IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 21, 2019, accepted November 13, 2019, date of publication November 20, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954540

Spatio-Temporal Unity Networking for Video Anomaly Detection

YUANYUAN LI, YIHENG CAI, JIAQI LIU, SHINAN LANG, AND XINFENG ZHANG
College of Information and Communications Engineering, Beijing University of Technology, Beijing 100124, China

Corresponding author: Yiheng Cai (caiyiheng@bjut.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1703302 and in part by the Science and Technology Project of Beijing Municipal Education Commission under Grant KM201710005028 and Grant KM201910005027.

ABSTRACT Anomaly detection in video surveillance is challenging due to the variety of anomaly types and definitions, which limit the use of supervised techniques. As such, auto-encoder structures, a type of classical unsupervised method, have recently been utilized in this field. These structures consist of an encoder followed by a decoder and are typically adopted to restructure a current input frame or predict a future frame. However, regardless of whether a 2D or 3D autoencoder structure is adopted, only single-scale information from the previous layer is typically used in the decoding process. This can result in a loss of detail that could potentially be used to predict or reconstruct video frames. As such, this study proposes a novel spatio–temporal U-Net for frame prediction using normal events and abnormality detection using prediction error. This framework combines the benefits of U-Nets in representing spatial information with the capabilities of ConvLSTM for modeling temporal motion data. In addition, we propose a new regular score function, consisting of a prediction error for not only the current frame but also future frames, to further improve the accuracy of anomaly detection. Extensive experiments on common anomaly datasets, including UCSD (98 video clips in total) and CUHK Avenue (30 video clips in total), validated the performance of the proposed technique and we achieved 96.5% AUC for the Ped2 dataset, which is much better than existing autoencoder-based and U-Net-based methods.

INDEX TERMS Anomaly detection, U-Net, ConvLSTM, video.

I. INTRODUCTION

High-precision anomaly detection remains a challenging but important task in automated video surveillance. In contrast to other computer vision problems, abnormal events occur far less often than regular events, producing a serious imbalance between positive and negative samples. In addition, abnormal events are often unbounded and occur with sparse yet unpredictable regularity. As a result, training sets often include only normal events, which complicates the use of conventional machine learning algorithms. The majority of conventional approaches use a training model to represent regular events in video sequences, considering any outliers to be abnormal events [1]. Some studies have focused on the extraction of hand-crafted features, used to precisely represent appearance and motion in video frames. These algorithms include HOG/HOF [2], the 3D spatio-temporal gradient [3], and optical flow features [4]. However, these hand-crafted features only provide a limited representation of complex motion, restricting the use of conventional machine learning.

Deep learning has successfully been used in a variety of computer vision tasks, including video anomaly detection. Among those video anomaly detection methods based on deep learning, autoencoder structures based on CNNs have been widely used to analyze video data. These structures consist of an encoder followed by a decoder, trained only with positive samples, which are used to restructure the current frame or predict a future frame. When abnormal events occur, the restructure error or prediction error in the autoencoder is higher than that of normal events. As a result, these unsupervised methods are highly suitable for video anomaly detection problems in which only positive samples are available in the training set. However, regardless of whether a 2D [5] or 3D [6] autoencoder structure is adopted, only single scale information from the previous layer is used in the decoding

The associate editor coordinating the review of this manuscript and approving it for publication was Huazhu Fu.

process, leading to a loss of detailed information that could be used to predict or reconstruct video frames.

To resolve this issue, Liu *et al.* [7] proposed using a U-Net with multi-scale information, instead of an autoencoder, to predict future frames. Inspired by this approach, we propose a novel spatio-temporal U-Net for video anomaly detection. This technique combines U-Net [7] with ConvLSTM which has advantages in modeling temporal information in time series data [9]. In addition, RGB differences from motion loss were used to replace the optical flow motion loss used by Liu *et al.* This resulted in a similar accuracy but significantly reduced the runtime required for motion loss calculation. Finally, unlike existing methods, the proposed technique calculates regular scores from not only the current video frame, but also the future frames, which produced higher detection accuracy. Our contributions are as follows:

1. A novel spatio-temporal U-Net for frame prediction is proposed. The new framework combines the benefits of U-Nets in representing spatial information [8] with the capabilities of ConvLSTM for modeling temporal motion data, which makes it more suitable for modeling image sequences.
2. The RGB differences are first used as motion loss to replace the optical flow motion loss used by Liu *et al.* [7] which significantly reduced the runtime required for motion loss calculation.
3. Unlike existing methods, a new regular score function including not only the current video frame but also the future frames is proposed which produced a higher detection accuracy.

