Multi-Task Learning based Video Anomaly Detection with Attention

Mohammad Baradaran
Université Laval, Canada
mohammad.baradaran.1@ulaval.ca

Robert Bergevin
Université Laval, Canada
robert.bergevin@gel.ulaval.ca

Abstract

Multi-task learning based video anomaly detection methods combine multiple proxy tasks in different branches to detect video anomalies in different situations. Most existing methods suffer from one of these shortcomings: I) Combination of proxy tasks in their methods is not in a complementary and explainable way. II) Class of the object is not effectively considered. III) All motion anomaly cases are not covered. IV) Context information is not engaged in anomaly detection. To address these shortcomings, we propose a novel multi-task learning based method that combines complementary proxy tasks to better consider the motion and appearance features. In one branch, motivated by the abilities of the semantic segmentation and future frame prediction tasks, we combine them into a novel task of future semantic segmentation prediction to learn normal object classes and consistent motion patterns, and to detect respective anomalies simultaneously. In the second branch, we leverage optical flow magnitude estimation for motion anomaly detection and we propose an attention mechanism to engage context information in normal motion modeling and to detect motion anomalies with attention to object parts, the direction of motion, and the distance of the objects from the camera. Our qualitative results show that the proposed method considers the object class effectively and learns motion with attention to the aforementioned determinant factors which results in precise motion modeling and better motion anomaly detection. Additionally, quantitative results show the superiority of our method compared with state-of-the-art methods.

1. Introduction

With the growth of surveillance cameras, automatic analysis of video content is called for. Generally, the aim of this analysis is to detect anomalous events (*i.e*. unfamiliar or unexpected events in a given context [10, 20]) in the video which may demand instant action. Due to the rarity and diversity of anomalous events, adequate training anomaly samples are typically not available for supervised training. Hence, researchers in the field dedicated more interest to semi-supervised approaches, in which normals are learned via a proxy task (i.e., a task that indirectly helps to achieve the target goal), and anomalies are detected by finding the deviations from normalities. For example, reconstruction of current frames or prediction of masked frames are popular proxy tasks in video anomaly detection (VAD), in which the trained models on normals show a worse reconstruction or prediction result for anomalies, and the error of the estimation determines the anomaly score. Researchers have employed different proxy tasks in multiple branches to consider different modalities (mostly appearance and motion) in their approaches. Different proxy tasks are meant to be complementary to each other and consequently are combined towards higher performance. For example, Nguyen and Meunier [47] proposed a two-stream network in which one stream models the appearance features and detects appearance-based anomalies while the other one models motion patterns and looks for motion anomalies. Multiple similar strategies have been proposed and each work proposes a different combination of proxy tasks with different anomaly score fusion strategies [7, 16, 20, 26, 47, 57]. Recently, researchers (*e.g*. [11, 20, 29]) proposed to add more proxy tasks (*i.e*. multi-task learning based methods) to cover more spatio-temporal patterns. The key question in multi-task learning based methods is how many/what proxy tasks to choose in order to be complementary and to increase performance. Generally, adding more proxy tasks may result in better performance; however, it adds to the computational load and running time. Hence, the design goal is to propose the least number of complementary tasks, considering their abilities and shortcomings in the detection of several anomaly types, in order to cover all necessary attributes. It is worth noting that explainable anomaly detection requires having a strong explanation behind choosing each proxy task.

Although recent methods attain better results, they still either do not consider motion patterns thoroughly or do not explicitly analyze the class of the object for anomaly detection. To address these shortcomings, inspired by [20], we propose an improved multi-task learning based VAD

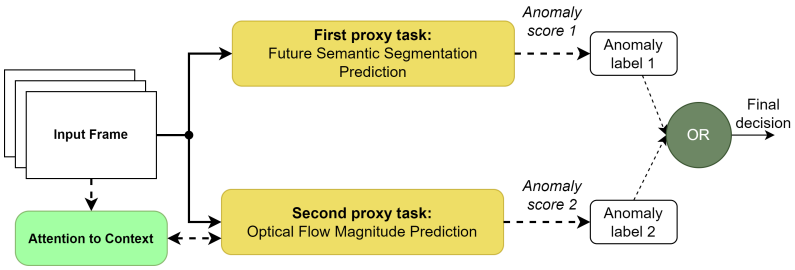

Figure 1. A general view of our proposed method.

method. Different from [20] and following the success of the semantic segmentation task in considering the object class in VAD [7], we benefit from the semantic segmentation proxy task's ability, in an improved way in the appearance branch. Additionally, contrarily to [20], which performs anomaly detection at the object level, we propose a holistic VAD method, avoiding their drawback of losing location information.

Nguyen *et al*. [47] proposed to estimate optical flow (OF) from a single frame to model motion patterns and to detect related anomalies. However, estimating OF from a single frame could be confusing for the motion network. To overcome this problem, Baradaran and Bergevin [7] proposed using optical flow magnitude (OFM) estimation for each object to detect motion anomalies. Although their method addresses the mentioned problem, their method neglects the motion direction information in motion estimation. Hence, their method does not effectively detect the anomalies which are due to sudden direction changes (such as in fighting, jumping, etc). Besides, to make a correspondence between each object and its motion magnitude (i.e., pixel-based object displacement through frames) some important factors, such as motion direction, object part, and distance of the object from the camera were not taken into account.

