Attention-Driven Loss for Anomaly Detection in Video Surveillance

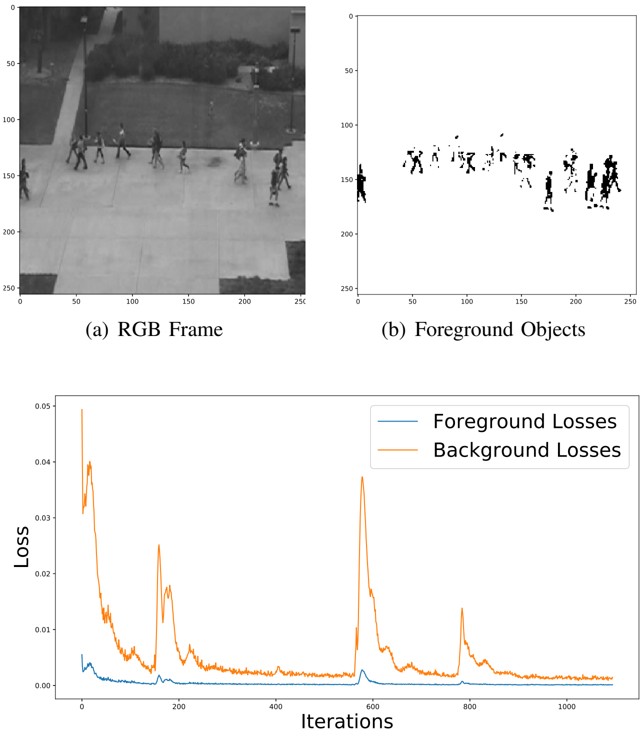
Joey Tianyi Zhou, Le Zhang, Zhiwen Fang, Jiawei Du, Xi Peng, Xiao Yang

Abstract*—Recent video anomaly detection methods focus on reconstructing or predicting frames. Under this umbrella, the long-standing inter-class data-imbalance problem resorts to the imbalance between foreground and stationary background objects in video anomaly detection and this has been less investigated by existing solutions. Naively optimizing the reconstructing loss yields a biased optimization towards background reconstruction rather than the objects of interest in the foreground. To solve this, we proposed a simple yet effective solution, termed attention-driven loss to alleviate the foreground-background imbalance problem in anomaly detection. Specifically, we compute a single mask map that summarizes the frame evolution of moving foreground regions and suppresses the background in the training video clips. After that, we construct an attention map through the combination of the mask map and background to give different weights to the foreground and background region respectively. The proposed attention-driven loss is independent of backbone networks and can be easily augmented in most existing anomaly detection models. Augmented with attention-driven loss, the model is able to achieve AUC 86.0% on Avenue, 83.9% on Ped1, 96% on Ped2 datasets. Extensive experimental results and ablation studies further validate the effectiveness of our model.

***Index Terms*—anomaly detection, deep learning, attention.**

I. Introduction

Proliferation of cameras, availability of cheap storage, and rapid developments in computer hardware has spurred the rise in automatic analysis of videos in which anomaly detection plays an inevitable role. Despite several decades of research, video anomaly detection remains challenging as the candidate pool of anomalies is often unbounded. Specifically, it is more difficult to define all possible negative (anomaly) samples, which is different from conventional classification or detection tasks [17]. Apart from this, it is label-intensive to collect sufficient anomalies due to its rarity. Due to these issues, most popular methods for video anomaly detection simplify the data collection procedure, and only use the videos of normal events as training data to learn a model. In the test set, they target at discovering the abnormal event which would do not conform the learned model [34], [21], [18].



(a) RGB Frame

(b) Foreground Objects

(c) Foreground Training Loss vs Background Training loss

Fig. 1. Foreground Objects v.s. Background Objects

The efforts to address the anomaly detection could be divided into two groups. The first one is to detect the testing sample with a reconstruction model and identify anomalies with higher reconstruction errors [5], [7], [12]. For example, sparse coding based methods [7], [21], [19], [44], [27] first learn a dictionary from a training data set that only consists of normal events and then discover the abnormal events that cannot be exactly reconstructed by a few of atoms of the learned dictionary. To enjoy the representative capacity of neural networks, some recent works tried to marry deep learning and anomaly detection such as Recurrent Neural Networks (RNNs) [5], 3D-CNNs [12], [29], fully convolutional neural

This work was supported by the National Natural Science Foundation of China (Grant No. 61502187, No. 61702182, No. 61806135, No. 61625204, and No. 61836006), by programmatic grant no. A18A1b0045 from the Singapore government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), by the Fundamental Research Funds for the Central Universities under Grant YJ201949. This paper was recommended by Associate Editor Hongying Meng, (Corresponding author: Le Zhang.)

Joey Tianyi Zhou, and Jiawei Du are with Institute of High Performance Computing (IHPC), the Agency for Science, Technology and Research (A*STAR), Singapore.

Le Zhang is with Institute for Infocomm Research (I2R), the Agency for Science, Technology and Research (A*STAR), Singapore.

Zhiwen Fang is with Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou, China. He is also with School of Energy and Mechanical-electronic Engineering, Hunan University of Humanities, Science and Technology, Loudi, China.

Xi Peng is with the College of Computer Science, Sichuan Univerisity, China.

Yang Xiao is National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology.

networks (FCNs) [30], and generative adversarial networks (GANs) [32], [1], [31]. Different from the first group, the second group is to estimate the prediction errors of next frame through a prediction model over the training data rather than reconstructing the current frame. Some recent works employ Convolutional Long Short-Term Memory (Conv-LSTM) networks [25] or U-Net [18]. These methods are able to predict the evolution of a video sequence from a small number of input frames and show a certain level of improvement over previous reconstruction models.

Most existing solutions employ the vanilla loss function (e.g. l_1 and l_2) to estimate either the reconstruction error or the prediction error. Under this umbrella, a natural assumption is that all regions in the scene, including the stationary background and moving foreground objects, contribute equally. Unfortunately, such an assumption may be sub-optimal since the stationary background is less important than the foreground that contains object of interests or moving objects. As shown in Figure 1(a), one could empirically find that the main elements in anomaly detection are moving person/objects rather than stationary backgrounds.