The remainder of this paper is organized as follows. Section II briefly discusses related work in anomaly detection. Section III provides a detailed description of our proposed framework. Section IV includes an experimental validation of our method applied to two major public datasets and Section V concludes the paper.

II. RELATED WORK

Video anomaly detection methods can be categorized as manual feature-based or deep learning network-based models. Most feature models include three steps: 1) extracting appearance and motion features, 2) training a model to represent regular events in video sequences, and 3) identifying outliers (anomalous events) in the trained model. Leyva *et al.* [2] extracted foreground occupancy and optical flow features, including optical flow energy and an HOF descriptor. They also established GMM, dictionary, and Markov models for anomaly detection. SanMiguel *et al.* [10] developed novel feature descriptors, including target size, shape, and speed. Anomalies were then detected using a two-state Markov chain. Wang *et al.* [11] proposed ULGP-OF features consisting of local gradient and optical flow information. An extreme learning machine (ELM) was then used to detect anomalies. However, because the manual features are designed artificially according to specific abnormal objects or videos scenarios—which is a limited representation—traditional machine learning algorithms are not suitable for anomaly detection in complex video surveillance scenes.

Deep learning methods exhibit obvious advantages in feature extraction and have proven to be highly effective for a variety of video analysis tasks. In addition, unsupervised deep learning models based on autoencoder structures have successfully been used for anomaly detection. Hasan *et al.* [12] extracted manual features and constructed a fully convolutional autoencoder to learn both local features and classifiers. However, because only convolution is used for feature extraction, this structure lacked ability to model temporal information in a video sequence. As such, Luo *et al.* [13] and Medel *et al.* [14] added convolutional long-term and short-term memory (ConvLSTM) layers to the autoencoder. This approach not only considered spatial features but also modeled temporal features from the input data, making it more suitable for video analysis. Zhao *et al.* [6] constructed a spatio-temporal convolution autoencoder that extracted features from both spatial and temporal dimensions by performing 3D convolutions. However, both 2D and 3D autoencoders use only single scale information from the previous layer in the decoding process, which can lead to a loss of detail. Thus, Liu *et al.* [7] proposed the use of a U-Net to predict future frames. This skip connection between high- and low-level layers with the same resolution allows multi-scale feature information which contains both high-level semantic information and low-level surface information to be used in the decoding process. However, single U-Nets also lack the modeling of temporal features. Based on the context provided by previous research [7], [13], [14], we propose adding the ConvLSTM layers to the U-Net to construct spatio-temporal U-Net for improved video anomaly detection.

III. APPORCH

As shown in Fig. 1, our framework can be divided into two parts: future frame prediction and abnormal event detection. The first step involves training a generator model to predict the next frame. This was done using a GAN module that included both discriminative and generator networks to increase the accuracy of predicted frames [7]. The original generator network was then replaced with our proposed spatio-temporal U-Net. This approach not only uses multi-scale spatial information in the decoding process, but also models temporal data between frames. The comprehensive application of both spatial and temporal information makes this module more suitable for video processing. In addition, a new loss function called RGB differences was introduced to further constrain the training process. This function represents motion loss in a fashion similar to optical flow but significantly reduces the required calculation time, compared to optical flow extraction. In the second step, PSNR was used to assess the quality of the prediction (I^*t+1), by comparing it with the real future frame ($It+1$) [7], [15]. We propose a new function based on PSNR to calculate the regular score of all test frames. In contrast to existing functions,