We propose a new method to address the previous issues. The proposed method leverages two different attention mechanisms. It takes advantage of a spatial and channel attention network and applies it to feature maps of the mid-layers in the encoder, which helps the network consider object parts (hands, feet, etc) for motion magnitude estimation. Moreover, a new attention network is designed that helps estimate the motion magnitude for each object with attention to its distance from the camera and the direction of its motion (details in Sec. 3). Finally, future frame prediction is leveraged, as another proxy task, to find sudden motion changes. In order to reduce the network size, the semantic segmentation and future frame prediction tasks are combined into a novel task of future semantic segmentation prediction to be performed by a single network.

In summary, our contributions are:

- A novel multi-task learning based video anomaly detection method that combines three complementary proxy tasks, "future frame prediction", "semantic segmentation", and "optical flow magnitude prediction", in an explainable way, to more generally consider appearance and motion features for anomaly detection.
- A combination of semantic segmentation and future frame prediction tasks into a novel proxy task to find both appearance and motion anomalies. To the best of our knowledge, this is the first work that introduces the future semantic segmentation prediction proxy task for video anomaly detection.
- A novel attention network to estimate precise motion magnitude for an object with attention to its motion direction and its distance from the camera. This introduces a novel way to engage context information in modeling normals and to detect respective anomalies. We also employ a spatial and channel attention mechanism in the backbone of the motion estimation branch to boost meaningful features and generate estimations specific to different object parts.

A general view of our proposed method is illustrated in Fig. 1.

2. Previous work

Researchers have formulated video anomaly detection via various proxy tasks especially frame reconstruction [1, 3, 13, 18, 19, 22–25, 28, 30, 37, 38, 40, 43, 45, 48, 51, 52, 59, 60, 64] or prediction tasks [4, 5, 12, 15, 32–34, 36, 39, 42, 44, 46, 54, 56, 58, 62, 63, 65, 67], assuming that the unsupervised network (*e.g*. A UNet) trained on normals generates a higher reconstruction/prediction error for anomalies. However, all previously mentioned methods consider low-level features (color, intensity, etc) for anomaly detection and do not explicitly consider the class of objects for their evaluation. Object-centric VAD methods [9, 17, 21, 26, 27, 53, 55, 61] detect and crop objects out of frame (by a pre-trained object detector) but they only consider low-level features in training and inference. Inspired by [8, 31], Baradaran and Bergevin [7] proposed a knowledge distillation based VAD that uses semantic segmentation as the proxy task and hence is able to explicitly consider the class of the objects for VAD.

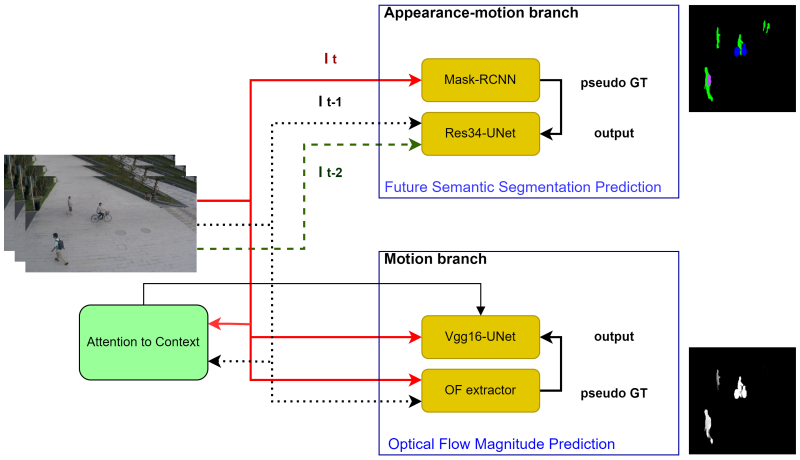

Figure 2. The pipeline of our proposed method. Mask-RCNN and OF extractors (as teachers respectively for the appearance-motion and the motion branches) provide the required pseudo-GT for training their students. During the inference stage, the output of each student network (estimation) is compared to its related pseudo-GT (expectation) to calculate the respective anomaly score. Best viewed in color.

Baradaran and Bergevin [6] report that single-branch approaches (such as [13, 24]) do not effectively cover all motion cases and are usually dominated by appearance features. Hence, to tackle the shortcomings of one-stream methods, researchers [7, 16, 26, 47] have proposed two-stream VAD methods, to detect motion and appearance effectively in separate branches. They mostly tackle the motion anomaly detection problem by reconstructing motion features (*e.g.* optical flow features, two-image gradients, etc.) [35]. One of the most noticeable related works is proposed by Neygun and Meunier [47] which formulates motion learning as a translation from the input frame to its corresponding optical flow map, trying to consider the correspondence between objects and their motion for the motion anomaly detection. Baradaran and Bergevin [7] proposed to translate the input frame to its optical flow magnitude (considering only the magnitude of motion), as they reported that the network can be confused while predicting the complete optical flow from a single frame. Although their approach addresses the issue of confusion and learns the correspondence between each object and its motion magnitude to detect related anomalies, their method neglects the direction information. Hence their method may fail in the detection of anomalies that are due to sudden direction changes. Moreover, they may fail in precise motion magnitude prediction since the perceived motion in frames is also a function of factors such as the distance from the camera (i.e. the same motion magnitude looks smaller farther) and also the direction of motion (objects moving parallel to the camera look faster than the same motion away from the camera), and these essential factors have not been taken to account in their method.