More importantly, the stationary background also prevents effective training optimization. To better understanding this, we plot the training loss of the latest anomaly detection method [18] on each frame of Ped2 [22]. As shown in Figure 1(c), the background takes up a large portion of the overall training loss while the foreground region of interest occupies a small part of the overall training loss. Given such an imbalanced dataset, the loss functions treats all regions equally, which will be dominated by the background with a large training loss. As a result, the model will be "lose the focus" and is less prioritised to reconstruct or predict the pixels of person/objects during optimization. Such a foreground-background imbalance problem, unfortunately, has been less touched in existing works, to the best of our knowledge. To alleviate such a learning bias, we propose an **attention-driven loss** to regularize the network training. The proposed loss essentially alleviates the data bias issue and guides the network to focus more on the regions of interest (ROI) in the scene. Rather than manually providing the ROIs for constructing attention map, we compute it directly from the training data. Extensive experimental results on the benchmark datasets show that, with the proposed loss, the performance of anomaly detection is significantly improved and reaching state-of-the-arts. Our contributions are summarized as follows:

1) The experiment results show the limitation of equally treating foreground and background objects when we reconstruct the frames which is largely ignored in existing studies. To the best of our knowledge, it could be the first work to formally investigate this foreground-background imbalance problem in deep learning for video anomaly detection.
2) To better supervise the network training for anomaly detection, we propose a data-dependent attention-driven loss, which essentially helps the model focus more on sparse foreground objects of interest rather than stationary background in the optimization.
3) We first explain the working mechanism of the proposed loss from the mathematics perspective. Furthermore, a series of experiments are conducted to further demonstrate the superiority of the proposed attention-driven loss over different baselines.
4) The attention-driven loss also serves a complementary piece to existing anomaly detection models. It is easy to be implemented and incorporated into existing or future more advanced architectures to improve their performance.

II. Related work

Due to the space limitation, we cannot to give an exhaustive review on all anomaly detection developments. Instead, we will briefly introduce the most related threads of works including anomaly detection in videos, learning with attention, and imbalanced data learning.

A. Anomaly Detection in Videos

For anomaly detection, early hand-crafted feature based works usually utilize low-level trajectory features to represent the regular motion patterns [35], [39], [42], [41]. To capture the motion and appearance, hand-crafted spatial-temporal features, such as histogram of oriented flows (HOF) [9] and histogram of oriented gradients (HOG) [8] are widely used. Based on these spatial-temporal features, some researches take one step further to model the motion and content pattern. Some typical works include Markov random filed (MRF) [43], mixture of probabilistic PCA (MPPCA) [16], and Gaussian mixture model [22]. Recently, deep learning have earned a great success in anomaly detection [12], [5], [21], [45]. To learn a better representation for scenes, various convolutional neural networks (CNN) based methods have been proposed, e.g., 3D convolutional (3D-CNNs) [12], [29], fully convolutional neural networks (FCNs), [30] and generative adversarial networks (GANs) [32], [1], [18]. With the CNN features, [5], [20], [25] proposed using long short term memory (LSTM) network and [21] proposed using Recurrent Neural Networks (RNN) to capture long-term dependency among frames.

By taking the advantages of both convolutional neural networks (CNN) and long short term memory (LSTM), [5] proposed a convolutional LSTM Auto-Encoder (ConvLSTM-AE) to simultaneously capture normal appearance and motion patterns. Considering the temporally coherent anomaly probability, a sparse coding based method within a stacked RNN framework is introduced to model the normal patterns in [21]. Different from reconstructing the frame, Liu et al. [18] utilize a predictive model which employs the prediction error as a proxy to identify the anomaly probability. Due to the extreme difficulty of the task, it is hard for a single auto-encoder to simultaneously handle all different patterns in videos. Moreover, none of the mentioned approaches exploit spatial and temporal information separately. To overcome the aforementioned limitation, multiple auto-encoders for learning feature representations of both content and motion patterns are designed [40]. Nevertheless, the auto-encoders are only used to extract the independent features rather than the end-to-end trainable module.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2962229, IEEE Transactions on Circuits and Systems for Video Technology

B. Learning with Attention

Attention model enables the neural network dynamically shift/select attributes so that the overall decision making is more reliable, which was first used in neural machine translation [2] and other natural language processing tasks [26], [37], [46]. Recently, attention based approaches also saw heavy usage in a variety of vision tasks such as segmentation [4], image classification [38], and so on.

Different from existing works that use attention model as a trainable module in the intermediate feature layers, our proposed attention-driven loss plays an regularization role to supervise the model training. On the other side, it is also different from saliency detection [6] in which the ground-truth saliency map needs to be pre-defined. In our method, the attention-driven map could be automatically computed from raw data without additional manual inputs.

C. Learning with Imbalanced Data

Data Imbalance issue ubiquitously exists in various machine learning tasks such as image annotation [23], object detection [17], semantic segmentation [15], sequence labeling [47] and so on. Most existing learning algorithms produce an inductive learning bias towards the frequent (i.e., majority) classes if training data are imbalanced, thus resulting in poor minority class recognition performance. This long standing data imbalance issue in these tasks usually refers to inter-class data-imbalance problem. A simple approach to alleviating class imbalance in learning is to re-sample the training data by offline data augmentation [13] or balanced sampling [13]. More recently, cost-sensitive learning [14] attracts more and more attention thanks to its effectiveness. For example, the focal loss [17] is designed to address the one-stage object detection issue of which there is an extreme imbalance distribution between the different object classes.

Although workable in certain scenarios, it is not straightforward to apply those concepts to video anomaly detection due to its unsupervised nature, i.e., we have no supervised information on abnormal events during training. To the best of our knowledge, we are also the first to systematically study the foreground-background imbalance problem for deep learning based video anomaly detection.

III. Attention-Driven Loss

In this section, we introduce the concept of Attention-Driven loss for anomaly detection which is a standard RGB attention output map to summarizes the appearance and dynamics of a whole video sequence .

