TAM-Net: Temporal Enhanced Appearance-to-Motion Generative Network for Video Anomaly Detection

Xiangli Ji, Bairong Li, Yuesheng Zhu
Communication and Information Security Laboratory
Shenzhen Graduate School of Peking University
Shenzhen, China
{Jxiangli, lbairong, zhuys}@pku.edu.cn

***Abstract*—Video anomaly detection is a challenging task due to the diversity of anomaly. Existing GAN-based approaches model normal motion pattern through transforming a single image to optical flow map, which tends to learn the mapping between two adjacent frames instead of motion evolution in normal scenes. Therefore, this paper proposes a Temporal enhanced Appearance-to-Motion generative Network (TAM-Net) to model evolution of appearance and motion for normal events. In the motion generative branch, the corresponding optical flow map is generated by a ConvLSTM-based generative adversarial network from consecutive frames to learn normal motion pattern. In order to learn appearance pattern, consecutive frames are reconstructed by a auto-encoder in the reconstruction branch. Temporal encoded features of consecutive frames are shared by these two branches to represent changes of normal appearance along with time. By modeling spatio-temporal evolution of normal events, our network can effectively highlight abnormal regions with high generation errors of the predicted optical flow map and reconstructed frame. Experimental results on three independent datasets, UCSD Ped1, Ped2 and Avenue, demonstrate the competitive performance of the proposed method with the other approaches.**

***Index Terms*—Appearance-to-Motion, Generative Adversarial Network, Temporal Encoded Features, Video Anomaly Detection**

I. Introduction

Anomaly detection in videos, which is crucial for video surveillance and scene understanding, has drawn more and more attention recently. However, this task faces two extremely challenging problems. First, abnormal events are rare which results in unbalance between normal samples and abnormal samples. Second, anomaly is unbounded and highly diverse. Therefore, most methods firstly learn representations of regular activities from normal videos under an unsupervised way, then discriminate the outliers as the anomalies.

A category of anomaly detection approaches is based on auto-encoder, which learns to reconstruct input images in normal situations, and uses the reconstruction error as an indicator of an anomaly. 3D Convolutional auto-encoder [1] and Convolutional Long Short Term Memory (ConvLSTM) based auto-encoder [2] are proposed to learn regularity among the appearance and motion patterns. However, these methods are based on the assumption that abnormal events correspond to larger reconstruction errors, which may not always hold. The auto-encoder is likely to reconstruct the abnormal data well resulting in missed discrimination [3], [4].

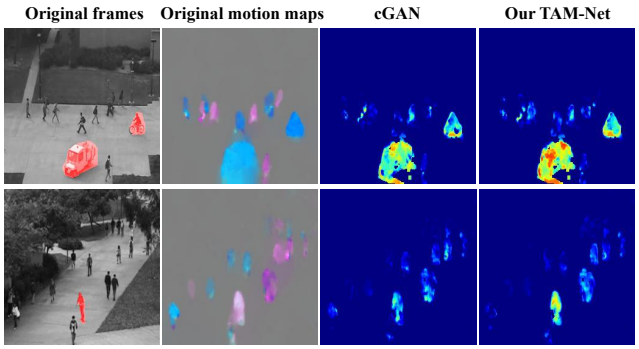


Fig. 1. Generation errors of cGAN and our TAM-Net on abnormal optical flow. Our network can better learn motion representation in normal scenes, and obtain greater generation error when facing with abnormal event.

Instead of auto-encoder, another category of anomaly detection approaches is based on Generative Adversarial Network (GAN) [5]–[7], [23], which learns to generate regular information, usually optical flows and frames, in normal situations. Similarly, regions with high generation error are detected as anomalies here. Ravanbakhsh *et al.* [5], [6] train two conditional GANs (cGANs) [8] to transform raw-pixel frames to corresponding optical flows and vice versa. In this way, the generator in cGAN can learn sufficiently informative representations of normal data. Based on this approach, Vu *et al.* [7] introduce multilevel representations (MLAD), where multiple cGANs are trained to generate level-wise representations of appearance and motion respectively. However, these methods based on GAN are not sufficient to character the motion in videos, which may also generate optical flow maps of abnormal events well. Generating optical flow from a single frame makes the network tend to learn the mapping between current frame and next frame, instead of regular motion of normal activities. As shown in third column of Fig. 1, generation errors of optical flow maps are small on the abnormal regions. Moreover, another cGAN that transforms

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

optical flow to corresponding raw-pixel frame has a limited improvement, which is redundant.

Thus, this paper proposes a temporal enhanced appearance-to-motion generative network (TAM-Net) for video anomaly detection, which effectively utilizes temporal information to excavate regularity in normal scenes and model evolution of appearance and motion. Our TAM-Net consists of a motion generative branch and a reconstruction branch. First, a content encoder is adopted to extract appearance features of each frame. A ConvLSTM further encodes appearance features of the input frame sequence, and temporal encoded features are shared by these two branches. In order to learn motion representation of normal scenes, the optical flow map is generated by a motion decoder under adversarial learning framework in the motion generative branch. At the same time, the frame sequence is reconstructed by a content decoder to learn normal appearance pattern in the reconstruction branch. The whole TAM-Net can be trained end-to-end. To summarize, the main contributions of this work are as follows:

- The regular motion pattern of normal activities is learned by the ConvLSTM-based GAN from consecutive frames in motion generative branch. Moreover, temporal encoded features can be obtained to effectively represent change of normal appearance along with time.
- A content decoder is also introduced to model normal appearance pattern under unsupervised learning framework.