Inspired by the success of multi-task learning based methods in considering different aspects essential for anomaly detection, we propose an improved multi-task learning based VAD method with complementary proxy tasks to overcome the aforementioned shortcomings and to cover appearance and motion anomalies more effectively. Motivated by the success of future frame prediction and semantic segmentation prediction tasks, respectively in detecting sudden motion changes and object class aware appearance anomalies, we combine them into a single task and introduce the future semantic segmentation prediction as a novel proxy task for video anomaly detection. We also design attentive layers which learn motion considering essential context information and estimate motion with attention to the object part, motion direction, and distance of the object.

3. Method

We propose a multi-task learning based video anomaly detection method that leverages three self-supervised proxy tasks abilities in two separate branches in order to model normal patterns and consequently detect anomalies. The pipeline of the proposed method is illustrated in Fig. 2 and described in detail in the following.

3.1. Multi-task learning

Inspired by [20] we propose a multi-task learning based VAD method, which leverages three proxy tasks in two branches for video anomaly detection. The first branch (named appearance-motion branch) combines two different tasks (semantic segmentation and future frame prediction) to model appearance and motion simultaneously. The second branch (*i.e.* the motion branch) is in charge of learning the correspondence between each normal object and its normal motion magnitude, with attention to its distance from the camera, motion direction, and its body parts. In this

way, all three tasks are complementary to each other, each trying to find anomalies for which the other tasks may be sub-optimal.

3.2. The appearance-motion branch

Experiments in [7] show that leveraging semantic segmentation as a proxy task helps to effectively find the appearance anomalies considering the class of objects. Moreover, our experiments (related information is provided in Sec. 4) show that future frame prediction is a suitable task to detect sudden motion changes (*e.g*. direction changes, acceleration) since a network trained to predict the future frame of two consecutive normal frames fails to predict the precise location of the objects having sudden motion changes (such as in fighting, jumping, etc). To leverage the abilities of both tasks, the first branch of our method combines two different tasks of semantic segmentation and future frame prediction in a new single task and aims to predict the semantic segmentation map of the future frame by observing two consecutive frames. In this way, not only does it learn the class of the normal objects in the frame during the training, but it also learns the normal evolution between two normal frames. This branch follows a teacher-student strategy for anomaly detection. During training, a student (resnet34-UNet in our method) gets two consecutive frames and learns to generate the semantic segmentation map of the future frame (i.e., the next frame), assuming that at inference time, the prediction error would be higher for anomalies. The pseudo Ground-Truth (GT) for training the student network is generated by Mask-RCNN which is trained on MS COCO.

3.3. The motion branch

Although the first branch partially considers motion anomalies, it does not cover all motion cases. Hence, in the second branch, we employ the idea proposed in [7] which is to learn a normal motion magnitude for each object by translating an input frame to its optical flow magnitude map. This branch also follows a teacher-student strategy where the student (a vgg16-UNet here) learns to translate an input frame to its optical flow magnitude map, generated by a pre-trained optical flow extractor (considering its past frame) as a pseudo-GT. In this way, the student network learns the correspondence between each object and its normal motion magnitude during training, assuming that it will make an imprecise motion estimation for objects moving faster/slower than their normal motions. However, the original method has some challenges as follows: 1) some objects (such as humans) do not have a constant motion magnitude in all their body parts. For example, hands and feet usually have a larger motion magnitude compared to the chest and head. This factor has not been taken into account in [7]. 2) The motion magnitude perceived in frames (pixel displacement of objects) is a function of some variables such as motion direction and object distance from the camera. Objects moving parallel to the camera show a larger motion (i.e, generates a larger optical flow magnitude) compared to objects moving away/towards from/to the camera. Hence motion modeling and estimation of motion magnitude without attention to these factors would result in an imprecise prediction.

In our motion branch (Fig. 3), we employ two different attention mechanisms to address the mentioned shortcomings and engage essential factors in motion modeling to address the shortcomings in [7]. We employ a spatial and channel attention network in the main network, to dedicate more attention to special body parts (such as feet and hands) and we also design a new attention network to help the network make predictions with attention to supplementary information (such as motion direction and relative distance information). The details of the attention mechanisms are provided next.

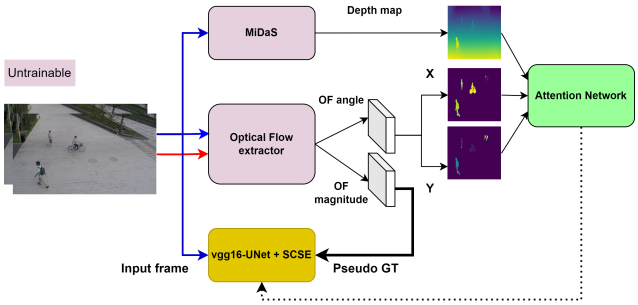

Figure 3. The motion branch in detail. Best viewed in color.

3.4. Spatial and channel attention

By visualizing the generated feature maps of the basic vgg16-UNet in the motion branch, we observed that the encoder of the vgg16-UNet (trained semi-supervisedly) generates different levels of features through different layers and the mid-layers generate feature maps that are activated for different object parts. Hence, we applied the spatial and channel attention (SCSE) mechanism proposed in [2] on the feature maps of mid-layers to help the network dedicate more attention to different body parts.