A. RGB Attention Map

As discussed in Introduction, the domination of the stationary background loss forces the training process focus on the optimization of the stationary background rather than the object of interest. To alleviate such a overwhelmingness, one simple solution is manually defining the ROIs and assigning them with a larger weight than that of background regions. Through the collaboration of project agreement, one global security company help to sketch ROIs in existing Ped1, Ped2 [22], and Avenue [19] benchmark datasets after watching the whole training videos. Specifically, we define the constructed attention map $\mathbf{A} \in R^{L\times W\times 3}$ for the input RGB frames $\{\mathbf{I}^t\}_{t=1}^T \in R^{L\times W\times 3}$ as follows,

$$\mathbf{A} = \mathbf{M} + \mathbf{B}, \tag{1}$$

where $\mathbf{M} \in \{0,1\}^{L\times W\times 3}$ is a binary mask matrix defined by security experts and $\mathbf{B} \in R^{L\times W\times 3}$ denotes the background weight. In the experiment, we set $\mathbf{B}$ to be a constant matrix with all the elements of 0.1. The attention map $\mathbf{A}$ for anomaly detection benchmarks are depicted in Figure 3, where the white region denotes the ROI and the grey region denotes the background region. The annotation map constructed by human expert indicates that those white regions are usually the regions of the moving object, and thus should attract more attentions when monitoring the scenes.

Unfortunately, this kind of manually annotated binary mask may be sub-optimal since all the RGB channels share the same weight and the boundary for ROI is hard to define. To solve this problem, we propose inferring the RGB attention map in a data-driven way. More specifically, a single attention map is computed to summarize the video and capture the evolution of the video frames at the same time, while averaging away background pixels and background motion patterns and focusing on the acting objects such as humans along frames.

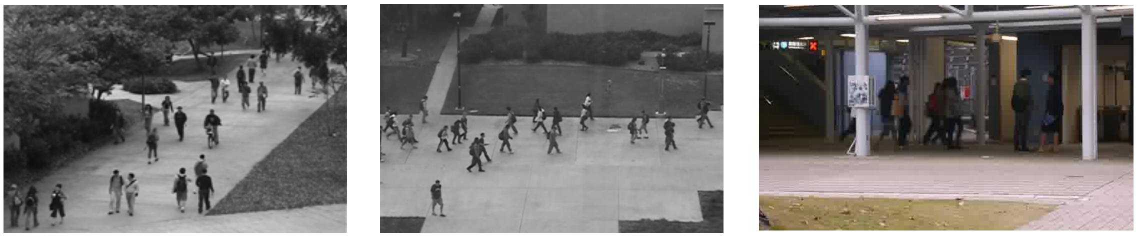
Motivated by dynamic image [3], we propose learning the mask map $\mathbf{M}$ along a set of frames $\mathbf{I}^1, \mathbf{I}^2, \cdots, \mathbf{I}^T$, through the following simple but effective objective,

$$\begin{aligned}\min_{\mathbf{M}} \frac{2}{T(T-1)} \sum_{t_1 \geq t_2} \max\left(0, 1 - S(t_1|\mathbf{M}) + S(t_2|\mathbf{M})\right) \\ + \frac{\lambda}{2}\|\mathbf{M}\|^2\end{aligned} \tag{2}$$

where $S(t|\mathbf{M}) = \langle \mathbf{M}, \mathbf{V}^t\rangle_F$ denotes the ranking score associated with the time-step t and $\langle\cdot\rangle_F$ denotes the Frobenius inner product. $\mathbf{V}^t = \frac{1}{t}\sum_{\tau=1}^t \mathbf{I}^t$ denotes the average frame within t time-steps. Equation 2 tries to compress the sequence of all ordered frames $\mathbf{I}^t$ into a single static image $\mathbf{M}$. The objective loss is then averaged over all the frames with satisfying $t_1 \geq t_2$, namely in total $\frac{2}{T(T-1)}$ frames.

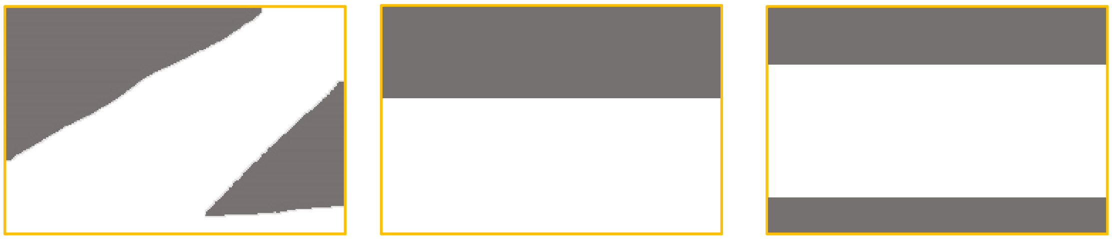
B. Augmenting With Attention-Driven Loss

In the above objective, the first term is used to constrain the ranking loss with a unit margin for any $\{t_1, t_2\}$, i.e., $\forall\{t_1, t_2\}$, if $t_1 \geq t_2$, then $S(t_1|\mathbf{M}) \geq S(t_2|\mathbf{M}) + 1$. In other words, the inner product between the learned $\mathbf{M}$ and $\mathbf{V}^t$ should become larger along timestamps. In this way, the spatial-temporal dynamic evolution information could be captured because the optimal attention vector $\mathbf{M}^*$ reflects the appearance order of frames. It is interesting to note that the objective in Equation 2 is irrelevant to the stationary background. Specifically, we decompose the t-th frame by $\mathbf{I}^t = \mathbf{I}_b^t + \mathbf{I}_f^t$, where $\mathbf{I}_b^t$ and $\mathbf{I}_f^t$ denote the background and foreground, respectively. With the decomposition, $\mathbf{V}^t = \mathbf{V}_b^t + \mathbf{V}_f^t = \frac{1}{t}\sum_{\tau=1}^t(\mathbf{I}_b^t + \mathbf{I}_f^t)$. Since the background is always stationary considering our video surveillance setting, then $\mathbf{V}_b^t = \sum_{\tau=1}^t \mathbf{I}_b^\tau$ is a constant. By



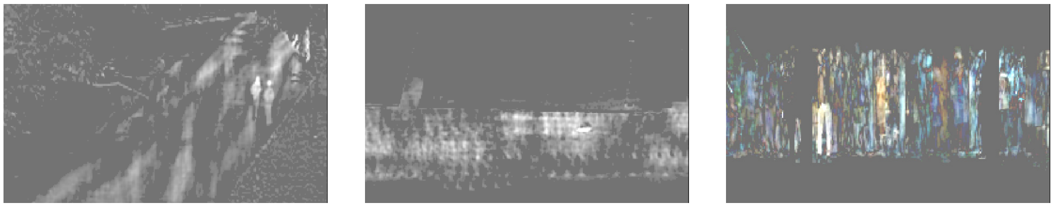
(a) Ped1-Frame (b) Ped2-Frame (c) Avenue-Frame

Fig. 2. Example RGB Frames of Benchmarks



(a) Ped1 (b) Ped2 (c) Avenue

Fig. 3. RGB Attention Map by Security Expert: the white region denotes the region of interest and the black region denotes the background regions.



