of a sequence of frames, using a deconvolutional network to reconstruct those frames. However, 3D convolution has proved to be unable to properly encode motion (Ji et al., 2012, Tran et al., 2015). A Convolutional Neural Network (CNN) and ConvLSTM were integrated with an autoencoder in ConvLSTM-AE (Luo et al., 2017a) to learn the regularity of appearance and motion for ordinary moments. Although LSTMs and Recurrent Neural Networks (RNNs) are effective, they are difficult to interpret; hence, several works focused on adapting sparse coding techniques and interpretable RNNs to anomaly detection (Luo et al., 2017b, Wisdom et al., 2016). As Liu et al. (2018) denoted, autoencoder-based approaches are at times able to accurately reconstruct abnormal frames based on the provided inputs, leading to missing their detection. To deal with this drawback, a memory module was added to the autoencoder by Gong et al. (2019), creating MemAE, a memory-augmented autoencoder.

Prediction-based methods use the previous frames to compute a prediction of the following one. This method was introduced by Liu et al. (2018); strategies to impose consistency on the generated images were imposed, by applying intensity and gradient constraints. Taking inspiration from the *cloze test* used in language understanding, Yu et al. (2020) proposed the prediction of erased patches of incomplete video events, fully exploiting temporal information in the video. Unlike these approaches, (Chen et al., 2022) aimed to explore the information contained in the anterior and posterior snippets of a given frame within a video. For that purpose, it modelled the relationship between appearance and motion through a multi-modal discriminator; the discriminator was fed the concatenation of an erased patch and its motion to learn to classify fake and real pairs. The temporal relationships in the video sequence were also considered.

Georgescu et al. (2021) proposed some alterations to middle-frame prediction (Lee et al., 2019), innovating by learning the discrimination of moving objects, referred to as the arrow of time. Additionally, it studied motion irregularity prediction and model distillation, the latter being an adaptation of Bergmann et al. (2020). This approach was inspired by the object-centric perspective of Ionescu et al. (2019), which employed an object detector on each frame, applying a convolutional autoencoder to learn deep unsupervised representations for a one-versus-rest classification. Several works continued the exploration of this research (Doshi & Yilmaz, 2020a, 2020b, Yu et al., 2020). The main limitations of semi-supervised approaches are the lack of consideration for the diversity of normal patterns and the ability of deep learning techniques to correctly recreate abnormal video frames based on already abnormal inputs. To this end, Park et al. (2020) proposed a memory module that updates items in the memory, while assuring that these represent prototypical patterns of normal data. Similarly, Cai et al. (2021) attempted to assure appearance and motion consistency through modality memory pools.

2.2. Weakly-supervised methods

Weakly-supervised strategies are considered to be a feasible method due to their competitive performance. Sultani et al. (2018) introduced the use of video-level labels in the tasks of anomaly detection in videos by presenting UCF-Crime, a large-scale video dataset for training and testing weakly-supervised anomaly detection approaches. Along with this strategy, Sultani et al. (2018) proposed a deep Multiple Instance Learning (MIL) ranking framework to detect anomalies. Several papers followed the MIL framework, suggesting improvements to the method. The inner-bag score gap regularization was introduced by Zhang et al. (2019); Wan et al. (2020) proposed a dynamic MIL-loss and centre-guided regularization. Additionally, Zhu and Newsam (2019), in an encoder-based approach, suggested an attention-based MIL model capable of encoding motion-aware features by using an autoencoder based on optical flow. To unify the representation learning and anomaly score learning, a temporal feature ranking loss was presented by Tian et al. (2021). Self-attention mechanisms were proposed to reduce the false alarm rates of these detectors (Li, Cai, et al., 2022, Zhang et al., 2023).

Zhong et al. (2019) denoted that the methods that used MIL suffered from error propagation throughout the training. To tackle this problem, Zhong et al. (2019) reformulated the task as a binary classification under a noisy label problem and suggested the use of a Graph Convolution Neural (GCN) network to correct low-confidence anomaly scores, replacing them with high-confidence ones, i.e., clear the label noise. Li, Liu, et al. (2022) proposed another approach capable of addressing the shortcomings of MIL-based methods. The authors of this paper used a Multi-Sequence Learning (MSL) method, opting for choosing the sequence with the highest sum of anomaly scores instead of the

Table 1
Brief summary of anomaly frame distribution in the analysed anomaly detection datasets.

Dataset	Number of Frames			Scenes	Anomalies	
	Normal	Abnormal	Total		Types	Total
CUHK Avenue (Lu et al., 2013)	26832	3820	30652	1	5	47
ShanghaiTech (Luo et al., 2017b)	300308	17090	317398	13	11	130
UCF-Crime (Sultani et al., 2018)	-	-	13741393	1900	13	-

instance with the highest score, reducing the probability of incorrect selection. Instead of using video-level labels as pseudo-labels, Feng et al. (2021) suggested the use of the learnt pseudo-labels to optimize the feature encoder; a similar approach was employed by Thakare et al. (2023). Ullah et al. (2021) developed an approach that aims to reduce the processing time required for deep anomaly detection. For this purpose, the features extracted from the sequence of frames were fed to a Bi-directional Long Short-term Memory (BD-LSTM) model, which differs from a regular LSTM by depending not only on the previous frames but also on the upcoming ones.

2.3. Datasets

In general, research on anomaly detection in video sequences has focused intensively on the analysis of video surveillance footage of pedestrians and crowds. As a result, most of the available datasets relate to these types of scenarios. Nonetheless, new datasets have attempted to cover new areas, such as violent and criminal behaviours, and surveillance of streets shared by pedestrians and vehicles. The three datasets summarized in Table 1 are worth highlighting because of their relevance to benchmarking state-of-the-art methods:

- **CUHK Avenue** (Lu et al., 2013) was acquired using a stationary video camera in the CUHK campus avenue. It has 16 training video samples and 21 test video samples. The abnormal behaviours represented in the scenes show people littering items, walking on the grass, and throwing or abandoning objects in the background. However, this dataset possesses severe limitations regarding its single-scene representation, lack of abnormality diversity, and amount of sequences;
- **ShanghaiTech Campus** (Luo et al., 2017b) took advantage of multiple surveillance cameras with different view angles installed at different spots, to capture real events at a university campus. ShanghaiTech has challenging light conditions and camera angles. It contains 130 abnormal events of 13 different types and annotations for pixel-level ground truth of abnormal events;
- **UCF-Crime** (Sultani et al., 2018) was developed as a new large-scale dataset to evaluate video anomaly detection. It is composed of 1900 untrimmed videos of real-world surveillance footage, extracted from the internet, with an average length of 4 minutes each. It includes 13 types of anomalous events with a high impact on public safety, such as abuse, burglary, shoplifting and shooting.

3. Selected methods for evaluation

Executing a meaningful analysis of different methods is a laborious task and, limited by the available resources, such as source code and pretrained models. The selection of the four benchmarked methods, FFP, MLEP, Sultani, and RTFM, was driven by two main criteria: (1) they are all open-source, making it possible for others to reproduce the experiments and extend the research; (2) these methods are relevant landmarks in the video anomaly detection area, having been widely used in previous works and achieving competitive performance. Therefore, benchmarking these methods provides valuable insights into the state-of-the-art in video anomaly detection and enable fair comparisons among them.

Table 2 lists some details on the code shared by the authors of some of the surveyed methods. Every repository was verified to confirm its accessibility and if it contained sufficient information to be properly implemented without major adjustments. Some additional information was also gathered, such as the Machine Learning framework that was used, mainly TensorFlow (Abadi et al., 2015) and PyTorch (Paszke et al., 2019).

3.1. Future frame prediction

The work of Liu et al. (2018) was an important landmark in deep anomaly detection, as it represented the first work that leveraged video prediction for anomaly detection, instead of simply reconstructing the present frame. Future Frame Prediction (FFP) still depended on the assumption that normal events are predictable while abnormal ones are unpredictable if the network that predicts the future frames is trained on normal data only. Therefore, FFP employed the typical semi-supervised approach to anomaly detection, relying on the assumption that only normal data is available in the training set, not only because anomalies are rare and unbounded, but also due to the effort required to label them. Furthermore, Liu et al. (2018) observed that spatial constraints were not efficient enough to predict video frames, as these lacked temporal information. Therefore, optical flow constraints were introduced, enforcing the optical flow between predicted frames to be close to their optical flow ground truth, thus representing temporal and motion information with more detail. Moreover, FFP's source code is open source, and the pre-trained models that were used were also made available by the authors. Hence, this work represents a relevant method to be evaluated in more detail by our proposed benchmark.

In general terms, the networks used for frame generation in existing work (Mathieu et al., 2015) usually contain two modules: an encoder that extracts features by gradually reducing the spatial resolution; a decoder that gradually increases the spatial resolution to construct a frame. However, an architecture like this faces the gradient vanishing problem and information imbalance in each layer. To avoid these drawbacks, U-Net was proposed and added shortcuts between the high-level and low-level layers with the same resolution, resulting in the suppression of gradient vanishing and resulting in information symmetry. The original U-Net architecture was slightly modified in this implementation, by keeping the output resolution unchanged for each two convolution layers. The final architecture of this method is shown in Fig. 1.

To evaluate the anomaly score, the difference between the predicted frame $\hat{I}$ and its ground truth I is leveraged. Mathieu et al. (2015) showed that Peak Signal to Noise Ratio (PSNR) is a better way to assess image quality; a high PSNR score is associated to a frame that is more likely to be normal, while a low score is linked with a significant difference between the prediction and ground truth, thus making it more likely to be abnormal.

3.2. MLEP

Liu et al. (2019) proposed a Margin Learning Embedded Prediction (MLEP) framework, to expand a typical prediction-based framework for video anomaly detection, such as FFP, providing it with a large margin constraint for the open-set supervised anomaly detection setting. This weakly-supervised approach to anomaly detection is particularly relevant as it illustrated that using abnormal events during the training process could aid the detection of similar occurrences. The final

Table 2
Available source code of the surveyed methods.