As shown in Fig. 1, our network can highlight abnormal regions with high generation errors of the motion on abnormal frames. Experiments on three independent datasets demonstrate the competitive performance of our approach compared with other methods at frame-level and pixel-level criterion.

II. Related Work

A. Trajectory-based Methods

Early methods based on trajectory features [9]–[11] are proposed to learn normal pattern in a particular scene, and then recognize unusual behaviour patterns based on the learned model. Trajectory-based methods usually contains three main stages. First, object tracking algorithms are used to extract trajectory-based features of foreground objects, such as flow vectors and control points. Then a statistical model is constructed to learn regular patterns in normal scenarios. Finally, activities that deviate from the learned model are discriminated as anomalies. However, performance of these trajectory-based methods significantly degrades in complex scenes with occlusions and dense crowds, because trajectory features rely on the output of object tracking algorithms.

B. Hand-crafted Features Based Methods

Hand-crafted features based detection approaches use usually low-level spatial-temporal features to learn normal patterns, such as histogram of oriented gradients (HOG) [12], histogram of oriented flows (HOF) [13], 3D spatio-temporal gradient and dense spatial-temporal interest points [14]. Kim and Grauman [15] use a Mixture of Probabilistic Principal Component Analyzers (MPPCA) to learn atomic motion patterns and introduce a space-time Markov Ranbdom Field (MRF) model to detect abnormal activities. Kratz and Nishino [16] propose a HMM-based approach that models the variations of local spatio-temporal motion patterns. Mehran *et al*. [17] propose a social force model based on optical flow to model the normal behavior of the crowd. Mahadevan *et al*. [18] joint model appearance and dynamics of the scene by using mixtures of dynamic textures. Besides, sparse coding is also usually used to encode the patterns of normal activities. Cong *et al*. [19] introduce a sparse coding model based on multi-scale histograms of optical flow and use the sparse reconstruction cost (SRC) to measure the normalness of the testing samples. Compared to trajectory-based methods, these methods based on hand-crafted features are more robust for anomaly detection in complex scenes. However, due to the diversity of abnormal activities, these hand-crafted features are difficult to define a priori and are insufficient to represent appearance and motion in videos.

C. Deep Learning Based Methods

Deep learning approaches have recently achieved successes in various computer vision tasks and many deep learning based approaches are proposed to deal with the anomaly detection in videos. Based on auto-encoder, a category of anomaly detection works [1], [2], [20] learns to reconstruct the normal training data and uses the reconstruction error as an indicator of an anomaly. Zhao *et al*. [20] propose a spatio-temporal auto-encoder (STAE) based on 3D convolutions to reconstruct frames and predict future frames. Different from STAE, our approach integrates convLSTM with cGAN to predict the future optical flow from a frame sequence, which model explicitly evolution of appearance and motion. Because of the great generalization of deep auto-encoder, these methods may also reconstruct the abnormal videos well.

Ionescu *et al*. [21] use the object-centric convolutional auto-encoders to encode motion and appearance and train a one-versus-rest abnormal event classifier to discriminate anomalies. Liu *et al*. [22] propose to predict a future frame with high quality for the anomaly detection problem. There are other detection approaches [5]–[7] that apply the conditional Generative Adversarial Networks (cGANs) to generate frames and optical flow respectively for anomaly detection. However, the models that generate frames and optical flows are redundant. Nguyen and Meunier *et al*. [23] combine a convolutional autoencoder and an image translation model to learn a correspondence between appearance and motion. Compared with this method, our approach can better model regularity of appearance and motion by generating an optical flow from the consecutive frames. There are other methods by multiple-instance learning [24] or supervised learning under noisy labels [25] for anomaly detection.

III. Proposed Method

Temporal information is extremely important to identify anomalies for video anomaly detection. Existing GAN-based