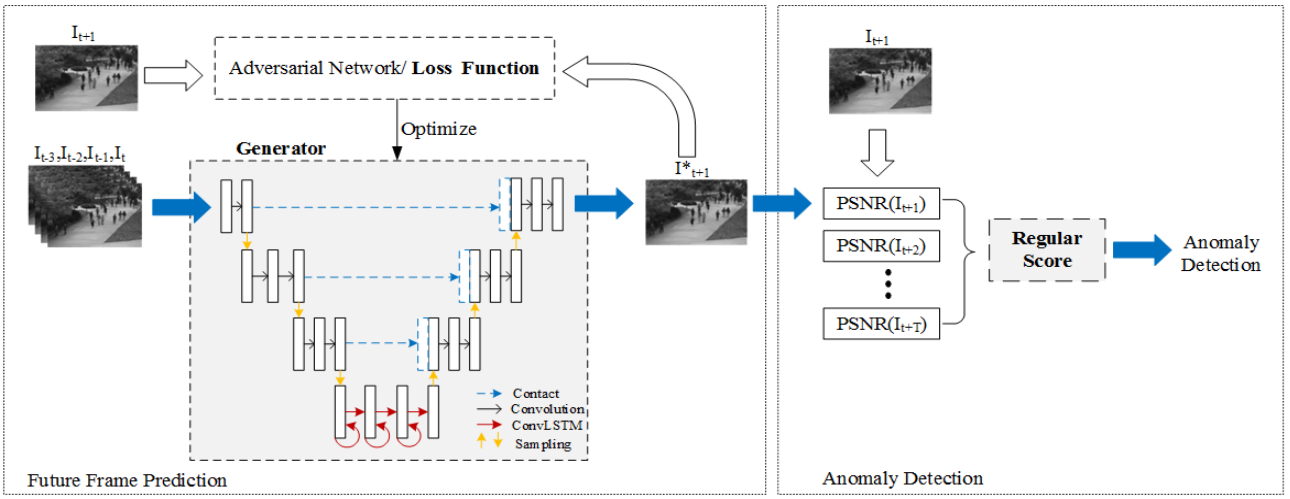


FIGURE 1. The workflow for our proposed framework.

this proposed technique considers the correlation between frames. This includes not only current (PSNR (I_{t+1})) but also future frame prediction errors (PSNR (I_{t+2}), PSNR (I_{t+3}), . . ., PSNR (I_{t+T})).

Fig. 1 shows the workflow for our proposed technique. In the training process, the framework input consists of continuous training frames and the output is comprised of predicted future frames generated by the proposed spatio-temporal U-Net. A discriminative network and loss function were used to optimize the spatio-temporal U-Net and generate realistic future frames. During the testing process, for the input continuous testing frames, our trained spatio-temporal U-Net predicts its subsequent future frame. The prediction error and regular score were determined by the Euclidean distance between predicted and actual future frames. Higher regular scores for test frames corresponded to lower anomaly probabilities. The following subsection discusses this proposed framework in detail.

A. FEARURE FRAME PREDICTION

1) SPATIO-TEMPORAL U-NET

In contrast to conventional autoencoder structures, the U-Net proposed by Ronneberger *et al.* [16] included a skip connection between high- and low-level layers with the same resolution, achieving a combination of high-level semantic features and low-level surface information. Although single U-Nets produce accurate spatial modeling results, they lack an ability to model temporal features. Inspired by the work of Liu *et al.* [7] and the success of ConvLSTM, we propose a novel spatio-temporal U-Net for modeling changes in sequential data [17], [18]. Since videos contain both spatial and temporal motion information, a ConvLSTM layer was added to a conventional U-Net. This approach combines the spatial advantages of a U-Net and the temporal advantages of ConvLSTM.

This framework, shown in Fig. 2, consists of three critical components. First, the encoder includes a series of convolution layers and maximum pooling layers, used to

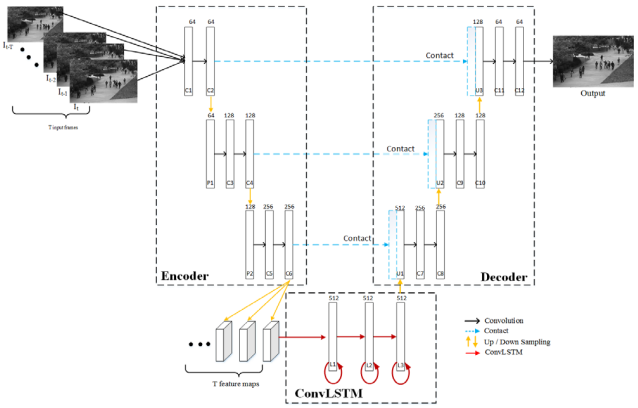


FIGURE 2. The proposed spatio-temporal U-Net. The encoder and decoder included stacking convolution layers with 64, 128, and 256 feature maps. The kernel size of all convolution operations was set to 3 × 3. The size of the max pooling layers and up-pooling layers was set to 2 × 2. ConvLSTM layers were stacked with 512 feature maps (3 × 3 kernel size).

extract important spatial information from input frames. Three ConvLSTM layers were then added at the end of the encoding process to memorize semantic feature changes after pre-encoding. Inputting pre-encoded features improved ConvLSTM efficiency by reducing interference and redundancy. The decoder also included a series of un-pooling layers and convolution layers used to decode feature maps and predict future frames. Unlike in autoencoders, a contact operation was applied between the encoder and decoder to help repair target details.