Year	Model	Availability	ML Framework
2018	FFP	GitHub Repository (Liu, 2018)	TensorFlow (Abadi et al., 2015)
	Sultani	GitHub Repository (Kosman, 2021)	PyTorch (Paszke et al., 2019)
2019	MLEP	GitHub Repository (Liu, 2019)	TensorFlow (Abadi et al., 2015)
2021	RTFM	GitHub Repository (Tian, 2021)	PyTorch (Paszke et al., 2019)

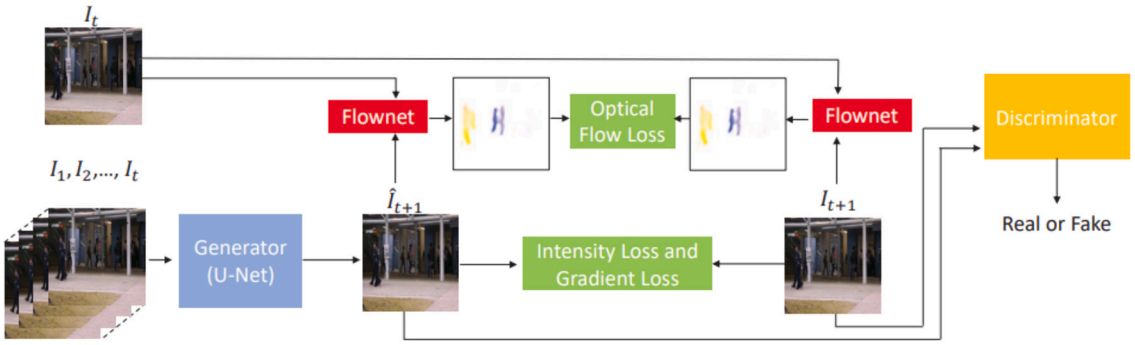


Fig. 1. Architecture of FFP, as originally presented by Liu et al. (2018).

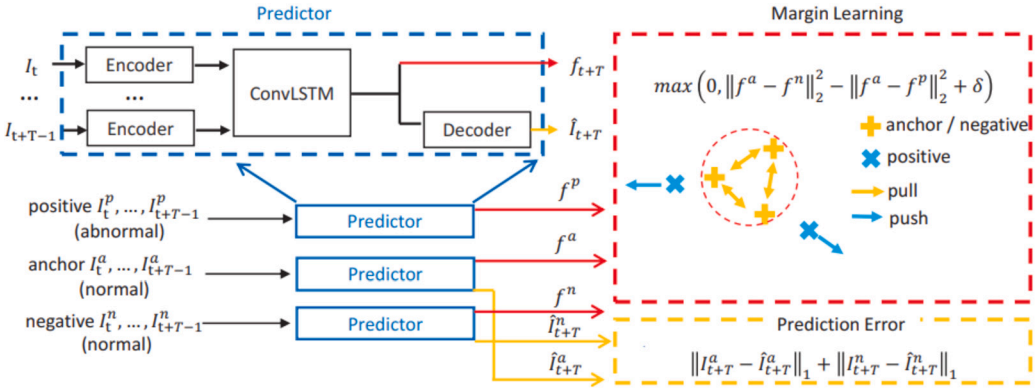


Fig. 2. Architecture of MLEP, as originally presented by Liu et al. (2019).

result is a video prediction framework that favours the prediction of normal events and distorts the prediction of abnormal ones, enlarging the margin between both. The proposed approach is summarized in Fig. 2. Additionally, the source code for this work is open source, and some of the pretrained models were made available in the official repository.

The existing networks that are used for video prediction can be split into three main categories: U-Net, as the one used in FFP (Liu et al., 2018), with its shortcuts that are often linked with an undesired prediction of abnormal events; a traditional Autoencoder (Mathieu et al., 2015) without the shortcuts between layers, which is not capable of properly encoding motion information; a Convolutional LSTM (Villegas et al., 2017), capable of using historical motion information, thus making it a proper candidate to provide information regarding anomalies in the training set. Luo et al. (2017a) have shown that combining two-dimensional convolutions with a Convolutional LSTM can encode both spatial and temporal information for action recognition. MLEP proposes to use a similar scheme, in which the first T frames of a sequence are encoded individually, and these features are fed into a Convolutional LSTM. Its output is fed into a decoder that produces the predicted frame corresponding to the instant T+1.

It is important to consider that only a few abnormal events are available to be used as part of the training set, and that many types of anomalies that could appear in the future are unseen. Hence, it is not sufficient to minimize the errors in the prediction of normal events, enlarging the margins between abnormal and normal events in the feature space is required. Liu et al. (2019) utilize margin learning for that purpose. Abnormal events can be annotated in two separate ways: either through video-level annotations, which indicate that a video contains an anomaly but does not specify its location; or through frame-level annotations, which indicate if a specific frame is normal or abnormal. MLEP can handle annotations of both types and mixtures of the two. For video-level annotation, a prediction network trained with only normal data is used to predict a normality score for each frame. The frames from abnormal videos that have a normalized score greater than 0.5 are added to the candidates of normal snippets, as a higher score indicates normality, similarly to FFP. On the other hand, if their scores are lower than that threshold, they are considered possible anomalies. For frame-level annotations, the provided labels are used instead. To illustrate the effectiveness of the proposed method, Liu et al. also trained a semi-supervised version of the proposed model, which served as the baseline against which the increased performance could be measured.

3.3. Sultani et al.

Sultani et al. (2018) proposed to learn anomalies by exploring both normal and abnormal sequences in surveillance videos. A Multiple Instance Learning scheme was utilised to avoid the need of providing frame-level labels of the clips. Instead, video-level training labels are leveraged to automatically learn a deep anomaly ranking model capable of predicting low anomaly scores for normal segments and high anomaly scores for abnormal ones. Moreover, sparsity and temporal

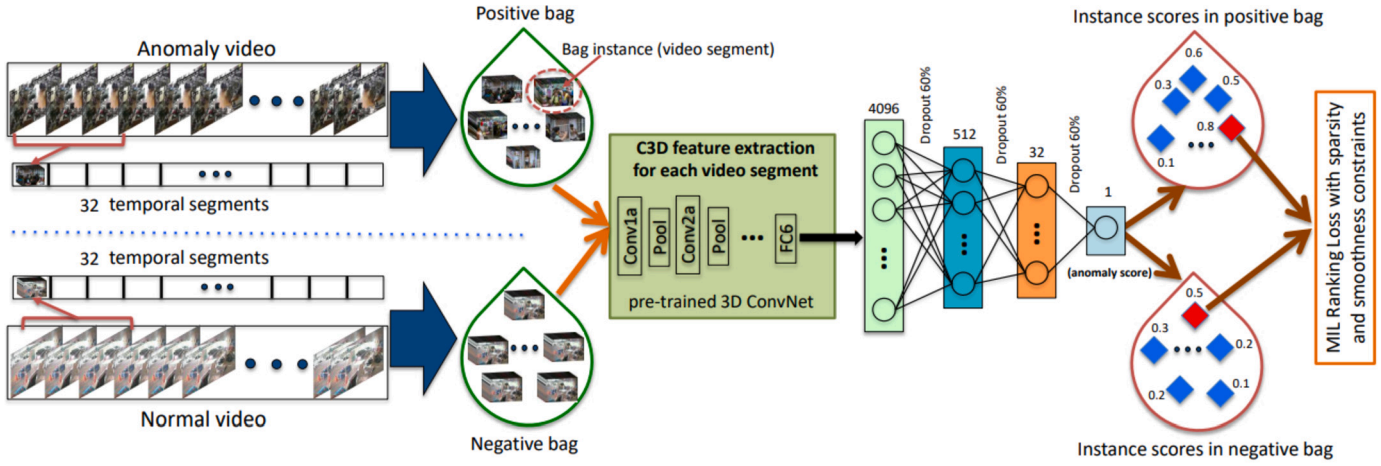


Fig. 3. Architecture of the network, as originally presented by Sultani et al. (2018).

constraints are introduced in the ranking loss function to improve the detection of anomalies during the training process. The experimental results showed significant improvements when compared to semi-supervised methods. These results will be fully explored in the benchmark. The source code for this project is open source and it consisted of a significant shift in deep anomaly detection for video sequences, thus making it an important method to analyse with more detail. The architecture of the proposed network is illustrated in Fig. 3.

Firstly, each video was divided into 32 non-overlapping temporal segments and the features that represent these video segments were used as bag instances. Given each video segment, the 3D convolution features for every 16-frame video clip in that segment were extracted, followed by an ℓ_2 normalization. To obtain the features that represent the video segment, the average of all 16-frame clip features within that segment was used. C3D (Tran et al., 2015) was the feature extractor that was used, due to its computational efficiency and ability to capture appearance and motion dynamics. As per the official recommendations, the features were extracted from layer `fc6`. These features were then input into a three-layer fully connected neural network that used them to produce an anomaly score for each segment.

3.4. RTFM

Tian et al. (2021) proposed Robust Temporal Feature Magnitude learning (RTFM) to address some of the issues that were found to be limiting the performance of the models that employed MIL to train a deep anomaly detection model. Firstly, the top anomaly score in an abnormal bag may not correspond to an abnormal snippet, and if a video contains more than one abnormal instance, more abnormal snippets per video are disregarded in the ranking process. Additionally, the normal snippets from normal sequences might not represent challenging situations, raising questions about training convergence. RTFM relies on the top-k instance MIL (Li & Vasconcelos, 2015) that trains the classifier using k instances with top classification scores from the abnormal and normal videos to address some of the aforementioned limitations. However, the proposed formulation assumes that the feature magnitude of the abnormal snippets is larger than the one verified in the normal snippets, instead of assuming the separability between the predicted anomaly scores. The source code for this project is open source, and several pretrained models are provided by the authors. Furthermore, its state-of-the-art performance makes it a very relevant method to evaluate in this benchmark. The architecture of RTFM is shown in Fig. 4.

