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/TMM.2019.2950530, IEEE Transactions on Multimedia

Learning Normal Patterns via Adversarial Attention-based Autoencoder for Abnormal Event Detection in Videos

Hao Song, Che Sun, Xinxiao Wu, Member, IEEE, Mei Chen, Member, IEEE, and Yunde Jia, Member, IEEE

***Abstract*—Automatically detecting anomalies in videos is a challenging problem due to non-deterministic definitions of abnormal events and lack of sufficient training data. To address these issues, we propose an autoencoder coupled with attention model to discover normal patterns in videos via adversarial learning. Abnormal events are detected by diverging them from the normal patterns with the reconstruction error produced by the autoencoder. To this end, we build an end-to-end trainable adversarial attention-based autoencoder network, called Ada-Net, to make the reconstructed frames indistinguishable from original frames. The Ada-Net combines an autoencoder network and a GAN model that is used to benefit enhancing the reconstruction ability of the autoencoder. To further improve the reconstruction performance, we integrate an attention model into the decoder to dynamically select informative parts of encoding features for decoding. The attenion mechanism is helpful to preserving important information for learning intrinsic normal patterns. Evaluations on four challenging datasets, including the Subway, the UCSD Pedestrian, the CUHK Avenue, and the ShanghaiTech datasets, demonstrate the effectiveness of the proposed method.**

***Index Terms*—Abnormal event detection, Ada-Net, attention mechanism, generative adversarial network.**

I. Introduction

In recent years, detecting abnormal events in videos has attracted growing attentions from both academia and industry [23], [2], [22], [57], [8]. It still remains a challenging problem due to the low resolution, complex and crowded scenes, unpredictability of individual appearance, and irregular pedestrian motion trajectories in videos [41], [28], [5]. Some early approaches detect abnormal events by classifying them into specified event categories [65]. However, with the rapid growth of surveillance videos, the categories of abnormal events are usually non-deterministic. In addition, since the abnormal events rarely happen in the real world, it is difficult to collect sufficient data for training robust classifiers.

Many recent methods approach this problem by extracting normal patterns from training videos and detecting abnormalies as events deviated from normal patterns [51], [61], [17], [7]. Due to the outperformance of deep learning on various visual tasks [19], [63], [16], [4], a variety of deep neural networks have been proposed for anomaly detection in an unsupervised learning way [34], [28]. Hasan *et al.* [17] proposed a fully convolutional autoencoder to learn the regular dynamics in long videos. In their method, the network just processes the frames into different channels, without effectively modeling the temporal relationship between sequential frames. Chong *et al.* [7] presented a spatiatemporal deep architecture to learn the regular patterns. The network encodes the frames into sptial feature representations with spatial convolution and then learns the temporal evolution of the spatial features with convolutional LSTM, and the Euclidean distance is used to compute the reconstruction error between the input frames and the reconstructed frames.

In this paper, we propose an attention-based autoencoder to discover normal patterns via an adversarial learning strategy. The adversarial learning is employed to make the reconstructed frames indistinguishable from the original frames, which improves the reconstruction performance. The attention mechanism is leveraged to automatically select the important information for effective decoding. Specifically, we build an adversarial attention-based autoencoder network, called Ada-Net, which is trained in an end-to-end manner without any supervision in the training data. The Ada-Net consists of an attention-based autoencoder network and a generative adversarial network (GAN), as illustrated in Figure 1. Inspired by the idea of the GAN, we introduce an adversarial loss as a regularization to train the autoencoder for reconstructing the frames. A discriminator is designed to distinguish the reconstructed frames frome the original frames and the decoder in the autoencoder is treated as a generator to generate the reconstructed frames. In adversarial learning, the decoder is trained to maximally confuse the discriminator so that the discriminator loses the ability of classifying the original and reconstructed frames.

To effectively reconstruct the frames at the pixel level, the encoder in the autoencoder is constructed by spatial convolutional layers to capture the spatial structure within frames and by a stack of convolutional LSTMs [55] to explore the temporal information between frames. Accordingly, the decoder consists of a stack of attention-based convolutional LSTMs and spatial de-convolutional layers. A local attention model is integrated into the convolutional LSTMs in the decoder to select more relevant parts of the feature maps for decoding. Softly associated with the encoding feature maps and the last decoded hidden state, the weights of all the

This work was supported in part by the Natural Science Foundation of China under Grants No.61673062.

Hao Song, Che Sun, Xinxiao Wu, and Yunde Jia are with the Beijing Lab of Intelligent Information Technology, the School of Computer Science, Beijing Institute of Technology, Beijing 100081, China. Email: songhao, sunche, wuxinxiao, jiayunde@bit.edu.cn. Mei Chen is with the Department of Electrical and Computer Engineering, State University of New York at Albany, NY 12222, USA. Email: meichen@albany.edu.

Hao Song and Che Sun contributed equally. Xinxiao Wu is the corresponding author.

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/TMM.2019.2950530, IEEE Transactions on Multimedia

feature map regions are dynamically calculated to represent the importance of the corresponding regions to the decoding.

Extensive experiments on four challenging datasets have demonstrated that our method can achieve best or competitive results compared with state-of-the-art methods. The contributions of this paper are summarized as follows:

- We propose an Adversarial Attention-based Autoencoder to learn normal patterns for abnormal event detection in an unsupervised way. To this end, a novel deep neural network called Ada-Net is presented by combining an attention-based autoencoder and a GAN model, which can be trained in an end-to-end manner.
- In contrast to traditional measurements of the reconstruction error such as Euclidean distance, we introduce an adversarial loss with respect to the frame discriminator to make the reconstructed frames indistinguishable from the original frames, which improves the reconstruction accuracy of the autoencoder.
- To maintain the important information for learning intrinsic normal patterns, we propose an attention-based convolutional LSTMs to softly select the more relevant feature map regions for reconstructing the frames at the pixel level.

II. Related Work

Recently, many researchers have focused on the abnormal event detection in surveillance videos [44], [42], [27], [40], [33]. The traditional methods of abnormal event detection can be roughly divided into supervised learning [65], [6], [64] and unsupervised learning [66], [39], [10], [54], [3], [17], [36], [7].

A. Supervised learning based anomaly detection.

Some work [65], [6] treats the abnormal event detection as a binary classification problem (normal and abnormal). Zhou *et al.* [65] presented spatial-temporal Convolutional Neural Networks to capture the spatial and temporal information by performing spatial-temporal convolutions. Zhao *et al.* [64] used the spatio-temporal feature and non-negative locality-constrained linear coding to generate high-level representations of videos for abnormal event detection. Sultani *et al.* [47] held that the training data of normal and abnormal events can help a detection system learn better. Thus, they formulated a weakly-supervised learning approach and built a new dataset containing abnormal training data from the Internet. Different from these methods, our method focuses on detecting abnormal events in an unsupervised way to overcome the ambiguous definition of abnormal classes and the limited number of training videos.