More specifically, the input mode for our network differs from existing algorithms that typically stack T consecutive frames and input them to an autoencoder or U-Net [7], [12]. Instead, we individually input T frames to a convolution encoder. This approach prevents the collapse of temporal information after the first convolution layer, since 2D operations only consider spatial data. Conventionally, the T input frames are connected to each channel in the first output feature map, which rarely preserves temporal information [6]. To solve this problem, we input T frames to

the convolution encoder and produced T feature maps containing only high-level spatial information. The ConvLSTM layers then allowed these T feature maps to be modeled in temporal space, preserving both spatial and temporal information in consecutive input frames.

As shown in Fig. 2, T consecutive frames: $I_{t-T}, \ldots, I_{t-2}, I_{t-1}, I_t$, were input to a convolution encoder to produce T feature maps, the changes in which were memorized by ConvLSTM layers. The output, including both high-level spatial features and primary temporal features, was input to the decoder to predict future frames. As the prediction of spatial information in the next frame is primarily based on the last frame (I_t), we contacted adjacent feature maps for the T frames acquired from the encoder process. These feature maps had the same resolution as those obtained from the decoder step. The application of multi-scale information in the decoding process allowed the proposed network to predict realistic future frames. Additionally, we set T to 4 according to previous reports [7], [12], which verified the validity of this value experimentally.

2) LOSS FUNCTION

Loss functions are used to estimate differences between predicted and real values, with a smaller function indicating higher robustness. Appropriate loss can constrain a model and improve detection results. As in previous studies [7], [15], intensity, gradient, and motion loss were included in the proposed framework to ensure appearance and motion accuracy for predicted frames. These quantities were defined as follows:

$$L_{int}(I^*, I) = \left\| I^* - I \right\|_2 \tag{1}$$

$$L_{gd}(I^*, I) = \sum_{i,j} \left\| \left| I^*_{i,j} - I^*_{i-1,j} \right| - \left| I_{i,j} - I_{i-1,j} \right| \right\|_1 + \left\| \left| I^*_{i,j} - I^*_{i,j-1} \right| - \left| I_{i,j} - I_{i,j-1} \right| \right\|_1 \tag{2}$$

$$L_{op}(I^*, I) = \left\| f(I^*_{t+1}, I_t) - f(I_{t+1}, I_t) \right\|_1 \tag{3}$$

where I^* represents a generated frame and I represents the corresponding ground truth. The term f represents optical flow estimation, $f(I^*_{t+1}, I_t)$ represents optical flow between the generated frame I^*_{t+1} and the previous real frame I_t, $f(I_{t+1}, I_t)$ represents the optical flow between the ground truth frame I_{t+1} and its previous real frame I_t, $\|\|_1$ and $\|\|_2$ represent the $\ell 1$- and $\ell 2$-norm, respectively. L_{int} is the intensity loss that guarantees the similarity of all pixels in RGB space, L_{gd} is the gradient loss that can sharpen the generated frame, and L_{op} is the motion loss that guarantees the correctness of motion prediction.

Inspired by Wang *et al.* [23], who explored the use of different input patterns to improve discrimination for deep action recognition and proposed RGB differences as a type of input; we replaced optical flow with RGB gradients as a new type of motion loss. Since optical flow is primarily derived from the partial derivation of pixel intensity with respect to time, the ability of optical flow to represent motion could be learned from variations in RGB values. This approach significantly reduced the runtime required for optical flow extraction. Motion losses calculated from RGB differences were defined as follows:

$$L_{rgb}(I^*, I) = \left\| Diff(I^*_{t+1}, I_t) - Diff(I_{t+1}, I_t) \right\|_1$$
$$Diff(Y^*, Y) = Y^* - Y \tag{4}$$

where $Diff(I^*_{t+1}, I_t)$ represents RGB differences between the generated frame I^*_{t+1} and its previous real frame I_t. The term $Diff(I_{t+1}, I_t)$ represents RGB differences between the ground truth frame I_{t+1} and its previous real frame I_t, which calculated the difference of each pixel in the corresponding channel. L_{rgb} is the new motion loss which replaced L_{op} to guarantee the correctness of motion prediction.

3) ADVERSARIAL NETWORKS

In addition to the loss functions discussed above, a variation of the conventional generative adversarial network (GAN) was used to constrain the training process, produce realistic frames, and improve model performance [7], [18], [19]. We utilized a patch discriminator, in which each output scalar corresponded to an input image patch. The resulting discriminator output scalar was classified as either class 0 or class 1, where 0 indicated the input image to be a generated frame and 1 indicated a real frame. The goal of the proposed spatio-temporal U-Net was to generate a realistic prediction frame that was not categorized as class 0 by the discriminator. A mean square error loss function was imposed in the course of adversarial training as follows:

$$L_{adv}(I^*) = \sum_{i,j} \frac{1}{2} L_{MSE}(D(I^*)_{i,j}, 1)$$
$$L_{MSE}(Y^*, Y) = (Y^* - Y)^2 \tag{5}$$