The feature extraction process resembles the one applied by Sultani et al. (2018), with each video being divided into 32 non-overlapping segments. The features from every 16-frame sequence in each segment were then extracted and averaged, using two different feature extractors: C3D (Tran et al., 2015) and I3D (Carreira & Zisserman, 2017). For C3D, these features were extracted from layer `fc6`, while for I3D layer `mix5c` was chosen. According to notes from the authors found in the official GitHub Repository (Tian, 2021), a ten-crop augmentation was performed on the frames, i.e., the image was cropped into four corners and the central crop plus the flipped version of these. Hence, each segment is represented by a set of ten features, one from each crop.

3.5. Score normalization requirements

In FFP and MLEP, the anomaly score is not directly predicted as a value in the range [0,1] as it happens in the method proposed by Sultani et al. and in RTFM; in the latter, a lower score is more likely to represent a normal scene, whilst a higher score is linked to abnormal instances. Instead, FFP and MLEP evaluated anomaly by calculating the Peak Signal-to-Noise Ratio between the predicted frame and the ground truth one, as shown in Equation (1). A higher PSNR indicates that a frame is more likely to be normal than a frame that achieved a lower PSNR.

$$PSNR(I,\hat{I}) = 10\log_{10}\frac{[max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i-\hat{I}_i)^2} \tag{1}$$

As these results were difficult to interpret, a normalization strategy was proposed by the authors of FFP and MLEP. In FFP, after calculating the anomaly scores for each frame in a specific testing video, the scores were normalized to the range [0,1] in that specific video, as shown in Equation (2). In this Equation, A represents the normalized anomaly score for a given frame i contained in a video v.

$$A_{v,i} = -\left(\frac{PSNR_{v,i} - \min(PSNR_v)}{\max(PSNR_v) - \min(PSNR_v)}\right) + 1 \tag{2}$$

This strategy raises several questions about the applicability of this method. Firstly, it artificially introduces large score gaps between frames in a video without considering the magnitude of the original PSNR score gaps. To illustrate this concept, one could assume that two videos are being analysed, one containing only normal frames and another one containing both normal and abnormal frames. Moreover, in this example, the maximum PSNR score is 40 in both videos, while the minimum score is 37 in the normal video and 20 in the one that contains anomalous frames. Normalizing the scores with the proposed method results in relabelling both frames with the score 0. Semi-supervised strategies evaluate the performance of their models using only videos that contain anomalies. Therefore, this normalization relies on the assumption that at least one abnormal frame exists. If we choose any valid threshold lower than 1, at least one frame will be detected. Furthermore, defining a practical threshold becomes impossible, since the normalization is not generic, it is specific to each video. This raises

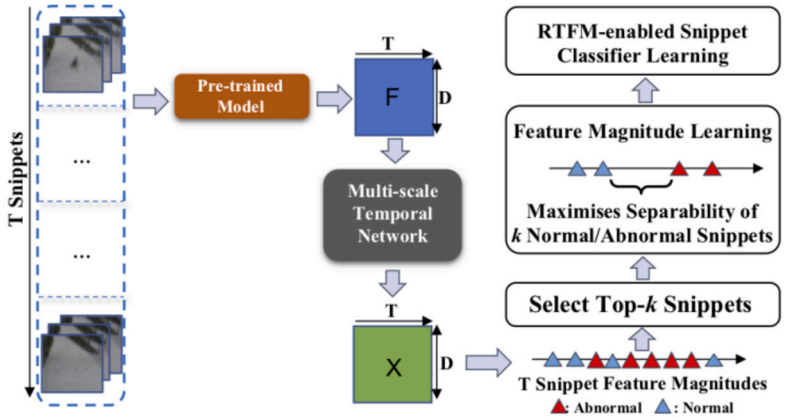


Fig. 4. Architecture of RTFM, as originally presented by Tian et al. (2021).

questions about the applicability of this method to real-world scenarios that were explored in more detail in our benchmark.

As far as MLEP is concerned, the normalization process is not applied to each video individually. It is instead applied on a per-scene basis, i.e., to all the frames captured under the same single surveillance camera, at a certain angle and point of view, as shown in Equation (3). In this Equation, A represents the normalized anomaly score for a given frame i contained in a video v and corresponding to a scene s. Despite being an improvement over FFP's normalization strategy, it assumes that the system cannot be applied to different backgrounds and scenes without, at least, requiring a calibration process, i.e., the expected range of PSNR scores is different for every scene.

$$A_{v,i} = -\left(\frac{PSNR_{v,i} - \min(PSNR_s)}{\max(PSNR_s) - \min(PSNR_s)}\right) + 1 \tag{3}$$

RTFM and the method proposed by Sultani et al. do not apply any form of normalization to the scores generated for the analysed frames. Therefore, these methods were designed to achieve a scene-independent anomaly detector, theoretically making them more robust than FFP and MLEP to new scenarios of application.

4. Methodology

This section describes in detail the methods and techniques in the benchmark. It clearly presents the metrics, leveraged data, and analysis structure. Furthermore, some potential limitations or biases in previous studies were addressed.

4.1. Metrics

The traditional method for evaluating and comparing deep anomaly detection methods relies heavily on the Receiver Operation Characteristic (ROC) curve, which illustrates the relation between True Positive Rate (TPR) and False Positive Rate (FPR) (Sultani et al., 2018, Liu et al., 2018, 2019, Tian et al., 2021). Evaluating all the points in the ROC curve would be inefficient, hence the area under the ROC curve provides an aggregate measure of performance across all possible classification thresholds. This metric can be easily interpreted as the probability that the model ranks a random positive example more highly than a random negative example. Although this metric provides some insight into the capabilities of the model, the false positive rate is an insufficient indicator for the false alarm rate. As Equation (4), where TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, demonstrates, FPR represents the ratio of negative samples that were mislabelled. However, for a highly unbalanced dataset, for instance, one with 10% of positive samples and 90% of negative ones, even if a perfect TPR was to be achieved, a typically low FPR of 11.1% means that the detector produced as many false alarms as correct ones.

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{TN + FP} \tag{4}$$

As the ROC curve and the corresponding AUC appear insufficient to attain a deep analysis of the effectiveness of certain models, an expansion of the typical benchmark of deep anomaly detection in videos is proposed. The Precision-Recall curve was leveraged to comprehend the impact of false alarms on the performance of the detector. As Equation (5) shows, Precision is a metric that quantifies the number of correct positive predictions made, by determining the ratio of True Positives amongst the samples that were labelled as positive. Similarly to the ROC curve, the area under the Precision-Recall curve is used as a scalar to compare different curves without applying an extensive analysis to a large set of thresholds. Additionally, the F1-score, the harmonic mean of the precision and recall, was also implemented as a way to unify both Precision and Recall as a function of the threshold. These metrics are seldom used to evaluate anomaly detection in videos but are common in benchmarks for other safety-critical applications, such as malicious attack detection in peer-to-peer smart grid platforms (Maseer et al., 2021). Accuracy generally describes how the model performs across all classes and is often used to evaluate the performance of classifiers. However, Equation (6) demonstrates that the traditional accuracy is not reliable for unbalanced datasets (Machado et al., 2022); balanced accuracy could have also been explored as an alternative metric capable of dealing with these unbalanced datasets (Jiang et al., 2014). If the previous example is considered, a model that classified all the samples as negative would achieve an accuracy of 90%, although it failed to detect any anomalies. Furthermore, a heuristic is proposed to automatically define an optimal threshold for the analysed methods, as the one that produces the highest F1-score in the testing set. With this heuristic, it was possible to generate a best-case scenario to draw a direct comparison between different methods. Additionally, defining and evaluating a strategy for automatically setting the optimal threshold is a desired feature that could support an easier deployment of these models.

$$Recall = TPR, \qquad Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F_1 = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}, \qquad Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

4.2. Datasets

The datasets used to train and test anomaly detectors for videos are made up of both normal and abnormal videos. Normal videos contain only normal frames, whereas abnormal videos contain both normal frames and frames that represent anomalies. The methods analysed in this benchmark present three strategies for splitting these videos between the training and testing sets:

- **Semi-Supervised Dataset:** during the training process, the datasets created for semi-supervised methods only use regular videos. As a

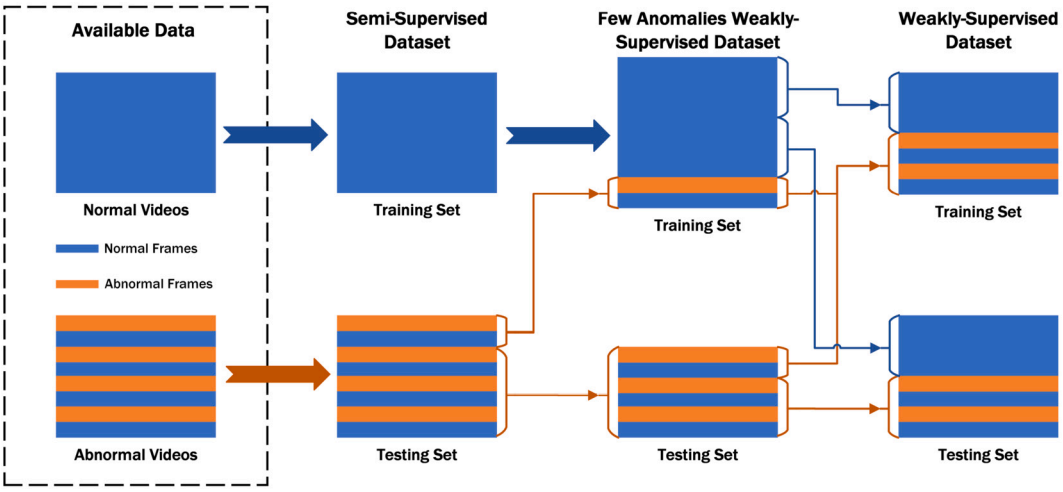


Fig. 5. Visual representation of the splits employed by the different types of Deep Anomaly Detection methods.

Table 3
Distribution of normal and abnormal frames in the testing sets of the datasets used in this benchmark.