B. Unsupervised learning based anomaly detection

1) Traditional learning methods: In contrast to supervised learning, many other methods train the detection models with little or even no supervision. These methods usually resort to learning normal motion patterns in videos and recognize abnormal events by diverging them from the normal patterns. In [66], [39], [10], [54], [3], the trajectories are extracted in advance for moving objects to represent the regular patterns. By analyzing the motion patterns of normal trajectories, abnormal events are identified as ones which do not match the normal motion patterns. Cui *et al.* [10] tracked interest points and proposed interaction energy potential to model the relationship among a group people for exploring the normal/abnormal patterns. Besides, the methods of sparse coding [49], [62], [9], [32] are widely utilized in anomaly detection and represent the regular patterns by a linear combination of basis with sparsity.

2) Deep learning methods: With the success of deep learning on image classification [24], [45], many researchers pay attention to solving the abnormal detection problem with deep networks by reconstructing videos [7], [13]. Hasan *et al.* [17] presented an autoencoder to effectively learn the regular dynamics in long-duration videos which is applied to identify irregularity. Xu *et al.* [56] proposed an appearance and motion deepNet to automatically learn feature representations, and utilized one-class SVM to predict abnormal events. Medel and Savakis [36] introduced a composite convolutional LSTM network to predict the evolution of a video sequence and detect anomalous video segments using a regularity evaluation algorithm at the output of the LSTM. Chong and Tay [7] used a spatio-temporal architecture for anomaly detection that consists of two components, one is for spatial feature representation, and the other is for learning the temporal evolution. Wang *et al.* [52] proposed a self-adaptive strategy to predict normal events and detected abnormal events by a two-stage unsupervised method without any priorly knowing of normal events. Jamadand *et al.* [21] proposed a video prediction framework called PredGAN for abnormal event detection. Ionescu *et al.* [20] adopted an unmasking technique to the abnormal event detection task for detecting abnormal events without training sequences. Different from these deep networks, our Ada-Net introduces an attention mechanism into the autoencoder to automatically select important informative parts for reconstructing normal patterns in videos. Benefiting from the good performance of the Generative Adversarial Network (GAN), Liu *et al.* [31] introduced a U-Net encoder-decoder with skip connections as the generator coupled with a patch-based discriminator to predict future frames for anomaly detection. Different from [31], our method applies attention mechanism into the convolutional LSTMs as the generator of GAN to learn normal patterns of events. The attention model is able to automatically discover different importance of different parts of normal patterns for effective decoding by learning weights of different regions of different feature maps.

III. Adversarial Attention-based Autoencoder Network

A. Overview

The Adversarial Attention-based Autoencoder Network (Ada-Net) consists of two components: an attention-based autoencoder network and a GAN model, as shown in Figure 1. The decoder in the autoencoder is treated as a generator of the GAN.

Our network encodes the normal patterns in videos with a small reconstruction error. Given a sequence of video frames

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/TMM.2019.2950530, IEEE Transactions on Multimedia

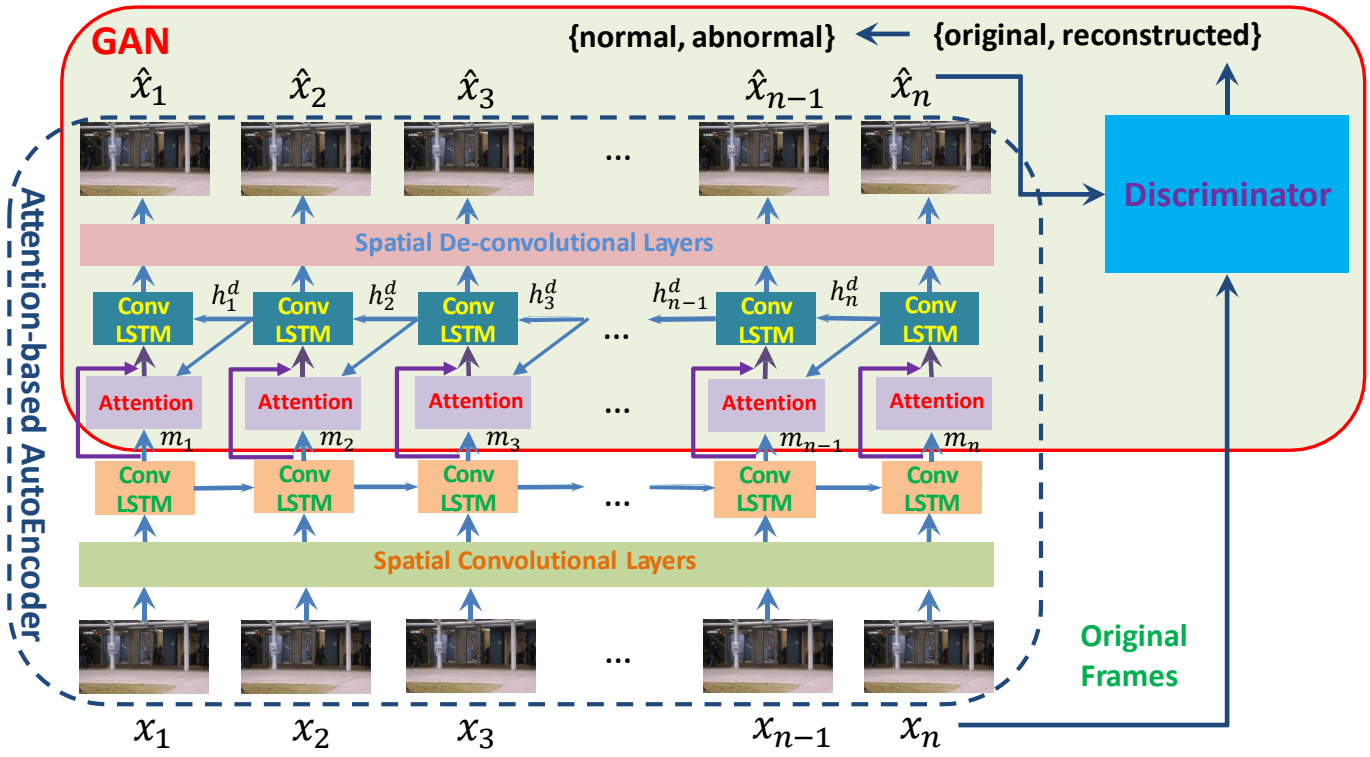


Fig. 1. The architecture of the adversarial attention-based autoencoder network (Ada-Net). The Ada-Net consists of an attention-based autoencoder and a GAN model. The decoder in the autoencoder is treated as the generator in the GAN.