where D denotes a discriminative network that attempts to identify frames generated by our predictive network as class 1 (genuine labels). $D(I^*)_{i,j}$ is the output of the discriminative network for frame I^* and i, j denote spatial patch indexes. L_{adv} is the adversarial loss that constrains the output frame of the predictive network to be as similar to the genuine frame as possible. Our proposed object function can therefore be defined as follows:

$$L = w_{int} L_{int} + w_{gd} L_{gd} + w_{rgb} L_{rgb} + w_{\text{adv}} L_{adv} \tag{6}$$

where w is the weight of each sub-loss. We set w_{int}, w_{gd}, w_{rgb}, and w_{adv} to 1, 1, 2, and 0.05, respectively, according to previous reports [7], which verified the validity of these values experimentally.

B. REGULAR SCORES

After training the model to represent regular events in video sequences, anomalies could be detected using the difference between a predicted frame and its ground truth. We propose a novel regular score function based on peak

signal-to-noise-ratio (PSNR) for anomaly detection [20]:

$$PSNR(I^*, I) = 10log_{10}\frac{[max_{I^*}]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i - I_i^*)^2} \tag{7}$$

Here, max_{I^*} represents the maximum value of image intensities, N represents the total number of pixels, and i represents pixel index. PSNR can be normalized as:

$$psnr(I_t^*) = \frac{PSNR(I_t, I_t^*) - min_{PSNR}}{max_{PSNR} - min_{PSNR}} \tag{8}$$

where I_t^* represents a prediction of the t^{th} frame and I_t is the corresponding ground truth. The terms min_{PSNR} and max_{PSNR} are the minimum and maximum values of the PSNR in every frame of each test video.

In previous studies, the regular score is typically calculated from the prediction error of the current frame: $scores(I_t^*) = psnr(I_t^*)$ [6], [7], [12]–[14]. Larger regular scores indicate lower abnormality probability. However, in the early stages of anomaly occurrence, abnormal targets only appear in the corners of a video scene. The regular score of such a frame is still high compared with that of frames in the middle of an abnormal occurrence because the target occupies very few pixels. As a result, the predicted error for early-sequence abnormal frames, calculated using the difference between a predicted frame and its ground truth, is comparable to a frame calculated under normal conditions. Therefore, if only the prediction error from the current frame is considered in the regular score, the performance of the proposed method would be limited to the boundary where the abnormal event occurs. To solve this problem, we propose a novel regular score function based on PSNR in both the current frame and future frames:

$$scores(I_t^*) = \frac{1}{T^2}\sum_{i=0}^{T}(T-i)psnr(I_{t+i}^*) \tag{9}$$

Here, the t^{th} frame regular score $scores(I_t^*)$ consists of the current normalized PSNR $psnr(I_t^*)$ and the normalized PSNR $(psnr(I_{t+i}^*), i = 0, 1, \ldots, T)$ in T future frames. As such, if the current frame is in the initial stages of an abnormal occurrence with high PSNR, the regular score will be lowered by the PSNR of several future frames in the middle of abnormal events. The value of T here is same as the value of inputted consecutive T frames of prediction network. The value of T is 4 in our work.

IV. EXPERIMENTAL VALIDATION

This section evaluates the proposed method using publicly available anomaly detection datasets, including the CUHK Avenue [21] and UCSD Pedestrian dataset [22].

A. DATA

The UCSD pedestrian set includes two different scenes: Ped1 and Ped2, in which non-pedestrian targets are considered abnormalities. Ped1 includes 34 short training clips and 36 testing clips. Ped 2 includes 16 training and 12 testing clips. The CUHK avenue dataset was collected on Campus Avenue at the Chinese University of Hong Kong. It contains 14 unusual events such as running, throwing objects, and loitering. The set includes 16 training videos and 21 testing videos.

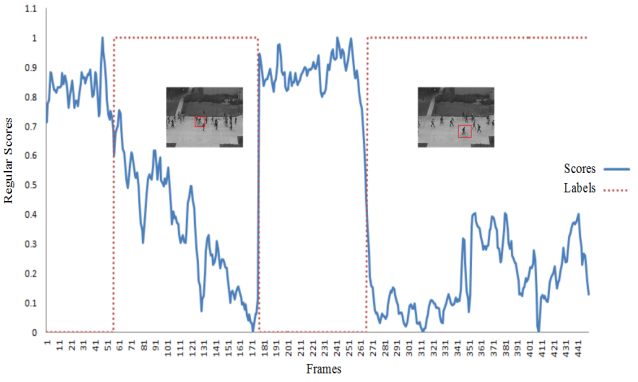


FIGURE 3. The regular score for a video sequence from the Ped2 dataset, which decreased when anomalous events occurred. The red dotted line represents the ground truth abnormal frames.