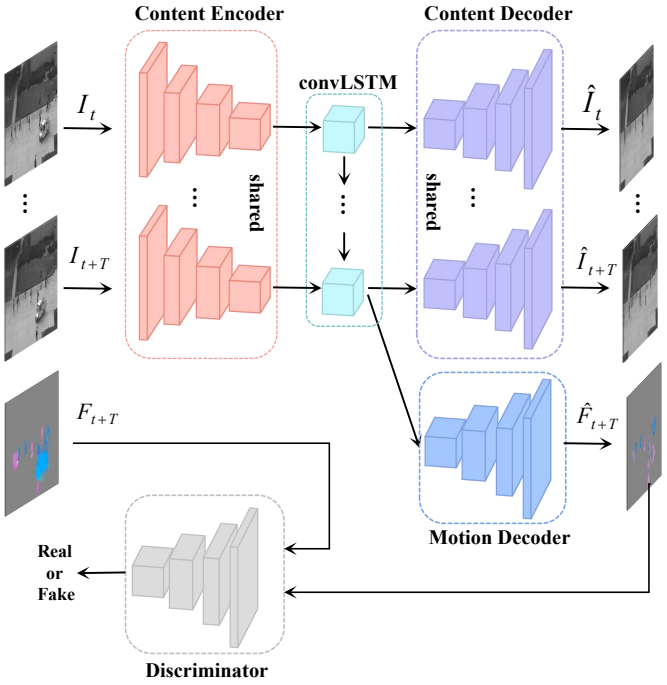


Fig. 2. Overview of our TAM-Net. The whole network consists of a motion generative branch (blue) and a reconstruction branch (purple), which share the same content encoder (red). The normal optical flow map $\hat{F}_{t+T}$ between frame I_{t+T} and I_{t+T+1} is predicted by the motion decoder, and the input frames are reconstructed by the content decoder. The convLSTM is used to better extract temporal encoded features for both decoders. Moreover, adversarial learning is also adopted to generate optical flow here.

approaches [5], [7], [23]usually learn to generate the corresponding optical flow from a raw-pixel frame in normal situations. While learning the representation of motion from a single frame constraints the construction of regular motion pattern in normal videos. In order to better excavate temporal information when modeling motion pattern, we integrate a ConvLSTM with a GAN to generate optical flow map from consecutive frames. In this way, our motion generative branch can effectively utilize temporal encoded features to learn regularity of motion in normal scenes. Moreover, regular appearance pattern can also be modeled with shared temporal encoded features by our reconstruction branch. The proposed TAM-Net is trained on normal data end-to-end. In the testing phase, the fused generation errors of the generated optical flow and the reconstructed frame, are used as the indicators of anomaly. The architecture of temporal enhanced appearance-to-motion generative network is shown in Fig. 2. Each component of our network is described as follows.

A. Temporal Content Encoder

Given a normal video, a snippet with T frames is sampled consecutively, defined as $\{I_1, I_2, ..., I_T\}$. The content encoder extracts appearance features by reducing gradually the spatial resolution of feature maps. We denote the mapping function of the content encoder as $x_t = f_c(I_t)$. This content encoder is constructed by a sequence of blocks, which contains a 2D convolutional layer, a batch-normalization layer and a leaky-ReLU activation. The parameters of first six blocks are same as cGAN's generator [8]. In order to preserve spatial information, convolutional layers with 3×3 kernels and 1×1 strides are used in the last two blocks. For each frame of the input sequence, the parameters of content encoder are shared.

LSTM has shown strong capability to model the sequential data. As a variant of LSTM, ConvLSTM [26] replaces the matrix multiplication with convolutional operation for the calculation of the three gates and the memory cell. Compared with LSTM, ConvLSTM captures spatio-temporal correlations better for the sequential data. Therefore, we introduce a ConvLSTM to memorize changes of appearance and motion information. The mapping function of the ConvLSTM is denoted as $h_T = f_l(f_c(I_1), f_c(I_2), ..., f_c(I_T))$.

B. Frame Reconstruction Branch

Through introducing temporal information, our content decoder reconstructs the input images, which helps to model appearance pattern in normal scenes. Moreover, temporal encoded features can be extracted effectively in this process, which represent changes of appearance patterns along with time. The content decoder contains a sequence of blocks, which contains a deconvolutional layer, a batch-normalization layer, a ReLU activation and a dropout layer. The mapping function of content decoder is defined as $\hat{I}_t = f_{cd}(h_t)$.

In order to make the reconstructed frames close to the real frames, $L2$ loss and gradient difference loss are adopted, which is defined as follows:

$$L_{recon} = \sum_{t=1}^{T} L_{\text{int}}(I_t, \hat{I}_t) + L_{grand}(I_t, \hat{I}_t) \tag{1}$$

where $L2$ loss guarantees the similarity of reconstructed frames and real frames in RGB space, which is given by:

$$L_{\text{int}}(I_t, \hat{I}_t) = \left\| I_t - \hat{I}_t \right\|_2^2 \tag{2}$$

Because the output images are blurred if $L2$ loss is only used, the gradient difference loss is also added, which can sharpen the reconstructed frames [22]. The formula of this loss is shown as follows:

$$L_{grand}(I_t, \hat{I}_t) = \sum_{d \in (x,y)} \left\| |g_d(I_t)| - \left| g_d(\hat{I}_t) \right| \right\|_1 \tag{3}$$

C. Motion Generative Branch

Optical flow map F_T between frame I_T and I_{T+1} contains three channels, which consists of the xy displacements and magnitude. The motion generative branch is used to generate the normal optical flow map $\hat{F}_T$ between I_T and I_{T+1} using the shared temporal encoded features. The motion generative branch consists of two modules: a motion decoder and a discriminator.

The motion decoder is also constructed by a sequence of blocks, which is same as the content decoder. The spatial resolution of hidden state h_T increases gradually and channel number of feature maps reduces at the same time. This motion decoder outputs the predicted optical flow map $\hat{F}_T$. We denote

the mapping function of decoder as $\hat{F}_T = f_{md}(h_T)$. Since the low-level features of the content encoder contain the edge and texture informations, the skip connections between content encoder and motion decoder are employed to enhance the prediction of optical flow's details. In order to simplify the architecture, the skip connections are leaved out in the Fig. 2.

