An Approach to Detect Anomaly in Video Using Deep Generative Network

SAVATH SAYPADITH (Graduate Student Member, IEEE), AND TAKAO ONOYE (Senior Member, IEEE)
Graduate School of Information Science and Technology, Osaka University, Suita 565-0871, Japan

Corresponding author: Takao Onoye (onoye@ist.osaka-u.ac.jp)

This work was supported in part by the AUN/SEED-Net Project of Japan International Cooperation Agency (JICA).

ABSTRACT Anomaly detection in the video has recently gained attention due to its importance in the intelligent surveillance system. Even though the performance of the state-of-art methods has been competitive in the benchmark dataset, the trade-off between the computational resource and the accuracy of the anomaly detection should be considered. In this paper, we present a framework to detect anomalies in video. We proposed a “multi-scale U-Net” network architecture, the unsupervised learning for anomaly detection in video based on generative adversarial network (GAN) structure. Shortcut Inception Modules (SIMs) and residual skip connection are employed to the generator network to increase the ability of the training and testing of the neural network. An asymmetric convolution has been applied instead of traditional convolution layers to decrease the number of training parameters without performance penalty in terms of detection accuracy. In the training phase, the generator network was trained to generate the normal events and attempt to make the generated image and the ground truth to be similar. A multi-scale U-Net kept useful features of an image that were lost during training caused by the convolution operator. The generator network is trained by minimizing the reconstruction error on the normal data and then using the reconstruction error as an indicator of anomalies in the testing phase. Our proposed framework has been evaluated on three benchmark datasets, including UCSD pedestrian, CHUK Avenue, and ShanghaiTech. As a result, the proposed framework surpasses the state-of-the-art learning-based methods on all these datasets, which achieved 95.7%, 86.9%, and 73.0% in terms of AUC. Moreover, the numbers of training and testing parameters in our framework are reduced compared to the baseline network architecture, while the detection accuracy is still improved.

INDEX TERMS Inception module, video anomaly detection, generative network.

I. INTRODUCTION

Currently, videos from Closed-Circuit Television (CCTV) cameras are rapidly generated every minute in accordance with an increasing number of cameras in public places in order to increase the efficiency, safety, and security due to criminal and terrorist attacks. CCTV cameras are also attributed to monitor areas such as shopping malls, hotels, streets, banks, and government buildings. However, the monitoring proficiency of anomaly events in hundred surveillance cameras using human labor is ambitious. To overcome this problem, developing intelligent computer vision algorithms to automatically detect events in a video scene is a viable solution. Anomaly detection in the video has attracted ones in many research fields, especially in the computer vision field. The term “anomaly” refers to an event that is different from normality. The challenging to detect an anomaly event is to distinguish the pattern of object movement, i.e. normal or anomaly, since the video scene captured by surveillance cameras may incur movement over the time. Real-world anomaly events are complicated and it is difficult to define every specific event. Although anomaly detection algorithms have reached the accuracy level under certain condition, the algorithm may still be affected by the external and internal variation such as the illumination, direction of movement object, motion velocity, occlusion and similar object motion.

Video data is high dimensional data containing noise, high variations, and interactions, making the analysis and defining the anomaly event in the scene more challenging. There are many successful researches in the field of action recognition [1]–[5] attempt to label the video where events and actions are clearly defined. However, human labeling of
a real-world video scene in every event type is extremely costly or impossible. The training video could also not cover every anomaly event due to some anomaly events rarely or never occurred. Three issues impact the challenge of detecting an anomaly in the video: 1) anomalies are diverse and highly contextual. For example, people running at a park are considered normal events, while running in the bank would be considered an anomaly [6]–[8]. 2) The computational complexity and cost of processing across a video frame due to the spatial and temporal dimensions of the video data. In practice, when an anomaly occurred, the algorithm should be able to alarm on time with an acceptable accuracy rate and false alarm rate. 3) Environmental conditions are diverse such as illumination, crowd density, object shadow, object occlusions, complex background, etc.

Among related works, several anomaly detection approaches based on either a convolution autoencoder (ConvAE) [9]–[11] or a U-Net [12] are performed in a different way to detect an anomaly. These approaches learn the normal patterns from training videos and then detecting abnormal events which would not correspond to the learned model [10], [13]–[15]. Deep learning has been proposed for anomaly detection [9], [10], [12]–[19] including supervised and unsupervised learning. The method based on deep learning has improved the accuracy as well as reducing the false alarm rate. Additionally, Generative Adversarial Network (GAN) [20] has been used to generate a realistic image based on a given input image sequence [12]. However, a generator network in the anomaly detection framework still lost some features during training which may affect the accuracy of the detection.