Dataset	Semi-Supervised			Few Anomalies			Weakly-Supervised		
	Norm.	Abn.	Total	Norm.	Abn.	Total	Norm.	Abn.	Total
Avenue	68.4%	31.6%	15684	64.8%	35.2%	11751	-	-	-
ShanghaiTech	57.5%	42.5%	40791	56.5%	43.5%	32609	94.4%	5.6%	142912
UCF-Crime	-	-	-	-	-	-	92.4%	7.6%	1114144

Table 4
Overview of the approaches employed by the benchmarked models for leveraging the available datasets.

Model	Datasets			Input Type		Shape
	Avenue	Shanghai	UCF	RGB	Features	
FFP (Liu, 2018)	✓	✓		✓		(5,256,256,3)
MLEP (Liu, 2019)	✓	✓		✓		(5,256,256,3)
Sultani (Kosman, 2021)		✓	✓		✓	(1,4096)
RTFM (Tian, 2021)		✓	✓		✓	(10,2048)

result, the enforced data split is rather simple: normal videos comprise the training set, while abnormal videos comprise the testing set. Because CUHK Avenue and ShanghaiTech Campus were originally intended to be used with semi-supervised methods such as FFP, their original separation follows this process;

- **Few Anomalies Weakly-Supervised Dataset:** MLEP's approach, which resembles a typical weakly-supervised strategy, altered both datasets. The training set still includes all of the normal videos, but it also includes a few anomalous videos to train the detector. As a result, the testing set contains fewer videos, though the normal/abnormal frame ratio remains similar to the semi-supervised splits;
- **Weakly-Supervised Dataset:** datasets for training and evaluating weakly-supervised methods are essentially different from the previous splits. Normal and abnormal videos are included in both the training and testing datasets. As a result, normal frames dominate the training and testing sets, raising concerns about the applicability of some metrics to highly unbalanced datasets.

Fig. 5 depicts the process of splitting the datasets. Furthermore, for the testing sets of the three datasets used in this benchmark, the distribution of normal and abnormal frames for the three types of methods is summarized in Table 3. The training set is irrelevant because the benchmark only looked at the set used to evaluate the methods.

Table 4 presents information on how the four different benchmarked models use the datasets. Specifically, two of the models, FFP and MLEP, use the RGB frames directly to make predictions. On the other hand, the Sultani and RTFM models use extracted features with certain shapes to make predictions. These features can be obtained by following the instructions and download links provided in the official GitHub repositories for these models, that are also indicated in this table. There is less flexibility in already having the extracted features, as it is not possible to know exactly what is being analysed. Therefore, it becomes important to also assess the soundness of the feature extraction in future work.

4.3. Benchmark structure

The proposed benchmark was divided into two main stages. Firstly, the pretrained models are evaluated in the datasets they were trained on. Secondly, a cross-dataset evaluation was performed by applying the models trained on a certain dataset to a different one. This evaluation strategy is summarized in Table 5 which also presents the available models and the type of distribution used for each case. Table 5 presents the naming scheme that was used in the following Sections of the benchmark to refer to the models. A model's name is represented by the method on which it is based (FFP, MLEP, Sultani or RTFM) and the dataset it was trained on (Avenue, Shanghai or UCF). One of the MLEP models trained on CUHK Avenue leveraged temporal annotations in the training process, for which a *T* was added to its name; another exploited video-tuned labels, hence a *VT* was added to its name.

In Subsection 3.5, several questions were raised about the influence that the score normalization strategies employed by FFP (Equation (2)) and MLEP (Equation (3)) could have on the metrics used for evaluating the models. It was suggested that employing these strategies could

Table 5
Matrix representing the two stages of the proposed benchmark and the type of data distribution used for each case. The first stage is identified in blue and the second one in green. Abbreviations: SS = Semi-Supervised; FA = Few Anomalies; WS = Weakly-Supervised; T = Model leveraging Temporal Annotations; VT = Model leveraging Video-Tuned Labels.

Method	Model	Datasets		
		Avenue	Shanghai	UCF
FFP	Avenue	SS	SS	—
	Shanghai	SS	SS	—
MLEP	Avenue	SS	SS	—
	Shanghai	SS	SS	—
	Avenue T	FA	FA	—
	Avenue VT	FA	FA	—
Sultani	Shanghai	—	WS	WS
	UCF	—	WS	WS
RTFM	Shanghai	—	WS	WS
	UCF	—	WS	WS

mask the limitations that these models theoretically possess in offering scene-independent and coherent anomaly scores without requiring undesired calibration processes. To assess the possible influence of both normalization processes, an alternative normalization process was proposed in Equation (7), representing the anomaly score A for each frame i. It was implemented in the models following Algorithm 1. This process enables the study of a scene-independent strategy while making the anomaly scores coherent with the ones produced by the other methods by normalizing them in the desired range but considering the global maximum and minimum PSNR scores. In the following Sections of the benchmark, FFP's normalization strategy was designated as 'video normalization' or 'per-video normalization'; MLEP's normalization strategy was renamed as 'scene normalization' or 'per-scene normalization'; the proposed strategy was named 'global normalization'.

Algorithm 1 Global Normalization Strategy Algorithm.

Require: $PSNR$
  $MAX \leftarrow maximum(PSNR)$
  $MIN \leftarrow minimum(PSNR)$
  $A \leftarrow 0$
  for $SCORE$ **in** $PSNR$ **do**
    $A \leftarrow 1 - (SCORE - MIN)/(MAX - MIN)$
  end for

$$A_i = -\left(\frac{PSNR_i - \min(PSNR)}{\max(PSNR) - \min(PSNR)}\right) + 1 \tag{7}$$

For each of these stages, the evaluation compared the results achieved by the models on a per-dataset basis. The ROC curve of each model was plotted, and its AUC was determined, to set the traditional evaluation process for such methods. Moreover, the Precision-Recall curve and the F1-Score curve were plotted, and the former AUC was determined. For the threshold value that maximised the F1-score for each model, the associated precision and recall scores were leveraged to draw a comparison between the effectiveness of the different models. When required, figures illustrating the comparison between the predicted anomaly score and the ground truth for a selection of frames were used to demonstrate some particularities in the performance of a certain model. For CUHK Avenue the first 5000 frames are shown; for ShanghaiTech Campus the first 7500 frames were chosen; UCF-Crime demonstrates the scores produced for the first 40000 frames. These selections of frames were chosen empirically: they covered videos in the training sets that contained anomalies; they represented well the behaviour indicated by the metrics; larger datasets required larger selections of frames to illustrate the achieved results.

The results presented in the following sections were obtained through the use of open-source code made available in the referenced GitHub repositories. The recommended data samples or extracted features were used as input to feed the pretrained models, resulting in the generation of the reported results. To visualize the obtained data, `Matplotlib` (Caswell et al., 2023) was utilized to generate the necessary plots. The use of these tools ensures the reproducibility of the results.

5. Intra-dataset evaluation results

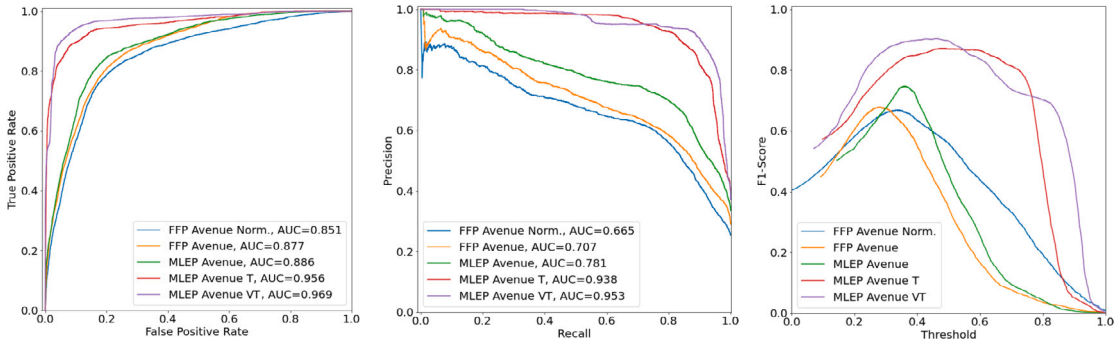
Intra-dataset evaluation is a common approach for assessing a model's ability to generalise to new, previously unseen data, and it entails testing the models on the test set of the dataset they were trained on. This section presents the results and analysis of the models performance on the available test sets.

5.1. CUHK avenue

FFP and MLEP showed promising results on CUHK Avenue, as shown by the results summarised in Table 6. Both models trained on normal data only achieved similar ROC curves, as Fig. 6a demonstrates. Despite this similarity, there is a clear advantage for the MLEP model in terms of Precision and Recall, with Fig. 6b showing that the area under this curve is 7.4 percentage points larger. The F1-score curves as illustrated in Fig. 6c show that higher F1-scores were achieved in MLEP when compared to FFP. In other words, it is possible to select an optimal threshold for the MLEP model such that both the Precision and Recall are greater than the ones achieved by the optimal threshold of the FFP model. The results contained in Table 6 show a notable improvement in the Precision score, with an increase of 9.98 percentage points, i.e., the rate of false alarms produced is lower. The consistent reduction of the number of false positives that were identified indicates that MLEP was able to learn normality better than FFP. Hence, the ConvLSTM-based predictor of MLEP was able to introduce significant improvements when compared to the one present in FFP.

By comparing the results achieved with FFP's per-video normalization with the ones achieved with the proposed global normalization, it is possible to conclude that the original normalization strategy did not result in an improvement for this dataset. There is even a marginal improvement in the performance when the per-video normalization was not applied. However, the improvement is not large enough to draw significant conclusions, besides per-video normalization not being significant for this evaluation. These results were expected since CUHK Avenue contains a single scene, consisting of the same camera position and angle and recording the same setting with similar light conditions. The scores per frame in Fig. 7 demonstrate that the scores generated with the per-video normalization and global normalization are very similar. Although the score gap increased when the per-video normalization strategy was applied, the noise spikes also increased, which could explain the slight difference in performance.