B. EVALUATION METRIC

As the model was trained using only regular data, video sequences consisting of regular events exhibited a higher score than anomalous sequences. As shown in Fig. 3, the regular score of Ped2 video clips decreased when anomalies occurred (i.e., bicycle intrusion). The red dotted line denotes a ground truth label, 0 denotes a normal frame, and 1 indicates an abnormal frame. Our proposed method was evaluated quantitatively using a series of common assessment metrics. ROC curves were produced by varying the threshold and calculating both the true positive rate (TPR) and false positive rate (FPR):

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

where TP represents true positive samples; TN represents true negative samples; FP represents false positive samples; FN represents false negative samples.

Conventional anomaly detection algorithms were compared to our method using area under ROC curve (AUC) and equal error rate (EER), the error occurring when the false positive rate is equal to the miss rate (i.e., FPR when FPR $=$ 1 $-$ TPR). In this study, higher AUC values and lower EER values indicated better anomaly detection performance.

C. ANOMALY EVENT DETECTION

Several experiments were conducted using the two common datasets to evaluate our proposed prediction network, RGB difference motion loss, and the regular score function. We also compared our proposed technique to multiple conventional algorithms [1], [5]–[7], [12], [13].

TABLE 1. AUC values for different prediction networks applied to the Ped1 and Ped2 datasets.

	Ped 1		Ped 2	
	AUC	EER	AUC	EER
U-Net	82.4%	23.56%	94.9%	11.97%
Spatio-Temporal U-Net	83.46%	22.36%	96.06%	8.89%

TABLE 2. AUC and EER values for different motion loss algorithms.

		Optical flow motion loss	RGB difference motion loss
AUC	Ped1	**83.46%**	83.30%
	Ped2	96.06%	**96.34%**
	Avenue	84.25%	**84.32%**
EER	Ped1	**22.36%**	23.01%
	Ped2	8.89%	**8.70%**
	Avenue	21.79%	**21.54%**

TABLE 3. Required runtimes for different motion loss algorithms.

	Optical flow motion loss	RGB difference motion loss
Ped1	0.0041 (s/batch)	0.5328 (s/batch)
Ped2	0.0041 (s/batch)	0.5296 (s/batch)
Avenue	0.0041 (s/batch)	0.5377 (s/batch)

1) EFFECT OF PREDICTION FRAMEWORK

The performance of the proposed spatio-temporal U-Net was evaluated through a comparison with a conventional U-Net, keeping all other conditions consistent. As shown in Table 1, our proposed technique achieved a higher AUC (∼1% for Ped1 and Ped2) and a lower EER (∼2% for Ped1 and Ped2) than a single U-Net-based prediction network. This suggests the extraction of spatial and temporal information is more suitable for anomaly detection.

2) EFFECT OF RGB DIFFERENCES ON MOTION LOSS

The performance of RGB difference motion loss, in terms of regression accuracy and training speed, was evaluated through a comparison with optical flow. As shown in Table 2, this adjusted motion loss constrained the model and achieved slightly better results for Ped2 and the Avenue datasets. Although the results of Ped1 dataset are slightly lower, the average time required for RGB difference calculation using batch data was reduced from 0.5333 (s/batch) to 0.0041 (s/batch) in all datasets, as shown in Table 3 (our framework was implemented with a NVIDIA GeForce GTX 1070 GPU and an Intel® Core TM i7-7700 CPU@ 3.60 GHz×8). This demonstrates that replacing optical flow with RGB differences can significantly reduce training time while maintaining model performance.

3) EFFECT OF PROPOSED REGULAR SCORES FUNCTION

The novel regular score function proposed in this study is based on the prediction error of both current and future video frames. It was evaluated by comparing its anomaly detection performance with that of other algorithms, including common score functions produced by either normalizing the prediction error of current frames or a combination of the prediction error from multiple frames.

TABLE 4. The gap(Δ_s), AUC AND EER values for different regular score functions.