In this paper, we proposed a framework for video anomaly detection using GAN structure. Spatial and temporal features are extract through the multi-scale U-Net architecture. In training, PatchGAN [21] has been utilized to distinguish the ground truth image and the output image from the generator. We also used the optical flow in training to optimize the training parameters. In the testing phase, a detector refers to the regularity score of an image. A lower score indicates anomaly situation, where errors between a generated frame based on the normal training data and the ground truth image are large. This paper is an extended version of work published in [22]. To summarize, our contributions in this paper are:

- We employ Shortcut Inception Module (SIM) and residual skip connection to the generator network called “multi-scale U-Net” to make the network learning higher-level features.
- We apply the idea of an asymmetric convolution layer and increase the width of the network architecture in order to attain both of small model size and high training efficiency.
- The proposed multi-scale U-Net reduces the parameter number of training and testing, while the anomaly detection accuracy still significantly improves.
- We evaluate our proposed framework with three benchmark datasets of different scene scenarios. Experiments on the benchmark datasets show the effectiveness of our proposed framework for video anomaly detection.

The rest of this paper is organized as follows. We provide an overview of related works of anomaly detection approaches in Section II. Then details of our proposed framework are described in Section III. The experimental results are provided in Section IV followed by our final conclusion in Section V.

## II. RELATED WORKS

Anomaly detection approaches can be classified into two main categories including a hand-crafted feature and a learning-based method. Existing hand-crafted features based method [14] has been proposed to learn a set of sparse combinations to model normal events in a video scene. However, hand-crafted features based methods require prior knowledge to define specific parameters for any possible abnormal patterns, which is difficult to adapt to huge variations of different video scenes in real-time anomaly detection.

Learning-based methods currently achieve significant performance in a wide range of computer vision applications, which is improving the accuracy and reducing false alarm rate of the detection and recognition. Motion features are required to model the object movement in a video. A 2D convolution layer will output an image that loses temporal feature of video signal. Only 3D convolution can extract the temporal feature and output in volume. Xu, Dan, et al. [17] presented a novel Appearance and Motion DeepNet (AMDN). Ionescu et al. [15] presented abnormal event detection based on a two-stage outlier elimination algorithm. The algorithm eliminates the outlier using k-mean clustering and classifies by training a one-class SVM. Therefore, Sultani et al. [19] proposed a framework for training anomaly and normal videos using Multiple Instance Learning (MIL) by dividing the videos and video segments into instance and bags. The deep anomaly ranking model has been used to predict high anomaly scores, which reflects an anomaly event. In these approaches, learning features are extracted in two streams (spatial and temporal), which may consume considerable time for training and testing. Fan et al. [18] proposed a video anomaly detection method based on a supervised learning approach called Gaussian Mixture Variational Autoencoder. The assumption is that the anomaly sample does not belong to any Gaussian component of a Gaussian Mixture Model (GMM). Recurrent neural network (RNN) and long short term memory are applied in anomaly detection to model temporal pattern. Another work [9] leverages a convolutional LSTM based on Auto-Encoder (ConvLSTM-AE) to model both appearance and motion information. Another work proposed in [16] iteratively updates sparse coefficients via a stacked RNN to detect anomalies in videos.

A generative network is one of learning-based methods proposed to generate more realistic datasets in anomaly detection. The generative network aims to infer the data distribution to generate new images that could belong to the...
same set as training data. Chong and Tay proposed end-to-end architecture for learning video representation [11], which included two main components, one for the spatial component and the other for the temporal component. The structure of this network architecture is based on an autoencoder that aims to reconstruct the input image. Liu, et al. proposed a framework for anomaly detection based on GAN [12]. To generate a more realistic future frame, a U-Net was used as a primary prediction network (a.k.a. a generator network). A motion feature is used in training by enforcing the optical flow between predicted images and ground truth images to be consistent. A skip connection is applied in each layer of the U-Net architecture [23] to improve the quality of the reconstruction image. At the end of the training phase, the discriminator network was used to distinguish an image created by the generator from the ground truth image. However, some features are lost in these generator networks during training due to convolution operators of each layer. To resolve this issue, we presented a framework that consists of a multi-scale generator network and residual skip connection to make a network learn higher-level features of images.