Generative adversarial networks (GANs) have shown strong capability in image translation and video generation task [8], [27]. In order to generate a realistic optical flow map, we also introduce a GAN model, which consists of a generator G and a discriminator D. The G and D are trained by adversarial learning with alternative update manner. In order to avoid the problem of mode collapse, we adopt the conditional GAN [8], which learns a mapping from observed image x and random noise z to y. In our approach, The content encoder, ConvLSTM and motion decoder are treated as G. For D, we use a patch discriminator that penalizes structure at the scale of patches. The parameters of this discriminator are same as cGAN [8]. The training step of the motion generative subnetwork is as follows:

a) Training D: D takes two images as input: the pair $\{I, F\}$ or the pair $\{I, \hat{F}\}$, and tries to classify if each patch in the images is real or fake. During training D, the weights of G are fixed. The loss function of D is defined as:

$$\begin{aligned} L_{adv}^{D}(I, F, \hat{F}) =& \frac{1}{2} \sum_{i,j} -\log D(I, F)_{i,j} \\ &+ \frac{1}{2} \sum_{i,j} -\log[1 - D(I, \hat{F})_{i,j}] \end{aligned} \tag{4}$$

where i, j indicate the indexes of spatial patches.

b) Training G: G is trained to generate the corresponding optical flow map $\hat{F}$ that is as similar as possible to the real F. Therefore, the loss function of G consists of L_{adv}^{G} and L_1 loss, is defined as:

$$L_G(I, F, \hat{F}) = L_{adv}^{G}(I, \hat{F}) + \lambda_f L_1(F, \hat{F}) \tag{5}$$

where λ_f is a loss trade-off parameter and L_{adv}^{G} loss asks the generator G to fool the D, which is define as follows:

$$L_{adv}^{G}(I, \hat{F}) = \sum_{i,j} -\log D(I, \hat{F})_{i,j} \tag{6}$$

$L1$ loss encourages the generated $\hat{F}$ to be near the ground truth F in $L1$ sense and be less blurring, which is given by:

$$L_1 = \left\| F - \hat{F} \right\|_1 \tag{7}$$

D. Objective Function

To summarize, Our model is trained by end-to-end. The whole used losses are combined into a objective function when generator G is trained, which is given by:

$$\begin{aligned} L_g =& L_{adv}^{G}(I_T, \hat{F}_T) \\ &+ \lambda_f L_1(F, \hat{F}) \\ &+ \lambda_r L_{recon}(\{I_1, ..., I_T\}) \end{aligned} \tag{8}$$

where λ_f and λ_r are the loss trade-off parameters.

When discriminator D is trained, the used objective function is defined:

$$L_d = L_{adv}^{D}(I_T, F_T, \hat{F}_T) \tag{9}$$

E. Anomaly Detection on Testing Data

At testing time, only the content decoder and generator G including the content encoder, the ConvLSTM and the motion decoder are used to detect anomalies. Given a testing video with N frames, N snippets with T consecutive frames are sampled and the stride size is 1. We pad the video in head with first frame so that the first $T-1$ snippets have the same length T. $s_i = \{I_{i-T+1}, I_{i-T+2}, ..., I_i\}$ indicates the i^{th} snippet, and F_i denotes the optical flow map between frame I_i and I_{i+1}.

Our generator G takes s_i as input, and generates optical flow map $\hat{F}_i$ between frame I_i and I_{i+1} and reconstructs the last frame $\hat{I}_i$. The generation error between $\hat{F}_i$ and F_i is defined as $\Delta F_i = F_i - \hat{F}_i$. The reconstruction error is defined as $\Delta I_i = I_i - \hat{I}_i$. Then, ΔF_i and ΔI_i are normalized into $[0, 1]$ for each channel. The normalized optical flow error map is defined as follows:

$$\Delta \bar{F}_i^c(x, y) = \Delta F_i^c(x, y) / m_{F,i}^c \tag{10}$$

where x, y denotes the position of pixel in error map and $m_{F,i}^c$ is the maximum value of all position in the generation error maps for c^{th} channel. The normalized reconstruction error map is given by:

$$\Delta \bar{I}_i^c(x, y) = \Delta I_i^c(x, y) / m_{I,i}^c \tag{11}$$

Finally, a abnormality map is obtained by fusing $\Delta \bar{F}_i$ and $\Delta \bar{I}_i$, which is defined as $e_i = \Delta \bar{F}_i + \alpha \Delta \bar{I}_i$, which is used as the indicator of anomaly.

IV. Experiments

In this section, we evaluate the proposed network on three anomaly detection datasets, including CUHK Avenue dataset [28], UCSD Pedestrian Ped1 dataset and Ped2 dataset [18]. The evaluation is performed at frame-level and pixel-level.

A. Datasets

UCSD [18]. The UCSD Pedestrian dataset consists of two subsets, Pedestrian 1 (Ped1) dataset and Pedestrian 2 (Ped2) dataset. The Ped1 dataset includes 34 training videos and 36 testing videos with 40 abnormal events. The resolution of each frame is 158×238 pixels. The Ped2 dataset is composed of 16 training videos and 12 testing videos with 12 abnormal events. The resolution of each frame is 240×360 pixels. The definition of anomaly for Ped1 is the same as Ped2, including bicycles, skate-boards, wheelchairs and vehicles crossing pedestrian areas.