		Regular score based on current frame	Regular score based on future frames (proposed)
Δ_s	Ped1	0.2523	**0.2691**
	Ped2	0.4918	**0.5024**
	Avenue	0.2721	**0.2873**
AUC	Ped1	83.46%	**83.82%**
	Ped2	96.34%	**96.56%**
	Avenue	84.32%	**84.59%**
EER	Ped1	22.36%	22.36%
	Ped2	8.70%	8.70%
	Avenue	21.77%	**21.54%**

In addition to AUC and EER, we utilized gaps between the average scores of normal and abnormal frames (denoted as Δ_s) to further illustrate the quantitative effects of our proposed function. Larger Δ_s values indicate the framework is more capable of distinguishing normal and abnormal patterns. As shown in Table 4, our proposed function increased the AUC by ∼0.3% with almost the same EER. The Δ_s increased by ∼0.02 in the Ped1, Ped2, and Avenue datasets.

4) COMPARISON WITH EXITING METHODS

A comparison of our approach with several conventional methods, including an autoencoder and a U-Net, is shown for the Ped1, Ped2, and Avenue datasets in Table 5. The proposed approach achieved an AUC of 96.5% and an EER of 8.7% for Ped2, outperforming comparable algorithms. Results for the Avenue dataset (AUC of 84.5%) were only slightly below the best performing model, which demonstrates the effectiveness of our method. AUC and EER values for our technique applied to Ped1 were 83.8% and 22.3%, respectively. This was higher than every other methodology except that of Zhao *et al.* [6], which produced a lower AUC for both the Ped2 and Avenue datasets. The limited performance of our method for Ped1 may be due to an inability to distinguish non-obvious abnormalities filtered out during the encoding process, as these structures would not have been modeled in temporal space or detected after decoding. The anomalies ignored by our method were detected by Zhao *et al.*, who used 3D convolutions to extract spatial and temporal information directly from original video frames. The 2D convolution operations used in our method were much simpler than complex 3D convolutions in terms of the algorithmic complexity. In addition, our method outperformed other models for the Ped2 and Avenue datasets.

Fig. 4 shows similar abnormal events (skateboarder and cyclist) in Ped1 and Ped2 for the same video scenario. It is evident that spatial characteristics for the same event differ quite significantly between the two sets, due to unique camera angles. Compared with Ped1, captured at a downward

TABLE 5. AUC and EER values for different algorithms applied on the PRD1, Ped2 and avenue datasets.

	Ped1		Ped2		Avenue	
	AUC	EER	AUC	EER	AUC	EER
Conv-AE [5]	58.5%	43.1%	84.7%	24.5%	77.2%	27%
Conv-AE [12]	81%	27.9%	90%	21.7%	70.2%	25.1%
CovnLSTM-AE [1]	74%	NA	81%	NA	84%	NA
Conv-AE+CovnLSTM [13]	75.5%	NA	88.1%	NA	77%	NA
3D-ConvAE [6]	**92.3%**	**15.8%**	91.2%	16.7%	80.9%	24.4%
U-Net [7]	83.1%	NA	95.4%	NA	**84.9%**	NA
The proposed approach	83.8%	22.3%	**96.5%**	**8.7%**	84.5%	**21.5%**

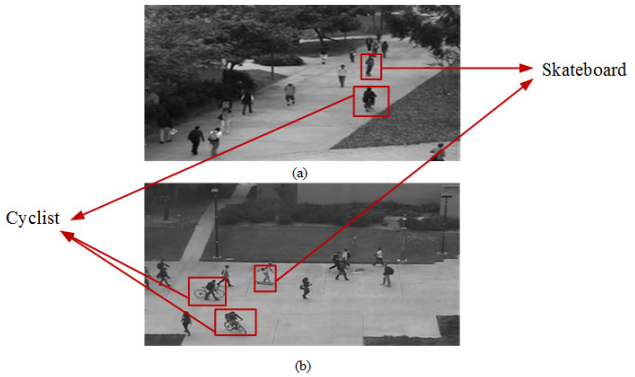


FIGURE 4. The same abnormal events (skateboard and cyclists) in the (a) Ped1 and (b) Ped2 datasets. Ped2 offers a less restricted view.

angle, Ped2 exhibits clearer spatial characteristics that are beneficial for predicting realistic future frames. For example, the skateboard is clearly visible under the skateboarder's feet in Fig. 4(b), but is difficult to see in Fig. 4(a) due to the angle. After analyzing the experimental data, we found that our method often failed to perform well in frames with these types of ambiguous spatial features. As such, future study is merited to improve the detection of obscure anomalous targets, by focusing on corresponding spatial characteristics or improving target tracking.