$X = \{x_1, x_2, \cdots, x_{N_t}\}$ where N_t is the number of frames, we first build the spatial convolutional layers to learn the spatial structures within the frames. With the effectiveness of LSTM in modeling the temporal relationships between sequential frames, we introduce a stack of convolutional LSTMs to encode the frames into their corresponding feature maps $M = \{m_1, m_2, \cdots, m_{N_t}\}$ where m_t represents the feature maps of the input frames x_t. In decoding, we design a stack of attention-based convolutinal LSTMs to generate the reconstructed feature maps of the input frames. We then apply the de-convolutional spatial layers to produce the reconstructed sequential frames $\hat{X} = \{\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{N_t}\}$. In the GAN, the generator (i.e., the decoder) aims to reconstruct the video sequences to confuse the discriminator, and the discriminator tries to distinguish the original X and the reconstructed $\hat{X}$.

B. Network Architecture

1) Encoder: To reconstruct the sequential frames at the pixel level, motivated by the traditional networks [24], [45], [48], we build two spatial convolutional layers to model the spatial structures within the input frames. The spatial convolution operation tries to maintain the spatial relationships between pixels by learning image features using small squares of the input data. With the spatial convolutional layers, an input frame can be effectively encoding into informative feature maps. Then, a stack of convolutional LSTMs (ConvLSTM) is employed to capture the temporal information between the sequential frames.

Long Short-Term Memory (LSTM) [18] is capable of learning long-term dependencies on sequential data and has been successfully applied to various visual tasks [30], [14], [29], [25]. Different from the traditional LSTM, the ConvLSTM replaces all the input-to-state and state-to-state with the convolution operations, which has been successfully utilized in the task of video prediction [38]. In this way, the ConvLSTM introduces fewer parameters and generates more descriptive spatial feature maps. Accordingly, each cell in the ConvLSTM can be computed as follows:

$$\begin{aligned}
\hat{C}_t &= \tanh(W_c \odot x_t + U_c \odot h_{t-1} + b_c), \\
i_t &= \sigma(W_i \odot x_t + U_i \odot h_{t-1} + b_i), \\
f_t &= \sigma(W_f \odot x_t + U_f \odot h_{t-1} + b_f), \\
o_t &= \sigma(W_o \odot x_t + U_o \odot h_{t-1} + b_o), \\
C_t &= f_t * C_{t-1} + i_t * \hat{C}_t, \\
h_t &= o_t * \tanh(C_t),
\end{aligned} \tag{1}$$

where i_t, f_t and o_t are the input, forget and output gates at the time-step t, respectively. C_t is the cell memory and h_t corresponds to its hidden state. $\odot$ stands for the convolution operation. W_c, W_i, W_f, W_o, U_c, U_i, U_f, U_o are the weight metrics, and b_c, b_i, b_f, b_o are the bias of ConvLSTM. x_t is the input and $*$ represents the element-wise multiplication.

2) Attention-based Decoder: In the decoding phase, we propose an attention-based convolutional LSTM to decode the feature maps generated by the encoder. Let $M = \{m_1, m_2, \cdots, m_N\}$ be the encoding feature maps of input frame sequences and f^d be the decoding function of the

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/TMM.2019.2950530, IEEE Transactions on Multimedia

attention-based LSTM. The hidden state h_t^d of the decoder at time t can be computed by

$$h_t^d = f^d(h_{t-1}^d, m_t^a). \tag{2}$$

Here h_{t-1}^d stands for the previously decoding hidden state of the frame $t-1$ and m_t^a represents the attention-based feature maps of the input frame t with the size of $K \times K \times D$ where D is the number of filters.

We introduce attention maps to automatically select the most informative regions for decoding, since the attention maps are able to represent the contributions of all the fields in the feature maps. To achieve this goal, a soft attention mechanism is leveraged to automatically compute the weights of different regions in the feature maps. The calculation of weights is determined by the encoding feature maps as well as the last decoding hidden state, characterized by a corresponding indicator u_t:

$$u_t = W_u^T \odot f(W_d \odot h_{t-1}^d + W_e \odot m_t + b_a), \tag{3}$$

where W_u, W_d, and W_e are the model parameters and b_a is the bias weight. f is an activation function. The value $f(W_d \odot h_{t-1}^d + W_e \odot m_t + b_a)$ reflects the matching degree of h_{t-1}^d and m_t. W_u is used to convert the matching degree to the "spatial attention distribution" u_t in the feature map m_t. Let α_t represent the final attention maps on the feature maps of the input frame t, which are normalized by

$$\alpha_t^{ijk} = \frac{\exp(u_t^{ijk})}{\sum_l \exp(u_t^{ijl})}. \tag{4}$$

The attention value α_t^{ijk} indicates the weight of the region (i, j, k) in the feature maps where i and j respectively represent the horizontal and vertical positions of the attention map, and k indicates the position of the filters. From Eq. (4), we calculate the attention value α^{ijk} to represent the important weight at the position (i, j) across the k-th filters. Specifically, at the k-th filter, if α^{ijk} is larger, it means that the location (i, j) of the encoding feature map is more important for the decoder; otherwise, it means that the location (i, j) of the encoding feature map is less important. Accordingly, the attention-based feature maps of the input frame t can be given by

$$m_t^a = \alpha_t * m_t. \tag{5}$$

The attention maps α_t indicate the importance of the encoding feature maps in decoding the next state h_t^d. With the attention mechanism in the decoder, the encoder does not need to encode all the information in the input sequence of frames, and the attention model can help dynamically select the informative region of encoding feature maps in the decoder. Then we append two spatial de-convolutional layers to generate the reconstructed frames.

3) Generative Adversarial Network: Generative Adversarial Network (GAN) [15] has been used in several domains such as Natural Language Processing [26], [59] and Image Reconstruction [60], [58], which has demonstrated its effectiveness on guiding the reconstructions. The GAN framework consists of two competing neural networks: a generative model G and a discriminative model D. G and D compete with each other in a two-player min-max game. The generator G aims to produce realistic samples to confuse the discriminator D while the discriminator D tries to distinguish the generated samples from real data correctly. More formally, G and D can be trained jointly via solving

$$\min_G \max_D [\mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))], \tag{6}$$

where $\mathbf{x}$ is the true data sample, $\mathbf{z}$ is the input of the generator G, and $\mathbb{E}$ is the empirical estimate of the expected value of the probability.

In our GAN model, the generator generates the reconstructed frames which are similar to the original frames while the discriminator tries to differentiate them correctly. Compared with traditional methods of measuring the reconstruction errors, our method of adversarial learning between a generator and a discriminator is beneficial to further improving the reconstruction accuracy.

Particularly, the generator of our network is constructed by the decoder in the autoencoder, and the discriminator consists of a stack of ConvLSTMs with one fully-connected layer. Each ConvLSTM contains two layers with 32 filters, and the kernel size is 28×28 with the stride of 3. The input to the discriminator is the original sequential frames or the reconstructed sequential frames. The output is a binary output, i.e., original or reconstructed.