(a) Ped1 (b) Ped2 (c) Avenue

Fig. 4. Our RGB Attention Map.

substituting this into $S(t|\mathbf{M})$ and cancel out the constant $\mathbf{V}_b^t$, we could obtain

$$\begin{aligned} &\max\left(0, 1 - S(t_1|\mathbf{M}) + S(t_2|\mathbf{M})\right) \\ =&\max\left(0, 1 - \left\langle \mathbf{M}, \mathbf{V}_f^{t_1} \right\rangle_F + \left\langle \mathbf{M}, \mathbf{V}_f^{t_2} \right\rangle_F\right), \end{aligned} \tag{3}$$

which suggests that the optimal solution in Equation 2 is irrelevant to the stationary background. In other words, the learned mask map could suppress the effects of the stationary background as validated in Figure 4.

To efficiently solve Equation 2, we adopt a gradient-based optimization. Starting with $\mathbf{M} = \mathbf{0}$, the first approximated solution obtained by the gradient descent is

$$\mathbf{M}^* = \mathbf{0} - \eta \nabla F(\mathbf{M})|_{\mathbf{M}=\mathbf{0}} \propto \nabla F(\mathbf{M})|_{\mathbf{M}=\mathbf{0}} \tag{4}$$

for any $\eta \geq 0$, where $F(\mathbf{M}) = \frac{\lambda}{2}\|\mathbf{M}\|^2 + \frac{2}{T(T-1)} \sum_{t_1 \geq t_2} \max\left(0, 1 - S(t_1|\mathbf{M}) + S(t_2|\mathbf{M})\right)$ and

$$\begin{aligned} \nabla F(\mathbf{0}) \propto & \sum_{t_1 > t_2} \nabla \max\{0, 1 - S(t_1|\mathbf{m}) + S(t_2|\mathbf{m})\}|_{\mathbf{m}=\mathbf{0}} \\ =& \sum_{t_1 > t_2} \left\langle \mathbf{m}, \mathbf{V}^{t_1} - \mathbf{V}^{t_2} \right\rangle \\ =& \sum_{t_1 > t_2} \mathbf{V}^{t_1} - \mathbf{V}^{t_2}. \end{aligned} \tag{5}$$

Then $\mathbf{M}$ could be rewritten as follows,

$$\begin{aligned} \mathbf{M}^* \propto & \sum_{t_1 > t_2} \mathbf{V}^{t_1} - \mathbf{V}^{t_2} \\ =& \sum_{t_1 > t_2} \left[\frac{1}{t_1} \sum_i^{t_1} \mathbf{I}^i - \frac{1}{t_2} \sum_j^{t_2} \mathbf{I}^j \right] \\ =& \sum_{t=1}^{T} \alpha(t) \mathbf{I}^t \end{aligned} \tag{6}$$

where the coefficient $\alpha(t)$ is given by $\alpha(t) = 2(T - t + 1) - (T+1)(H_T - H_{t-1})$, and $H_t = \sum_{i=1}^t 1/t$ is the t-th Harmonic number. Although the mask map $\mathbf{M}$ is able to memorize the

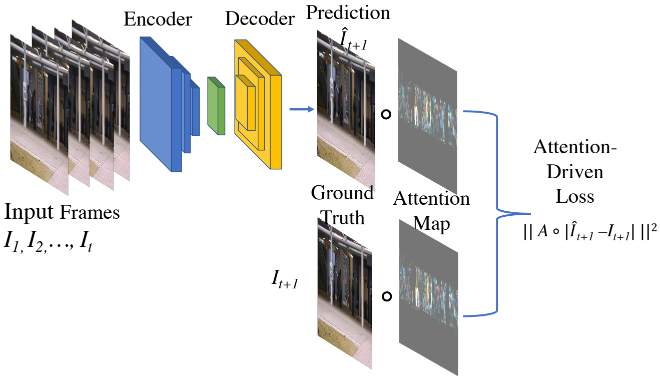


Fig. 5. Augmenting Anomaly Detection Models With Attention-Driven Loss.

temporal spatial information, we normalize it into $[0, 1]$ for weighting and avoiding the possible scale problem via

$$\mathbf{A} = |\mathbf{M}|/(\max(\mathbf{M}) - \min(\mathbf{M})) + \mathbf{B}, \tag{7}$$

where $\max(\mathbf{M})$ and $\min(\mathbf{M})$ denote the largest and smallest value of $\mathbf{M}$. The operator $|\cdot|$ denotes the operation of taking the absolute value in element-wise. $\mathbf{B}$ is the background region weight which cannot be zero because the background still contains some important stationary information in some scenarios. For example, some persons leave unattended bags or items on the public areas. We show the resulting $\mathbf{A}$ in Figure 4. The pixels in the attention map appear to focus on the identity and motion of the salient pedestrians in videos, indicating that they may contain the information necessary to perform anomaly detection in video sequences.

The network commonly used for reconstruction model and prediction model in existing work [18], [12] usually contains two modules: 1) an encoder which extracts features by gradually reducing the spatial resolution; and 2) a decoder which gradually recovers the frame by increasing the spatial resolution. In this paper, we use the prediction model as a showcase to demonstrate the effect of the attention-driven loss. Following the suggestion given by [18], we employed U-Net [28] as the encoder and decoder to avoid gradient vanishing problem and information imbalance in each layer. An overview architecture is illustrated in Figure 5.

1) Attention-Driven Content Loss: The content loss is used to guarantee the predicted frame being close to its ground-truth in RGB space. Here, we augment the content loss with the attention map $\mathbf{A}$, termed as *Attention-Driven Content Loss* (ACL) to minimize the distance between a predicted frame and its ground-truth in intensity. In mathematical,

$$\ell_{acl} = \sum_t \|\mathbf{A} \circ (\hat{\mathbf{I}}^t - \mathbf{I}^t)\|_2^2, \tag{8}$$

where $\hat{\mathbf{I}}^t = \mathcal{G}(\mathbf{I}^1, \cdots, \mathbf{I}^{t-1})$ and $\mathbf{I}^t$ denote the predicted frame and its ground truth on the time stamp t, respectively. $\circ$ denotes the element-wise multiplication.