III. PROPOSED APPROACH

In this section, details of our proposed framework are described, as illustrated in Fig. 1. Firstly, given an input image sequence, a multi-scale U-Net is utilized as a generator network $G$ to extract spatial features. Fig. 2 shows the structure of the multi-scale U-Net. We employed SIM inside the multi-scale U-Net to make the network learning the feature in different scales. Instead of the traditional skip connection of the U-Net architecture, residual skip connections are applied in our architecture for the purpose of propagating spatial information which was lost during the convolutional operation from encoder to decoder. The use of these residual skip connections is beneficial for learning higher feature of an image. In the training phase, we enforce the optical flow of the generated image to be close to that of the ground truth image in order to optimize the network parameters. Further, we used PatchGAN [21] as a discriminator network $D$ into our framework to distinguish between the generated image and the ground truth image. We describe details of each part in subsections below.

A. FEATURE EXTRACTION USING INCEPTION ARCHITECTURE

U-Net is considered as one of convolutional neural network (CNN) architectures that was introduced for biomedical image segmentation [23]. The U-Net architecture mainly consists of two parts, the encoder and the decoder. The encoder captures the context of the image by extracting the feature to a small vector size, called latent vector. On the other hand, the decoder aims to extract the feature and recover image details from the vector, where the upsampling layer is applied to increase the size of the feature. Typically, the encoder involves a sequence of two consecutive $3 \times 3$ convolution layers followed by a max-pooling operation. As explained in [24], the sequence of two $3 \times 3$ convolutional layers actually resembles a $5 \times 5$ convolution operation with the same input size and output size. To improve the learning efficiency of the U-Net architecture with feature learning in different scales,
a viable way is to integrate $5 \times 5$ and $7 \times 7$ convolution operations in parallel to the $3 \times 3$ convolution operation. Another possible option for improving the detection performance is increasing the size of the network architecture in terms of the depth and the width [25], [26].

In this work, we used U-Net as our base network architecture for the generator network. Unlike [23], we modified and replaced original convolution layers by SIM and increased the width of the network to make it possible to learn the input image in higher-level features. Fig. 3 (c) illustrates a proposed SIM inspired by the idea of an inception module [24] (Fig. 3 (a)). As described above, we can replace the convolutional layers with the inception blocks. Although the performance gain can be expected in introducing the larger size of convolution operations such as $5 \times 5$ and $7 \times 7$, the parallel network structure consumes high computational capacity. In the same manner as [24], we factorize and stack the larger size of $5 \times 5$ and $7 \times 7$ convolution operators by a $3 \times 3$ convolution operation, the output of the last two $3 \times 3$ convolutions competent to the $5 \times 5$ and $7 \times 7$ convolutions approximately as shown in Fig. 3 (b). We also take an advantage of a feature concatenation to extract the feature from different scales [25]. We then add a shortcut connection with an additional $1 \times 1$ convolutional layer to add more non-linearity information to enhance the representation as well as reducing the network’s dimension without a performance penalty [27]. In [22], the traditional convolution layers are utilized in the inception module. Unlike [22], our proposed block uses the ideas of the asymmetric convolution operation, which aims to factorize a standard two-dimensional convolution kernel into two one-dimensional convolution kernels. For example, a $3 \times 3$ convolution is equivalent to a stacking of a $3 \times 1$ convolution followed by a $1 \times 3$ convolution, which reduces the size of the model and increasing the training efficiency [24], [28], [29].

Details of the proposed SIM are summarized in Table 1. We assign $W_j$ to control the number of filters used in our proposed module in each depth $j$. Inside the SIM, we assign $W_j$, $W_j$, and $W_j$ to each of three convolution layers respectively, as this combination achieved the best results in our preliminary experiment. We compute $W_j$ as follows.

$$W_j = \alpha \times N_j,$$

where $N_j$ is the number of filters in the corresponding depth $j$ of the multi-scale U-Net, and $\alpha$ is a scalar coefficient. Typically, the filters should be gradually increased to prevent the memory usage of the earlier depth from rising the deeper network. Therefore, the number of filters $N_j$ of the network architecture in depth $j$ is set to $2^{5-j}$. We selected $\alpha = 1.5$ as it keeps the number of parameters slightly below that of the original U-Net.

