BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection

Sangmin Lee, *Student Member, IEEE*, Hak Gu Kim, *Member, IEEE*, and Yong Man Ro, *Senior Member, IEEE*

***Abstract*— Abnormal event detection is an important task in video surveillance systems. In this paper, we propose novel bidirectional multi-scale aggregation networks (BMAN) for abnormal event detection. The proposed BMAN learns spatio-temporal patterns of normal events to detect deviations from the learned normal patterns as abnormalities. The BMAN consists of two main parts: an inter-frame predictor and an appearance-motion joint detector. The inter-frame predictor is devised to encode normal patterns, which generates an inter-frame using bidirectional multi-scale aggregation based on attention. With the feature aggregation, robustness for object scale variations and complex motions is achieved in normal pattern encoding. Based on the encoded normal patterns, abnormal events are detected by the appearance-motion joint detector in which both appearance and motion characteristics of scenes are considered. Comprehensive experiments are performed, and the results show that the proposed method outperforms the existing state-of-the-art methods. The resulting abnormal event detection is interpretable on the visual basis of where the detected events occur. Further, we validate the effectiveness of the proposed network designs by conducting ablation study and feature visualization.**

***Index Terms*— Video analysis, abnormal event detection, normal pattern encoding, multi-scale.**

I. Introduction

RECENTLY, surveillance videos are being acquired in various environments such as car black boxes, industrial factories, and CCTVs in public places due to concerns over security and safety. Such videos are intended to detect meaningful moments (*i.e.*, abnormal events) like accidents, process errors, and crimes. However, it is labor-intensive and time-consuming for people to manually check all the video sequences to find abnormal events because most of the acquired scenes are normal and meaningless. Therefore, automatic abnormal event detection is needed to reduce labor and time resources.

For detecting abnormal events, an intuitive approach is to learn abnormal patterns directly. However, there are some problems in modeling abnormal patterns. First, it is difficult to obtain sufficient abnormal data for modeling abnormal patterns because abnormal events occur sporadically. Second, the definition of abnormal events is not bounded. Thus, it is difficult to model all possible abnormal events in various environments. One reasonable solution is to utilize normal data that can easily be obtained. By modeling normal patterns, events that deviate from the modeled normal patterns can be considered as abnormal ones.

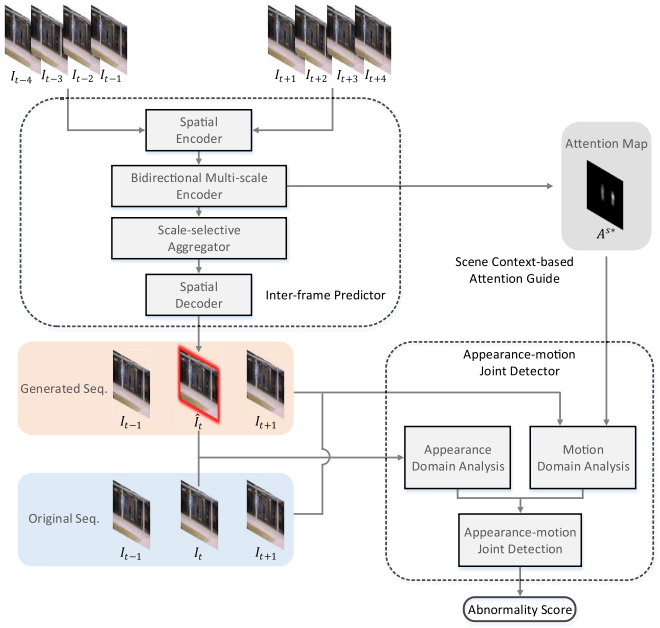

Fig. 1. Overview of the BMAN. An inter-frame predictor takes previous and later video sequences to generate an inter-frame and an attention map. An appearance-motion joint detector receives the attention map, the generated sequence, and the original sequence to output an abnormality score.

Some previous works have proposed abnormal event detection based on normal pattern modeling. Trajectory-based methods [1]–[4] usually employ tracking algorithms to extract dynamic information. Based on the tracking results, statistical modeling methods are applied to acquire normal patterns. However, since tracking algorithms are not robust to occluded and crowded scenes, trajectory-based methods could be vulnerable to such conditions [5]. Other common abnormal event detection methods utilize local low-level features [6]–[11]. These methods model spatio-temporal patterns by utilizing low-level features such as HOF [12] and HOG [13]. However, they require prior knowledge to design appropriate features for various events [5].

In recent years, generative model-based methods using deep learning have achieved state-of-the-art performances in

Manuscript received November 21, 2018; revised May 21, 2019 and August 11, 2019; accepted October 3, 2019. Date of publication October 24, 2019; date of current version January 10, 2020. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korea Government (MSIT) (No. 2017-0-00780, Development of VR sickness reduction technique for enhanced sensitivity broadcasting). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giacomo Boracchi. *(Corresponding author: Yong Man Ro.)*

The authors are with the Image and Video Systems Laboratory, School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea (e-mail: sangmin.lee@kaist.ac.kr; hgkim0331@kaist.ac.kr; ymro@ee.kaist.ac.kr).

Digital Object Identifier 10.1109/TIP.2019.2948286

abnormal event detection [14]–[21]. Since the generative models are trained to create frames for normal events, they could not properly create frames for abnormal events at testing time. Differences between the original frames and the generated frames can be used to discriminate abnormal events from normal events. However, the previous deep generative networks are not sufficiently capable of learning spatio-temporal features of complex events, aggregating unidirectional or single-scale features. Thus, the previous deep generative networks are not fully capable of capturing complex events.

In this paper, we propose novel bidirectional multi-scale aggregation networks (BMAN) in which normal spatio-temporal features are learned to detect abnormal events. The proposed BMAN consists of two main parts: an inter-frame predictor and an appearance-motion joint detector (see Fig. 1). The inter-frame predictor receives previous and later frames to create an inter-frame. Since the inter-frame predictor is trained to generate target inter-frames for normal scenes only, it could not generate inter-frames clearly for abnormal scenes at testing time. Thus, abnormal events can be discriminated from normal events by analyzing the differences between the generated frames and the corresponding original frames. The appearance-motion joint detector receives a generated sequence and an original sequence. It outputs an abnormality score considering both appearance domain analysis and motion domain analysis.

To properly distinguish abnormal events from normal events, it is required to effectively learn the spatio-temporal characteristics of normal events. The proposed work focuses on encoding multi-scale spatio-temporal features to address object scale variations in abnormal event detection. In surveillance environments, an object scale could have large variations due to the actual object size and the distance from the camera. We do not know at which scale abnormal events would occur. Therefore, multi-scale encoding is needed in abnormal event detection. The proposed model learns normal patterns of various scales using the multi-scale encoding. Abnormal events can be compared with the normal patterns learned at different scales, which makes it possible to detect abnormal events with various scales. To the best of our knowledge, the proposed method is a first attempt to handle multi-scale encoding with fully learnable neural networks in abnormal event detection.

The proposed inter-frame predictor is composed of four subparts: (i) a spatial encoder for representing the spatial features of frames, (ii) a bidirectional multi-scale encoder for extracting bidirectional multi-scale motion features, (iii) a scale-selective aggregator for encoding which scale is important for a target scene, and (iv) a spatial decoder for generating an inter-frame. In addition to the multi-scale encoding, it is worth focusing on specific regions in abnormal event detection because actual events occur locally in surveillance environments. To this end, attention encoding is devised to effectively focus on movement regions by considering the spatio-temporal scene context in the inter-frame predictor. The resulting attention maps are utilized as feature refiners in the inter-frame predictor and also as a supplementary motion guide in the appearance-motion joint detector. To the best of our knowledge, this is a first work to deal with attention encoding with fully learnable neural networks in abnormal event detection.

Based on the normal patterns encoded with the inter-frame predictor, an appearance-motion joint detector is devised to detect abnormal events by analyzing both appearance and motion characteristics of target scenes. The appearance domain analysis uses pixel-level differences between a generated frame and an original frame while the motion domain analysis uses flow-level differences between a generated sequence and an original sequence. For the motion domain analysis, we adopt a learnable motion domain [22] that can be trained with normal videos in an unsupervised way. Furthermore, the scene context-based attention from the inter-frame predictor is utilized as a supplementary motion guide in the motion domain analysis. Finally, appearance-motion joint detection is achieved by combining two domain analysis results. Different from the previous works with deep generative models, the proposed method detects abnormal events on both the appearance and the motion domains without utilizing previously designed handcrafted features. The proposed appearance-motion fusion is constructed as a fully learnable scheme with normal data, which is more suitable for detecting abnormal events. Analysis in each detection domain works complementarily to improve the detection performance.

We evaluate the performance of the proposed method in terms of both quantitative and qualitative aspects with various public datasets. Extensive experimental results show that the proposed method outperforms the existing state-of-the-art methods and has interpretability by presenting the visual basis of where the detected events occur. In addition, we verify the effectiveness of the network designs through ablation study. In particular, the effectiveness of the multi-scale encoding is analyzed through attention feature visualization.