A shorter testing set was leveraged for the evaluation of the models trained with a few abnormal sequences of CUHK Avenue, as indicated in Table 3. Nonetheless, the data distribution is close to the original one, thus the results are comparable. The ROC curves of both models show a large increase in performance when compared to those trained using normal data only. The same trend is displayed by the Precision-Recall curves and the F1-Score curves in Figs. 6b and 6c, respectively. When compared to the optimal threshold of the best-performing model that leverages normal data only, the use of temporal information results in an increase of 8.1 percentage points in the Recall score and 15.9 percentage points for the Precision. The video-tuning strategy was the best performer, with an increase of 8.6 percentage points in the Recall and 22.2 percentage points for the Precision. These results are compiled in Table 6 and can be explained by Fig. 8, which demonstrates that the gap in between the scores of normal and abnormal instances is large for both models; however, this difference appears to be slightly larger



(a) ROC curves. (b) Precision-Recall curves. (c) F1-score curves.

Fig. 6. Curves of the proposed benchmark metrics obtained when testing the FFP and MLEP models trained on CUHK Avenue on the corresponding dataset.

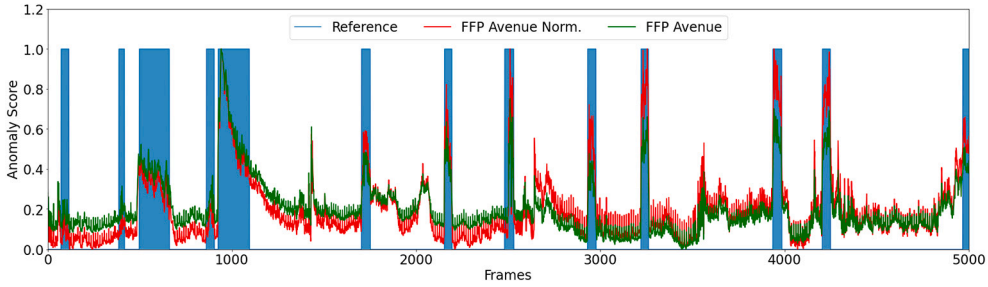


Fig. 7. Sample of the anomaly scores produced by the FFP model trained on CUHK Avenue for the first 5000 frames of the corresponding dataset.

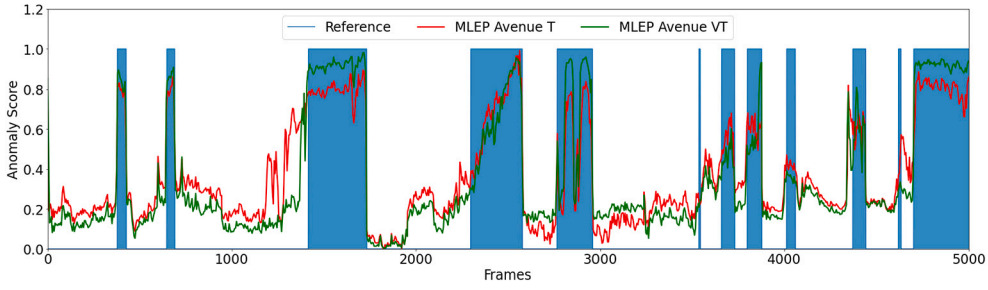


Fig. 8. Sample of the anomaly scores produced by the MLEP models trained on CUHK Avenue with some anomalies for the first 5000 frames of the corresponding dataset.

Table 6
Summary and comparison of the results achieved on CUHK Avenue for the benchmark of the models trained on the corresponding dataset.

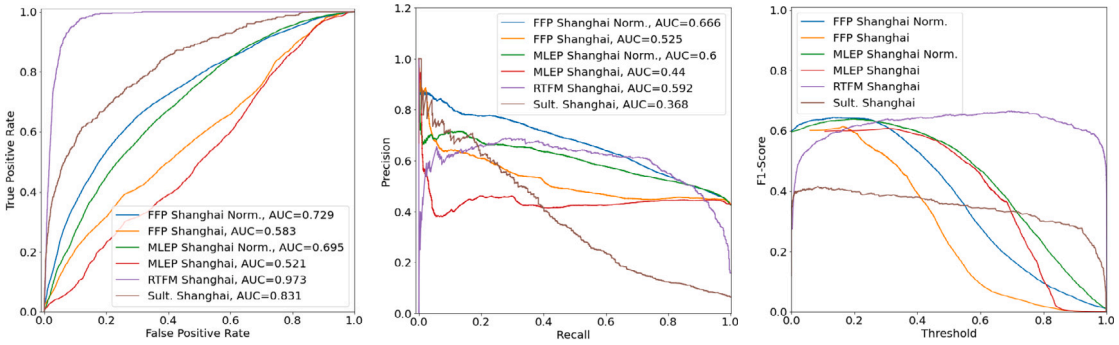
Model	Video Normalization					Global Normalization				
	AUC		Top	Prec.	Rec.	AUC		Top	Prec.	Rec.
	ROC	P-R	F1			ROC	P-R	F1		
FFP Avenue	0.851	0.665	0.669	0.599	0.757	0.877	0.707	0.678	0.603	0.775
MLEP Avenue	—	—	—	—	—	0.886	0.781	0.746	0.701	0.798
MLEP Avenue T	—	—	—	—	—	0.956	0.938	0.871	0.864	0.879
MLEP Avenue VT	—	—	—	—	—	**0.969**	**0.953**	**0.903**	**0.923**	**0.884**

for the model that leveraged video-tuned labels (VT). Therefore, the false alarm rate was highly reduced and more anomalies were correctly identified.

5.2. ShanghaiTech campus

The performance of the FFP and MLEP models on ShanghaiTech was underwhelming when compared to the results achieved by their CUHK Avenue counterparts, as Table 7 demonstrates. Both models trained using normal data only achieved similar ROC curves when their scores were normalized, as Fig. 9a demonstrates, with a slight advantage for the FFP model. The same tendency is observed in the Precision-Recall curves in Fig. 9b and for the F1-Scores in Fig. 9c. The optimal threshold produced by the proposed heuristic revealed that both models were able to correctly identify a large portion of the anomalies, 81.7% for FFP and 88.7% for MLEP, at the cost of producing many false alarms, with Precision scores of 53.2% for the former and 50.1% for the latter. The performance of these methods is even less competent when the data distribution of the testing set present in Table 3 is analysed. This table shows that 42.5% of the frames present in the training set correspond to abnormal instances, thus classifying all the instances as abnormal would produce a recall of 100% and a precision of 42.5%.

When a per-video or per-scene normalization was not applied, it was not possible to grant a consistent performance, as the area under



(a) ROC curves. (b) Precision-Recall curves. (c) F1-score curves.

Fig. 9. Curves of the proposed benchmark metrics obtained when testing the FFP, MLEP, Sultani and RTFM models trained on ShanghaiTech Campus on the corresponding dataset.

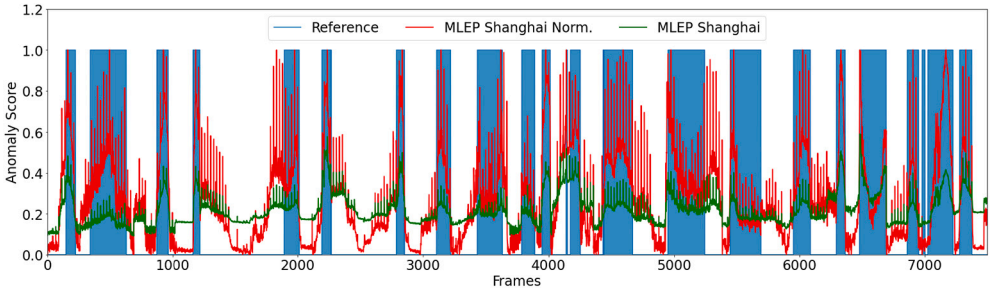


Fig. 10. Sample of the anomaly scores produced by the FFP model trained on ShanghaiTech Campus for the first 7500 frames of the corresponding dataset.

the ROC curves in Fig. 9a show and the results in Table 7 reiterates. Furthermore, Fig. 9b shows that the precision-recall curves of both models asymptotically converged to the classification of all data instances as abnormal, and the heuristic for producing an optimal threshold generated similar results. To illustrate the effects produced by the score normalization, the scores per frame for the FFP model are displayed in Fig. 10, demonstrating that it is impossible to discern anomalies from normality by looking at the generated scores when a per-video normalization was not applied. Additionally, even when the normalization was applied, the score gaps were low and prone to noise. It would have been relevant to test the improvements produced by the inclusion of temporal annotations and self-labelling in ShanghaiTech but the MLEP models were not provided and the source code that was made available by the authors was not compatible with their training.

The ROC curve in Fig. 9a corresponding to the model trained on ShanghaiTech and based on the architecture proposed by Sultani et al. shows that the area under this curve is much higher than the ones attained in FFP and MLEP, even if we consider the results obtained with a per video normalization strategy. However, directly comparing these results is not straightforward. Firstly, the testing sets are far too different, and the one that Sultani uses contains videos that were originally allocated by Luo et al. (2017a) to the training set. This results in a highly unbalanced distribution, as Table 3 demonstrates, containing 94.4% of normal frames and 5.6% of abnormal ones. This is significantly different from the 57.5% of normal frames and 42.5% of abnormal ones in the original distribution.

Although diluting abnormal instances in a pool of normality poses a more challenging evaluation scenario, it renders the ROC curve useless for defining an operating point and evaluating performance. The number of normal samples is vastly larger; hence, a low false positive rate might not be compatible with a high precision score. The precision-recall curve in Fig. 9b and the F1-score curve in Fig. 9c are consistent with this premise; despite what the area under the ROC curve, the reference benchmark, indicated, the precision scores are not higher than the ones attained by FFP and MLEP. On the other hand, the RTFM model generates satisfactory results in terms of the precision and recall achieved by the proposed heuristic, when compared to the ones achieved by the FFP and MLEP models in Table 7. Considering the more challenging data distribution and the fact that a normalization strategy was not applied, these results are indicators of a robust model. Nonetheless, an AUC of 97.3% for the ROC curve is not a good indicator of performance, especially when compared to what was achieved by the MLEP video-tuned model with a lower score in CUHK Avenue. As Fig. 11 shows, the anomaly scores per frame produced by Sultani et al. and RTFM illustrate two tendencies: the detection of anomalies in the former is not coherent, as it cannot consistently produce high scores for the identified anomalous segments; the latter struggles with the detection of the starting and end points of such events, generating false positives.