To train the GAN, the adversarial loss $\mathcal{L}_{adv}$ is defined as

$$\mathcal{L}_{adv} = \mathbb{E}_X[\log(D_{\theta_a}(X))] + \mathbb{E}_M[\log(1 - D_{\theta_a}(G_{\theta_d}(M)))], \tag{7}$$

where $D_{\theta_a}(\cdot)$ denotes the discriminator in the GAN with the parameter θ_a. $G_{\theta_d}(\cdot)$ denotes the generator (i.e., the decoder in the autoencoder) with the parameter θ_d. X represents the input video with the sequential frames $x_1, x_2, ..., x_{N_t}$. M represents the feature maps of X by the encoder.

C. Learning

In order to train the autoencoder, a reconstruction loss is introduced which is based on the Euclidean distance of the input sequences of frames and the reconstructed sequences of frames from the output of the decoder, formulated as

$$\mathcal{L}_{rec} = \frac{1}{2N} \sum_i ||X_i - \hat{X}_i||_2^2 + \gamma(||\theta_e||_2^2 + ||\theta_d||_2^2), \tag{8}$$

where X_i and $\hat{X}_i$ indicate the i-th input sequence of frames (i.e., i-th input video) and the corresponding reconstructed sequence of frames, respectively. N is the size of the mini batch. Suppose $F_{\theta_e}(\cdot)$ represents the encoder in the autoencoder with the parameter θ_e, then $M_i = F_{\theta_e}(X_i)$ represents the output feature maps of the input X_i. By the decoder $G_{\theta_d}(\cdot)$, the reconstructed sequence of frames is given by $\hat{X}_i = G_{\theta_d}(M_i)$. The regularization terms of $||\theta_e||_2^2$ and $||\theta_d||_2^2$ are introduced to prevent the parameter learning from overfitting. γ is a hyper-parameter to balance the reconstruction error and the regularization.

We train θ_e and θ_d using the reconstruction loss $\mathcal{L}_{rec}$, and update θ_d as well as θ_a using the adversarial loss $\mathcal{L}_{adv}$. An iterative learning algorithm is designed to jointly optimize θ_e,

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/TMM.2019.2950530, IEEE Transactions on Multimedia

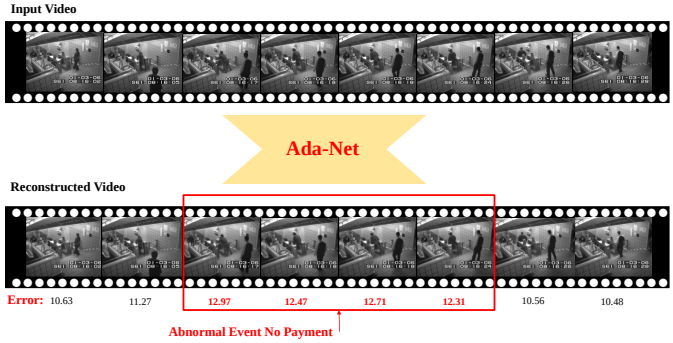


Fig. 2. A video is reconstructed with the proposed Ada-Net. The frames with high reconstruction errors will be detected as the abnormal event.

θ_d and θ_a:
1. Update θ_e by minimizing $\mathcal{L}_{rec}$;
2. Update θ_d by minimizing ($\mathcal{L}_{rec}$ + $\mathcal{L}_{adv}$);
3. Update θ_a by maximizing $\mathcal{L}_{adv}$.
The training procedure of the Ada-Net is summarized in Algorithm 1.

Algorithm 1 The training procedure of the Ada-Net.

Input: The training sequences of frames $\{X_1, X_2, \cdots, X_N\}$ where N is the size of the mini-batch;
Output: The parameters $\{\theta_e, \theta_d, \theta_a\}$ of the Ada-Net.
1: Initialize the parameters $\{\theta_e, \theta_d, \theta_a\}$.
2: **for** iteration number **do**
3: Encode $\{X_1, X_2, \cdots, X_N\}$ to $\{M_1, M_2, \cdots, M_N\}$ using the encoder.
4: Decode $\{M_1, M_2, \cdots, M_N\}$ to $\{\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_N\}$ by the attention-based decoder.
5: Update the parameters $\{\theta_e, \theta_d, \theta_a\}$:
6: $\theta_e \leftarrow$ -$\nabla(\mathcal{L}_{rec})$.
7: $\theta_d \leftarrow$ -$\nabla(\mathcal{L}_{rec}$ + $L_{adv})$.
8: $\theta_a \leftarrow$ +$\nabla(\mathcal{L}_{adv})$.
9: **end for**

D. Anomaly Score

In the detection procedure as shown in Figure 2, with one forward pass, the reconstruction error e_t of all the pixel values in frame t is computed by the Euclidean distance between the input frame and the reconstructed frame. Then the calculated Euclidean distances of all the frames are normalized to the range of [0,1). Finally, the anomaly score s_t of the frame t can be given by

$$s_t = \frac{e_t - \min_t e_t}{\max_t e_t}. \quad (9)$$

IV. Experiments

A. Datasets

We evaluate our method on four challenging benchmarks: the Subway [2], the UCSD [35], the Avenue [32], and the ShanghaiTech [34] datasets. All the training videos in the experiments are normal events.

The **Subway** dataset contains two scenarios: the entrance (1 hour 36 minutes with 144249 frames) and exit (43 minutes with 64900 frames). The abnormal events include walking in the wrong direction, no payment, loitering, irregular interactions between people, and miscellaneous. The first 15 minutes of both the entrance and exit videos are used for training and the rest of videos are used for testing.

The **UCSD** dataset consists of two sub-datasets: Ped1 and Ped2, which record the pedestrian walkways. The Ped1 dataset contains 34 and 36 video clips in the training and testing sets, respectively. Each video clip consists of 200 frames with the size of 158×238. The Ped2 dataset has 16 training and 12 testing video clips with different numbers of frames. Anomalies of these two datasets can be summarized as: carts, cars, the person skating or bicycling among pedestrians.

The **Avenue** dataset has 16 training and 21 testing video clips with 35240 frames, totally. Each video clip lasts about 2 minutes long. The anomalies include running, walking in opposite direction, throwing objects and loitering.

The **ShanghaiTech** dataset contains 13 scenes with complex light conditions and various viewpoints. This dataset has 130 abnormal event and over $270,000$ training frames.

Figure 3 shows several examples of the abnormal events of these four datasets.

B. Evaluation Metric

We apply the ROC (Receiver Operating Characteristic) curve and the corresponding AUC (Area Under Curve) as the evaluation metrics, which are commonly used in the abnormal event detection task. Moreover, the EER (Equal Error Rate) is introduced to evaluate the equal probability of miss-classifying a positive or negative sample in ROC curve. In this paper, all the evaluations are based on the frame level, where we compute the anomaly scores of frames.

C. Implementation

In the encoder of our Ada-Net, we use two spatial convolution layers. The number of filters are 128 and 64, respectively. The kernel size of the first layer is 11×11 with the stride of 4. The kernel size of the second layer is 5×5 with the stride of 2. The ConvLSTM contains two layers with 32 filters, and the kernel size is 28×28 with the stride of 3. The decoder has the same parameter settings of the encoder. In the attention layer, the kernel size of W_u, W_d and W_e are set to 1×1, containing 32 filters.