The major contributions of this paper are as follows.

1) To the best of our knowledge, the proposed method is a first attempt to deal with multi-scale and attention schemes with fully learnable neural networks in abnormal event detection. Utilizing bidirectional multi-scale and attention encoding, the proposed model robustly learns normal patterns including object scale variations and complex motions.
2) Based on the learned normal patterns, the appearance-motion joint detector is devised to detect abnormal events by analyzing appearance and motion characteristics of target scenes. Analyses in appearance and motion domains work complementarily for detecting abnormal events. In addition, the resulting detection is interpretable on the visual basis of where the detected events occur at each scene.

II. Related Work

A. Trajectory-Based Methods

Trajectory-based abnormal event detection methods utilize the dynamic information of objects usually relying on the

tracking algorithms. Makris and Ellis [1] propose a semantic activity-based Bayesian approach with motion tracking. Hu *et al.* [2] introduce a method for learning statistical motion patterns with multi-object tracking. Jiang *et al.* [3] propose a method based on hidden markov model (HMM) with trajectory clustering. Mo *et al.* [4] exploit a trajectory-based joint sparse reconstruction framework. The trajectory-based methods show satisfactory performances in simple scenes containing few objects without occlusions. However, those methods do not guarantee satisfactory results in complex environments where tracking algorithms do not work robustly. Unlike these previous works, our method does not depend on tracking algorithms and could be more robust to complex and occluded scenes.

B. Local Low-Level Feature-Based Methods

This approach models normal patterns with spatio-temporal characteristics using locally extracted low-level features. Zaharescu and Wildes [6] adopt distributions of spatio-temporal oriented energy to model behavior. Wang and Snoussi [7] utilize histograms of the orientation of optical flow (HOF) to represent motion information. Kaltsa *et al.* [8] introduce a dynamic representation on histograms of oriented swarms (HOS). Cheng *et al.* [9] present a hierarchical framework based on gaussian process regression (GPR). Colque *et al.* [10] propose a spatio-temporal descriptor, histograms of optical flow orientation and magnitude and entropy (HOFME). Leyva *et al.* [11] exploit a compact set of features with optical flow and foreground occupancy information. Moreover, there is an approach utilizing both trajectory analysis and pixel-level analysis [23]. Unlike these previous works, our method does not require any prior knowledge when designing features on various events. The proposed model is based on a data-driven approach and is learned by specific normal data without utilizing any prior handcrafted features.

C. Deep Learning-Based Methods

There are many neural network architectures for handling video data in deep learning field. A recurrent neural network (RNN), which recurrently receives sequential inputs has been generally adopted to deal with sequential features of video data. A long short-term memory (LSTM) [24], a type of the RNN, is proposed to solve the vanishing gradients problem in long-term encoding by utilizing input, output, and forget gate units. In [25], a convolutional LSTM (ConvLSTM) is proposed by modifying fully connected layers with convolutional layers in LSTM to capture spatio-temporal features. In terms of generation tasks for video analysis, a convolutional auto-encoder structure [26] has been utilized alone or with the RNN. The convolutional auto-encoder contains encoding and decoding configurations with convolutions and deconvolutions, which enables to analyze video data. Recently, generative adversarial networks (GAN) framework [27] is adopted to generate data close to real distribution. In the GAN, the generator tries to deceive the discriminator by generating more realistic data, and the discriminator tries to distinguish generated data from real data. Through this GAN process, the generator creates data more similar to the real distribution.

In recent years, deep learning has shown more effective performances than traditional methods in various computer vision tasks [27]–[29]. In the abnormal event detection task, several methods using deep learning have been proposed to achieve effective performances. The deep learning-based abnormal event detection methods usually adopt a generative model to learn normal patterns in an unsupervised way. These methods utilize the assumption that if a generative model is trained to create frames for normal scenes only, then it cannot properly create frames for abnormal scenes. Therefore, differences between original frames and generated frames can be used to detect abnormal events. Hasan *et al.* [14] utilize the convolutional auto-encoder to reconstruct video frames for learning temporal regularity. Medel and Savakis [15] employ the ConvLSTM to predict future frames for abnormal event detection. Then, architectures that combine the convolutional auto-encoder with the ConvLSTM are proposed to reconstruct frames for detecting abnormal events [16], [17]. More recently, the GAN is adopted in the abnormal event detection task. GAN-based abnormal event detection methods include spatio-temporal adversarial networks [18], future prediction networks [19], and image-to-image translation networks [20], [21]. These methods detect abnormal events by predicting frames or changing domains with adversarial learning.

Although the image-to-image translation model [20], [21] tries to detect abnormal events in motion domains, it is difficult to consider actual motion characteristics in such motion domains because only a single frame is used as input to generate the motion domain map. Overall, the previous deep learning-based works adopt simple feature aggregation networks, which are not sufficiently capable of learning complex spatio-temporal patterns. Compared to the previous deep learning-based methods, our work focuses on effective spatio-temporal feature encoding with bidirectional multi-scale feature aggregation and appearance-motion joint detection.

III. Proposed Method

A. Inter-Frame Predictor

Fig. 2 shows the network configuration of the proposed inter-frame predictor. The inter-frame predictor is trained with normal videos to generate a clear frame for normal scenes only. Therefore, the inter-frame predictor could not generate inter-frames properly for abnormal scenes at testing time. In the inter-frame predictor, the context of the surrounding sequence can be utilized to determine whether the target frame is abnormal. When there are normal events involving complex motions such as actions with legs or arms, it is possible to predict the frame better for such normal events by using the inter-frame predictor because the previous and later contexts of the target inter-frame are utilized together. To consider the previous and later contexts, the inter-frame predictor encodes bidirectional spatio-temporal features in a multi-scale manner and aggregate those bidirectional features in scale-wise. With the bidirectional configuration, the inter-frame predictor

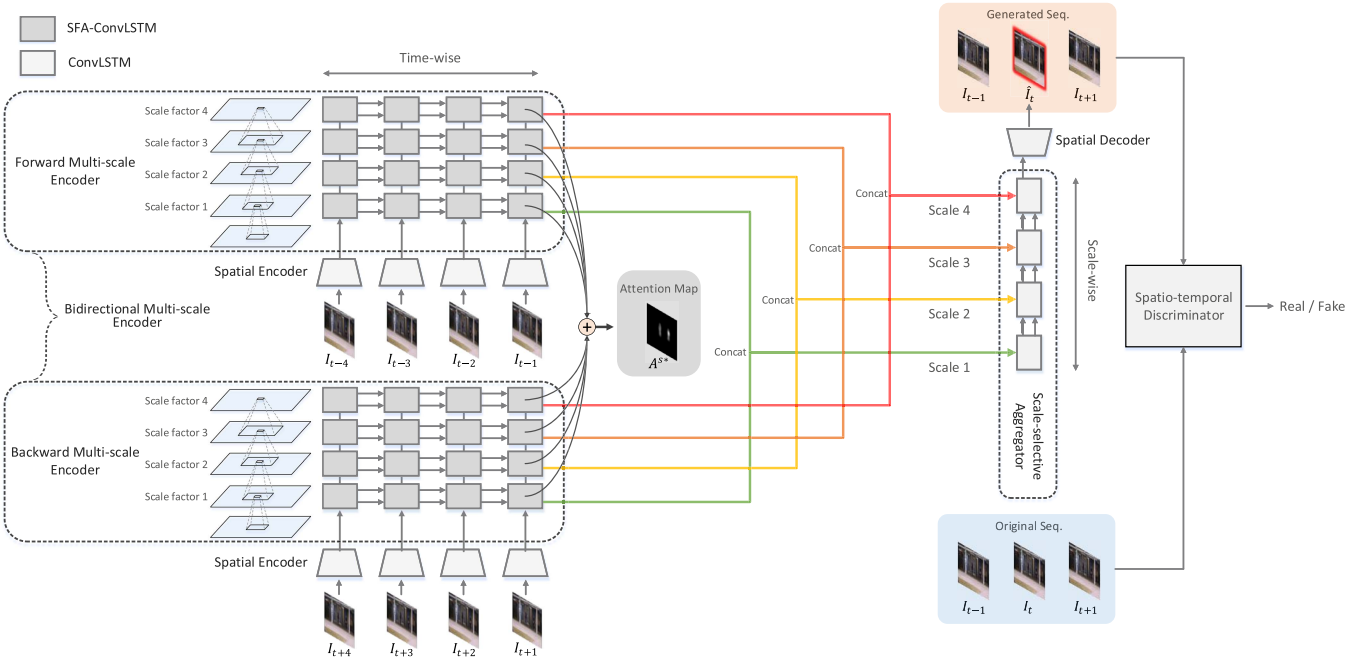

Fig. 2. Network configuration of the inter-frame predictor that consists of four sub-parts: a spatial decoder, a bidirectional multi-scale encoder, a scale-selective aggregator, and a spatial decoder. The inter-frame predictor is trained with a spatio-temporal discriminator in an adversarial way for normal pattern encoding. At testing time, the inter-frame predictor outputs the inter-frame and the attention map without the spatio-temporal discriminator.

can effectively address normal events that include complex motions. However, it could be difficult to precisely generate the frame for the normal events containing complex motions by simply predicting the future only based on the previous context, which could lead to false alarms in abnormal event detection.