3.5. Attention to distance and direction

Normality is defined in context. Hence the essential point and challenge in semi-supervised anomaly detection methods is to model normal patterns precisely and to find anomalies by measuring deviations from normals. Previous methods do not consider important factors (such as direction and distance from camera) for motion modeling (and estimation). To provide attention to these factors and obtain a precise motion model for objects, we designed another attention network, as illustrated in Fig. 4. The network uses

the direction and distance information as inputs and generates an attention map to apply to attentive layers.

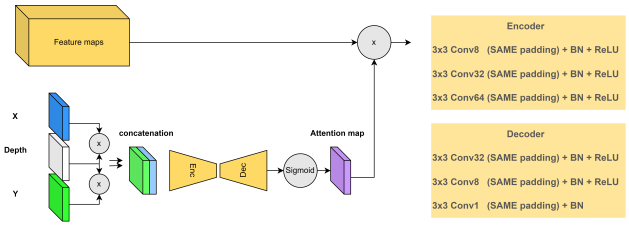

Figure 4. The proposed attention network. The generated attention map is applied to feature maps to provide attention to supplementary information (i.e., depth and direction). Best viewed in color.

In the motion branch, the teacher network extracts the optical flow of two consecutive frames (I_{t-1}, I_t) and provides the magnitude of the optical flow as the pseudo-GT to be utilized by the student to map the input frame (I_t) to its OFM. The generated OF features by the teacher (denoted as OF in Eq. (1)) encompasses the direction information in addition to magnitude information. In Eq. (1) `Mag` stands for the magnitude of the optical flow and `Ang` stands for the angle of the motion relative to the horizontal axis in the frame. The `Ang` features can be supplied to our attention network as direction features. However, we calculate the Cosine and Sine of `Ang` to normalize it and to generate two different features: motion parallel to the camera (denoted as X in Eq. (2) and Fig. 4) and motion towards/away to/from the camera (denoted as Y in Eq. (3) and Fig. 4).

$$Mag, Ang = OF(I_{t-1}, I_t) \tag{1}$$

$$X = |Cos(Ang)| \tag{2}$$

$$Y = |Sin(Ang)| \tag{3}$$

Since we do not have the actual information about the object's distance from the camera, we extract the depth maps of the frames to represent the relative distance information of the objects to the camera. We use MiDaS [49,50] to estimate the relative depth maps of input frames. MiDaS ensures high-quality depth map generation for a wide range of inputs since it is pre-trained on 10 different datasets using multi-objective optimization. We use the hybrid version of the method to balance precision and execution time. Figure 4 shows how the extracted informations are combined and processed inside the attention network to generate the attention map.

3.6. Inference

At the inference stage, we provide both normal and abnormal frames to each branch and we compare the estimation of each branch with their respective expectations (*i.e.* pseudo-GT generated by teachers of each branch) to calculate the anomaly map of that frame. For each of the two branches, we calculate the sum of activations in the anomaly map as the anomaly score S(t) of that branch for the frame (Eq. (4)).

$$S(t) = \Sigma|Out_{student}(I_t) - Out_{teacher}(I_t)| \tag{4}$$

In Eq. (4), $Out_{student}(I_t)$ and $Out_{teacher}(I_t)$ respectively denote the estimations and expectations of a given branch. Summation is done over all pixels in the anomaly map (*i.e.* the difference between student and teacher outputs).

In our experiments, we found some false positives which were due to two reasons: 1) false detections or misdetections by Mask-RCNN which produce large activations in the anomaly map. 2) jumps between frames which are apparently due to recording or saving issues. These jumps generate false large motions between some frames. The mentioned false positives produce sudden jumps/falls in the anomaly score of some frames. However, considering the frame rate of the video, we assume that adjacent frames should have a similar anomaly score. Hence, to relax the anomaly scores (*i.e.* temporal denoising), Savitzky–Golay filter (Eq. (5)) [14] is applied on the anomaly scores.

$$S_r(t) = \frac{1}{N}\sum_{i=-w}^{i=w} \alpha^{S(t+i)} \tag{5}$$

In this equation, $S_r(t)$ represents the relaxed anomaly score generated from noisy anomaly scores $S(t)$. N is the normalizing factor and α and w are the convolutional coefficients and window size respectively.

Finally, as the final decision, we flag a frame as an anomaly if and only if the anomaly score S(t) of either branch 1 or branch 2 (or both) is larger than a predefined threshold. As the networks of each branch have been trained with normal frames, we expect to observe a considerable difference between estimations and expectations if the input frame contains any anomaly specific to that branch. These anomaly maps are expected to contain activations at the position of the anomalies in the frame.

4. Experiments and results

We trained and evaluated the performance of our proposed method and the effectiveness of each contribution on the ShangahiTech Campus [32] and UCSD-Ped2 [41] datasets. The details of the experiments, qualitative and quantitative results, and a comparison with state-of-the-art approaches appear next.

4.1. Datasets

ShanghaiTech Campus and UCSD-Ped2 datasets are two of the benchmark datasets popularly used to evaluate

semi-supervised VAD methods. They provide only normal frames in their training subsets and both normal and abnormal frames in their test subsets, along with frame-based and pixel-based annotations. The definition of normality and anomaly is similar in both datasets. People walking on the sidewalk (could be carrying bags or backpacks) are considered normal, however, the presence of some previously unseen objects (such as bikes, bicycles, cars, etc) or some previously unseen motion patterns (such as running, chasing, fighting, riding, etc) are considered as anomalies. Compared to UCSD-Ped2, ShanghaiTech Campus is a more complex dataset, since it has multiple different scenes (13 scenes) and a larger number of anomalies. On the other hand, low resolution and gray scale frames make the UCSD-Ped2 dataset challenging and prone to failures for the segmentation task.