5.3. UCF-crime

The results for UCF-Crime reiterate the inadequacy of the ROC curve as a tool to evaluate the performance of anomaly detection methods for highly unbalanced testing sets. UCF-Crime's testing set comprises 92.4% of normal frames and 7.6% of abnormal ones, as Table 3 shows. The areas under the ROC curves displayed in Fig. 12a do not reflect the ability of the models to discern normality from abnormality. The Precision-Recall curves in Fig. 12b reveal that the models produce very low precision scores for acceptable recall scores, i.e., detecting at least half of the existing anomalies. The results in Table 8 produced by the proposed optimal threshold show that the RTFM model performs vastly better than the Sultani one. Nonetheless, it is still not capable of detecting more than 65.5% of the existing anomalies. Moreover, to achieve that recall rate, more than 75% of the produced alarms were false.

The anomaly scores per frame in Fig. 13 indicate that the model proposed by Sultani et al. did not appear to correctly and consistently identify both normality and abnormality. As for RTFM, the results reveal that the model frequently failed to classify normal instances as such. The main deficiency of this model was not its ability to produce

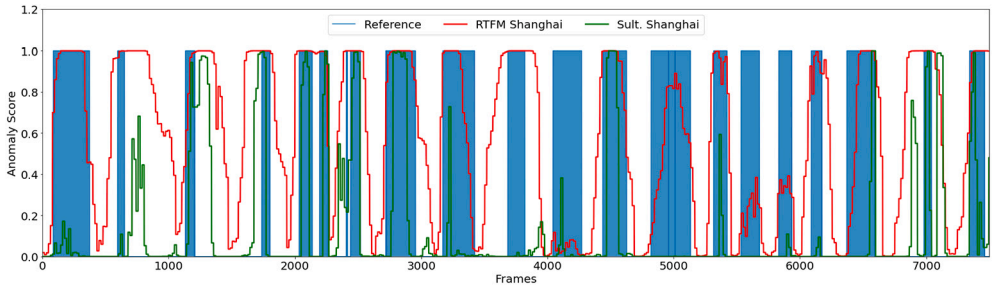
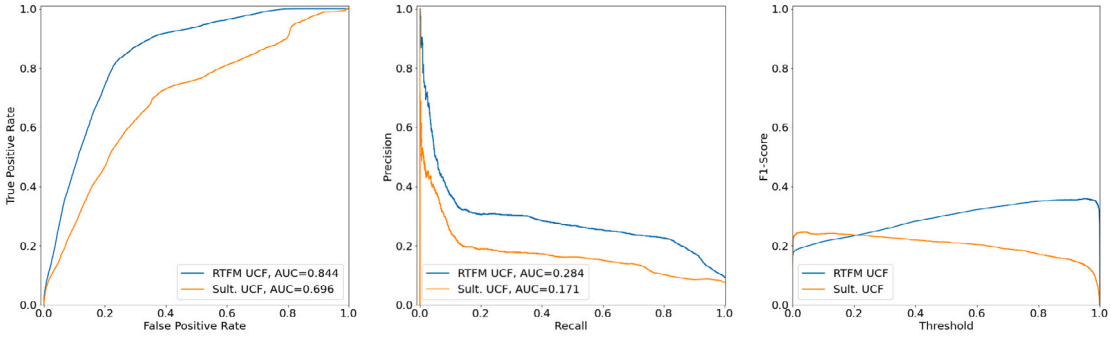


Fig. 11. Sample of the anomaly scores produced by the Sultani and RTFM models trained on ShanghaiTech Campus for the first 7500 frames of the corresponding dataset.

Table 7
Summary and comparison of the results achieved on ShanghaiTech Campus for the benchmark of the models trained on the corresponding dataset.

Model	Video/Scene Normalization					Global Normalization				
	AUC		Top	Prec.	Rec.	AUC		Top	Prec.	Rec.
	ROC	P-R	F1			ROC	P-R	F1		
FFP Shanghai	0.729	**0.666**	**0.644**	0.532	0.817	0.583	0.525	0.614	0.447	**0.979**
MLEP Shanghai	0.695	0.600	0.639	0.501	0.881	0.521	0.440	0.608	0.443	0.967
Sult. Shanghai	—	—	—	—	—	0.831	0.368	0.416	0.454	0.383
RTFM Shanghai	—	—	—	—	—	**0.973**	0.592	0.665	**0.605**	0.740



(a) ROC curves. (b) Precision-Recall curves. (c) F1-score curves.

Fig. 12. Curves of the proposed benchmark metrics obtained when testing the models trained on UCF-Crime on the corresponding dataset.

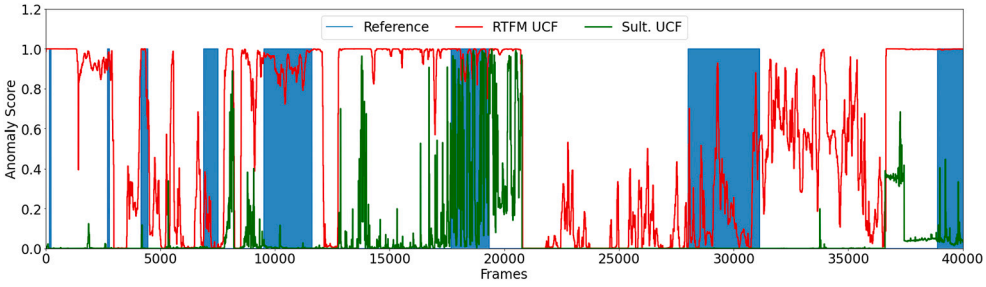


Fig. 13. Sample of the anomaly scores produced by the Sultani and RTFM models trained on UCF-Crime for the first 40000 frames of the corresponding dataset.

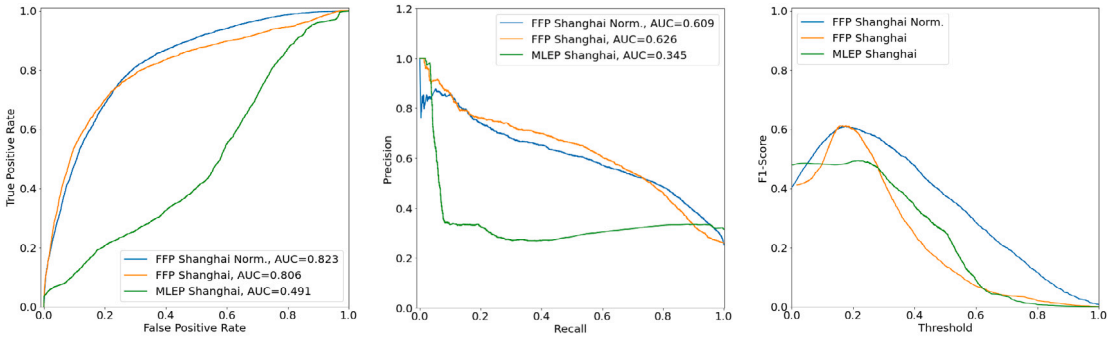
Table 8
Summary and comparison of the results achieved on UCF-Crime for the benchmark of the models trained on the corresponding dataset.

Model	AUC		Top	$t_{Precision}$	Recall
	ROC	P-R	F1		
RTFM UCF	**0.844**	**0.284**	**0.359**	**0.247**	**0.655**
Sult. UCF	0.696	0.171	0.247	0.162	0.521

high scores for abnormal instances, but rather producing low scores for normal ones. UCF-Crime, in a different approach from other anomaly detection datasets, limits abnormality to a set of 13 actions but does not define boundaries for normality. The model recognized these actions as Fig. 13 shows; however, as it never learnt a consistent pattern for normality, the scores for such sequences remain unpredictable.

6. Cross-dataset evaluation results

In the cross-dataset analysis, the results were evaluated and grouped based on the testing dataset that was used. By evaluating and grouping the results in this way, it is possible to understand the strengths and weaknesses of applying the various models to the different datasets; as



(a) ROC curves. (b) Precision-Recall curves. (c) F1-score curves.

Fig. 14. Curves of the proposed benchmark metrics obtained when testing on CUHK Avenue the FFP and MLEP models trained on ShanghaiTech Campus.

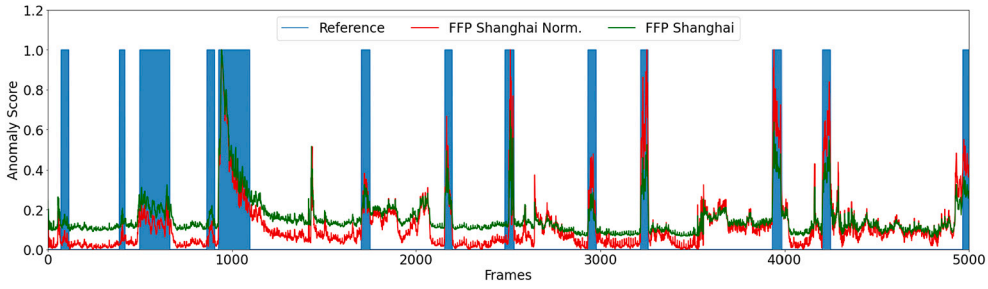


Fig. 15. Sample of the anomaly scores produced by the FFP models trained on ShanghaiTech Campus for the first 5000 frames of CUHK Avenue.

Table 9
Summary and comparison of the results achieved on CUHK Avenue for the benchmark of the models trained on ShanghaiTech Campus.

Model	Video Normalization					Global Normalization				
	AUC		Top	Prec.	Rec.	AUC		Top	Prec.	Rec.
	ROC	P-R	F1			ROC	P-R	F1		
FFP Shanghai	**0.823**	0.609	0.609	0.510	0.758	0.806	**0.626**	**0.612**	**0.545**	0.698
MLEP Shanghai	—	—	—	—	—	0.491	0.345	0.494	0.333	**0.953**

it mirrors the structure of the intra-dataset evaluation, it is also easier to draw comparisons to the models that were trained and tested on the same dataset.