2) Attention-Driven Gradient Difference Loss: Although the content loss captures the major content in images, blurry predictions are always achieved [24]. To sharpen the image prediction, we directly penalize the differences of image gradient predictions in the generative loss function with the attention map, termed as attention-driven gradient difference loss (AGDL), namely,

$$\ell_{agdl} = \sum_t \sum_{i,j} \left\| \mathbf{A} \circ (|\hat{\mathbf{I}}_{i,j}^t - \hat{\mathbf{I}}_{i-1,j}^t| - |\mathbf{I}_{i,j}^t - \mathbf{I}_{i-1,j}^t|) \right\|_1 \tag{9}$$

$$+ \left\| \mathbf{A} \circ (|\hat{\mathbf{I}}_{i,j}^t - \hat{\mathbf{I}}_{i,j-1}^t| - |\mathbf{I}_{i,j}^t - \mathbf{I}_{i,j-1}^t|) \right\|_1, \tag{10}$$

where i and j denote the spatial index of a video frame, and $|\cdot|$ computes the absolute value.

IV. Adversarial Learning

Adversarial learning [11] has demonstrated success in improving the generation of image and video. Specifically, a discriminative network $\mathcal{D}$ is used to estimate the probability that a sample comes from the dataset rather than being produced by a generative model $\mathcal{G}$. The two models are simultaneously trained so that $\mathcal{G}$ could generate frames that are hard to classify by $\mathcal{D}$, and meanwhile $\mathcal{D}$ learns to discriminate the frames generated by $\mathcal{G}$ from the ground-truth. In theory, when $\mathcal{G}$ is well trained, it would be impossible for $\mathcal{D}$ to perform better than chance. As the problem has been cast as frame prediction task, the system trained with objectives formulated in previous section directly may not be the optimal choice. In this work, we construct $\mathcal{G}$ using the U-Net and utilize a patch discriminator. In the discriminator, each output scalar of $\mathcal{D}$ corresponds to a patch of an input image. The training schedule is detailed in this following subsection.

A. Training of the Generator $\mathcal{G}$

With the fixed weights of $\mathcal{D}$, the goal of training $\mathcal{G}$ is to generate frames where $\mathcal{D}$ classify them into class 1. With this goal, the following adversarial loss is computed as below:

$$\ell_{adv}^{\mathcal{G}} = \sum_t \sum_{i,j} (\mathcal{D}(\hat{\mathbf{I}}^t)_{i,j} - 1)^2, \tag{11}$$

where i and j denote the spatial index of a video frame.

With the above, we could obtain the overall loss function for generator $\ell_{\mathcal{G}}$ as follows,

$$\ell_{\mathcal{G}} = \lambda_{acl}\ell_{acl} + \lambda_{agdl}\ell_{agdl} + \lambda_{of}\ell_{of} + \lambda_{adv}\ell_{adv}^{\mathcal{G}}, \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2962229, IEEE Transactions on Circuits and Systems for Video Technology

where ℓ_{of} represents the optical flow loss which could evaluate the coherence of motion and serve as an complementary piece to the RGB based loss. Note that, previous studies [21], [18] have shown its superiority in performance improvement. Here we implement ℓ_{of} through Flownet [10] with all the fixed parameters. The tradeoff parameters are used to balance different loss terms. We follow the suggestion of the parameter setting in [18], and experimentally found that $\lambda_{acl} = 1$, $\lambda_{agdl} = 1$, $\lambda_{of} = 2$, and $\lambda_{adv} = 0.05$ usually lead to a good performance across datasets.

B. Training of the Discriminator $\mathcal{D}$

With the fixed $\mathcal{G}$, the discriminator $\mathcal{D}$ aims to classify $\mathbf{I}^t$ into class 1 (genuine sample) and $\hat{\mathbf{I}}^t = \mathcal{G}(\mathbf{I}^1, \cdots, \mathbf{I}^{t-1})$ into class 0 (fake sample), respectively. The loss function for training $\mathcal{D}$ is as follows,

$$\ell_{\mathcal{D}} = \sum_t \sum_{i,j} \frac{1}{2}\{(\mathcal{D}(\hat{\mathbf{I}}^t)_{i,j} - 0)^2 + (\mathcal{D}(\mathbf{I}^t)_{i,j} - 1)^2\}, \quad (13)$$

V. Experiments

In this section, we evaluate our proposed method on three anomaly detection benchmarks, including the CUHK Avenue dataset [19], the UCSD Pedestrian 1 and the UCSD Pedestrian 2 [22]. To train the network, we follow the setting used in [18]. Specifically, the intensity of pixels in all frames are normalized into [-1, 1] and the size of each frame is rescaled to 256 × 256. We used a random clip of five sequential frames as the input and set the mini-batch size as four. The best coefficient in Equation 12 is chosen for different datasets. All the elements in the background weight matrix $\mathbf{B}$ are set to 0.1.

A. Datasets

In this section, we briefly introduce the datasets used in our experiments and illustrate some image samples in Fig. 2.

- CUHK Avenue dataset contains 16 training videos and 21 testing videos with a total of 47 abnormal events, including throwing objects, loitering, and running. The size of people may change with the positions and angles of camera.
- The UCSD dataset contains two parts: The UCSD Pedestrian 1 (Ped1) dataset and the UCSD Pedestrian 2 (Ped2) dataset. The Ped1 dataset includes 34 training videos and 36 testing videos (40 irregular events). All of these abnormal cases are about vehicles such as bicycles and cars. The Ped2 dataset contains 16 training videos and 12 testing videos with 12 abnormal events. The definition of anomaly for Ped2 is the same with Ped1. Usually different methods are evaluated on these two parts separately.

B. Testing Metric

In testing phase, we use the prediction error to measure the difference between frame I^{t-1} and frame I^t for anomaly prediction. A lot of studies have shown that Peak Signal to Noise Ratio (PSNR) [24] could be better to estimate the reconstruction quality [18] for each frame than Euclidean distance. The metric is defined as follows:

$$PSNR_t = 10\log_{10} h_t, \quad (14)$$

where

$$h_t = \frac{[\max_{\hat{I}^t}]^2}{\frac{1}{N}\sum\limits_{i,j}(I^t(i,j) - \hat{I}^t(i,j))^2}, \quad (15)$$

where i and j represent the spatial index of $\hat{I}^t$ and I^t, respectively. $[\max_{\hat{I}^t}]$ is the maximum value of $\hat{I}^t$, and N is the number of pixels. Lower PSNR$_t$ for the t-th frame indicates that it is more likely to be anomaly. Following [24], [18], the scores of all frames are normalized into $[0, 1]$ after getting the PSNR score of each frame in each testing video via:

$$S_t = \frac{PSNR_t - \min_t PSNR_t}{\max_t PSNR_t - \min_t PSNR_t}. \quad (16)$$

where $\min_t PSNR_t$ and $\max_t PSNR_t$ represent the minimum and the maximum PSNR values in a sequence, respectively. Therefore, for a given threshold, we can distinguish whether a frame is normal or abnormal according to its score S_t.