V. CONCLUSION

We proposed a novel spatio-temporal framework for anomaly detection by combing a U-Net with ConvLSTM to process video sequences. A new regular score function was also developed to improve the accuracy of anomaly detection. Qualitative analysis and quantitative comparisons were performed using the USDC (Ped1 and Ped2) and Avenue datasets, with results demonstrating that our proposed framework performed well on both, which achieved 96.5% AUC for Ped2 and 84.5% AUC for Avenue. The proposed method achieved the highest AUC on Ped2 and relatively high results on Ped1 and Avenue. Although the AUC measured on ped1 was slightly lower than previously achieved with a 3D autoencoder, our 2D convolution operations were simpler than complex 3D convolutions in terms of the algorithmic complexity. Future research will focus on improving the proposed framework by detecting non-obvious anomalies.

REFERENCES

[1] B. Kiran, T. Dilip, and P. Ranjith, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," Nov. 2018, *arXiv:1801.03149*. [Online]. Available: https://arxiv.org/abs/1801.03149

[2] 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.

[3] F. Yachuang, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017.

[4] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Oct. 2017.

[5] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," *Pattern Recognit. Lett.*, vol. 105, pp. 13–22, Apr. 2018.

[6] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proc. MM*, Mountain View, CA, USA, 2017, pp. 1933–1941.

[7] 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)*, Jun. 2018, pp. 6536–6545.

[8] V. Zyuzin, P. Sergey, A. Mukhtarov, T. Chumarnaya, O. Solovyova, A. Bobkova, and V. Myasnikov, "Identification of the left ventricle endocardial border on two-dimensional ultrasound images using the convolutional neural network Unet," in *Proc. Ural Symp. Biomed. Eng., Radioelectron. Inf. Technol. (USBEREIT)*, Yekaterinburg, Russia, May 2018, pp. 76–78.

[9] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[10] J. C. SanMiguel, J. M. Martínez, and L. Caro-Campos, "Object-size invariant anomaly detection in video-surveillance," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Madrid, Spain, 2017, pp. 1–6.

[11] S. Wang, E. Zhu, and J. Yin, "Video anomaly detection based on ULGP-OF descriptor and one-class ELM," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Vancouver, BC, Canada, Jul. 2016, pp. 2630–2637.

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

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

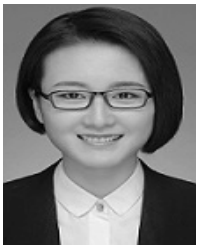
[14] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2016, pp. 1–27.

[15] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2015, *arXiv:1511.05440*. [Online]. Available: https://arxiv.org/abs/1511.05440

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2015, pp. 234–241.

[17] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *Comput. Sci.*, vol. 58, no. 11, pp. 2415–2422, 2015.

[18] 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. NIPS*, 2015, pp. 802–810.
[19] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *CoRR*, vol. abs/1701.00160, pp. 1–57, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1701.00160
[20] P. Gupta, P. Srivastava, S. Bhardwaj, V. Bhateja, "A modified PSNR metric based on HVS for quality assessment of color images," in *Proc. Commun. Ind. Appl. (ICCIA)*, Dec. 2011, pp. 1–4.
[21] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 2720–2727.
[22] V. Mahadevan, W. Li, and V. Bhalodia, "Anomaly detection in crowded scenes," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1975–1981.
[23] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.



YUANYUAN LI received the B.S. degree in communication engineering from the School of Information and Communications Engineering, Beijing University of Technology, Beijing, China, in 2017, where she is currently pursuing the M.S. degree in information and communication engineering. Her research interests include video analysis and image retrieval.

YIHENG CAI received the Ph.D. degree in pattern recognition and intelligent system from the Beijing University of Technology, China, in 2006. She is currently an Associate Professor with the School of Information and Communications Engineering, Beijing University of Technology. She has published over 60 articles in journals, book chapters, and conferences. Her research interests include image processing and pattern recognition. She is a member of the China Computer Federation.

JIAQI LIU received the B.S. degree in communication engineering from the School of Information, North China University of Technology, in 2018. He is currently pursuing the M.S. degree in information and communication engineering with the School of Information and Communications, Beijing University of Technology, Beijing, China. His research interest includes video analysis.

SHINAN LANG received the B.S. degree from the Beijing University of Science and Technology, in 2010, and the Ph.D. degree in engineering, in July 2015. She was escorted to the Institute of Electronics, Chinese Academy of Sciences, in 2010. In 2015, she joined the School of Information and Communication Engineering, Ministry of Informatics, Beijing University of Technology. She is currently an Associate Professor with the Beijing University of Technology.

XINFENG ZHANG received the B.S. degree in chemical machinery from Fuzhou University, the M.S. degree in fault diagnosis of mechanical equipment from the China University of Petroleum, Beijing, and the Ph.D. degree in pattern recognition and intelligent system from the Beijing University of Technology, China. He is currently an Associate Professor with the Beijing University of Technology. His research interests include image processing and machine learning.