6.1. CUHK avenue

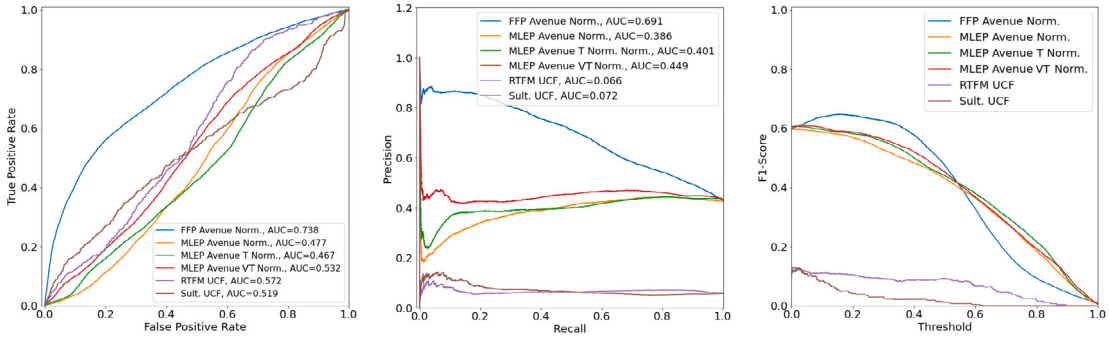
The FFP model trained on ShanghaiTech achieved ROC curves that are near the ones that characterise the modes trained on CUHK Avenue, as Fig. 14a shows. These results are valid for the scores that were normalized on a per-video basis and for the globally normalized ones. The analysis of the Precision-Recall curves in Fig. 14b and the F1-Score curves in Fig. 14c reveals that the model struggles with false alarms more than the model specifically trained on CUHK Avenue, which performed significantly better as shown in Figs. 6b and 6c. The results for the optimal threshold in Table 9 sustain this fact when compared to their counterparts in Table 6: with FFP's per-video normalization, Precision dropped 8.9 percentage points while Recall remained similar; the proposed global normalization strategy resulted in a decrease of 5.8 percentage points in the Precision score and 7.7 percentage points in the Recall. The analysis of the anomaly scores per frame in Fig. 15 demonstrates that although the distinction between normal and abnormal scores is clear, the score gap is smaller than the one in Fig. 7, which explains the performance decrease.

MLEP replaced the U-Net as the generator of the Generative Adversarial Network. Instead, it used three separate modules: an encoder, a Convolutional LSTM to extract temporal relations, and a decoder to output the frame prediction. The ROC curve for the MLEP model in Fig. 14a reveals that the proposed method cannot apply a model trained on a certain dataset to a different setting, at least with the referenced architecture for the generator. The other proposed metrics sustain this fact, revealing that the model does not distinguish normal and abnormal instances. These results were expected, as when the U-Net was replaced, the shortcut connections were lost, ending the direct connection between the encoder and decoder. The training process results in a bias of the encoder and decoder towards the dataset that was used. Hence, MLEP loses its generalization capabilities.

6.2. ShanghaiTech campus

Due to a large number of different models, for the ones that applied a per-video or a per-scene normalization strategy, only these results are displayed in Fig. 16. For FFP, the normalized results achieved by the model trained on CUHK Avenue and compiled in Table 10 rival those achieved by the model trained on ShanghaiTech Campus, producing very similar outcomes. Similarly, when a global normalization was not applied, the model could not sustain a consistent level of performance. As Fig. 16 shows, this was the only model that could achieve a reasonable performance.

For the MLEP-based models, the explanation for their inability to be applied to different datasets relies on their architecture, as it was previously explored. However, Sultani and RTFM could theoretically demonstrate some robustness in terms of scene independency. However, the pool of anomalies that the model was trained to recognize on UCF-Crime does not resemble the one that can be found in ShanghaiTech

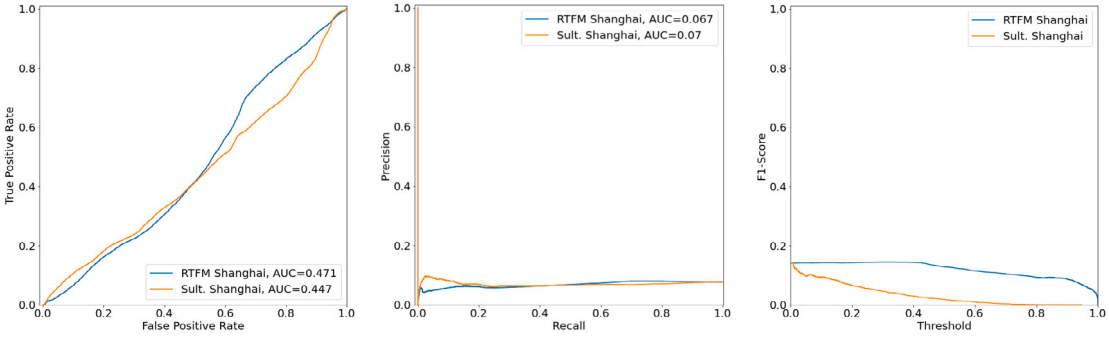


(a) ROC curves. (b) Precision-Recall curves. (c) F1-score curves.

Fig. 16. Curves of the proposed benchmark metrics obtained when testing on ShanghaiTech Campus the FFP, MLEP, Sultani and RTFM models trained on CUHK Avenue and UCF-Crime.

Table 10
Summary and comparison of the results achieved on ShanghaiTech Campus for the benchmark of the models trained on CUHK Avenue and UCF-Crime.

Model	Video Normalization					Global Normalization				
	AUC		Top	Prec.	Rec.	AUC		Top	Prec.	Rec.
	ROC	P-R	F1			ROC	P-R	F1		
FFP Avenue	**0.738**	**0.691**	**0.648**	**0.538**	0.812	0.485	0.430	0.609	0.438	0.998
MLEP Avenue	0.477	0.386	0.597	0.426	0.996	0.391	0.346	0.598	0.429	0.983
MLEP Avenue T	0.467	0.401	0.606	0.435	**1.000**	0.396	0.355	0.609	0.441	0.987
MLEP Avenue VT	0.532	0.449	0.608	0.443	0.972	0.432	0.388	0.611	0.441	0.992
RTFM UCF	—	—	—	—	—	0.491	0.345	0.130	0.070	0.862
Sult. UCF	—	—	—	—	—	0.519	0.072	0.128	0.109	0.156



(a) ROC curves. (b) Precision-Recall curves. (c) F1-score curves.

Fig. 17. Curves of the proposed benchmark metrics obtained when testing on UCF-Crime the Sultani and RTFM models trained on ShanghaiTech Campus.

Table 11
Summary and comparison of the results achieved on UCF-Crime for the benchmark of the models trained on ShanghaiTech Campus.

Model	AUC		Top	Precision	Recall
	ROC	P-R	F1		
RTFM Shanghai	**0.471**	0.067	**0.144**	**0.079**	0.782
Sult. Shanghai	0.447	**0.070**	0.143	0.077	**0.987**

Campus. Therefore, the models were not able to discern normal from abnormal sequences, as they did not learn the concept of normality and abnormality contextualized in this dataset.

6.3. UCF-crime

The results displayed in Fig. 17 show that neither model was capable of meaningfully discerning normal from abnormal instances. The reason for their lack of performance was explored in the previous analysis and is connected to the pool of anomalies that the models were trained to recognize during the training process. The behaviours that constitute an anomaly on ShanghaiTech are not the same that represent an anomaly in UCF-Crime, thus rendering the detector useless in such a context. (See Table 11.)

7. Discussion

The final goal of training an anomaly detection model is to implement it. For this purpose, during the training process, the model must learn how to operate in several conditions and scenarios, i.e., an optimal model would be universal. However, one could set a more realistic goal: to achieve a model that can generalize well in a context but in different scenarios, for instance, the detection of pedestrian anomalies in different areas of a city. To evaluate this capability, models trained on one dataset were applied to a different dataset. Additionally, on methods that normalize scores on a per-sequence or per-video bases, such as FFP and MLEP, a global normalization process was also evaluated.

FFP was able to generate competitive results in the cross-dataset test, although the global normalization strategy identified significant defi-

ciencies in the applicability of a single model to several settings, i.e., new camera positions and scenarios. The same can be said of MLEP, as the absence of a direct link between the encoder and decoder rendered the models useless when applied to other datasets rather than the ones they were trained on. It would be interesting to test a generator based on a U-Net, as the FFP results indicate that it is a reliable approach to near a competent generalization. However, the training process of methods trained on normal data only demonstrates an advantage when compared to models that learn to recognize certain events as anomalous, which is their ability to deal with novelty. As these models are bounded on normalcy only, when confronted with a novel anomaly, they remain unable to reconstruct those events.

The typical architecture of weakly-supervised methods is advantageous in terms of generalization, as it has been shown that they can generate anomaly scores independent of the camera position and the background setting. Moreover, the training process of semi-supervised methods is computationally more expensive, as it typically relies on adversarial strategies for frame prediction. The state-of-the-art benchmark techniques sustained that weakly-supervised methods could outperform prediction-based methods by a large margin. However, our benchmark found that the area under the ROC curve, the reference metric for comparing deep anomaly detection methods, is not a good metric for evaluating the performance of an anomaly detector, especially for highly imbalanced testing sets. In addition, the area under the precision-recall curve provides some insight into the effects of the dataset distribution. It could be leveraged as a scalar for comparing the effectiveness of different methods. The proposed heuristic for automatically setting a threshold yielded a satisfactory compromise between the achieved recall and precision scores. The results compiled in this benchmark show that the use of the area under the ROC curve as the reference metric led to the misconception that, in general terms, weakly-supervised methods always perform better. Since semi-supervised methods learn a model of normality from familiar events, by training the models with normal data only, this formulation results in a model that lacks knowledge of abnormal examples, but that is not biased towards the detection of a limited set of events. Weakly-supervised strategies propose a closed-set evaluation scenario, where training and test anomalies belong to the same action categories. This approach fails to consider the importance of novelty. Therefore, although improvements such as the ones proposed by Liu et al. (2019) in MLEP, and that robust scene-independent solutions should be explored, semi-supervised approaches are globally better at detecting abnormal patterns than weakly-supervised ones.