In the training process, we initialize the parameters of the autoencoder with the parameters of a pre-trained recurrent autoencoder model trained on feature sequences from original sequential frames. In the reconstruction process by using the autoencoder, it is shown in [46] that a decoder LSTM which attempts to reconstruct the reverse sequence is easier to train. So we reconstruct the video sequence in the reverse order with the attention-based decoder LSTM. Note that we have the similar order of the sequential frames in calculating the reconstruction errors. To effectively train the discriminator, we append a prior uniform distribution to the input to regularize the learning of the discriminator. The Ada-Net is trained with

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/TMM.2019.2950530, IEEE Transactions on Multimedia

Fig. 3. The typical examples of the abnormal events from the Avenue, Subway, ShanghaiTech and UCSD datasets.

TABLE I
ABNORMAL EVENT DETECTION RESULTS IN TERMS OF FRAME-LEVEL AUC AND EER ON THE AVENUE DATASET.

Method	AUC	EER
Del Giorno *et. al* [11]	78.3%	-
Tudor *et. al*l [50]	80.6%	-
Lu *et. al* [32]	80.9%	-
Chong *et. al* [7]	80.3%	20.7%
Hasan *et. al* [17]	70.2%	25.1%
Ionescu *et. al* [20]	80.6%	-
Luo *et. al* [34]	81.7%	-
Liu *et. al* [31]	84.9%	-
Wang *et. al* [52]	85.3%	23.9%
Ours (Ada-Net)	**89.2%**	**17.6%**

TABLE II
ABNORMAL EVENT DETECTION RESULTS IN TERMS OF FRAME-LEVEL AUC AND EER ON THE SUBWAY DATASET.

Method	Subway (Entrance)		Subway (Exit)	
	AUC	EER	AUC	EER
Mehran *et. al* [37]	67.5%	31.0%	55.6%	42.0%
Wang *et. al* [53]	81.6%	22.8%	84.9%	17.8%
Xu *et. al* [56]	-	-	87.9%	**6.8%**
Hasan *et. al* [17]	**94.3%**	26.0%	80.7%	9.9%
Ionescu *et. al* [20]	70.6%	-	85.7%	-
Wang *et. al* [52]	-	-	84.5%	21.4%
Chong *et. al* [7]	84.7%	23.7%	94.0%	9.5%
Ours (Ada-Net)	90.2%	**22.67%**	**94.6%**	9.3%

adam optimizer with default parameters, and the Ada-Net network is implemented using the TensorFlow toolkit [1].

D. Quantitative Analysis

1) Results on the Avenue Dataset: Table I shows the quantitative comparison of our approach with several state-of-the-art methods on the Avenue dataset in terms of Area Under the Curve (AUC) and Equal Error Rate (EER). The performances of the compared methods are taken from the original papers. Although there are many works about detecting anomalies, some of them do not evaluate on the Avenue dataset [37], [56], and several other methods do not report the result of EER [32], [11]. Compared with the deep network based methods [17], [7], [34], [31], [52], our Ada-Net at least gains an improvement of 3.9% on the AUC evaluation and an improvement of 3.1% on the EER evaluation, which clearly validates the effectiveness of the proposed Ada-Net.

2) Results on the Subway Dataset: Table II provides the comparison results between our method and several state-of-the-art methods on the Subway dataset in terms of frame-level AUC and EER. Compared with [53], [37], we obtain better AUC results on both the entrance and exit sub-datasets, which indicates that our deep network can find more informative normal patterns than the traditional methods which are based on the hand-crafted features. Compared with the method [17], our method achieves comparable results on the entrance videos. But on the exit videos, our method yields much better results than [17] on the exit videos, which demonstrates the good stability of our method in different scenarios. For the evaluation of EER on the entrance videos, our approach outperforms all the compared methods. This verifies that our method can perform more precise detection of the abnormal events. For the exit gate videos, our Ada-Net achieves the best result on the AUC evaluation. Compared with [56], our method has higher EER. The possible reason is that [56] employs the one-class SVM after feature learning by the autoencoder, which further improves the precision of detection especially on the more complex exit gate videos.

3) Results on the UCSD Dataset: We also report the results on the UCSD Ped1 and Ped2 datasets in Table III. Our method is better than most of the compared methods, which clearly demonstrates the effectiveness of combining attention mechanism and adversarial learning strategy for reconstruction. The methods of [56] and [52] works better than our method, probably due to that [56] and [52] both use the background subtraction technology to extract the local regions for anomaly detection which can improve the performance, but our method does not require any preprocessing techniques such as background subtraction or region detections. To further reduce the EER values, we follow the operation in [43], dividing the video frames in the UCSD dataset into $4 \times 4 = 16$ patches of the same size. Then we appropriately reduce the size of filters in the Ada-Net to train the model and identify abnormal events. The results of our method with 16 patches are shown in Table III. Without the complex background subtraction technology, we achieved the comparable result by simply dividing each frame into patches.

4) Results on the ShanghaiTech Dataset: We also evaluate our method on the ShanghaiTech dataset, as shown in Table IV. The ShanghaiTech dataset is a newly proposed dataset, which contains many pedestrians in a scene. Com-

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/TMM.2019.2950530, IEEE Transactions on Multimedia

TABLE III
ABNORMAL EVENT DETECTION RESULTS IN TERMS OF FRAME-LEVEL AUC AND EER ON THE UCSD DATASET.

Method	UCSD (Ped1)		UCSD (Ped2)	
	AUC	EER	AUC	EER
Adam *et. al* [2]	77.1%	38.0%	-	42.0%
Kim *et. al* [23]	59.0%	-	69.3%	-
Mehran *et. al* [37]	67.5%	31.0%	55.6%	42.0%
Mahadevan *et. al* [35]	74.2%	32.0%	61.3%	36.0%
Wang *et. al* [53]	72.7%	33.1%	87.5%	20.0%
Xu *et. al* [56]	**92.1%**	16%	90.8%	17%
Hasan *et. al* [17]	81.0%	27.9%	90.0%	21.7%
Ionescu *et. al* [20]	68.4%	-	82.2%	-
Chong *et. al* [7]	89.9%	**12.5%**	87.4%	12.0%
Liu *et. al* [31]	83.1%	-	95.4%	-
Wang *et. al* [52]	77.8%	29.2%	**96.4%**	**8.9%**
Ours (Ada-Net)	90.4%	15.8%	90.3%	15.5%
Ours (with 16 patches)	90.5%	**11.9%**	90.7%	11.5%

TABLE IV
ABNORMAL EVENT DETECTION RESULTS IN TERMS OF FRAME-LEVEL AUC ON THE SHANGHAITECH DATASET.

Method	AUC	EER
Hasan *et. al* [17]	60.85%	-
Luo *et. al* [34]	68.00%	-
Ours (Ada-Net)	**70.00%**	36.5%

pared with [17], [34], our method performs best with the grains of 5.8% and 1.9%, respectively, on the AUC evaluation.