The proposed inter-frame predictor consists of a spatial encoder, a bidirectional multi-scale encoder, a scale-selective aggregator, and a spatial decoder (see Fig. 2). With the spatial encoder, latent spatial features are encoded to represent the visual information of each frame. The spatial features of the previous four frames and the later four frames are fed into the forward and the backward multi-scale encoders, respectively, to encode the spatio-temporal features in both directions. Each multi-scale encoder consists of a stack of four SFA-ConvLSTM that is described in sub-section B. By adjusting a scale factor of each SFA-ConvLSTM, the differences between the receptive fields of each scale are adjusted to encode various scales of motion information. Each SFA-ConvLSTM receives cell states and hidden states obtained from the previous step SFA-ConvLSTMs. Then, the hidden states resulting from the last step of each SFA-ConvLSTM are concatenated in each scale pair to combine forward and backward features. The scale-selective aggregator that consists of the typical ConvLSTM [25] receives the concatenated hidden states as sequential inputs. The scale-selective aggregator focuses on which scale information is important at which part according to LSTM gate mechanism [30]. As shown in Fig. 2, the scale-selective aggregator outputs a hidden state, a spatio-temporal latent feature of the inter-frame. Finally, the inter-frame is generated by the spatial decoder. In the process of generating the inter-frame, the scene context-based attention maps from the last step of the multi-scale encoding are obtained. The obtained maps are utilized as feature refiners in the inter-frame predictor and used later in the appearance-motion joint detector at testing time. More details on the attention scheme are covered in sub-sections B and C.

The proposed bidirectional prediction requires the four next frames of the target frame to determine whether or not the target frame is abnormal. It is possible to apply the proposed model to real applications by allocating a buffer for the next four frames. It results in a detection delay of 4 frames, which corresponds to only 0.16 sec for 25 fps videos and 0.06 sec for 60 fps videos. Such marginal delay for the target frame could also happen in existing abnormal event detection methods [6], [15], [31], [32] that use future frames.

We additionally employ the spatio-temporal discriminator to help the predictor learn the spatio-temporal features of normal patterns at training time. The spatio-temporal discriminator consists of 3D convolutional layers (*i.e.*, 3D CNN). The 3D CNN has one more dimension for temporal encoding than the typical 2D CNN. The 3D CNN can reliably determine whether the sequence is real or fake by considering both spatial and temporal characteristics of scenes [33]. The discriminator outputs a 4×4 probability map for exploiting the local adversarial loss [34]. We average the loss values over 4×4 patches for training. Consecutive frames including the generated inter-frame are considered as a fake sequence (generated sequence) while a real sequence (original sequence) contains the real inter-frame (original inter-frame). Through the adversarial learning [27], the spatio-temporal discriminator

TABLE I
NETWORK DETAILS OF THE PREDICTOR AND THE DISCRIMINATOR

Inter-frame Predictor			Spatio-temporal Discriminator		
Layer	**Filter/ Stride**	**Output Size ($h\times w\times c$)**	**Layer**	**Filter/ Stride**	**Output Size ($l\times h\times w\times c$)**
Conv1	3×3/ (2, 2)	128×128×32	3D Conv1	2×3×3/ (1, 2, 2)	2×128×128×64
Conv2	3×3/ (1, 1)	128×128×32			
Conv3	3×3/ (2, 2)	64×64×64	3D Conv2	2×3×3/ (1, 2, 2)	1×64×64×128
Conv4	3×3/ (1, 1)	64×64×64			
Forward multi-scale encoder (Scale 1-4)	3×3/ (1, 1)	64×64×64	3D Conv3	1×3×3/ (1, 2, 2)	1×32×32×256
Backward multi-scale encoder (Scale 1-4)	3×3/ (1, 1)	64×64×64	3D Conv4	1×3×3/ (1, 2, 2)	1×16×16×512
Scale selective aggregator	3×3/ (1, 1)	64×64×128	3D Conv5	1×3×3/ (1, 2, 2)	1×8×8×1024
DeConv1	3×3/ (1, 1)	64×64×64	3D Conv6	1×3×3/ (1, 2, 2)	1×4×4×2048
DeConv2	3×3/ (2, 2)	128×128×32			
DeConv3	3×3/ (1, 1)	128×128×32	3D Conv7	1×3×3/ (1, 1, 1)	1×4×4×1
DeConv4	3×3/ (2, 2)	256×256×3			

is trained to determine whether the input sequence is real or fake. The inter-frame predictor is trained to generate the inter-frame that can fool the discriminator. Note that the discriminator is only utilized at training time. The network details of the inter-frame predictor and the spatio-temporal discriminator are seen in Table I.

To generate a desirable inter-frame, we utilize a pixel-wise loss and a generative adversarial loss for the inter-frame predictor. By minimizing the pixel-wise loss between the t-th real frame $I_t \in \mathbb{R}^{256\times256\times3}$ and the t-th generated frame $\hat{I}_t \in \mathbb{R}^{256\times256\times3}$, the generated frame becomes similar to the real frame at the pixel-level. Let $P_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{\phi}}$ denote the predictor function with parameter $\boldsymbol{\theta}$ and the discriminator function with parameter $\boldsymbol{\phi}$, respectively. The pixel-wise loss ℓ_{pixel} can be written as

$$\ell_{\text{pixel}}(\boldsymbol{\theta}; t) = \| P_{\boldsymbol{\theta}}(I_{t-4}, \ldots, I_{t-1}, I_{t+1}, \ldots, I_{t+4}) - I_t \|_2^2. \quad (1)$$

In addition, by minimizing the generative adversarial loss, the inter-frame predictor can force the discriminator to fail to classify the generated sequence as fake. Let $\hat{S}_t \in \mathbb{R}^{3\times256\times256\times3}$ denote the fake sequence including the generated frame $\hat{I}_t$ and $S_t \in \mathbb{R}^{3\times256\times256\times3}$ denote the real sequence with the real frame I_t. The generative adversarial loss function ℓ_{Gadv} can be written as

$$\ell_{\text{Gadv}}(\boldsymbol{\theta}; t) = -\log(D_{\boldsymbol{\phi}}(\hat{S}_t)), \quad (2)$$

where $D_{\boldsymbol{\phi}}(.)$ is the probability of the input being real. Finally, the total objective loss $\mathcal{L}_P(\boldsymbol{\theta})$ for the predictor is defined as a combination of (1) and (2).

$$\mathcal{L}_P(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t\in batch} \ell_{\text{pixel}}(\boldsymbol{\theta}; t) + \lambda_L \ell_{\text{Gadv}}(\boldsymbol{\theta}; t), \quad (3)$$

where N is a mini batch size, λ_L is a hyper-parameter to balance the pixel-wise loss and the generative adversarial loss.

A discriminative adversarial loss is designed for training the discriminator. By minimizing the discriminative adversarial loss, the discriminator tries to distinguish between the real sequence and the fake sequence. The objective loss for the discriminator can be written as

$$\mathcal{L}_D(\boldsymbol{\phi}) = \frac{1}{N} \sum_{t\in batch} -\log(1 - D_{\boldsymbol{\phi}}(\hat{S}_t)) - \log(D_{\boldsymbol{\phi}}(S_t)), \quad (4)$$

where the first term, $-\log(1 - D_{\boldsymbol{\phi}}(\hat{\mathbf{S}}_t))$, allows the discriminator to classify the fake sequence as fake. The second term, $-\log(D_{\boldsymbol{\phi}}(\mathbf{S}_t))$, allows the discriminator to classify the real sequence as real.

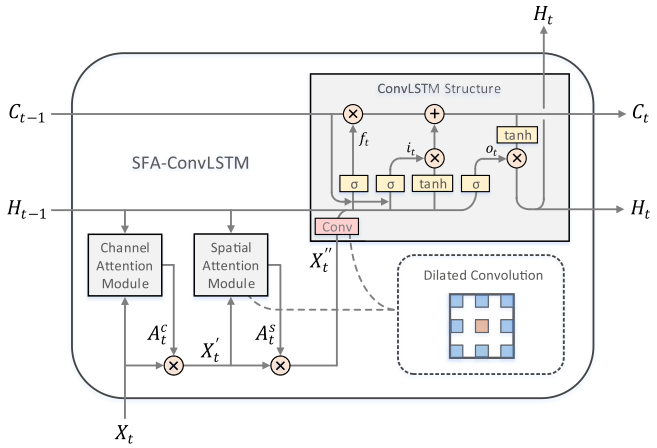

Fig. 3. Network structure of the SFA-ConvLSTM that consists of attention modules and a ConvLSTM structure. Note that dotted box shows a convolutional kernel with scale factor = 2 (dilation rate = 2) used in the spatial attention module and the ConvLSTM structure.