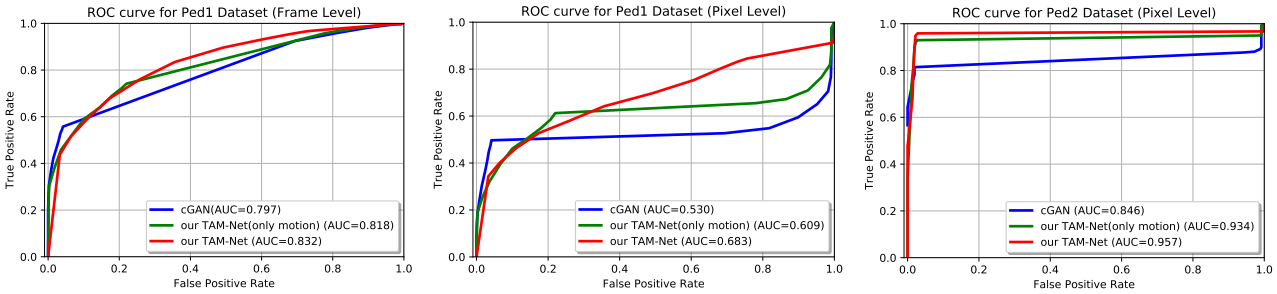
Avenue [28]. The Avenue dataset consists of 16 training videos and 21 testing videos with 47 abnormal events. The resolution of each frame is 360×640 pixels. For each testing frame, a pixel-level mask is provided as ground-truth position of anomaly. In this scene, abnormal activities usually include throwing objects, loitering and running.

TABLE I

ABLATION STUDY OF THE PROPOSED MOTION GENERATIVE SUBNETWORK ON THREE DATASETS

Model	Ped2				Ped1				Avenue			
	frame		pixel		frame		pixel		frame		pixel	
	AUC*	EER*	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
cGAN	94.6	8.8	84.6	8.7	79.7	26.2	53.0	26.8	70.7	33.6	32.7	50.7
motion generative branch	97.1	4.7	93.4	5.1	81.8	24.7	60.9	24.8	76.2	30.6	46.7	38.3
whole TAM-Net	**98.1**	**3.1**	**95.7**	**3.3**	**83.2**	**24.7**	**68.3**	**23.0**	**78.3**	**25.9**	**56.2**	**37.7**

*Higher AUC and lower EER indicate better performance.



(a) frame-level ROC curves on Ped1 dataset (b) pixel-level ROC curves on Ped1 dataset (c) pixel-level ROC curves on Ped2 dataset

Fig. 3. ROC curves for UCSD Ped1 and Ped2

B. Evaluation Metric

Following the literature [29], two criteria of frame-level and pixel-level are used to evaluate an anomaly detection method. A frame that contains anomalies is denoted as a positive, otherwise a negative.

Frame-level criterion. At the frame-level evaluation, an abnormality label is assigned to the testing frame if any pixel is detected as an anomaly. Therefore, one frame is true-positive if its ground truth mask contains abnormality and it is assigned an abnormality label. The pairs of the true-positive rate (TPR) and the false-positive rate (FPR) are computed according to different confidence thresholds, and a Receiver Operating Characteristic (ROC) curve is drawn using these pairs of TPR and FPR. The Area Under Curve (AUC) and Equal Error Rate (EER) are also used to evaluate performance.

Pixel-level criterion. Compared with frame-level criterion, pixel-level criterion is much stricter and more rigorous. A frame is a true-positive if the area of the detected abnormal pixels overlaps with its ground-truth area by at least 40%. A frame is a false-positive if any of its pixels is detected as anomalous and it is negative. For the pixel-level evaluation, AUC and EER are also used to evaluate performance.

C. Implementation Details

Our proposed network is end-to-end trained using training videos of UCSD Ped1, Ped2 and Avenue datasets under unsupervised learning framework. All frames are resized to 256×256 pixels as the inputs of our network. The optical flow maps of videos are calculated using the method in [30]. λ_f is set to 100 and λ_r is set to 0.001. The Adam based Stochastic Gradient Descent algorithm [31] is adopted to train the whole network and the mini-batch size is 1. The learning rate are initially set to 2×10^{-4} for generator G, content decoder and discriminator D and then reduced by a factor of 10 after every 20 epochs. We implement our system using pytorch, and training is executed on a machine with 32G memory, NVIDIA Titan Xp GPU.

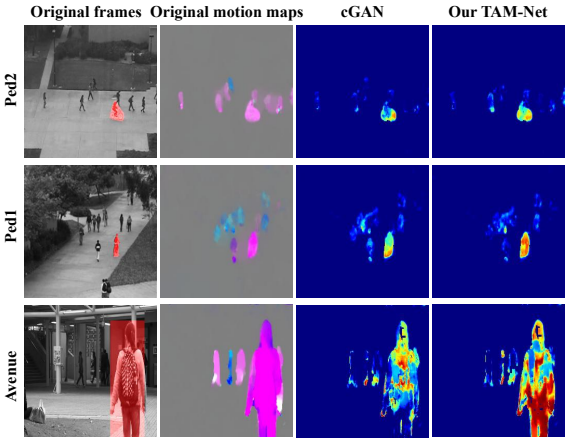


Fig. 4. Generation error maps of cGAN and our motion generative subnetwork on the abnormal frames of Ped1, Ped2 and Avenue.