4.2. Evaluation metric

Following the state-of-the-art (SOTA) methods in the field, we provide our quantitative evaluation by measuring the frame-level AUC (Area Under Curve). This curve is plotted by registering multiple True Positive Rate (TPR) and False Positive Rate (FPR) of the method by changing the anomaly score threshold from min to max. A higher AUC indicates better performance.

4.3. Implementation details

Input frames in our experiments are resized to 256*256 for each branch. For a faster convergence, we initialize encoders of both student networks (res34-UNet and vgg16-Unet) with parameters of the networks trained on Imagenet. The learning rates of both branches are initialized to 0.001 and are halved every 10 epochs. We trained networks of both branches with the patch-based MSE loss (Eq. (6)) [7] (dividing the input frame into 16 patches) by the Adam optimizer. We employed Mask-RCNN (pretrained on MS COCO) as the teacher of the appearance branch and the Farneback algorithm from the OpenCV library as the teacher for pseudo-GT optical flow feature extraction from two consecutive frames. Finally, to discard the background and to bring more attention to the foreground objects, we mask each extracted optical flow map with the corresponding semantic segmentation map.

$$\text{Patch-loss} = \max(Loss_i) \tag{6}$$

where:

$Loss_i$ is the MSE loss in i_{th} patch in the frame.

4.4. Future frame prediction

To explore the abilities of future frame prediction proxy tasks in more detail, we conducted a preliminary experiment. We trained resnet-UNet to estimate future frames by observing two consecutive frames. The qualitative results

Figure 5. (a): Predicted future frame. (b): Actual future frame. (c): Difference between the predicted and actual future frame.

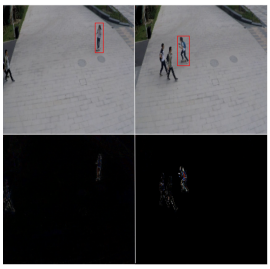

Figure 6. Future frame prediction task's ability in detecting fast motions vs detecting sudden direction changes. (a) Future frame prediction fails in the detection of abnormal motion with constant speed. (b) Future frame prediction task detects abnormal motions with sudden motion changes. The brightness of the anomaly maps has been increased slightly to better show the activations.

(Figure 5) demonstrate the ability of the future frame prediction proxy task in finding the sudden motion changes. As seen in this figure, the estimated and the real future frames are different in the position of chasing persons (hence producing larger activations in the anomaly map). However, the network is able to generate a precise prediction, at the position of objects with normal motion. It is worth noting that, as this proxy task is trained on normal speed motions it can find abnormal fast motions as well. However, due to the high capacity of CNNs, the future prediction ability can be generalized to these anomalies too. In other words, the CNN can predict the precise position of fast objects if they move monotonically fast through frames and don't show sudden changes in direction. In Figure 6(a), a skate rider travels fast but uniformly through adjacent frames, hence the CNN does not produce a high anomaly activation in that position. However, as she suddenly puts her foot on the ground to accelerate (Figure 6(b)), the generated anomaly activation is comparatively higher. In our proposed method, this shortcoming is handled by the OFM prediction proxy task.

4.5. Qualitative evaluation

To qualitatively analyze the effectiveness of the proposed method, we observe the anomaly maps for multiple normal and abnormal frames. Figures 7 and 8 present qualitative results for the appearance-motion and motion branches, respectively. Figure 7 contains multiple samples of anomalous frames (top row) and corresponding anomaly maps (bottom row), generated by the appearance-motion branch.

Figure 7. Qualitative results for the appearance-motion branch. Top: Input frames. Bottom: Anomaly maps. Anomalies are indicated by red bounding boxes.

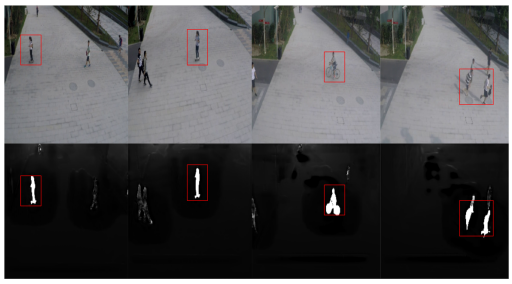

Figure 8. Qualitative results for the motion branch. Top: Input frames. Bottom: Anomaly maps. Anomalies are indicated by red bounding boxes.

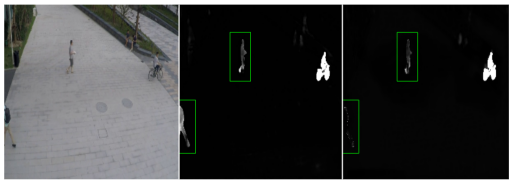

Figure 9. Attention in motion estimation. (a) Input frame. (b) Anomaly map without attention. (c) Anomaly map with attention. Normal moving objects are indicated by green bounding boxes.