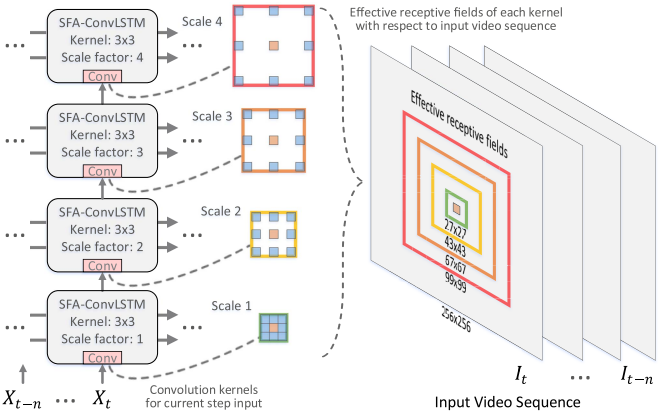

Fig. 4. Illustration of the effective receptive fields that each scale SFA-ConvLSTM covers. The SFA-ConvLSTMs in the inter-frame predictor cover from 27×27 to 99×99 regions with respect to an input video sequence.

B. Scale Factor Attentive-ConvLSTM

The scale factor attentive-ConvLSTM (SFA-ConvLSTM) in the inter-frame predictor has two objectives. First, the

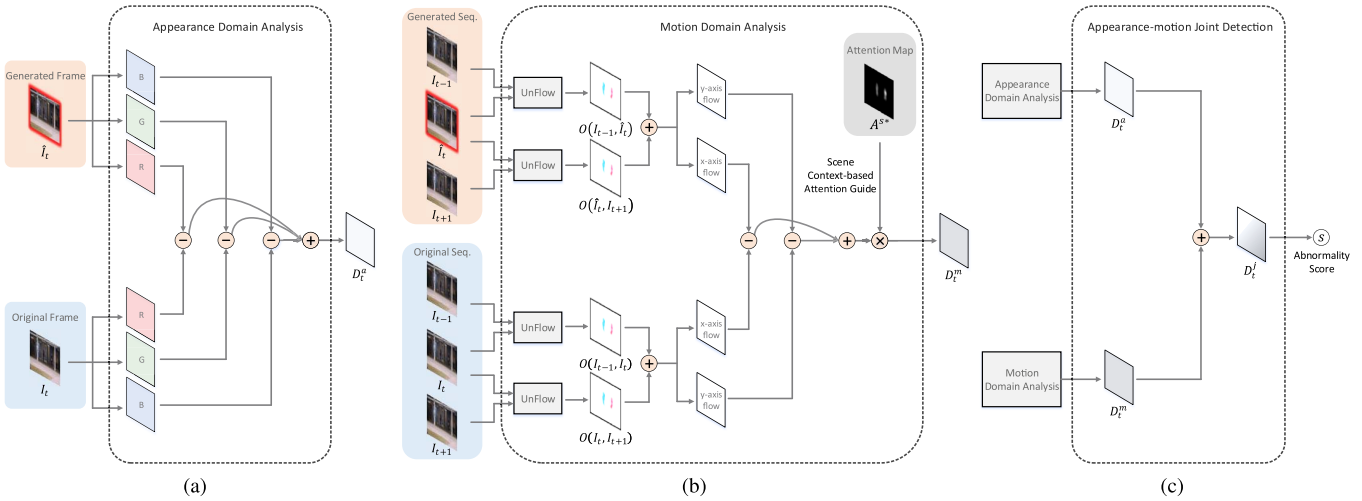

Fig. 5. Framework of the appearance-motion joint detector. (a) Appearance domain analysis with pixel-level comparison between a generated frame and an original frame. (b) Motion domain analysis with flow-level comparison between a generated sequence and an original sequence. (c) Appearance-motion joint detection that combines difference maps from the appearance domain analysis and the motion domain analysis.

SFA-ConvLSTM learns the motion information of various scales. By adjusting dilation rates of the convolutional kernels in the SFA-ConvLSTM, we can control the receptive fields of spatio-temporal encoding. Second, the SFA-ConvLSTM performs attention encoding to focus on movement regions in surveillance environments where events occur locally. The attention maps in the SFA-ConvLSTM refine the features in the inter-frame predictor to be able to focus on local parts considering scene context.

Fig. 3 shows the structure of the SFA-ConvLSTM. Each instance of attention is generated by encoding both the previous step hidden state H_{t-1} and the current step input X_t to account for the scene context according to an attention-based RNN [35]. Taking both H_{t-1} and X_t allows the normal pattern encoding to focus on the specific regions where the actual motion occurs by considering the spatio-temporal features in the SFA-ConvLSTM. A channel attention map A_t^c and a spatial attention map A_t^s are sequentially applied to the input X_t of the SFA-ConvLSTM. Finally, the refined input feature X_t'' is entered to the ConvLSTM structure. We adopt a parallel attention structure of max pooling and average pooling [36]. The proposed attention modules can be formulated as

$$\begin{aligned} A_t^c = \sigma(&MLP_H(AvgPool_s(H_{t-1})) \\ &+ MLP_H(MaxPool_s(H_{t-1})) \\ &+ MLP_X(AvgPool_s(X_t)) \\ &+ MLP_X(MaxPool_s(X_t))), \end{aligned} \tag{5}$$

$$X_t' = A_t^c \circ X_t, \tag{6}$$

$$\begin{aligned} A_t^s = \sigma(&W_H * [AvgPool_c(H_{t-1}); MaxPool_c(H_{t-1})] \\ &+ W_X^d * [AvgPool_c(X_t'); MaxPool_c(X_t')]), \end{aligned} \tag{7}$$

$$X_t'' = A_t^s \circ X_t', \tag{8}$$

where σ denotes the sigmoid function, $Pool_s$ and $Pool_c$ respectively represent the pooling along spatial and channel axes. W denotes the convolutional layer, and W^d denotes the dilated convolutional layer [37]. Each MLP, multi-layer perceptron, consists of two fully connected layers (4, 64) with reduction ratio 16 as in [36]. The hyperbolic tangent (tanh) is used as an activation function in the middle of the MLP.

The proposed attention maps are obtained by learning the parameters in an end-to-end fashion. The attention parameters in the SFA-ConvLSTM are concurrently learned to emphasize the regions needed for predicting a frame when the inter-frame predictor is trained to generate an inter-frame. The proposed attention maps are not only utilized as a supplementary motion guide in the motion domain analysis but also utilized as feature refiners in the inter-frame predictor. Since the scale factor is applied to the proposed attention module, attention maps can be generated suitably for each scale flow in multi-scale encoding through the parameter learning. As the feature refiners, the attention scheme helps the inter-frame predictor to effectively concentrate on the regions by considering local movements.

After applying the attention to X_t, the formulation of the remaining ConvLSTM structure can be represented as

$$i_t = \sigma(W_{xi}^d * X_t'' + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i), \tag{9}$$

$$f_t = \sigma(W_{xf}^d * X_t'' + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f), \tag{10}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc}^d * X_t'' + W_{hc} * H_{t-1} + b_c), \tag{11}$$

$$o_t = \sigma(W_{xo}^d * X_t'' + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o), \tag{12}$$

$$H_t = o_t \circ \tanh(C_t). \tag{13}$$

Unlike the typical ConvLSTM [25], the dilated convolution W^d is used in the proposed network to adjust the receptive field for the input without increasing the number of weight parameters. The scale factor of the SFA-ConvLSTM indicates the dilation rate used in the spatial attention module and the ConvLSTM structure. i_t, f_t, and o_t indicates input, forget, and output gates for controlling how much information to select. The cell state C_t has accumulated information and is further controlled by the output gate o_t to obtain H_t.

The output hidden state H_t represents the high-level feature in which the sequence up to the t-th frame is considered in the procedure of predicting the target inter-frame. Spatial encoding is performed through convolutional layers in the SFA-ConvLSTM and temporal encoding is conducted through stacks of the SFA-ConvLSTM in the temporal axis. Thus, the output H_t contains the spatio-temporal characteristics of the sequence up to the t-th frame.

In the proposed model, the SFA-ConvLSTM is applied to the forward multi-scale encoder and the backward multi-scale encoder parts in the inter-frame predictor. Each scale of the SFA-ConvLSTM encodes input video sequences with different effective receptive fields. The receptive field of the CNN indicates the region of an input frame that can be seen in one kernel at a time. Fig. 4 shows the effective receptive fields of the SFA-ConvLSTM for each scale. Each SFA-ConvLSTM is stacked in the inter-frame predictor. The regions of the kernels for each current step input are manipulated by the scale factor so that effective receptive fields of kernels with respect to the actual input video sequence are varied. As shown in Fig. 4, the kernels with the different scales could cover regions from 27×27 to 99×99 for input frames, which makes it possible to encode multi-scale motion features.

C. Appearance-Motion Joint Detector

At testing time, the appearance-motion joint detector receives the original sequence S_t and the generated sequence $\hat{S}_t$ to output an abnormality score. Fig. 5 shows the overall process of the appearance-motion joint detection. Abnormal event detection in the appearance domain uses the pixel-level difference between the original frame and the generated frame. Abnormal event detection in the motion domain uses the flow-level difference between the original sequence and the generated sequence containing the generated inter-frame.