D. Ablation Study

We analyze the contribution of the two key components in the proposed model: motion generative branch and temporal encoded features in this section.

Impact of the motion generative branch: In order to evaluate the performance of the motion generative branch, we

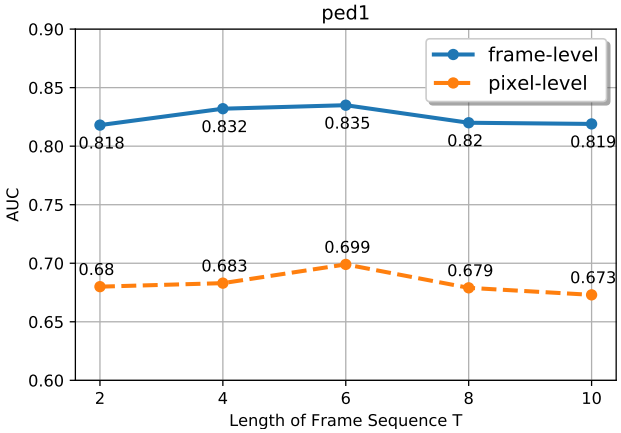
TABLE III
DETECTION RESULTS OF DIFFERENT METHODS ON UCSD PED1 DATASETS

Algothrim	Ped1 (frame-level)		Ped1 (pixel-level)	
	AUC	EER	AUC	EER
MPPCA [15]	67.4	35.6	21.4	76.8
Social force (SFM) [17]	68.8	36.5	37.5	59.1
MDT [18]	81.8	25.0	57.7	40.7
Unmasking [34]	68.4	-	52.4	-
MC2ST [35]	71.8	-	-	-
ConvLSTM-AE [2]	75.5	-	-	-
ConvAE [1]	81.0	27.9	-	-
MLAD(0+3) [7]	82.3	**23.5**	66.6	**22.7**
our TAM-Net	**83.5**	25.0	**69.9**	23.6

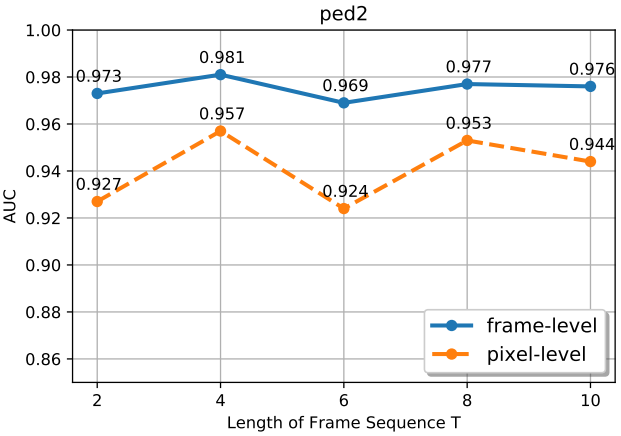
construct a baseline model based on cGAN [8], which learns to transform a raw-pixel frame to the corresponding optical flow. Abnormal regions are detected by the generation errors between generated optical flow maps and real optical flows following Ravanbakhsh [5] and Vu [7]. Then we evaluate the performance of the baseline cGAN generator, the proposed motion generative branch and the whole TAM-Net on UCSD Ped1, Ped2 and Avenue datasets respectively.

The results are shown as the Table I. Compared to the cGAN generator, our motion generative branch improves the AUC about 8.8%, 7.9% and 14% at pixel-level criterion on the Ped2, Ped1 and Avenue dataset respectively. Compared to the motion generative, the whole model improves the AUC about 2.3%, 7.9% and 9.5% at pixel-level criterion on these datasets respectively. Besides, our method obtains lower EER compared with cGAN. Fig. 3 shows the ROC curves of the three networks on UCSD Ped1 and Ped2 datasets. Some qualitative results of the baseline and our model are shown as Fig. 4. Compared with the results of cGAN generator, the generation errors of our model in abnormal optical flow regions are higher. From these results, we can obtain two conclusions: (1) our proposed motion generative branch based on ConvLSTM can effectively learn temporal regularity of motion in normal scenes using strong temporal encoded features; (2) the reconstruction branch can further help to identify anomaly by modeling common appearance patterns of normal events.

Impact of temporal encoded features: In order to evaluate the impact of temporal encoded features, we set length T of input sequence to 2, 4, 6, 8, 10 and train multiple TAM-Nets respectively. The results are shown as Fig. 5, where Fig. 5(a) is the AUC curves at frame-level and pixel-level criterion on UCSD Ped1 dataset and Fig. 5(b) is the AUC curves on UCSD Ped2 dataset. For the Ped1 dataset, the model that generates optical flow map from 6 consecutive frames achieves best performance among these models. For the Ped2, the model that takes 4 frames as the input achieves best performance. The reason for this phenomenon may be that a abnormal activity come to the camera or is away from the camera because of the localization of camera in the Ped1 dataset, which makes network predict motion from more frames.