As can be seen in the figure, anomaly maps contain higher activations at anomalous objects (bicycle and motorcycle in columns 1 and 2) or even at normal objects with sudden abnormal motions (legs and hands of a fighting man in column 3 and for running or chasing persons in columns 4 and 5).

Similar results can be observed in Fig. 8 for the motion branch. As can be seen, our method generates larger activations for objects with anomalous motions since the estimation of the motion branch is considerably different from its expectation for an anomaly.

4.5.1 Importance of attention

Figure 9 presents an anomalous frame (9a) and the generated anomaly map with and without applying the attention mechanisms. As can be noticed, the generated anomaly map contains weaker activations at the position of normal moving objects (green bounding boxes) in presence of attention compared to without attention. It demonstrates that the motions of normal objects are more precisely estimated when the attention mechanisms are active. For example, for the pedestrian close to the camera (the lower green bounding box in Figure 9b) the basic network does not take the distance information into account and estimates a smaller motion compared to its pseudo-GT which leads to a bigger difference between estimation and pseudo-GT (and hence a larger activation in the anomaly map compared to when the attention mechanisms are active (Figure 9c)). Additionally, we observe a larger activation at the feet of the normal moving object in the upper green bounding box when the attention module is not active. In both cases, the attention mechanisms reduce the likelihood of false anomaly detection.

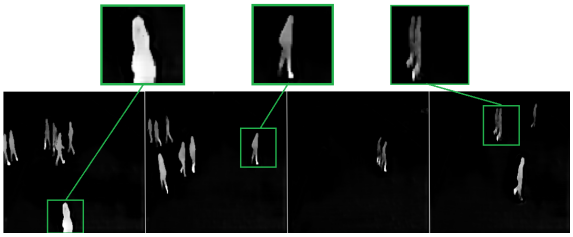

Figure 10. Estimated outputs of the motion branch with attention.

Figure 10 shows the estimations of the motion branch with attention. As can be observed, the network is aware of the object parts and estimates a larger motion for the feet, compared to other parts. Moreover, the estimated motion is larger for objects close to the camera.

4.6. Quantitative evaluation

Table 1 compares the performance of our proposed method with state-of-the-art (SOTA) holistic semi-supervised VAD methods on two benchmark datasets (ShanghaiTech Campus and UCSD-Ped2). The comparison is based on frame-level AUC. To have a fair comparison, we kept the settings of the original papers.

As can be seen, our method provides superior results compared to SOTA holistic semi-supervised VAD methods. Noticeably, the performance improvement (compared to previous methods (especially [7])) is more significant for the ShanghaiTech Campus compared to the UCSD-Ped2. This can be explained by the fact that precise motion modeling is more crucial in ShanghaiTech Campus as it has objects at different distances from the camera and motions in various directions, compared to UCSD-Ped2 which is limited in terms of motion directions and distances from the camera. Hence, our contributions in motion modeling (i.e. engaging context information by attention mechanisms) bring larger improvements for ShanghaiTech Campus, confirming our method successfully addresses the identified limitations of previous methods.

The above qualitative and quantitative results are sufficient to show our goal is attained. Nevertheless, it could

Method	ShanghaiTech Campus	UCSD Ped2
Hasan *et al.* [24]	60.9	90.0
Chong *et al.* [13]	N/A	87.4
Dong *et al.* [15]	73.7	95.6
Liu *et al.* [32]	72.8	95.4
Liu *et al.* [33]	N/A	87.5
Park *et al.* [48]	70.5	97.0
Gong *et al.* [23]	71.2	94.1
Nguyen *et al.* [47]	N/A	96.2
Tang *et al.* [57]	73.0	96.3
Baradaran *et al.* [7]	86.18	97.76
Georgescu *et al.* [20]	83.5	92.4
Luo *et al.* [37]	73.0	95.0
Yu *et al.* [66]	73.0	95.0
Ours	**89.1**	**97.8**

Table 1. Performance comparison (frame-level AUC) with SOTA methods. The best-performing method is denoted in boldface.

be interesting to further assess the generality of our method with experiments on other datasets and one possibility we investigated is the Avenue dataset, which is commonly used in the field. Although comparable with SOTA methods, the results obtained are not as meaningful as those in Table 1, for two main reasons: 1) concept of anomalies (based on the available annotations) in the Avenue dataset is not totally compatible with the goal and formulation of our method, resulting in a number of invalid true positives and false positives. 2) low camera height results in the generation of considerable scene occlusions with the ensuing difficulty in properly estimating the motion direction, which is needed to address the limitations of previous methods.

4.7. Ablation study

Tables 2, 3, and 4 confirm the effectiveness of each proposed contribution in increasing the obtained performance. As can be noticed in Table 2, all proxy tasks are complementary to each other and can detect more anomalies when combined. Most importantly, we observe that by adding the future frame prediction task to our method, it detects more motion anomalies compared to just using the OFM.

Table 3 demonstrates quantitatively the contribution of attention mechanisms in the proposed method. It is worth noting that in order to concentrate on the contribution of each attention mechanism in improving the performance of motion modeling and motion anomaly detection, we have conducted this ablation study only on the motion branch. As shown in the results, adding attention mechanisms to the motion branch results in higher performance.

Finally, we conducted another ablation study (Table 4) to analyze the effect of the position of attention maps on the performance of the motion branch. To concentrate on the importance of this factor, we deactivated the appearance-motion branch and the SCSE mechanism in this study. This table indicates that adding attention at any position in the network (encoder, decoder, skip connection, or the final layer) improves performance. However, a higher performance has been observed for the encoder, decoder, and final layer positions, compared to the skip connection.