C. Evaluation Metric

Based on Equation 16, we can estimate the score of the t-th frame and judge whether the abnormal event occurs. Given a fixed threshold, the frame can be recognized as an anomaly frame if its score is lower than the threshold. Obviously, higher threshold will lead to higher false negative ratio and lower one may produce more false alarms. Thus, the Area Under Curve (AUC) is a more suitable metric [22], [19], [21], which measures the performance by changing different thresholds.

D. Comparison with Existing Methods

In this section, we compare our method MESDnet with different state-of-the-art methods including: Conv-AE [12], ConvLSTM-AE [20], DeepAppearance [33], Unmasking [36], TSC [21], Stacked RNN [21] and Liu *et al.*. [18]. We listed the AUC performance of different methods on these datasets in Table I. For fair comparison, we also report our results with tuned hyper-parameters in this table. It could be seen that performance of our proposed model consistently outperforms other methods.

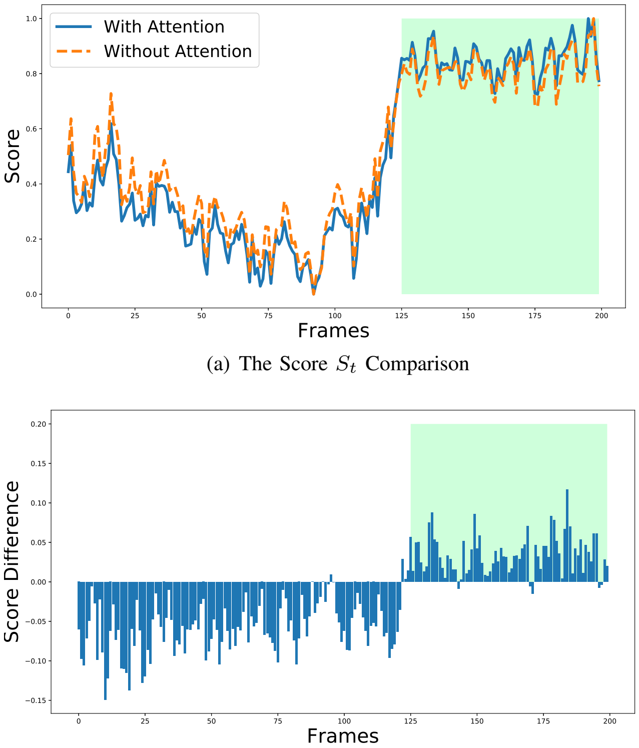
E. Ablation Studies on Attention-Driven Loss

In this section, we evaluate the effect of attention-driven loss in both qualitative and qualitative ways.

1) Evaluation of Different Attention-Driven Losses: In the first experiment, we conduct the ablation studies on two proposed attention-driven losses, i.e., ℓ_{acl} and ℓ_{agdl}. For ease of presentation, we denote their counterpart losses without augmenting attention map as ℓ_{cl} and ℓ_{gdl}. The results are summarized in Table II. Note that in this experiment, we discard the optical flow loss and adversarial loss and only focus on the contribution of attention-driven loss. From the results, one could observe that both two attention-driven loss could enhance the anomaly detection performance over their counterparts.

TABLE I
AUC COMPARISON WITH THE STAT-OF-THE-ART METHODS ON THE AVENUE, PED1, PED2. THE BEST RESULT IS MARKED IN BOLD.

	CUHK Avenue	UCSD Ped1	USCD Ped2
MPPCA [16]	N/A	59.0%	69.3%
MPPCA+SFA [22]	N/A	66.8%	61.3%
MDT [22]	N/A	81.8%	82.9%
3DConv-AE [12]	80.0%	75.0%	85.0%
ConvLSTM-AE [20]	77.0%	75.5%	88.1%
150FPS[19]	80.9%	N/A	N/A
DeepAppearance [33]	84.6%	N/A	N/A
Unmasking [36]	80.6%	68.4%	82.2%
TSC [21]	80.6%	N/A	91.0%
Stacked RNN [21]	81.7%	N/A	92.2%
AnoPred [18]	85.1%	83.1%	95.4%
Ours	**86.0%**	**83.9%**	**96.0%**



(a) The Score S_t Comparison

(b) Score Difference Between w and w/o Attention

Fig. 6. The scores gap between the normal frames and abnormal frames are enlarged by the attention-driven loss. Green shadow regions denote the abnormal frames.

2) Score Gap Analysis: Larger score gap usually means the model is with better performance to distinguish normal and abnormal patterns, while enjoying stronger robustness to noises. To investigate the score gap, we plot the irregularity score S_t for each frame $\mathbf{I}^t$ of Ped2 as a showcase in Figure 6(a). From the figure, it is interesting to observe that the scores of model with the attention-driven loss are always lower than that without the attention-driven loss for those normal frames. For the abnormal frames, the score is larger than that without the attention-driven loss. To better illustrate this, we also demonstrate the score difference between the base model with and without the attention-driven loss in Figure 6(b). The results indicate that the score gap between normal and abnormal events are enlarged by the attention-driven loss.

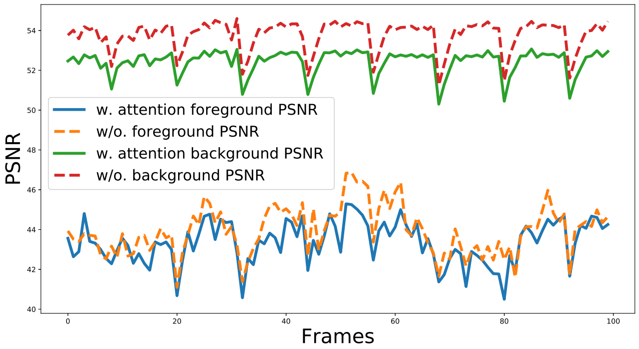


Fig. 7. PSNR$_t$ for Foreground and Background: Comparison Between w. and w/o. Attention-Driven Loss. Both the PSNR of background and foreground for inference are significantly dropped after applying the attention-driven loss.