(a) AUC curves on Ped1 dataset



(b) AUC curves on Ped2 dataset

Fig. 5. AUC scores on UCSD Ped1 and Ped2 obtained by selecting values for the length T of input sequences from the set $\{2, 4, 6, 8, 10\}$

E. Comparison with other Methods

On the UCSD Ped1, Ped2 and Avenue datasets, we compare our TAM-Net with other approaches, and report detection performance on frame-level and pixel-level respectively. In our experiments, we set the length of frame sequence to 4 for Ped2 and Avenue dataset, and 6 for Ped1 dataset.

Ped1. Since some works [5], [6], [9], [20] report results only a subset of 16 videos on the UCSD Ped1 dataset, we compare our method with other methods [1], [2], [7], [15], [17], [18] reporting results on all 36 testing videos. As shown in Table III, our model increases the AUC from 82.3% to 83.5% (about 1.2% improvement) on frame-level evaluation and from 66.6% to 69.9% (about 3.3% improvement) on pixel-level evaluation.

Ped2 and Avenue. We compare our method with different hand-crafted features based methods [15], [18], [32], auto encoder based methods [1], [2], [9], [20], and generative adversarial network based methods [5]–[7]. The results are shown as Table II. At both frame-level and pixel-level evaluation, our model outperforms these methods on Ped2 and Avenue, which demonstrates the superiority of our method.

TABLE II

DETECTION RESULTS OF DIFFERENT METHODS AT FRAME-LEVEL AND PIXEL-LEVEL CRITERIA ON THE UCSD PED2 AND AVENUE DATASETS

Algothrim	UCSD Ped2				Avenue			
	frame-level		pixel-level		frame-level		pixel-level	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER
OC-SVM [32]	61.01	44.43	26.27	26.47	71.66	33.87	33.16	47.55
GMM [32]	75.20	30.95	51.93	18.46	67.27	35.84	43.06	43.13
MDT [18]	76.5	27.9	52.2	43.2	-	-	-	-
MPPCA [15]	71.0	35.8	22.2	77.6	-	-	-	-
Social force [17]	70.2	35.0	21.7	72.4	-	-	-	-
ConvAE [1]	90.0	21.7	-	-	70.2	25.1	-	-
AMDN [36]	90.8	17.0	-	-	-	-	-	-
ConvLSTM-AE [2]	88.1	-	-	-	77.0	-	-	-
STAE-grayscale [20]	91.2	16.7	-	-	77.1	33.8	-	-
FRCN action [33]	92.2	13.9	89.1	15.9	-	-	-	-
GAN/gen [5]	93.5	14.0	-	-	-	-	-	-
GAN/dis [6]	95.5	11.0	-	-	-	-	-	-
MLAD(0+3) [7]	97.5	4.7	94.5	4.6	71.5	36.3	52.8	51.8
our TAM-Net	**98.1**	**3.3**	**95.7**	**3.3**	**78.3**	**25.9**	**56.2**	**37.7**

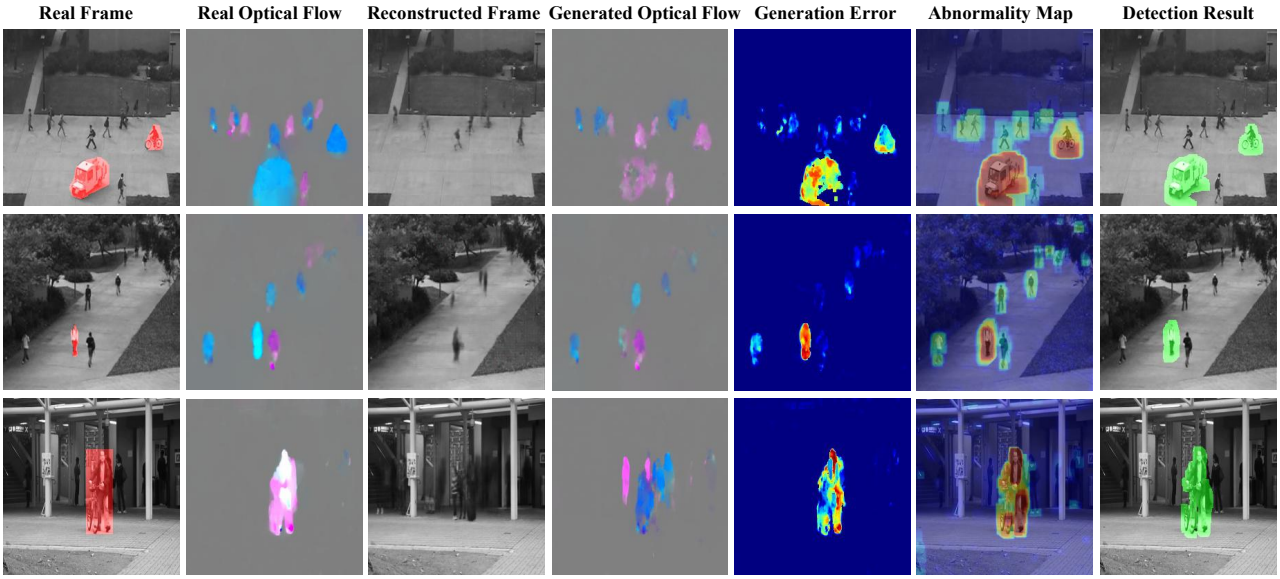


Fig. 6. A few detection examples of our method. The last column shows the detected abnormal regions when threshold is set to 0.8.