### B. RESIDUAL SKIP CONNECTION

The U-Net architecture [23] also proposed the idea of using the skip connections between the encoder after the max-pooling operation and the decoder before the deconvolution layer. The aim of the skip connection is to propagate the spatial information that lost in every convolution operation from the encoder and the decoder, which is beneficial in recovering the clean image. As the design of the U-Net architecture, the features from the encoder are supposed to be low-level features, and the features from the decoder are supposed to be higher level as they are computed in a deep network. Thus, the fusion of these sets of features from encoder-decoder could cause feature learning which affects the reconstruction output.

Following the deep residual network [30] that proposed the idea of using residual learning block, as shown in Fig. 4 (a),

![Figure 3](image-url)
we introduce residual skip connection blocks to our proposed generator network. As illustrated in Fig. 4 (b), the proposed residual skip connection block consists of an asymmetric convolution layer of a $3 \times 1$ convolution layer followed by a $1 \times 3$ convolution layer and a shortcut connection of a $1 \times 1$ convolution layer, which allows the network to learn additional information from the input. In the generator network, instead of concatenating the feature maps from the encoder to the decoder, we pass encoder features as the input through a chain of the residual skip connection block, and the output of the block is concatenated with the decoder features. A significant amount of image detail could be lost or corrupted using more convolution layers [31]. Therefore, the residual skip connections make it possible to keep useful features lost by the convolution operation, and it is beneficial to train the deep network while still having fewer parameters. We denoted a residual skip connection block as $RB_{i,j}$, where $i$ is the number of blocks used in each depth $j$. Basically, the feature maps of each layer in the encoder are likely to decrease as we down-sampling the image in every step using the max-pooling layer. Therefore, we gradually decrease the number of $RB_{i,j}$, which is $i = 4, 3, 2, 1$ in each depth $j$, respectively. The number of filters in the block is set to the same with $N_j$ in each depth $j$.

C. OBJECTIVE FUNCTIONS

At the training time, $G$ learns to map the ground truth image $I$ and the generated image $\hat{I}$ to be consistent. The intensity loss $L_{int}$ and the gradient loss $L_{grad}$ of two images along two spatial dimensions are used to minimize the reconstruction error between $I$ and $\hat{I}$, which can be computed as follows.

$$L_{int}(I, \hat{I}) = \|I - \hat{I}\|_2^2$$  \hspace{1cm} (2)

$$L_{grad}(I, \hat{I}) = \|\hat{I}_{i,j} - \hat{I}_{i-1,j}\|_1 + \|\hat{I}_{i,j} - \hat{I}_{i,j-1}\|_1 + \|I_{i,j} - I_{i-1,j}\|_1 + \|I_{i,j} - I_{i,j-1}\|_1.$$  \hspace{1cm} (3)

The optical flow loss is applied to capture the motion information and to optimize training parameters, we used a FlowNet [32] to estimate the optical flow. Following [9], we apply the $L1$ distance loss to calculate the motion penalty:

$$L_{\text{flow}}(\hat{F} - F) = \|\hat{F} - F\|_2,$$  \hspace{1cm} (4)

where $F$ is the ground truth optical flow estimated from two consecutive frame $I_t$ and $I_{t+1}$ and $\hat{F}$ is the output optical flow calculated by $I_t$ and the generated image $\hat{I}_{t+1}$.

In addition to the loss functions described above, we used an adversarial loss based on Generative Adversarial Network (GAN) [20] to constrain the training process and improve model performance [12]. Given an input image sequence, the proposed multi-scale generator $G$ is trained by the adversarial loss, which encourages the generator to generate a more realistic image. A discriminator network $D$ is used to optimize model parameters to make generated image $\hat{I}$ indistinguishable from the ground truth image $I$. We utilized PatchGAN [21] as a discriminator network. PatchGAN maps $\hat{I}$ to small patches, where a discriminator takes each individual patch and predicts whether a patch come from $I$ or $\hat{I}$. A discriminator outputs a scalar which classify the patch from $I$ as class 1 and the patch from $\hat{I}$ as class 0. The goal of training $G$ is to generate an image where $D$ classifies it into class 1. A mean square error loss function $L_{MSE}$ and adversarial loss $L_{adv}$ were used as objective functions which can be calculated as follows.