Generating a universal detector also proved complex. The results for UCF-Crime reveal a system that failed to learn normality patterns. The main limiting factor is not that the UCF-Crime dataset defined anomalies as a closed set; it is that it defined normalcy as an open set. This is a structural flaw in UCF-Crime that makes this dataset not appropriate for anomaly detection. An anomaly is an element that is different from what is expected; this set is unbounded and difficult to define. However, normality is not difficult to define because if it is expected, it must be known. Hence, traditional datasets limit normality to a closed set of actions that occur in a certain context. UCF-Crime takes the opposite approach and limits abnormality to a set of 13 actions. However, as there is not a defined context for the normal sequences, normality could be, as far as we know, any other actions besides the learnt 13 ones. Normality becomes unbounded and the scores of the model are unpredictable for those sequences. The dataset is challenging, but it is not fair, and it is not representative of real-world circumstances.

8. Conclusions

In this work, it could be concluded that the area under the ROC curve is not a good metric for evaluating the performance of an anomaly detector, particularly for highly unbalanced test sets. Furthermore, the area under the precision-recall curve sheds light on the impact of dataset distribution and it could be used as a scalar to compare the efficacy of various methods. The proposed heuristic for automatically defining a threshold revealed an acceptable compromise between the achieved recall and precision scores. The results compiled in this benchmark demonstrate that using the area under the ROC curve as the reference metric led to the misconception that weakly-supervised methods consistently outperform the semi-supervised ones.

The UCF-Crime dataset, commonly used as the reference benchmark for weakly-supervised methods, has a structural flaw that makes it an unsuitable dataset for anomaly detection. An anomaly is a feature that differs from what is expected. This set is illimitable and difficult to define. Normalcy, on the other hand, is not difficult to define because if it is expected, it must be known. That is why traditional datasets limit normalcy to a specific set of actions that occur in a particular context. UCF-Crime takes a different approach, limiting abnormality to a set of 13 actions. However, because there is no defined context for the normal sequences, normality becomes unbounded and the model's scores for these sequences are unpredictable.

9. Future work

In terms of future work, the implementation of method-agnostic solutions to reduce punctual false positives could prove essential to improve the performance of current methods without requiring any change in their architecture. One approach could be to apply a filter to the anomaly scores to mitigate instances that are likely to be false positives, thereby improving the overall accuracy of the system. Additionally, there is an opportunity to improve solutions based on MIL by leveraging temporal relationships within bags. By considering the temporal relationships between instances, the ranking process could achieve better results in selecting the potentially anomalous sequences within the instances contained in the bag. This strategy could be complemented with self-labelling techniques to improve the shortcomings of weakly-supervised approaches. Self-labelling has been shown to be effective in the context of MLEP and could also be a valuable approach to improve MIL.

Finally, an additional experiment could be conducted by rebuilding the current testing sets using only the videos that contain anomalies; these videos contain both normal and abnormal portions. This strategy could help to evaluate the ability of the methods to accurately localize the anomalous events within videos in the temporal domain, which is essential for a successful practical implementation of an anomaly detector.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was partially supported by projeto NEXUS – Investment project n.° 53 in the contexto of Agendas para a Inovação Empresarial (AAC n.° 02/C05-i01/2022) – a project suported by PRR – Plano de Recuperação e Resiliência and by NextGeneration EU European Funds.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. software available from https://www.tensorflow.org/.

Acsintoae, A., Florescu, A., Georgescu, M., Mare, T., Sumedrea, P., Ionescu, R. T., Khan, F. S., & Shah, M. (2022). Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4183–4192).

Caetano, F., Carvalho, P., & Cardoso, J. (2022). Deep anomaly detection for in-vehicle monitoring — an application-oriented review. *Applied Sciences*, *12*(19). https://doi.org/10.3390/app121910011. https://www.mdpi.com/2076-3417/12/19/10011.

Cai, R., Zhang, H., Liu, W., Gao, S., & Hao, Z. (2021). Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35* (pp. 938–946).

Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299–6308).

Caswell, T. A., Lee, A., de Andrade, E. S., Droettboom, M., Hoffmann, T., Klymak, J., Hunter, J., Firing, E., Stansby, D., Varoquaux, N., Nielsen, J. H., Root, B., May, R., Gustafsson, O., Elson, P., Seppänen, J. K., Lee, J.-J., Dale hannah, D., McDougall, D., ... Moad, C. (2023). matplotlib/matplotlib: Rel: v3.7.1. https://doi.org/10.5281/zenodo.7697899.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58.

Chen, C., Xie, Y., Lin, S., Yao, A., Jiang, G., Zhang, W., Qu, Y., Qiao, R., Ren, B., & Ma, L. (2022). Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In *Proceedings of the American association for artificial intelligence* (pp. 1–9).

Doshi, K., & Yilmaz, Y. (2020a). Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 934–935).

Doshi, K., & Yilmaz, Y. (2020b). Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 254–255).

Feng, J.-C., Hong, F.-T., & Zheng Mist, W.-S. (2021). Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*.

Georgescu, M.-I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021). Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12742–12752).

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Hengel, A. v. d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1705–1714).

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 733–742).

Ionescu, R. T., Khan, F. S., Georgescu, M.-I., & Shao, L. (2019). Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7842–7851).

Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.

Jiang, P., Missoum, S., & Chen, Z. (2014). Optimal SVM parameter selection for non-separable and unbalanced datasets. *Structural and Multidisciplinary Optimization*, *50*(4), 523–535.

Kosman, E. (2021). Pytorch implementation of real-world anomaly detection in surveillance videos. https://github.com/ekosman/AnomalyDetectionCVPR2018-Pytorch.

Lee, S., Kim, H. G., & Ro, Y. M. (2019). Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, *29*, 2395–2408.

Li, G., Cai, G., Zeng, X., & Zhao, R. (2022). Scale-aware spatio-temporal relation learning for video anomaly detection. In *Computer vision–ECCV 2022: 17th European conference, proceedings, Part IV*. Springer (pp. 333–350).

Li, S., Liu, F., & Jiao, L. (2022). Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *24*.

Li, W., & Vasconcelos, N. (2015). Multiple instance learning for soft bags via top instances. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4277–4285).

Liu, W. (2018). Future frame prediction for anomaly detection – a new baseline. https://github.com/StevenLiuWen/ano_pred_cvpr2018.

Liu, W. (2019). Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI 2019*. https://github.com/svip-lab/MLEP.

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6536–6545).

Liu, W., Luo, W., Li, Z., Zhao, P., & Gao, S. (2019). Margin learning embedded prediction for video anomaly detection with a few anomalies. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 3023–3030). ijcai.org.

Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 fps in Matlab. In *Proceedings of the IEEE international conference on computer vision* (pp. 2720–2727).

Luo, W., Liu, W., & Gao, S. (2017a). Remembering history with convolutional LSTM for anomaly detection. In *2017 IEEE international conference on multimedia and expo* (pp. 439–444). IEEE.

Luo, W., Liu, W., & Gao, S. (2017b). A revisit of sparse coding based anomaly detection in stacked RNN framework. In *Proceedings of the IEEE international conference on computer vision* (pp. 341–349).

Machado, A. P. F., Vargas, R. E. V., Ciarelli, P. M., & Munaro, C. J. (2022). Improving performance of one-class classifiers applied to anomaly detection in oil wells. *Journal of Petroleum Science & Engineering*, *218*, Article 110983. https://doi.org/10.1016/j.petrol.2022.110983. https://www.sciencedirect.com/science/article/pii/S0920410522008348.

Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly based intrusion detection systems in the cicids2017 dataset. *IEEE Access*, *9*, 22351–22370.

Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. arXiv preprint, arXiv:1511.05440.

Pang, G., Shen, C., Jin, H., & Hengel, A. v. d. (2019). Deep weakly-supervised anomaly detection. arXiv preprint, arXiv:1910.13601.

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, *54*(2), 1–38.

Park, H., Noh, J., & Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14372–14381).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library*. *Advances in neural information processing systems: Vol. 32*. Curran Associates, Inc. (pp. 8024–8035). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Sabokrou, M., Fathy, M., & Hoseini, M. (2016). Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, *52*(13), 1122–1124.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479–6488).

Thakare, K. V., Raghuwanshi, Y., Dogra, D. P., Choi, H., & Kim, I.-J. (2023). Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 5541–5550).

Tian, Y. (2021). Rtfm. https://github.com/tianyu0207/RTFM.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., & Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4975–4986).

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).

Ullah, W., Ullah, A., Haq, I. U., Muhammad, K., Sajjad, M., & Baik, S. W. (2021). CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimedia Tools and Applications*, *80*(11), 16979–16995.

Villegas, R., Yang, J., Hong, S., Lin, X., & Lee, H. (2017). Decomposing motion and content for natural video sequence prediction. arXiv preprint, arXiv:1706.08033.

Wan, B., Fang, Y., Xia, X., & Mei, J. (2020). Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo* (pp. 1–6). IEEE.

Wisdom, S., Powers, T., Pitton, J., & Atlas, L. (2016). Interpretable recurrent neural networks using sequential sparse recovery. arXiv preprint, arXiv:1611.07252.

Wu, P., & Liu, J. (2021). Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, *30*, 3513–3527. https://doi.org/10.1109/TIP.2021.3062192.

Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., & Kloft, M. (2020). Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 583–591).

Zhang, J., Qing, L., & Miao, J. (2019). Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE international conference on image processing* (pp. 4030–4034). IEEE.

Zhang, Q., Wei, H., Chen, J., Du, X., & Yu, J. (2023). Video anomaly detection based on attention mechanism. *Symmetry*, *15*(2). https://doi.org/10.3390/sym15020528. https://www.mdpi.com/2073-8994/15/2/528.

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., & Hua, X.-S. (2017). Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 1933–1941).

Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T. H., & Li, G. (2019). Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1237–1246).

Zhu, Y., & Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. arXiv preprint, arXiv:1907.10211.