5) Impacts of different components in Ada-Net: We also compare the contributions of different components in our Ada-Net, and the results are shown in Table V. "w/o GAN", "w/o attention" and "w/o attention & GAN" represent the method of removing the discriminator, the method of decoding without attention mechanism, and the method of reconstructing the frames without the attention-mechanism and the discriminator, respectively. It is interesting to observe that: (1) When discarding the discriminator, our method drops about 5%, which demonstrates the benefit of the GAN model in training autoencoder by using the adversarial learning, and (2) The performance is significantly improved by integrating the attention mechanism into the decoder, since the attenon model is capable of selecting the important and informative parts of the input feature maps for decoding.

6) Event Count: Following the settings in [17], to reduce the noise in the regularity score, we assume that the local minima within 50 frames belongs to the same abnormal event. The length of the abnormal event is reasonable because the anomaly should last at least about 2-3 seconds long to be meaningful.

Table VI shows the number of detected anomalies and false alarm on the three datasets. For both the Ped1 and Ped2 of the UCSD dataset, we achieved better results than [36], [17]. When the Ada-Net detects the same number of abnormal events, it produces less false alarms. For the Avenue dataset, the Ada-Net can detect the abnormal event more precisely, despite it generates more false alarms. For the subway dataset, we achieve better performance than other methods. The result demonstrates that the Ada-Net can determine the temporal region of the anomalies more accurately, which makes it more practical in real scenes.

E. Qualitative Results

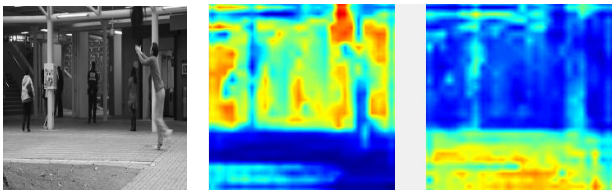
Figure 4 shows several examples of the detected abnormal events using the Ada-Net on the Subway Entrance and Exit gate dataset, and the Avenue dataset. The qualitative results demonstrate that our method can effectively detect anomalies in the crowded scenes. The detected abnormal events are "no payment", "wrong direction", "running" on the Subway Entrance gate video, "clean the wall" on the Subway Exit gate video, and "opposite direction", "throwing bag" on the Avenue dataset.

F. Visualizing Attention Maps

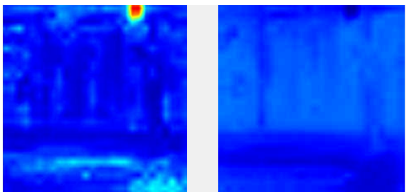
Figure 5, Figure 6, and Figure 7 show some of the learned filter responses of our model on the Avenue dataset, Subway (Entrance) dataset, and the UCSD (Ped1) dataset, respectively.

Figure 5(a) shows one gray image with an abnormal event: throwing a bag. Figure 5(b) visualizes two filter responses of the encoding feature maps m_t. Figure 5(c) visualizes two corresponding filter responses of the attention maps α. These two filters in the attention maps show opposite responses to the abnormal object-the bag in the human's hand. We can see that the response of the first filter is high (red color) while that of the second filter is low (blue color). The first filter of α acts on the corresponding filter of m_t, which can be described as the filter that focuses on the encoding feature map at the corresponding filter position of the abnormal object. The second filter works on other regions, ignoring the encoding feature map at the corresponding filter position of the abnormal object. All other activated filters show similar characteristics.

Similarly, the filters of the attention maps in Figure 6(c) focus on the abnormal object (the loitering people), and those in Figure 7(c) focus on the abnormal car in the pedestrian strip mostly.



(a) Input Image (b) Responses of Encoding Feature Maps



(c) Responses of Attention Feature Maps

Fig. 5. Filter responses of the encoding featurmaps and attention feature maps trained on the Avenue dataset.

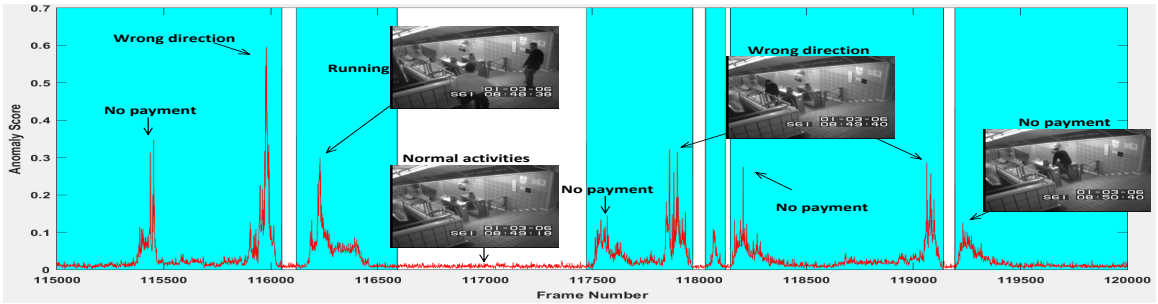
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2950530, IEEE Transactions on Multimedia

TABLE V
THE AUC RESULTS OF DIFFERENT COMPONENTS OF THE ADA-NET ON THE FOUR PUBLIC DATASETS.

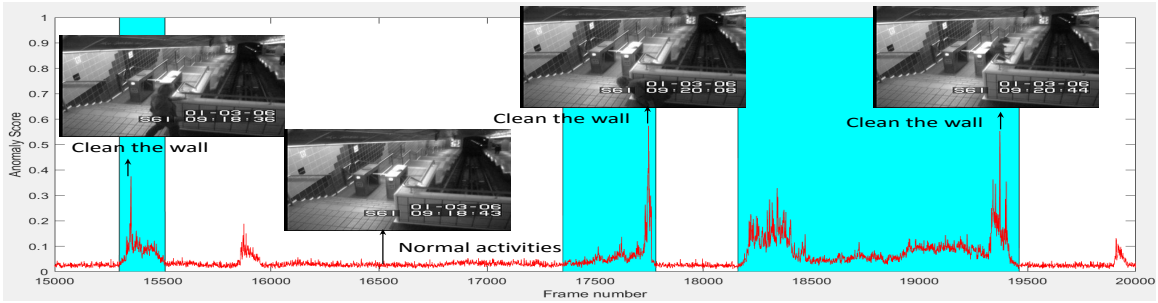
Method	Avenue		Entrance		Exit		Ped1		Ped2		ShanghaiTech	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
w/o GAN	85.3%	21.1%	81.7%	23.5%	93.0%	11.2%	88.9%	18.2%	89.7%	17.8%	68.4%	38.3%
w/o attention	82.4%	22.1%	82.1%	25.2%	92.9%	12.9%	87.7%	20.3%	87.4%	19.6%	64.8%	39.4%
w/o attention & GAN	81.2%	23.1%	80.7%	27.8%	91.7%	13.5%	87.1%	20.8%	85.6%	21.7%	62.5%	41.9%
Ours (Ada-Net)	89.2%	17.6%	90.2%	22.67%	94.6%	9.3%	90.4%	15.8%	90.3%	15.5%	70.0%	36.5%

TABLE VI
THE NUMBER OF DETECTED ABNORMAL EVENTS AND FALSE ALARM ON THE THREE PUBLIC DATASETS. GT STANDS FOR GROUDTRUTH VALUES OF EVENT COUNT.