$$L_{adv}(\hat{I}) = \sum_{m,n} \frac{1}{2} L_{MSE}(D_{\text{gen}}(I)_{m,n}, 1)$$  \hspace{1cm} (5)

$$L_{MSE}(I, \hat{I}) = (I - \hat{I})^2,$$  \hspace{1cm} (6)

where $D_{\text{gen}}(\cdot)$ is the output of the discriminator network of $\hat{I}$ and $m, n$ denote patch indexes. Finally, our proposed final objective function $L$ can be computed as follows.

$$L = \lambda_{\text{int}} L_{int} + \lambda_{\text{grad}} L_{\text{grad}} + \lambda_{\text{adv}} L_{adv} + \lambda_{\text{flow}} L_{\text{flow}},$$  \hspace{1cm} (7)

where $\lambda_{\text{int}}, \lambda_{\text{grad}}, \lambda_{\text{adv}},$ and $\lambda_{\text{flow}}$ are weights of each loss.

D. ANOMALY DETECTION USING REGULARITY SCORE

In testing, we computed the anomaly score in every frame of the testing video by measuring the similarity of the ground truth image and the generated image. In the same manner with [12], Peak Signal to Noise Ratio (PSNR) is utilized as the detection score in our framework. The PSNR calculates the image quality, where a low value of PSNR means that the image is likely to be abnormal. PSNR is defined as

$$\text{PSNR}(I_t, \hat{I}_t) = 10 \log_{10} \frac{[\text{max}_j]^2}{\sum_{i=0}^{N}(p_i(i) - \hat{p}_i(i))^2},$$  \hspace{1cm} (8)

where $\text{max}_j$ represents the maximum intensity value in a generated image $\hat{I}_t$, $p_i(i)$ and $\hat{p}_i(i)$ are pixel intensity of index $i$ in $I_t$ and $\hat{I}_t$, respectively. $N$ denotes the total number of pixels in the image. Then, we obtain a regularity score $R(t)$ for each frame $t$ in the video by normalizing the PSNR to the range of $[0, 1]$ as follows:

$$R(t) = \frac{\text{PSNR}(I_t, \hat{I}_t) - \text{min}_{PSNR}}{\text{max}_{PSNR} - \text{min}_{PSNR}},$$  \hspace{1cm} (9)

where the terms $\text{min}_{PSNR}$ and $\text{max}_{PSNR}$ are the minimum and maximum values of the PSNR in every frame of each test video. Finally, we detect anomaly event in an image based on a threshold of regularity score $R(t)$. 

![FIGURE 4. The comparison design of (a) the residual learning block [30] with (b) our proposed residual skip connection.](image-url)
IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETUP

Our implementation is based on Tensorflow framework [33] using Python 3.7. The network architecture was trained and tested on NVIDIA Geforce RTX 2080 GPU. Training is based on Adam optimizer and the batch-size is fixed to 4. In our training and testing, input images of the network architecture are resized to $256 \times 256$ of 3 color channels and normalize it to the range of $[-1, 1]$. To be consistent with [12], we have set 4 consecutive images as the input image sequence.

B. DATASETS

We evaluated and compared our framework with state-of-the-art learning based methods using three benchmark datasets for video anomaly detection, including the CUHK Avenue [14], the UCSD Pedestrian [13], and ShanghaiTech datasets [12]. The CUHK Avenue dataset contains 16 training and 21 testing videos with abnormal events such as abnormal objects, throwing, and strange actions. The UCSD Pedestrian dataset provided two subsets, Ped1 and Ped2. The Ped1 subset contains 34 training and 36 testing videos, and the Ped2 contains 16 training and 12 testing videos. The anomaly events in the UCSD dataset includes cars, scooters, wheelchairs, and bicycles. The ShanghaiTech dataset covers challenging scenarios for video anomaly due to large variations in appearance and viewpoint, which consists of 13 scenes of 330 training and 107 testing videos.

C. EVALUATION METRIC

Following the framework described in [12], we use a receiver operating characteristic (ROC) curve, Equal Error Rate (EER), and corresponding an area under the curve (AUC) to evaluate the detection performance for qualitative comparison. ROC is used to visualise the performance of a binary classifier by plotting the trade-off between the true positive rate (TPR) and the false positive rate (FPR) with varying its discrimination threshold. AUC used in most previous works [9]–[12], [16], [18] that measures the entire of two dimensional area of ROC, which provides an aggregate measure of performance across all possible classification thresholds. In this study, higher AUC values and lower EER values indicated better performance of anomaly detection.