TABLE II
EVALUATION OF DIFFERENT ATTENTION-DRIVEN LOSSES

	Avenue	Ped1	Ped2
ℓ_{cl}	82.0%	74.6%	83.3%
ℓ_{acl}	83.5%	77.1%	86.2%
$\ell_{acl} + \ell_{gdl}$	84.0%	77.6%	87.5%
$\ell_{acl} + \ell_{agdl}$	84.7%	79.2%	89.0%

From Figure 6(b), we also observe that the attention map helps the algorithm to improve the performance up to 15% in terms of normalized score range, which can be considered as a significant enhancement. On the other side, it implies a big difference in the practical use. For example, if we define the score threshold as 0.6, the 2nd, 8th, and 10th frame will be flagged as abnormal event by the system without the attention map. In contrast, training with the attention map will eliminate such false alarms effectively.

3) Analysis of Predicted Image Quality: As stated in Introduction, the state-of-the-art models without attention-driven loss is overwhelmed by the stationary background loss. In this experiment, we examine the prediction error of model trained with the attention-driven loss. We plot the PSNRs of the foreground and background of a segment of video clip in Ped2 in Figure 7.

From the figure, one could observe that both the PSNR of background and foreground for inference are significantly

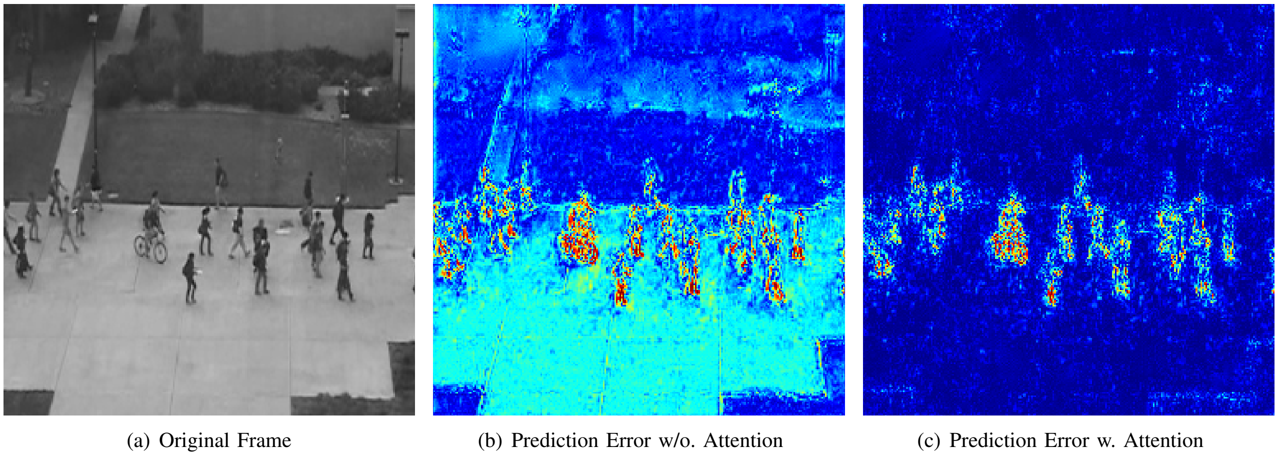

(a) Original Frame (b) Prediction Error w/o. Attention (c) Prediction Error w. Attention

Fig. 8. Visualization results of w. and w/o. attention-driven loss. For the the background area in (b) and (c), darker color indicates smaller error.

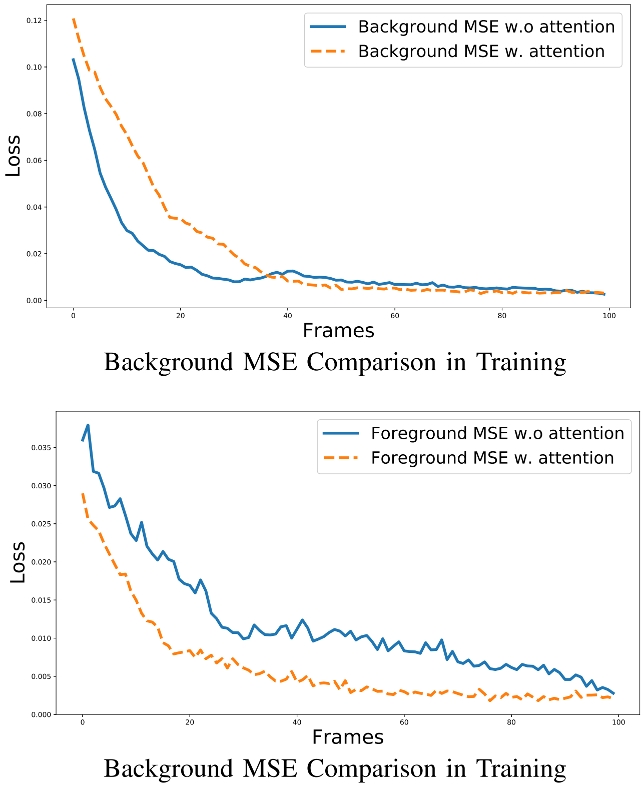


Fig. 9. Training augmented with attention-driven loss helps algorithm focus more on optimizing the foreground loss rather than the background loss.

dropped after applying the attention-driven loss. This may attribute to the adaptive selection merit of attention map.

We further give the visualization comparison in Figure 8. From the Figure 8(c), we observe that model with the attention-driven loss is able to suppress the error from background objects and focus on the prediction error from the moving objects in the foreground. In contrast, the model without the attention-driven loss (Figure 8(b)) is diverted to the background thus leading to large uninformative background reconstruction errors.

4) Impact on Training: In order to illustrate how the attention map assists in optimization, we also plot both background and foreground Mean Square Errors (MSE) for algorithm learning with/without attention-driven loss on each frame in Figure 9. From Figure 9(a), we observe that the algorithm without attention-driven loss focus on the minimization of background which leads a lower background MSE compared to the optimization with the proposed loss at the beginning. In Figure 9(b), one can see that attention-driven loss is able to significantly shift the optimization focus from the background to the foreground objects and a lower foreground MSE is achieved. These observations explains why the model could benefits from the proposed attention-driven loss.