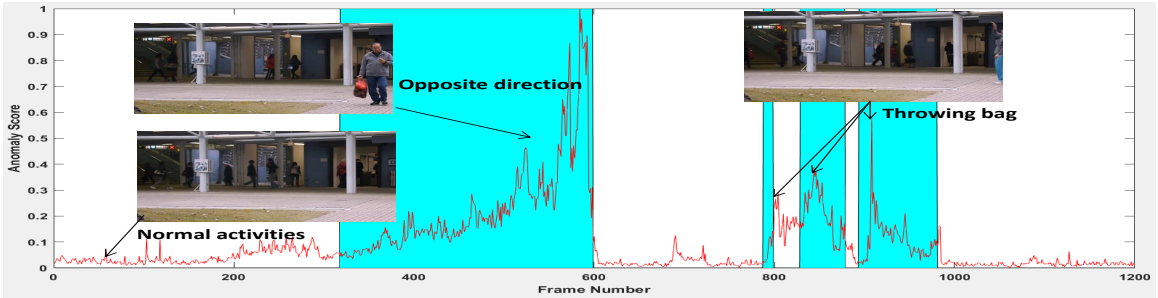
Method	True Positives/ False Alarm				
	UCSD Ped1 GT:40	UCSD Ped2 GT:12	Subway Entrance GT:66	Subway Exit GT:19	Avenue GT:47
Lu *et. al* [32]	-	-	57/4	19/2	-
Kim *et. al* [23]	-	-	56/3	18/0	-
Dutta *et. al* [12]	-	-	60/5	19/2	-
Zhao *et. al* [62]	-	-	60/5	19/2	-
Medel *et. al* [36]	40/7	12/1	62/14	19/37	40/2
Hasan *et. al* [17]	38/6	12/1	61/15	17/5	45/4
Chong *et. al* [7]	-	-	61/9	18/10	44/12
Ours(Ada-Net)	40/6	12/1	62/9	19/9	45/10



(a) Subway Entrance Dataset



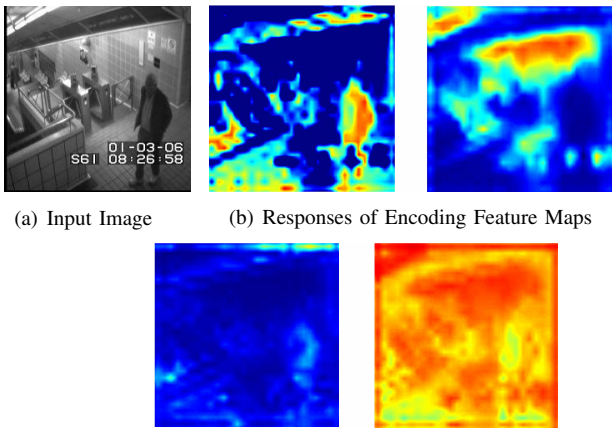
(b) Subway Exit Dataset



(c) Avenue Dataset

Fig. 4. The qualitative results of our Ada-Net on the Subway Entrance dataset and Avenue dataset. We list the computed anomaly scores of a portion of frames in the test video. The positive true abnormal events are "no payment", "wrong direction", "running", "clean the wall", "opposite direction", and "throwing bag".

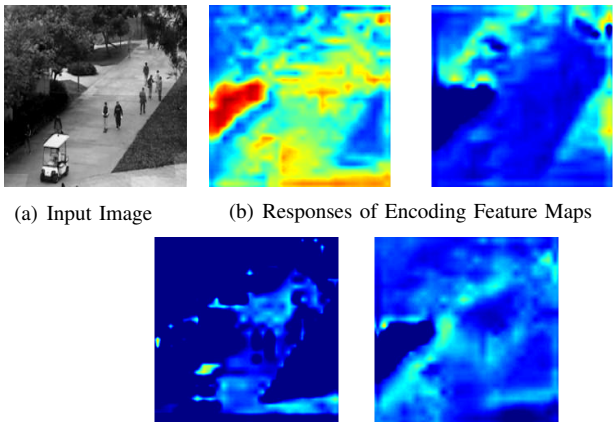
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2950530, IEEE Transactions on Multimedia



(a) Input Image (b) Responses of Encoding Feature Maps

(c) Responses of Attention Feature Maps

Fig. 6. Filter responses of the encoding featurmaps and attention feature maps trained on the Subway (Entrance) dataset.



(a) Input Image (b) Responses of Encoding Feature Maps

(c) Responses of Attention Feature Maps

Fig. 7. Filter responses of the encoding featurmaps and attention feature maps trained on the UCSD (Ped1) dataset.

V. Conclusion

In this work, we have presented an adversarial attention-based autoencoder (Ada-Net) that can discover normal patterns and detect abnormal events in videos. The adversarial learning strategy is used to replace the traditional reconstruction errors to enhance the reconstruction ability of the Ada-Net, and the attention mechanism helps the decoder reconstruct the original frames with more informative encoding feature maps, which can preserve important information for learning intrinsic normal patterns. Extensive experiments on four public datasets can validate the effectiveness of our method. In our future work, we will pay attention to both the temporal and spatial dimensions to increase the reconstruction accuracy.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, p. 555, 2008.

[3] N. Anjum and A. Cavallaro, "Multifeature object trajectory clustering for video analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1555–1564, 2008.

[4] Y. Bengio, P. Lamblin, P. Dan, and H. Larochelle, "Greedy layer-wise training of deep networks," in *International Conference on Neural Information Processing Systems*, 2006, pp. 153–160.

[5] S. Biswas and R. V. Babu, *Anomaly detection via short local trajectories*. Elsevier Science Publishers B. V., 2017.

[6] H. Chen, X. Zhao, T. Wang, M. Tan, and S. Sun, "Spatial-temporal context-aware abnormal event detection based on incremental sparse combination learning," in *Intelligent Control and Automation*, 2016, pp. 640–644.

[7] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," pp. 189–196, 2017.

[8] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Transactions on Multimedia*, 2018.

[9] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, no. 7, pp. 1851–1864, 2013.

[10] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3161–3167.

[11] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *European Conference on Computer Vision*. Springer, 2016, pp. 334–349.

[12] J. K. Dutta and B. Banerjee, "Online detection of abnormal events using incremental coding length," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 3755–3761.

[13] Y. Feng, Y. Yuan, and X. Lu, "Deep representation for abnormal event detection in crowded scenes," pp. 591–595, 2016.