The appearance-level difference map $D_t^a \in \mathbb{R}^{256\times256}$ in Fig. 5(a) is obtained from the pixel-level difference between the generated frame $\hat{I}_t$ and the original frame I_t. The pixel-level difference maps $D_t^{a,R}$, $D_t^{a,G}$, and $D_t^{a,B}$ for color channels are computed as

$$D_t^{a,R} = \left| \hat{I}_t^R - I_t^R \right|, \tag{14}$$

$$D_t^{a,G} = \left| \hat{I}_t^G - I_t^G \right|, \tag{15}$$

$$D_t^{a,B} = \left| \hat{I}_t^B - I_t^B \right|, \tag{16}$$

where I_t^R, I_t^G, and I_t^B denote R, G, and B channels for the original frame while $\hat{I}_t^R$, $\hat{I}_t^G$, and $\hat{I}_t^B$ denote R, G, and B channels for the generated frame. The average of the pixel-level difference maps is determined as the appearance-level difference map. The appearance-level difference map D_t^a can be written as

$$D_t^a = \frac{D_t^{a,R} + D_t^{a,G} + D_t^{a,B}}{3}. \tag{17}$$

We employ an unsupervised optical-flow extractor, UnFlow [22] for the motion domain analysis. The UnFlow is first trained alone to extract the optical-flow between two frames with corresponding normal data in an unsupervised way. By using the UnFlow trained with the normal data, the motion-level difference map $D_t^m \in \mathbb{R}^{256\times256}$ in Fig. 5(b) is obtained from the flow-level differences between the original sequence S_t (I_{t-1}, I_t, I_{t+1}) and the generated sequence $\hat{S}_t$ $(I_{t-1}, \hat{I}_t, I_{t+1})$ from the inter-frame predictor. For D_t^m, the flow-level difference maps $(D_t^{m,x}, D_t^{m,y})$ are computed from the x-axis and y-axis optical-flows between S_t and $\hat{S}_t$. Then, the two flow-level difference maps and the scene context-based spatial attention maps are aggregated to obtain the motion-level difference map D_t^m. D_t^m has higher values for abnormal events and is utilized to detect abnormal events. Note that scene context-based attention maps are obtained from the inter-frame predictor as shown in Fig. 2. The attention guide map A^{s*} utilized as the supplementary motion guide in the motion domain analysis can be written as

$$A^{s*} = \frac{1}{4}\sum_{k=1}^{4} \frac{A_{t-1}^{s,scale\,k} + A_{t+1}^{s,scale\,k}}{2}, \tag{18}$$

where $A_t^{s,scale\,k}$ denotes the spatial attention from scale k SFA-ConvLSTM in the inter-frame predictor. In other words, A^{s*} is the average map of the attention maps from the last step of each SFA-ConvLSTM. Using the attention map A^{s*} as the motion guide, the proposed motion-level difference map D_t^m can be defined as

$$D_t^{m,x} = \left| \left(\frac{O^x(I_{t-1}, \hat{I}_t) + O^x(\hat{I}_t, I_{t+1})}{2} \right) - \left(\frac{O^x(I_{t-1}, I_t) + O^x(I_t, I_{t+1})}{2} \right) \right|, \tag{19}$$

$$D_t^{m,y} = \left| \left(\frac{O^y(I_{t-1}, \hat{I}_t) + O^y(\hat{I}_t, I_{t+1})}{2} \right) - \left(\frac{O^y(I_{t-1}, I_t) + O^y(I_t, I_{t+1})}{2} \right) \right|, \tag{20}$$

$$D_t^m = A^{s*} \circ \left(\frac{D_t^{m,x} + D_t^{m,y}}{2} \right), \tag{21}$$

where O^x and O^y respectively denote optical flow maps for x-axis and y-axis extracted by the UnFlow network.

We perform the appearance-motion joint detection by combining the two difference maps into an appearance-motion joint difference map $D_t^j \in \mathbb{R}^{256\times256}$. Then, D_t^j is converted to a scalar value and normalized to an abnormality score $s(t)$. The proposed abnormality score $s(t)$ can be defined as

$$D_t^j = D_t^a + \lambda_D D_t^m, \tag{22}$$

$$\tilde{s}(t) = \| D_t^j \|_2^2, \tag{23}$$

$$s(t) = \frac{\tilde{s}(t) - \min_t \tilde{s}(t)}{\max_t \tilde{s}(t) - \min_t \tilde{s}(t)}, \tag{24}$$

where λ_D is a hyper-parameter to balance the appearance domain detection and the motion domain detection. Note that video sequences containing abnormal events have higher abnormality scores.

TABLE II
SUMMARY OF THE DATASETS FOR ABNORMAL EVENT DETECTION

Dataset	Training set #Video clips / #Frames	Testing set #Video clips / #Frames	#Abnormal events	#Scenes	Object scale variation
UCSD Ped2	16 clips / 2,550 frames	12 clips / 2,010 frames	12	1	
UMN	11 clips / 3,300 frames	11 clips / 4,439 frames	11	3	
Avenue	16 clips / 15,324 frames	21 clips / 15,328 frames	47	1	✓
ShanghaiTech	330 clips / 274,515 frames	107 clips / 42,883 frames	130	12	✓

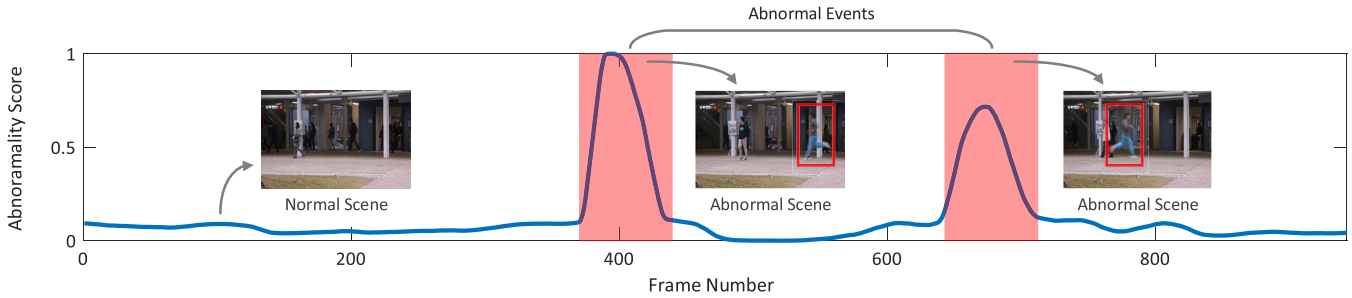

Fig. 6. Abnormality score graph for a testing video in the Avenue dataset. Red regions in the graph indicate frame-level ground truth of abnormal events.

IV. EXPERIMENTS

A. Datasets

To validate the proposed method, we conduct experiments with four public datasets: UCSD [38], UMN [39], Avenue [40], and ShanghaiTech [41] datasets. The training sets of these datasets contain only normal events while the testing sets contain both normal and abnormal events. The datasets used in the experiments are summarized in Table II.

1) UCSD Dataset: The UCSD Pedestrian dataset consists of two subsets, Ped1 and Ped2. We only use the Ped2 dataset because Ped1 has low resolution [42], frame corruption [18], and ambiguity of abnormal events. The UCSD Ped2 dataset consists of 16 video clips for training and 12 video clips for testing with a frame resolution of 360 × 240 pixels. There are 2,550 frames in the training set and 2,010 frames in the testing set. Abnormal events contain occurrences of bicycles, vehicles, and skateboards.

2) UMN Dataset: The UMN Unusual Crowd Activity dataset contains three different scenes with 11 video clips. We use the first 300 frames of each clip as a training set and the remaining as a testing set. To sum up, the training set contains 3,300 frames and the testing set contains 4,439 frames. The frame resolution is 320 × 240 pixels. Abnormal events include people running with panic.

3) Avenue Dataset: The Avenue dataset contains 16 video clips for training and 21 video clips for testing with frames of 640 × 360 pixels. The training set consists of 15,328 frames and the testing set consists of 15,324 frames. The Avenue dataset is more challenging than the UCSD and the UMN datasets because it contains motions of various scale objects and large size objects with complex motions. In addition, abnormal events include various kinds of behaviors such as dancing, throwing, and moving in the wrong direction.

4) ShanghaiTech Dataset: The ShanghaiTech is a large-scale dataset that includes 330 video clips for training and 107 video clips for testing. There are 274,515 frames in the training set and 42,883 frames in the testing set. The resolution of each frame is 856 × 480 pixels. The ShanghaiTech dataset is the most challenging dataset because it includes object scale variations with complex motions and 13 different scenes with complex light conditions. Abnormal events of the dataset contain various situations such as vehicles, personal behaviors, and interactions between two people such as fighting.