Tasks	Seg	OFM	Seg+Pred	Seg+OFM	Seg+OFM+Pred
AUC	76.3	79.19	80.61	88.21	**89.1**

Table 2. Contribution of each proxy task (Seg: semantic segmentation, OFM: optical flow magnitude, Pred: prediction) in increasing the performance.

Model	UNet	UNet+Att	UNet+Att+SCSE
AUC	69.7	78.14	**79.19**

Table 3. Contribution of each attention mechanism (Att: attention to supplementary information) in increasing the performance of the motion branch, SCSE: spatial and channel-wise attention [2]. In this ablation study only the motion branch is active (the appearance-motion branch is deactivated).

Position	No map	Encoder	Decoder	Skip connection	Final layer
AUC	69.7	77.91	**78.14**	72.07	77.52

Table 4. Position of attention map. This table shows how the position of attention map integration affects the performance of the motion branch. In this ablation study, only the motion branch with the context attention is active. (The appearance-motion branch and SCSE mechanism are deactivated).

5. Conclusion

We proposed an improved multi-task learning based video anomaly detection method, which introduces future semantic segmentation prediction as a novel proxy task for video anomaly detection and combines multiple complementary proxy tasks for a better consideration of appearance and motion anomalies. Moreover, we introduced a novel mechanism to improve the precision of motion modeling with attention to context. Experimental results show that adding each proxy task results in higher performance in terms of AUC. Importantly, experimental results confirm that our proposed idea of paying attention to both direction of object motion and the distance of the object from the camera introduces a new and meaningful way to engage context information in video anomaly detection and results in more precise motion estimation and likely fewer false detections. Our qualitative results show the explainability of estimations and detections and also the effectiveness of each contribution. Quantitative results on the ShanghaiTech Campus and UCSD-Ped2 datasets demonstrate the superior performance of our method, compared to SOTA methods.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–490, 2019. 2

[2] Roy AG, Navab N, and Wachinger C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Transactions on Medical Imaging*, pages 540–549, 2019. 4, 8

[3] S. Akcay, A. Atapour-Abarghouei, and T.P. Breckon. "ganomaly: Semi-supervised anomaly detection via adversarial training". *Jawahar C., Li H., Mori G., Schindler K. (eds) Computer Vision – ACCV. Lecture Notes in Computer Science*, 11363, 2018. 2

[4] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 2

[5] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 2

[6] Mohammad Baradaran and Robert Bergevin. A critical study on the recent deep learning based semi-supervised video anomaly detection methods. *arXiv preprint arXiv:2111.01604*, 2021, unpublished. 3

[7] Mohammad Baradaran and Robert Bergevin. Object class aware video anomaly detection through image translation. *19th Conference on Robots and Vision (CRV)*, 2022. 1, 2, 3, 4, 6, 7, 8

[8] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. "uninformed students: Student-teacher anomaly detection with discriminative latent embeddings". *CVPR2020*, 2020. 2

[9] K. M. Biradar, A. Gupta, M. Mandal, and S. K. Vipparthi. "challenges in time-stamp aware anomaly detection in traffic videos". *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 2

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), July 2009. 1

[11] Xingya Chang, Yuxin Zhang, Dingyu Xue, and Dongyue Chen. Multi-task learning for video anomaly detection. *Journal of Visual Communication and Image Representation*, 87:103547, 2022. 1

[12] D. Chen, P. Wang, L. Yue, Y. Zhang, and T. Jia. Anomaly detection in surveillance video based on bidirectional prediction. *Image and Vision Computing*, 98, 2020. 2

[13] Yong Shean Chong, Yong Haur Tay, Fengyu Cong, Andrew Leung, and Qinglai Wei. Abnormal event detection in videos using spatiotemporal autoencoder. In *Advances in Neural Networks - ISNN 2017*, pages 189–196, Cham, 2017. Springer International Publishing. 2, 3, 8

[14] Wu Chongke, Shao Sicong, Tunc Cihan, Satam Pratik, and Hariri Salim. An explainable and efficient deep learning framework for video anomaly detection. In *2021 Cluster computing*, pages 1–23, 2021. 5

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

[16] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4037–4042, 2020. 1, 3

[17] K. Doshi and Y. Yilmaz. "any-shot sequential anomaly detection in surveillance videos". *CVPR*, 2020. 2

[18] E. Duman and O. A. Erdem. Anomaly detection in videos using optical flow and convolutional autoencoder. *IEEE Access*, 7:183914–183923, 2019. 2

[19] T. Ganokratanaa and N. Sebe S. Aramvith. "unsupervised anomaly detection and localization based on deep spatiotemporal translation network". *IEEE Access*, 8:50312–50329, 2020. 2

[20] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12742–12752, June 2021. 1, 2, 3, 8

[21] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4505–4523, 2022. 2

[22] E. Gherbi, B. Hanczar, J. Janodet, and W. Klaudel. "an encoding adversarial network for anomaly detection". *Proceedings of The Eleventh Asian Conference on Machine Learning, in PMLR*, 101:188–203, 2019. 2

[23] Dong Gong, Lingqiao Liu, Le Vuong, Budhaditya Saha, Moussa Mansour, Svetha Venkatesh, and Anton Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, 10 2019. 2, 8