[14] R. D. Geest and T. Tuytelaars, "Modeling temporal structure with lstm for online action detection," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1549–1557.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[16] D. Guo, W. Li, and X. Fang, "Fully convolutional network for multiscale temporal action proposals," *IEEE Transactions on Multimedia*, 2018.

[17] M. Hasan, J. Choi, J. Neumann, A. K. Roychowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. Hou, X. Wu, Y. Sun, and Y. Jia, "Content-attention representation by factorized action-scene network for action recognition," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1537–1547, 2018.

[20] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 2914–2922.

[21] A. Jamadandi, S. Kottursheltar, and U. Mudenagudi, "Predgan-a deep multi-scale video prediction framework for detecting anomalies in videos," 2018.

[22] F. Jiang, Y. Wu, and A. K. Katsaggelos, "Detecting contextual anomalies of crowd motion in surveillance video," in *IEEE International Conference on Image Processing*, 2009, pp. 1113–1116.

[23] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2921–2928.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.

[25] C. Li, J. Cao, Z. Huang, L. Zhu, and H. T. Shen, "Leveraging weak semantic relevance for complex video event classification," in *IEEE International Conference on Computer Vision*, 2017, pp. 3667–3676.

[26] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and J. Dan, "Adversarial learning for neural dialogue generation," *arXiv preprint arXiv:1701.06547*, 2017.

[27] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Transactions on Circuits Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[28] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[29] J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[30] J. Liu, G. Wang, P. Hu, L. Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3671–3680.

[31] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - A new baseline," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 6536–6545.

[32] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.

[33] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017*, 2017, pp. 439–444.

[34] W. Luo, L. Wen, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *IEEE International Conference on Computer Vision*, 2017.

[35] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[36] J. R. Medel and A. E. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," *CoRR*, vol. abs/1612.00390, 2016.

[37] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.

[38] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *arXiv preprint arXiv:1511.06309*, 2015.

[39] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[40] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition: A review," *IEEE Transactions on Systems Man and Cybernetics Part C*, vol. 42, no. 6, pp. 865–878, 2012.

[41] H. Ren, W. Liu, S. I. Olsen, S. Escalera, and T. B. Moeslund, "Unsupervised behavior-specific dictionary learning for abnormal event detection." in *BMVC*, 2015, pp. 28–1.

[42] M. J. Roshtkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions ," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1436–1452, 2013.

[43] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 3379–3388.

[44] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2112–2119.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[46] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning*, 2015, pp. 843–852.

[47] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 6479–6488.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[49] X. Tang, S. Zhang, and H. Yao, "Sparse coding based motion attention for abnormal event detection," in *IEEE International Conference on Image Processing*, 2013, pp. 3602–3606.

[50] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2895–2903.

[51] C. Wang, H. Yao, and X. Sun, "Anomaly detection based on spatio-temporal sparse representation and visual attention analysis," *Multimedia Tools and Applications*, vol. 76, pp. 1–17, 2016.

[52] S. Wang, Y. Zeng, Q. Liu, C. Zhu, E. Zhu, and J. Yin, "Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, 2018, pp. 636–644.

[53] T. Wang and H. Snoussi, "Histograms of optical flow orientation for abnormal events detection," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*. IEEE, 2013, pp. 45–52.

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

[55] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[56] 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*, vol. 156, pp. 117–127, 2016.

[57] K. Xu, X. Jiang, and T. Sun, "Anomaly detection based on stacked sparse coding with intraframe classification strategy," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1062–1074, 2018.

[58] J. Yang, A. Kannan, D. Batra, and D. Parikh, "Lr-gan: Layered recursive generative adversarial networks for image generation," *arXiv preprint arXiv:1703.01560*, 2017.

[59] Z. Yang, W. Chen, F. Wang, and B. Xu, "Improving neural machine translation with conditional sequence generative adversarial nets," *arXiv preprint arXiv:1703.04887*, 2017.

[60] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *IEEE International Conference on Computer Vision*, 2017, pp. 2868–2876.

[61] D. Zhang, D. Gatica-Perez, S. Bengio, and I. Mccowan, "Semi-supervised adapted hmms for unusual event detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 611–618.

[62] B. Zhao, F. F. Li, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3313–3320.

[63] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.

[64] Y. Zhao, L. Zhou, K. Fu, and J. Yang, "Abnormal event detection using spatio-temporal feature and nonnegative locality-constrained linear coding," in *IEEE International Conference on Image Processing*, 2016, pp. 3354–3358.

[65] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing Image Communication*, vol. 47, pp. 358–368, 2016.

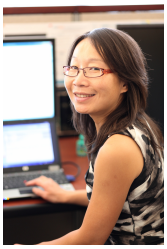
[66] S. Zhou, W. Shen, D. Zeng, and Z. Zhang, "Unusual event detection in crowded scenes by trajectory analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 1300–1304.

Hao Song received the B.S. degree from North China Electric Power University (NCEPU), Baoding, China, in 2012. He is currently pursuing the Ph.D. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, under the supervision of Prof. Y. Jia. His research interests include computer vision, machine learning and video retrieval.

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/TMM.2019.2950530, IEEE Transactions on Multimedia

Che Sun received the B.S. degree from Beijing Institute of Technology(BIT), Beijing, China, in 2017. He will pursuing the Ph.D. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, under the supervision of Assistant Prof. Yuwei Wu. His research interests include computer vision, machine learning.

Xinxiao Wu (M'09) is an Associate Professor in the School of Computer Science at the Beijing Institute of Technology. She received the B.A. degree in computer science from the Nanjing University of Information Science and Technology in 2005 and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. She was a post-doctoral research fellow at Nanyang Technological University, Singapore, from 2010 to 2011. Her current research interests include machine learning, computer vision, and video analysis and understanding.



Mei Chen is an Associate Professor in the Electrical and Computer Engineering Department at the State University of New York, Albany. She was the Intel Principal Investigator for the Intel Science and Technology Center on Embedded Computing hosted at Carnegie Mellon University. Previously she held researcher and research lead positions at Intel Labs, HP Labs, and SRI Sarnoff Corporation. Meis work in computer vision and biomedical image analysis were nominated finalists for 6 Best Paper Awards and won 3. She earned a Ph.D. in Robotics from the School of Computer Science, Carnegie Mellon University, and a M.S. and B.S. from Tsinghua University, Beijing, China.

Yunde Jia (M'11) received the B.S., M.S., and Ph.D. degrees from the Beijing Institute of Technology (BIT) in 1983, 1986, and 2000, respectively. He was a visiting scientist with the Robot Institute, Carnegie Mellon University (CMU), from 1995 to 1997. He is currently a Professor with the School of Computer Science, BIT, and the team head of BIT innovation on vision and media computing. He serves as the director of Beijing Lab of Intelligent Information Technology. He has authored over 300 publications in computer vision and media computing. In recent years, his interests have extended to vision-based HCI and HRI, intelligent robotics, and cognitive systems.