B. Implementation

The input frames are normalized to intensity of $[-1, 1]$ and resized to a 256 × 256 resolution. We use an Adam solver [43] to optimize the proposed BMAN with a learning rate of 0.0002 and a batch size of 5. At training time, first, only the inter-frame predictor is trained to minimize the pixel-wise loss ℓ_{pixel}. Then, the predictor and the discriminator are trained in an adversarial way to minimize $\mathcal{L}_P(\boldsymbol{\theta})$ and $\mathcal{L}_D(\boldsymbol{\phi})$ alternately. The predictor training hyper-parameter λ_L in (3) is set as 20. Batch normalization [44] is applied after the convolutions and the deconvolutions of the spatial encoder and decoder except for the last deconvolution. The exponential linear unit (ELU) [45] is used as the activation function on the predictor and the discriminator except for the last layer. At the end of the predictor and the discriminator, tanh and sigmoid are used as the activation functions respectively to match the intensity scale. The UnFlow-C architecture [22] is used as the flow extractor in the appearance-motion joint detector. Note that both the inter-frame predictor and the UnFlow network are trained with corresponding normal data. At testing time, we set the detection hyper-parameter λ_D as 0.2 in (22) for all datasets. The experiments are conducted on a server system

TABLE III
FRAME-LEVEL PERFORMANCE ON THE UCSD PED2

Method	Frame-level AUC (%)
MPPCA [48]	69.3
Social force (SF) [39]	55.6
MPPCA + SF [38]	61.3
MDT [38]	82.9
Saliency Detector [49]	87.7
HOFME [10]	87.5
AMDN [5]	90.8
Conv-AE [14]	90.0
ConvLSTM-AE [16]	88.1
S-RBM [50]	86.4
Optical flow-GAN [20]	93.5
STAE-grayscale [31]	91.2
STAE-optflow [31]	88.6
Unmasking [51]	82.2
Generic Knowledge [42]	92.2
Stacked RNN [41]	92.2
STAN [18]	96.5
PredictionNet [19]	95.4
Proposed method	**96.6**

TABLE IV
PIXEL-LEVEL PERFORMANCE ON THE UCSD PED2

Method	Pixel-level AUC (%)
MPPCA [48]	22.2
Social force (SF) [39]	21.7
MDT [38]	66.5
S-RBM [50]	72.1
Generic Knowledge [42]	**89.1**
Proposed method	86.7

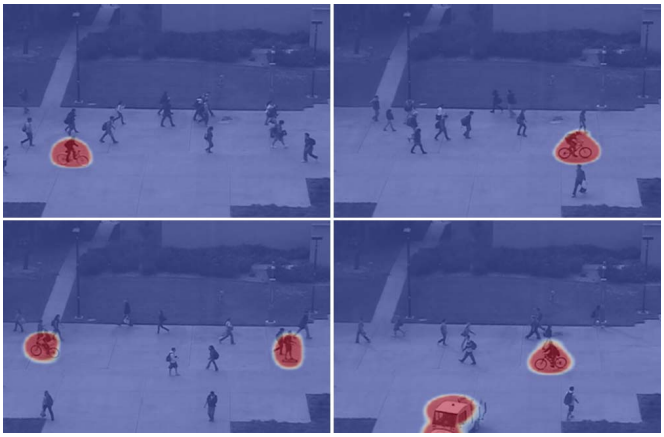

Fig. 7. Visualization results of abnormal events in the UCSD Ped2.

with Intel Xeon Scalable Silver 4114 CPU @ 2.20 GHz, 128 GB memory, and Nvidia TITAN XP GPU. We implement the proposed model in TensorFlow [46].

C. Evaluation

To evaluate the performance of the proposed method, two evaluation metrics are employed. First, a frame-level criterion [47] is used to validate how well the abnormal frames are detected. Fig. 6 shows an example of the abnormality score

TABLE V
FRAME-LEVEL PERFORMANCE ON THE UMN

Method	Frame-level AUC (%)			
	Scene1 Lawn	Scene2 Indoor	Scene3 Plaza	Overall
Optical Flow [39]	-	-	-	84.0
Social Force (SF) [39]	-	-	-	96.0
Sparse [53]	99.5	97.5	96.4	97.8
Saliency Detector [49]	-	-	-	93.8
Statistical Aggregates [54]	-	-	-	98.5
Hashing Filter [55]	99.2	98.3	98.7	98.7
Commotion [56]	99.9	97.9	98.8	98.9
Discriminative Framework [57]	-	-	-	91.0
Compact Feature [11]	-	-	-	88.3
Deep-cascade [58]	-	-	-	99.6
Optical flow-GAN [20]	-	-	-	99.0
Unmasking [51]	99.3	87.7	98.2	95.1
Proposed method	**99.9**	**99.8**	**99.3**	**99.6**

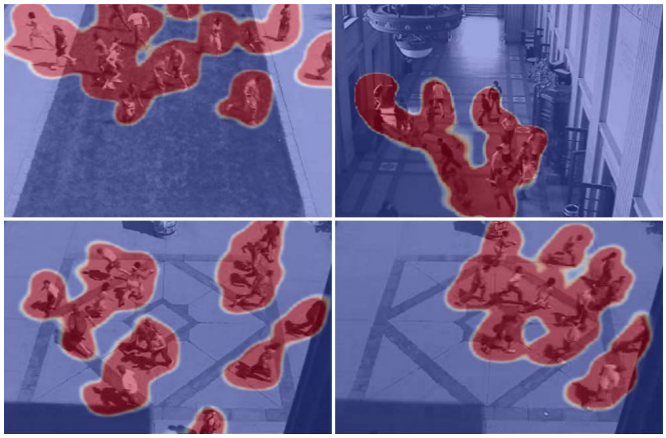

Fig. 8. Visualization results of abnormal events in the UMN.

curve. It can be seen that the abnormality score has low values for normal scenes and relatively high values when abnormal events occur. Based on the abnormality score, we obtain a frame-level ROC curve with the false positive rate and the true positive rate, by changing the threshold within the range of the score. We then obtain the AUC value of the ROC curve and compare it to the values of other state-of-the-art methods. Second, a pixel-level criterion [47] is adopted to evaluate how well the abnormal regions are localized. In the proposed model, the appearance-motion joint difference map D_t^j contains the local information of the detected events. However, activated pixels are scattered sporadically. Thus, we split the difference map into overlapping patches, and add the average of each patch to the image plane to cluster the nearby activated pixels for localization of the detected events. A true positive detection indicates that detected regions cover more than 40% of the abnormal ground truth pixels. Otherwise, it is considered as a false positive detection. The pixel-level ROC curve is obtained by changing the threshold. Then the AUC value of the pixel-level ROC curve is used for the performance comparison. The frame-level and pixel-level AUC values of the other methods are taken from each previous work and [5].

TABLE VI
FRAME-LEVEL PERFORMANCE ON THE AVENUE

Method	Frame-level AUC (%)
Detection at 150fps [40]	80.9
Discriminative Framework [57]	80.9
Conv-AE [14]	70.2
ConvLSTM-AE [16]	77.0
S-RBM [50]	78.8
STAE-grayscale [31]	77.1
STAE-optflow [31]	80.9
Deep Appearance Features [59]	84.6
Unmasking [51]	80.6
Stacked RNN [41]	81.7
STAN [18]	87.2
PredictionNet [19]	85.1
Narrowed Normality Clusters [60]	88.9
Proposed method	**90.0**

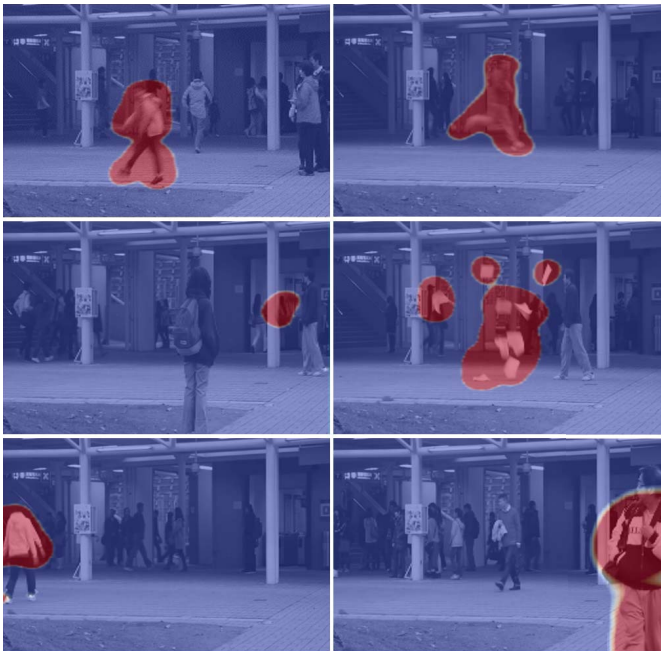

Fig. 9. Visualization results of abnormal events in the Avenue.