D. QUANTITATIVE RESULTS

We have compared our proposed with several anomaly detection methods based on deep learning, including autoencoders [10], [11], [18], LSTM [9], [16], and GAN [12]. Experimental results summarized in Table 2 demonstrated that our proposed framework achieved better AUC and EER in UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets. Specifically, compared to Liu et al. [12] that provides initial results for our proposed framework, our proposed achieves relative AUC and EER improvement over the original U-Net in all these datasets. The performance is slightly improved with respect to the ShanghaiTech dataset due to the fact that the ShanghaiTech is a larger-scale dataset that contains several types of anomalies and complicated movement. The performance in UCSD Ped1 is 85.3% whereas the best result of 94.9% is achieved by Fan et al. [18], who employed a two-stream network that combines the appearance and motion of anomalies which may incur considerable computational cost. The overall result illustrated that the design of our proposed framework is able to capture appearance and motion information to detect anomalies in real-world scene scenarios.

E. QUALITATIVE RESULTS

The qualitative results of our proposed framework on two testing videos in the CUHK Avenue and UCSD Ped2 dataset are illustrated in Fig. 5. We can see that the generated image tends to achieve a high regularity score while decreasing when the anomaly occurred (e.g., running, bicycle intrusion). We also show the output of the proposed generator network in Fig. 6. The result of the generated image and the image

| Methods          | CUHK Avenue AUC | UCSD Ped1 AUC | UCSD Ped2 AUC | ShanghaiTech AUC |
|------------------|-----------------|---------------|---------------|------------------|
| Chong et al. [11]| 80.3%           | 89.9%         | 87.4%         | -                |
| Conv-AE [10]     | 80.0%           | 81.9%         | 90.9%         | -                |
| ConvLSTM-AE [9]  | 77.0%           | 75.7%         | 88.1%         | -                |
| StackRNN [16]    | 81.7%           | -             | 92.2%         | 68.0%            |
| Fan et al. [18]  | 83.4%           | 94.9%         | 92.2%         | 72.8%            |
| Liu et al. [12]  | 84.9%           | 83.1%         | 95.4%         | 32.6%            |
| **Our Proposed** | **86.9%**       | **85.3%**     | **95.7%**     | **73.0%**        |

| Methods          | CUHK Avenue EER | UCSD Ped1 EER | UCSD Ped2 EER | ShanghaiTech EER |
|------------------|-----------------|---------------|---------------|------------------|
| Chong et al. [11]| 20.7%           | 27.9%         | 21.7%         | -                |
| Conv-AE [10]     | 25.1%           | 29.9%         | 21.7%         | -                |
| ConvLSTM-AE [9]  | 75.7%           | -             | -             | -                |
| StackRNN [16]    | 81.7%           | -             | 92.2%         | 68.0%            |
| Fan et al. [18]  | 22.7%           | 11.3%         | 12.6%         | -                |
| Liu et al. [12]  | 20.6%           | 23.5%         | 12.0%         | 72.8%            |
| **Our Proposed** | **20.2%**       | **23.6%**     | **12.0%**     | **32.3%**        |
the inception module and residual skip connection achieved the best detection performance compared to the original U-Net architecture.

G. RUNNING TIME ANALYSIS
A comparison of the parameter number of the original U-Net architecture and proposed ones are presented in Table 4, demonstrating that our multi-scale U-Net reduces the number of parameters for training and testing an anomaly detection while improving the accuracy. We also evaluated the computational cost of the proposed framework on the Shanghaitech dataset. The running times were measured on NVIDIA Geforce RTX 2080 Ti GPU with 24 GB of RAM. Our proposed framework takes 0.041 seconds per frame in averaged. Hence, it can run at 24 frames per second (fps) within the entire pipeline, which is on par or slightly better than using the baseline network architecture [23] that is able to run at about 22 fps.

V. CONCLUSION
In this work, we present a framework based on multi-scale U-Net architecture for anomaly detection in video. The inception modules are employed instead of using the traditional convolution layers that utilize in the original U-Net, making our multi-scale U-Net has ability to learn image features in different scales. The skip connections were replaced by our proposed residual skip connection including shortcut connections increase the ability to train a deeper network while still having fewer parameters. In the feature extraction part, an asymmetric convolution kernel is applied to reduce the number of network parameters while still keeping the detection accuracy. As a result in both qualitative and quantitative, our proposed framework based on multi-scale U-Net achieved better performance with a lightweight model and less memory usage compared to other learning-based anomaly detection approaches. However, the generator network is unable to distinguish an ambiguous anomalous object in a scene. For future work, we will explore on experimenting on applying two-stream inputs for feature extraction to our model to capture both the appearance and the motion of the object characteristic and to enhance the performance of the anomaly detection.