[24] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016. 2, 3, 8

[25] R. Hinami, T. Mei, and S. Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3639–3647, 2017. 2

[26] Radu Tudor Ionescu, Fahad Khan, Mariana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7834–7843, 06 2019. 1, 2, 3

[27] Z. Jianfei, Z. Yi, S. Pan, Y. Zhao, Z. Zhao, F. Su, and B. Zhuang. "unsupervised traffic anomaly detection using trajectories". *CVPR Workshops*, 2019. 2

[28] M. Kimura and T. Yanagihara. "anomaly detection using gans for visual inspection in noisy training data". *Carneiro G., You S. (eds) Computer Vision – ACCV Workshops. Lecture Notes in Computer Science*, 11367, 2018. 2

[29] Joo-Yeon Lee, Woo-Jeoung Nam, and Seong-Whan Lee. Multi-contextual predictions with vision transformer for video anomaly detection, 2022. 1

[30] S. Leroux, B. Li, and P. Simoens. Multi-branch neural networks for video anomaly detection in adverse lighting and weather conditions. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3027–3035, 2022. 2

[31] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2152–2161, 10 2019. 2

[32] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 8

[33] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *British Machine Vision Conference*, 2018. 2, 8

[34] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13568–13577, 2021. 2

[35] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13568–13577, 2021. 3

[36] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019. 2

[37] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, 2017. 2, 8

[38] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349, 2017. 2

[39] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future frame prediction network for video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[40] M. Fathy M. Sabokrou, M. Khalooei and E. Adeli. "adversarially learned one-class classifier for novelty detection". *Proc. CVPR*, 2018. 2

[41] 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, 2010. 5

[42] Jefferson Ryan Medel and Andreas Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks, 2016. 2

[43] M. Minderer, C. Sun, R. Villegas, F. Cole, K. Murphy, and H. Lee. "unsupervised learning of object structure and dynamics from videos". *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2

[44] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh. "learning regularity in skeleton trajectories for anomaly detection in videos". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11988–11996, 2019. 2

[45] M.G. Narasimhan and S.K. S. "dynamic video anomaly detection and localization using sparse denoising autoencoders". *Multimed Tools Appl*, 77:13173—13195, 2018. 2

[46] Khac-Tuan Nguyen, Dat-Thanh Dinh, Minh N. Do, and Minh-Triet Tran. Anomaly detection in traffic surveillance videos with gan-based future frame prediction. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 457–463, New York, NY, USA, 2020. Association for Computing Machinery. 2

[47] Trong Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1273–1283, 2019. 1, 2, 3, 8

[48] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14360–14369, 2020. 2, 8

[49] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 5

[50] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 5

[51] Manassés Ribeiro, André Eugênio Lazzaretti, and Heitor Silvério Lopes. A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters*, 105:13 – 22, 2018. Machine Learning and Applications in Artificial Intelligence. 2

[52] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13566–13576, 2022. 2

[53] Pankaj Raj Roy, Guillaume-Alexandre Bilodeau, and Lama Seoud. Local anomaly detection in videos using object-centric adversarial learning. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 219–234, Cham, 2021. Springer International Publishing. 2

[54] Guodong Shen, Yuqi Ouyang, and Victor Sanchez. Video anomaly detection via prediction network with enhanced spatio-temporal memory exchange. In *ICASSP 2022 - 2022*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3728–3732, 2022. 2

[55] L. Shine, A. Edison, and C. V. Jiji. "a comparative study of faster r-cnn models for anomaly detection in 2019 ai city challenge". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 306–314, 2019. 2

[56] N. Srivastava, E. Mansimov, , and R. Salakhutdinov. "unsupervised learning of video representations using lstms". *ICML*, 2015. 2

[57] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020. 1, 8

[58] Tuan-Hung Vu., Sebastien Ambellouis., Jacques Boonaert., and Abdelmalik Taleb-Ahmed. Anomaly detection in surveillance videos by future appearance-motion prediction. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 484–490. INSTICC, SciTePress, 2020. 2

[59] Bokun Wang and Caiqian Yang. Video anomaly detection based on convolutional recurrent autoencoder. *Sensors*, 22(12), 2022. 2

[60] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan. Abnormal event detection in videos using hybrid spatio-temporal autoencoder. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2276–2280, 2018. 2

[61] J. Wei, Y. Zhao J. Zhao, and Z. Zhao. "unsupervised anomaly detection for traffic surveillance based on background modeling". *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1290–1297, 2018. 2

[62] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao. "anopcn: Video anomaly detection via deep predictive coding network". *Proceedings of the 27th ACM International Conference on Multimedia (MM '19). Association for Computing Machinery*, pages 1805—1813, 2019. 2

[63] Z. Yiru, D. Bing, S. Chen, L. Yao, L. Hongtao, and H. Xian-Sheng. "spatio-temporal autoencoder for video anomaly detection". *ACM Multimedia Conference*, 2017. 2

[64] H. Zenati, M. Romain, C. Foo, B. Lecouat, and V. Chandrasekhar. "adversarially learned anomaly detection". *IEEE International Conference on Data Mining (ICDM)*, pages 727–736, 2018. 2

[65] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multispace for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3694–3706, 2021. 2

[66] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multispace for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3694–3706, 2021. 8

[67] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xiansheng Hua. Spatio-temporal autoencoder for video anomaly detection. *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 2