TABLE VII
FRAME-LEVEL PERFORMANCE ON THE SHANGHAITECH

Method	Frame-level AUC (%)
ConvLSTM-AE [16]	60.9
Stacked RNN [41]	68.0
PredictionNet [19]	72.8
Proposed method	**76.2**

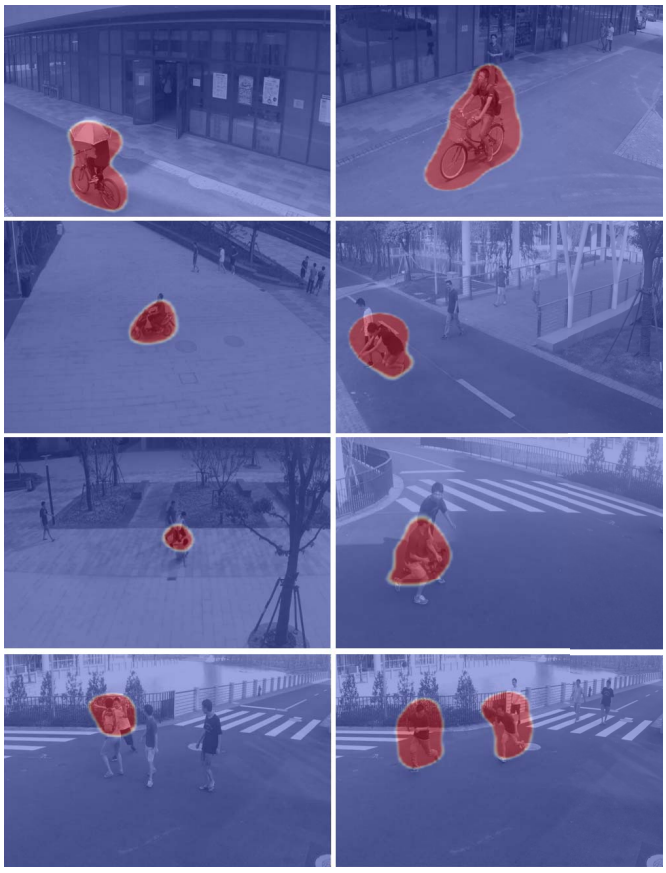

Fig. 10. Visualization results of abnormal events in the ShanghaiTech.

1) Results on UCSD Dataset: Table III shows the frame-level performance compared with non-deep learning-based methods [10], [38], [39], [48], [49] and deep learning-based state-of-the-art methods [5], [14], [16], [18]–[20], [31], [41], [42], [50], [51] on the UCSD Ped2 dataset. As shown in Table III, the proposed method outperforms other existing state-of-the-art methods. Table IV shows the pixel-level performance indicating how well localization is performed. The proposed method outperforms most of the other methods in the pixel-level performance, except for the Generic Knowledge [42]. Note that the Generic Knowledge model utilizes an object detection framework based on the Fast R-CNN [52], whereas our model does not employ any previously designed local instruction. Fig. 7 shows actual visualization results on the UCSD Ped2 dataset. The proposed model properly detects abnormal event regions such as bicycles, a skateboard, and an automobile. When two abnormal events occur simultaneously in a scene, they are also detected correctly as shown in the figure.

2) Results on UMN Dataset: The UMN dataset includes gray and color videos with scene variations. Frame-level performance on the UMN dataset is shown in Table V. As is clear from the table, the proposed model surpasses most of the other methods and is comparable to the existing state-of-the-art method in frame-level performance evaluation. In addition, performances for each individual scene have better results than with the existing methods. Note that the proposed model is trained at once without dividing the data by each scene, which shows that our model can be applied to various environments at the same time. Visualization results on the UMN dataset are shown in Fig. 8. The panic crowds are correctly localized in the lawn, the indoor, and the plaza scenes.

3) Results on Avenue Dataset: The Avenue dataset is more challenging than the UCSD and the UMN datasets because it includes the object scale variations with complex motions

TABLE VIII
EFFECTS OF THE NETWORK DESIGNS ON THE PERFORMANCE AND THE COMPUTATIONAL COST

Network Design			Frame-level AUC (%)				Computational Cost
Bidirectional multi-scale feature encoding	Adversarial learning	appearance-motion joint detection w/ attention	UCSD Ped2	UMN	Avenue	ShanghaiTech	Detection time per frame (sec)
✗	✗	✗	93.5	96.6	86.1	66.0	0.010
✓	✗	✗	96.3	99.4	89.3	73.3	0.028
✓	✓	✗	96.4	99.4	89.5	73.6	0.028
✓	✓	✓	**96.6**	**99.6**	**90.0**	**76.2**	0.038

TABLE IX
COMPARISON OF THE DIFFERENT DETECTION STREAMS

	Frame-level AUC (%)			
	UCSD Ped2	UMN	Avenue	ShanghaiTech
Detection with D_t^a (appearance)	96.5	99.4	89.7	75.7
Detection with D_t^m (motion)	86.1	88.4	77.3	71.2
Detection with D_t^j (appearance-motion joint)	**96.6**	**99.6**	**90.0**	**76.2**

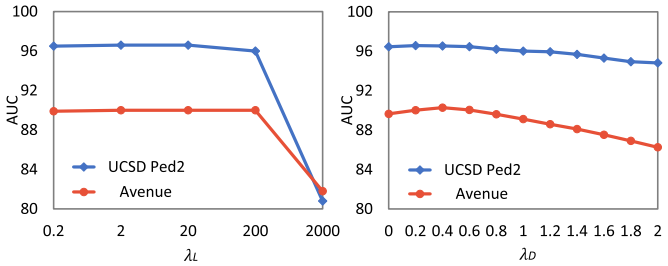

Fig. 11. Effects of the hyper-parameters λ_L and λ_D on the frame-level performance.

and frequent occlusions. Table VI shows the frame-level performance on the Avenue dataset. Comparison methods are classified into non-deep learning-based methods [40], [57] and deep learning-based methods [14], [16], [18], [19], [31], [41], [50], [51]. The proposed method outperforms the existing state-of-the-art methods. Visualization results are shown in Fig. 9. As the figure reveals, the proposed model properly detects abnormal events such as a jumping person, a running person, a thrown bag, fluttering papers, dancing people, and moving in the wrong direction.

4) Results on ShanghaiTech Dataset: The ShanhaiTech is the most challenging dataset containing object scale variations with complex motions, frequent occlusions, and scene variations. As shown in Table VII, the proposed model outperforms other deep learning-based state-of-the-art methods [16], [19], [41] in frame-level performance. Fig. 10 shows the visualization results of detected abnormal events. Various events including bicycles, an automobile, a jumping person, pushing, stealing, fighting, and chasing are correctly localized by the proposed method as confirmed by the figure. Despite various scene changes, each event is properly visualized. Note that the proposed method is a data-driven approach without utilizing any previously designed handcrafted features or pre-trained models.

D. Ablation Study

We analyze the impact of network designs by ablating them as shown in Table VIII. A baseline model consists of a spatial encoder, a forward direction encoder with typical ConvLSTM and a spatial decoder. A bidirectional multi-scale encoding model contains the scale-selective aggregator and the bidirectional multi-scale encoder without attention. An adversarial learning model utilizes spatio-temporal discriminator with the adversarial loss at training time. In addition, we conduct an ablation study by adjusting the presence of scene context-based attention and appearance-motion joint detection. Note that the appearance-motion joint detection is guided by the scene context-based attention. As shown in the table, each factor contributes to the performance of the model. In particular, the bidirectional multi-scale encoding and the appearance-motion joint detection with attention contribute significantly to the performances on the Avenue and the ShanghaiTech datasets. Note that the Avenue and the ShanghaiTech datasets include object scale variations with complex motion information. Results show that the proposed network designs are effective for the challenging datasets that are similar to real-world environments.

In terms of the computational cost, the average detection times are measured for the different network designs as shown in Table VIII. The models are implemented with a single Nvidia TITAN XP GPU and Tensorflow. The baseline model shows a detection time of 0.010 sec per frame. When a bidirectional multi-scale encoding structure is added, the detection time increases to 0.028 sec per frame. Note that the detection time does not increase when adversarial learning structure is added because the discriminator is not used at testing time. Finally, when appearance-motion joint detection with attention is added, the detection time is 0.038 sec per frame. The proposed final model performs the detection at 0.038 sec per frame, which corresponds to 26 fps. Other state-of-the-art methods have detection speeds of 25fps (with TITAN GPUs) in [19], 50fps (with N/A) in [41], and 2fps (with i7-2600 CPU) in [9].