5) Comparison of Attention Maps: Human vs Learned: In this experiment, we evaluate the proposed two attention maps on the state-of-the-art prediction model [18]: 1) the attention map defined by human (see Equation 1); and 2) the attention map computed from data (see Equation 7). The results are summarized in Table III. From the results, one could observe that the attention constructed by human is able to improve the performance in the Avenue and Ped1 dataset. However, it fails in Ped 2, which suggests that the learned attention map is a better choice.

TABLE III
COMPARISON OF ATTENTION MAPS: HUMAN VS OURS

	Avenue	Ped1	Ped2
Base	85.0%	83.1%	95.2%
+Human Attention	85.3%	83.6%	94.7%
+Our Attention	86.0%	83.9%	96.0%

6) The Versatility: In previous experiments, one could observe that the proposed attention-driven loss works well with the prediction model [18]. In order to further demonstrate the versatility of the proposed loss, we apply it to different network structures including 3DConv-AE [12] and ConvLSTM-AE [20]. Both two models are state-of-the-art reconstruction based paradigms. The results are summarized in Table IV which shows that the attention-driven loss is able to give a remarkable improvement over different architectures.

7) The Parameter Analysis: In the proposed attention loss (i.e., Equation 7), $\mathbf{B}$ is the background region weight which cannot be zero because the background still contains some important stationary information in some scenarios. For example, some persons leave unattended bags or items on the public

TABLE IV
GENERATLIZATION TO OTHER MODELS

	Avenue	Ped1	Ped2
3DConv-AE [12]	80.0%	75.0%	85.0%
+ Attention-Driven Loss	**82.5%**	**77.1%**	**87.2%**
ConvLSTM-AE [20]	77.0%	75.5%	88.1%
+ Attention-Driven Loss	**78.6%**	**76.8%**	**90.3%**

areas. Meanwhile, setting background weight to be zero results in the failure of the frame reconstruction. On the other hand, in principle, we want to suppress the background contribution into a small scale. In experiments, we also find that fixing $\mathbf{B}_{ij} = 0.1$ achieves good results across different datasets. We also conduct the parameter analysis of this parameter on Ped2 as the showcase in Table V.

TABLE V
PARAMETER ANALYSIS OF $\mathbf{B}$

$\mathbf{B}_{ij}$	0	0.1	0.2	0.3	0.4
AUC	92.0%	96.0%	95.2%	94.0%	93.5%

VI. CONCLUSION AND FUTURE WORK

In this paper, we investigated the problem of background reconstruction loss dominating the training loss for anomaly detection in videos. To solve this imbalance distribution problem, a simple but effective solution, called attention-driven loss, was introduced to learn generalizable features. Specifically, the human designed and learned from data attention maps were proposed. With them, we proposed two attention-driven losses for anomaly detection, i.e., attention-driven content loss and attention-driven gradient loss. The proposed method is independent from backbone networks and can be easily integrated into most existing models. Extensive experimental results and ablation studies also validate the effectiveness of our proposed model. In future, we would like to investigate how to extend the proposed attention-driven loss into a broader scenario of anomaly detection tasks including unsupervised anomaly detection and supervised anomaly detection.

REFERENCES

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. *arXiv preprint arXiv:1805.06725*, 2018. 2

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[3] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, 2016. 3

[4] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016. 3

[5] Y. S. Chong and Y. H. Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017. 1, 2

[6] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018. 3

[7] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3449 – 3456, 07 2011. 1

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 2

[9] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision*, pages 428–441, 2006. 2

[10] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 6

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 5

[12] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016. 1, 2, 5, 6, 7, 8, 9

[13] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008. 3

[14] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 3

[15] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 1–9, 2016. 3

[16] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, 2009. 2, 7

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 1, 3

[18] W. Liu, D. L. W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6, 7, 8

[19] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 1, 3, 6, 7

[20] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *Proceeding of the IEEE International Conference on Multimedia and Expo*, pages 439–444, 2017. 2, 6, 7, 8, 9

[21] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 1, 2, 6, 7

[22] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. 2, 3, 6, 7

[23] Q. Mao, I. W.-H. Tsang, and S. Gao. Objective-guided image annotation. *IEEE Transactions on Image Processing*, 22(4):1585–1597, 2013. 3

[24] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 5, 6

[25] J. R. Medel and A. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016. 2

[26] M. Rei, G. Crichton, and S. Pyysalo. Attending to characters in neural sequence labeling models. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, 2016. 3

[27] H. Ren, H. Pan, S. I. Olsen, and T. B. Moeslund. A comprehensive study of sparse codes on abnormality detection. *CoRR*, abs/1603.04026, 2016. 1

[28] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 5

[29] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette. Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017. 1, 2

[30] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 2018.

2

[31] M. Sabokrou, M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli. Avid: Adversarial visual irregularity detection. In *Proceedings of the Asian Conference on Computer Vision*, 2018. 2

[32] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Proceedings of the International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017. 2

[33] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe. Deep appearance features for abnormal behavior detection in video. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 779–789. Springer, 2017. 6, 7

[34] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Proceedings of the European Conference on Computer Vision*, pages 787–802. Springer, 2014. 1

[35] Q. Sun, H. Liu, and T. Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201, 2017. 2

[36] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2895–2903, 2017. 6, 7

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 3

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 3

[39] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2054–2060, 2010. 2

[40] D. Xu, Y. Yan, E. Ricci, and N. Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 2

[41] Y. Yuan, J. Fang, and Q. Wang. Online anomaly detection in crowd scenes via structure analysis. *IEEE Transactions on Cybernetics*, 45(3):548–561, 2015. 2

[42] Y. Yuan, D. Wang, and Q. Wang. Anomaly detection in traffic scenes via spatial-aware motion reconstruction. *IEEE Transactions on Intelligent Transportation Systems*, 18(5):1198–1209, 2017. 2

[43] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted hmms for unusual event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 611–618, 2005. 2

[44] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3320, 2011. 1

[45] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550, Oct 2019. 2

[46] J. T. Zhou, H. Zhang, D. Jin, and X. Peng. Dual adversarial transfer for sequence labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 3

[47] J. T. Zhou, H. Zhang, D. Jin, X. Peng, Y. Xiao, and Z. Cao. Roseq: Robust sequence labeling. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2019. 3