Qualitative results. Fig. 6 shows the some detection results of our proposed method on Ped1, Ped2 and Avenue datasets. The generation error map of motion is shown as the fifth column, and the sixth column is the fused abnormality map. The last column shows the detected abnormal pixels when the threshold is set to 0.8. We can observe that the TAM-Net can obtain greater generation error when facing with abnormal event to detect anomaly in videos.

V. CONCLUSION

In this paper, we propose a temporal enhanced appearance-to-motion generative network for video anomaly detection, which consists of a motion generative branch and reconstruction branch. In the motion generative branch, a ConvLSTM-based GAN learns regular motion pattern via generating optical flow map from the shared temporal encoded features. The abnormal pixels can be highlighted in the generation error maps of optical flow for abnormal events. Besides, a content decoder is integrated with motion generative branch to model appearance pattern, which can further improve the AUC values of anomaly detection. By modeling evolution of appearance and motion, our TAM-Net can effectively learn temporal regularity of normal events. Experiments on 3 datasets demonstrate the superiority of our TAM-Net over other methods.

ACKNOWLEDGMENT

This work was supported in part by the Shenzhen Municipal Development and Reform Commission (Disciplinary Develop-

ment Program for Data Science and Intelligent Computing), in part by R&D Program in Key Areas of Guangdong Province (2019B010137001), and in part by NSFC-Shenzhen Robot Jointed Founding (U1613215).

REFERENCES

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

[2] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in IEEE International Conference on Multimedia and Expo, July 2017, pp. 439–444.

[3] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in IEEE International Conference on Computer Vision, October 2019.

[4] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in International Conference on Learning Representations, 2018.

[5] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni,and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in IEEE International Conference on Image Processing, Sep. 2017, pp. 1577–1581.

[6] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," in IEEE Winter Conference on Applications of Computer Vision, Jan 2019, pp. 1896–1904.

[7] H. Vu, T. D. Nguyen, T. Le, W. Luo, and D. Phung, "Robust anomaly detection in videos using multilevel representations," in AAAI Conference on Artificial Intelligence, July 2019, pp. 5216–5223.

[8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in IEEE Conference on Computer Vision and Pattern Recognition, July 2017, pp. 5967–5976.

[9] F. Tung, J. S. Zelek, and D. A. Clausi, "Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance," Image and Vision Computing, vol. 29, no. 4, pp. 230–240, 2011.

[10] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in IEEE Conference on Computer Vision and Pattern Recognition, June 2008, pp.1–8.

[11] T. Zhang, H. Lu, and S. Z. Li, "Learning semantic scene models by object classification and trajectory clustering," in IEEE Conference on Computer Vision and Pattern Recognition, June 2009, pp. 1940–1947.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2005, pp. 886–893.

[13] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in European Conference on Computer Vision, May 2006, pp. 428–441.

[14] K. Cheng, Y. Chen, and W. Fang, "Video anomaly detection and localization using hierarchical feature representation and gaussian process regression," in IEEE Conference on Computer Vision and Pattern Recognition, June 2015, pp. 2909–2917.

[15] 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, June 2009, pp. 2921–2928.

[16] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in IEEE Conference on Computer Vision and Pattern Recognition, June 2009, pp.1446–1453.

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

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

[19] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in IEEE Conference on Computer Vision and Pattern Recognition, June 2011, pp. 3449–3456.

[20] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in ACM International Conference on Multimedia, Oct 2017, pp. 1933–1941.

[21] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in IEEE Conference on Computer Vision and Pattern Recognition, June 2019.

[22] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - a new baseline," in IEEE Conference on Computer Vision and Pattern Recognition, June 2018, pp. 6536–6545.

[23] T.-N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in IEEE International Conference on Computer Vision, October 2019.

[24] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in IEEE Conference on Computer Vision and Pattern Recognition, June 2018, pp. 6479–6488.

[25] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in IEEE Conference on Computer Vision and Pattern Recognition, June 2019.

[26] 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 International Conference on Neural Information Processing Systems, Dec 2015, pp. 802–810.

[27] S. Tulyakov, M. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in IEEE Conference on Computer Vision and Pattern Recognition, June 2018, pp. 1526–1535.

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

[29] 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, Jan 2014.

[30] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in European Conference on Computer Vision, 2004, pp. 25–36.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in CoRR,abs/1412.6980, 2004.

[32] H. Vu, T. D. Nguyen, A. Travers, S. Venkatesh, and D. Phung, "Energy-based localized anomaly detection in video surveillance," in Advances in Knowledge Discovery and Data Mining, 2017, pp. 641–653.

[33] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in IEEE International Conference on Computer Vision, Oct 2017, pp. 3639–3647.

[34] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in IEEE International Conference on Computer Vision, Oct 2017, pp. 2914–2922.

[35] Y. Liu, C. Li, and B. Pczos, "Classifier two sample test for video anomaly detections," in British Machine Vision Conference, Sep 2018.

[36] 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. 117–127,2017.