Table IX shows the ablation study for the three detection cases (appearance, motion, appearance-motion joint). Each detection is performed based on the final proposed inter-frame

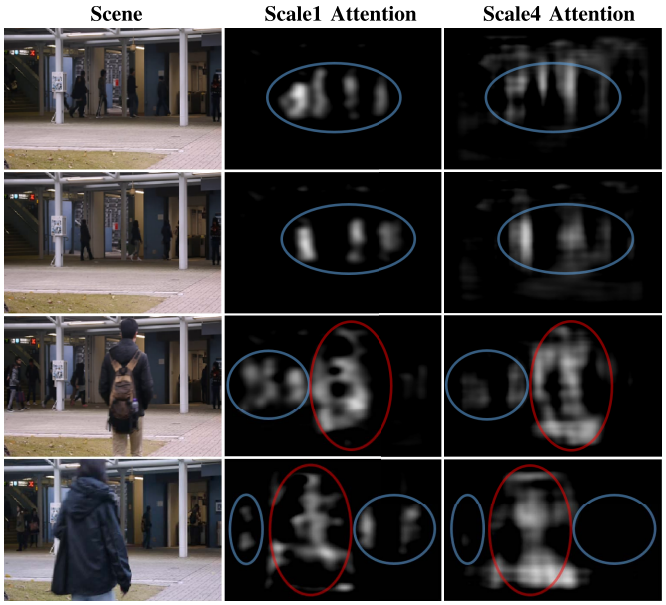

Fig. 12. Visualizations of attention maps from the different scale SFA-ConvLSTMs. Blue circles and red circles indicate small scale object regions and large scale object regions, respectively.

predictor. In the table, for the detection stream: 'appearance', the abnormality score is obtained from the appearance-level difference map D_t^a in (17). For the detection stream 'motion', the abnormality score is predicted from the motion-level difference map D_t^m in (21). 'Appearance-motion joint' indicates that the abnormality score is estimated from D_t^j, the joint form of D_t^a and D_t^m as in (22). As seen in the table, 'appearance-motion joint' achieves the highest performances for the all datasets. Particularly, for the ShanghaiTech which is the challenging dataset, the importance of the motion part can be seen in detection performances. Note that ShanghaiTech includes various types of complex abnormal motions with different camera views. In addition, we conduct a hypothesis testing [61] on the differences between 'appearance' and 'appearance-motion joint' performances. The hypothesis testing shows that the improvements from using the motion part are statistically significant at $P < 0.1$ for the UCSD Ped2 and at $P < 0.05$ for the UMN, the Avenue and the ShanghaiTech.

E. Effects of Hyper-Parameters

We conduct experiments to see the effects of the hyper-parameters on the performance. There are two hyper-parameters (λ_L, λ_D) in the proposed model. λ_L is the training hyper-parameter that balances the pixel-wise loss and the generative adversarial loss in (3). λ_D is the detection hyper-parameter that balances the appearance domain detection and the motion domain detection in (22). These hyper-parameters are determined experimentally. Fig. 11 shows the effects of λ_L and λ_D on the performances. First, the training hyper-parameter λ_L is varied with an exponential scale [62] for the UCSD Ped2 and the Avenue datasets. When λ_L is 20, high performances are obtained on both datasets. Second, the detection hyper-parameter λ_D is changed from 0 to 2 linearly. When λ_D is 0.2, the proposed model achieves good performances on both the UCSD Ped2 and the Avenue.

F. Visualizations for Multi-Scale Encoding

We visualize spatial attention feature maps to show that different scales of motion information are encoded in each scale stage of the SFA-ConvLSTM. The attention maps for scale1 and scale4 SFA-ConvLSTM are shown in Fig. 12. Blue circles and red circles represent small and large object motion regions, respectively. As shown in the figure, regions with motion occurrence are emphasized. The scale1 attention maps show clearer localized results for the small motion regions. On the other hand, the large motion regions are spotted more clearly in the scale4 attention maps than in the scale1 attention maps. Note that we do not use any supervised instruction to extract local attention information; rather, that is extracted in an unsupervised manner. The visualization results show that the earlier stage SFA-ConvLSTM tends to learn small motion information and the later stage tends to focus on large object motions with consideration of the global regions.

V. Conclusion

In this paper, we propose novel bidirectional multi-scale aggregation networks (BMAN) for abnormal event detection. The proposed method is a data-driven approach, which does not need prior designed handcrafted features or pre-trained models. The inter-frame predictor aggregates features in bidirectional multi-scale and attention aspects to effectively represent the spatio-temporal characteristics of normal events. Based on the learned normal patterns, abnormal events are detected by the proposed appearance-motion joint detector. The proposed method outperforms the existing state-of-the-art methods. In particular, it significantly surpasses other methods on the challenging datasets containing object scale variations and complex motions that are found in real-world environments. In addition, by visualizing the detected abnormal event regions, we could interpret how the proposed BMAN determines abnormal events for each scene. Our ablation study and feature visualization results demonstrate the effectiveness of the network configurations.

References

[1] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 397–408, Jun. 2005.
[2] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, Sep. 2006.
[3] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 907–913, Apr. 2009.
[4] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 631–645, Apr. 2014.
[5] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2016.
[6] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2010, pp. 563–576.

[7] T. Wang and H. Snoussi, "Histograms of optical flow orientation for visual abnormal events detection," in *Proc. Adv. Video Signal-Based Surveill. (AVSS)*, Sep. 2012, pp. 13–18.

[8] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2153–2166, Jul. 2015.

[9] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5288–5301, Dec. 2015.

[10] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017.

[11] R. Leyva, V. Sanchez, and C.-T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3463–3478, Jul. 2017.

[12] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2006, pp. 428–441.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[14] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 733–742.

[15] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*. [Online]. Available: https://arxiv.org/abs/1612.00390

[16] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.

[17] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw. (ISNN)*. Berlin, Germany: Springer, 2017, pp. 189–196.

[18] S. Lee, H. G. Kim, and Y. M. Ro, "STAN: Spatio-temporal adversarial networks for abnormal event detection," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2018, pp. 1323–1327.

[19] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 6536–6545.

[20] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.

[21] M. Baydoun *et al.*, "A multi-perspective approach to anomaly detection for self-aware embodied agents," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2018, pp. 6598–6602.

[22] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7251–7259.

[23] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.

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

[25] X. Shi, 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 *Proc. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 802–810.

[26] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[27] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 580–587.

[30] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 1997, pp. 473–479.

[31] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal AutoEncoder for video anomaly detection," in *Proc. ACM Multimedia (ACMMM)*, 2017, pp. 1933–1941.

[32] D. F. Atrevi, D. Vivet, and B. Emile, "Bayesian generative model based on color histogram of oriented phase and histogram of oriented optical flow for rare event detection in crowded scenes," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2018, pp. 3126–3130.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Jul. 2015, pp. 4489–4497.

[34] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2107–2116.

[35] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Adding attentiveness to the neurons in recurrent neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–151.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[37] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[38] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1975–1981.

[39] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 935–942.

[40] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2720–2727.

[41] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

[42] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3639–3647.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[45] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[46] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[47] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[48] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 2921–2928.

[49] Y. Wang, Q. Zhang, and B. Li, "Efficient unsupervised abnormal crowd activity detection based on a spatiotemporal saliency detector," in *Proc. Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[50] H. Vu, D. Phung, T. D. Nguyen, A. Trevors, and S. Venkatesh, "Energy-based models for video anomaly detection," in *Proc. Asia–Pacific Conf. Knowl. Discovery Data Mining (PAKDD)*, 2017, pp. 641–653.

[51] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2895–2903.

[52] R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440–1448.

[53] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3449–3456.

[54] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2112–2119.

[55] Y. Zhang, H. Lu, L. Zhang, X. Ruan, and S. Sakai, "Video anomaly detection based on locality sensitive hashing filters," *Pattern Recognit.*, vol. 59, pp. 302–311, Nov. 2016.

[56] H. Mousavi, M. Nabi, H. Kiani, A. Perina, and V. Murino, "Crowd motion monitoring using tracklet-based commotion measure," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2354–2358.
[57] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2016, pp. 334–349.
[58] 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 Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.
[59] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe, "Deep appearance features for abnormal behavior detection in video," in *Proc. Int. Conf. Image Anal. Process. (ICIAP)*. Berlin, Germany: Springer, 2017, pp. 779–789.
[60] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, "Detecting abnormal events in video using narrowed normality clusters," in *Proc. Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1951–1960.
[61] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, pp. 837–845, Sep. 1988.
[62] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 2694–2703.

Sangmin Lee received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include machine learning, image/video analysis, and human visual perception.

Hak Gu Kim received the B.S. and M.S. degrees from Inha University, Incheon, South Korea, in 2012 and 2014, respectively, and the Ph.D. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2019. He is currently a Post-Doctoral Researcher at École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. His current research interests include deep learning and machine learning in 2D/3D/VR image processing and computer vision, human visual perception, and medical image processing.

Yong Man Ro (S'85–M'92–SM'98) received the B.S. degree from Yonsei University, Seoul, South Korea, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He was a Researcher at Columbia University, a Visiting Researcher at the University of California at Irvine, Irvine, CA, USA, and a Research Fellow at the University of California at Berkeley, Berkeley, CA, USA. He was a Visiting Professor with the Department of Electrical and Computer Engineering, University of Toronto, Canada. He is currently a Professor with the Department of Electrical Engineering, KAIST. He established Image and Video Systems (IVY) Laboratory , KAIST, in 1997. Among the years, he has been conducting research in a wide spectrum of image and video systems research topics. His current research interests include deep learning, machine learning in computer vision and image processing (2D, 3D, VR), medical imaging, visual recognition, and visual quality assessment. He was a recipient of the Young Investigator Finalist Award of ISMRM in 1992, and the Year's Scientist Award (South Korea), in 2003. He served as a TPC in many international conferences, including the program chair and organized special sessions. He served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He serves as an Associate Editor for *Transactions on Data Hiding and Multimedia Security* (Springer-Verlag).