REFERENCES
[1] K. Liu, W. Liu, H. Ma, W. Huang, and X. Dong, “Generalized zero-shot learning for action recognition with web-scale video data,” World Wide Web, vol. 22, no. 2, pp. 807–824, Mar. 2019.
[2] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 4597–4605.
[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1725–1732.
[4] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.
[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 4489–4497.
[6] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, 2009.
[7] B. R. Kiran, D. M. Thomas, and R. Parakkal, “An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos,” J. Imag., vol. 4, no. 2, p. 36, 2018.
[8] R. Nayak, U. C. Pati, and S. K. Das, “A comprehensive review on deep learning-based methods for video anomaly detection,” Image Vis. Comput., vol. 106, Feb. 2021, Art. no. 104078.
[9] W. Luo, W. Liu, and S. Gao, “Remembering history with convolutional LSTM for anomaly detection,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2017, pp. 439–444.
[10] M. Hasan, J. Choi, 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.
[11] Y. S. Chong and Y. H. Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” in Proc. 14th Int. Symp. ISNN, Sapporo, Japan, Jun. 2017, pp. 189–196.
[12] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—A new baseline,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6536–6545.
[13] V. Mahadevan, W. Li, V. Bhadolia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1975–1981.
[14] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 FPS in MATLAB,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 2720–2727.
[15] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, “Detecting abnormal events in video using normalized abnormality clusters,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2019, pp. 1951–1960.
[16] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked RNN framework,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 341–349.
[17] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, “Learning deep representations of appearance and motion for abnormal event detection,” 2015, arXiv:1510.01555.
[18] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, “Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder,” Comput. Vis. Image Understand., vol. 195, Jun. 2020, Art. no. 102920.
[19] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE/CCF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6479–6488.
[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1125–1134.
[22] S. Saypadith and T. Onoye, “Video anomaly detection based on deep generative network,” in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2021, pp. 1–5.
[23] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Springer, 2015, pp. 234–241.
[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1, pp. 1–7.
[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 1–9.
[26] S. Zagoruyko and N. Komodakis, “Wide residual networks,” 2016, arXiv:1605.09283.
[27] X. Li, C. Xiao, H. Li, and W. Zhang, “EACNet: Enhanced asymmetric convolution for real-time semantic segmentation,” IEEE Signal Process. Lett., vol. 28, pp. 234–238, 2021.
[28] Y. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.
[29] X.-J. Mao, C. Shen, and Y.-B. Yang, “Image restoration using convolutional auto-encoders with symmetric skip connections,” 2016, arXiv:1606.08921.
[30] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, “FlowNet: Learning optical flow with convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 2758–2766.
[31] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI), 2016, pp. 265–283.
[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.
[33] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: Learning basic visual concepts with a constrained variational framework,” in Proc. ICLR, p. 6, vol. 2, no. 5, 2017.
[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2921–2929.
[35] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[36] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 802–810.
[37] D. P Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114.
[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2818–2826.

SAVATH SAYPADITH (Graduate Student Member, IEEE) received the B.Eng. degree in information technology from the National University of Laos, Vientiane, Laos, in 2014, and the M.Eng. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 2018. He is currently pursuing the Ph.D. degree in information system engineering with the Graduate School of Information Science and Technology, Osaka University, Japan. He was a Lecturer with the Faculty of Engineering, National University of Laos. His research interests include image and video processing, computer vision, and machine learning.

TAKAO ONOYE (Senior Member, IEEE) received the B.E. and M.E. degrees in electronic engineering and the Dr.Eng. degree in information systems engineering from Osaka University, Osaka, Japan, in 1991, 1993, and 1997, respectively. He was an Associate Professor with the Department of Communications and Computer Engineering, Kyoto University, Kyoto, Japan. Since 2003, he has been a Professor with the Department of Information Systems Engineering, Osaka University. He has published over 200 research papers in the field of VLSI design and multimedia signal processing in reputed journals and proceedings of international conferences. His current research interests include media-centric low-power architecture and its SoC implementation. Dr. Onoye has been served as a member of the CAS Society Board of Governors, since 2008. He is also a member of IEICE, IPSJ, and ITE-J.