OUKA
The University of Osaka
Institutional Knowledge Archive

Title	A Study on Anomaly Detection in Surveillance Video using Generative Adversarial Network
Author(s)	Saypadith, Savath
Citation	大阪大学, 2022, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/88145
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

A Study on Anomaly Detection in Surveillance Video using Generative Adversarial Network

Submitted to

Graduate School of Information Science and Technology

Osaka University

January 2022

Savath SAYPADITH

Publications

Journal Article

[J1] Saypadith, S., & Onoye, T. (2021). An Approach to Detect Anomaly in Video Using Deep Generative Network. *IEEE Access, 9,* 150903-150910.

International Conference Papers

[I1] Saypadith, S., & Onoye, T. (2021, May). Video Anomaly Detection based on Deep Generative Network. In *Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1-5). IEEE.

[I2] Saypadith, S., & Onoye, T. Joint Representation Learning for Video Anomaly Detection. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. *(Under review)*

Other Publications

[1] Saypadith, S., & Onoye, T. (2021) Abnormal Event Detection using Joint Spatiotemporal Learning. In *Proceedings of The 14th Regional Conference on Electrical and Electronics Engineering (RC-EEE)*

[2] Saypadith, S., & Aramvith, S. (2018, November). Real-time multiple face recognition using deep learning on embedded GPU system. In *Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1318-1324). IEEE.

[3] Saypadith, S., Ruangsang, W., & Aramvith, S. (2017, December). Optimized human detection on the embedded computer vision system. In *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*(pp. 1707-1711). IEEE.

[4] Vilaysouk, X., Saypadith, S., & Hashimoto, S. (2021). Semisupervised machine learning classification framework for material intensity parameters of residential buildings. *Journal of Industrial Ecology.*

[5] Bounnady, K., Sibounnavong, P., Chanthavong, K., & Saypadith, S. (2019, July). Smart Crop Cultivation Monitoring System by Using IoT. In *Proceedings of the 2019 5th International Conference on Engineering* (pp. 1-3). IEEE.

Summary

Videos from Closed-Circuit Television (CCTV) cameras are rapidly generated every minute in accordance with an increasing number of cameras either in public places or private places in order to increase the efficiency, safety, and security due to criminal and terrorist attacks. 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 recently gained attention due to its importance in the intelligent surveillance system. 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.

Even though the performance of the state-of-art methods has been competitive in the benchmark dataset, the trade-off between the processing time and the accuracy of the anomaly detection should be considered. This dissertation proposes a framework for detecting anomalies in video, which designs a "multi-scale U-Net" network architecture based on generative adversarial network (GAN) structure for unsupervised learning to detect anomaly in video. To improve the training and testing of the neural network, Shortcut Inception Modules (SIMs) and residual skip connections are used in the generator network. Instead of using traditional convolution layers, an asymmetric convolution was used to reduce the number of training parameters without impacting detection accuracy. 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. This dissertation evaluates the performance with three benchmark datasets including UCSD Pedestrian, CUHK Avenue and ShanghaiTech datasets. The experimental results demonstrate that the framework surpasses the state-of-the-art learning-based methods, which achieved 95.7%, 86.9%, and 73.0% in terms of AUC. The multi-scale U-Net reduces the number of network parameters by 22.6% compared to the original U-Net architecture. In average, the proposed architecture takes 0.041 seconds per frame. As a result, the complete pipeline can run at 24 frames per second (fps), which is on par or slightly better than the baseline network architecture, which can run at roughly 22 fps.

This dissertation also proposes a joint representation learning for video anomaly detection. The proposed architecture extracts features from the object appearance and their associate motion features via different encoders based on ResNet network architecture. The network architecture is designed to combine spatial and temporal features, which share the same decoder. Using a joint representation learning approach, the proposed architecture effectively learn both appearance and motion features to detect anomalies in various scene scenarios. The experiments on three benchmark datasets demonstrate the remarkable detection accuracy with respect to existing state-of-the-art methods, which achieve 96.5%, 86.9%, and 73.4% in UCSD Pedestrian, CHUK Avenue, and ShanghaiTech datasets, respectively.

Acknowledgments

This doctoral dissertation would not have been possible without the assistance, support, or guidance of the following people and organizations.

First of all, I would like to express my sincerest gratitude to my research supervisor Professor Takao Onoye of the Osaka University for providing me the opportunity and environment to study as a doctoral student in his laboratory. I would like to thank you for his tireless efforts, encouragement, supports and guidance throughout the entire course of my study. I also would like to express my appreciation to Associate Professor Ittetsu Taniguchi of Osaka University, who also support me and the precious suggestions during my doctoral course.

I would like to thank the members of my dissertation committee, Professor Yasushi Sakurai, Associate Professor Ittetsu Taniguchi, and Associate Professor Supavadee Aramvith, for giving critical reviews and their advice on this dissertation.

Next, I would like to thank other colleagues who belong or belonged to the Information System Synthesis Laboratory in Osaka University for daily discussions and their support: Mr. Kenshiro Kato, Mr. Buntueng Yana, Ms. Prasara Jakkaew, Mr. Liu Jin, Mr. Shun Kimura and Mr. Defang Zhao. I express my heartfelt thanks to all members of the Information System Synthesis Laboratory at Osaka University for having an interesting and comfortable time in the laboratory. I would like to thank the laboratory secretary Ms. Yuki Yoshida for her various support.

Finally, I would like to extend my gratitude to my parents (Saveng Saypadith and Nguyen Thi Hue), my sister and brother, other relatives and friends. They always support me and encourage me with their best wishes, inspiration, and suggestion. This research was partly supported by AUN/SEED-Net Project from the Japanese Government through Japan International Cooperation Agency (JICA).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation focuses on approaches to detect anomaly in surveillance video using generative neural network. This chapter describes the research background, objectives and scopes of this dissertation. Section 1.1 describes the background of the anomaly detection, followed by motivations and problem statements. Then, Section 1.2 presents research objectives and contributions of this research. Finally, the overall organization of this dissertation is summarized in Section 1.3

1.1 Motivation and Problem Statement

Surveillance cameras are crucial that makes to secure shield houses, offices, and surroundings. In recent years, the global surveillance camera market has a significant growth due to the increase in the adoption at various vanues such as hospitals, streets, banks, airports, shopping malls, universities, and home security. According to a new report issued by IHS Markit [1], over one billion surveillance cameras will be installed worldwide in the end of 2021. With the development of Artificial Intelligence (AI) technology, along with new innovations in advanced features of the surveillance systems, Closed Circuit Television (CCTV) cameras are becoming more automated as they are able to analyse, detect, classify, and track predefinded objects in the video scene. The report of [2,3] shows that CCTV is associated with a significant reduction in crime. The findings of the study also revealed evidence of large decreases in crime in other settings, notably residential regions. Global adoption of AI surveillance is increasing

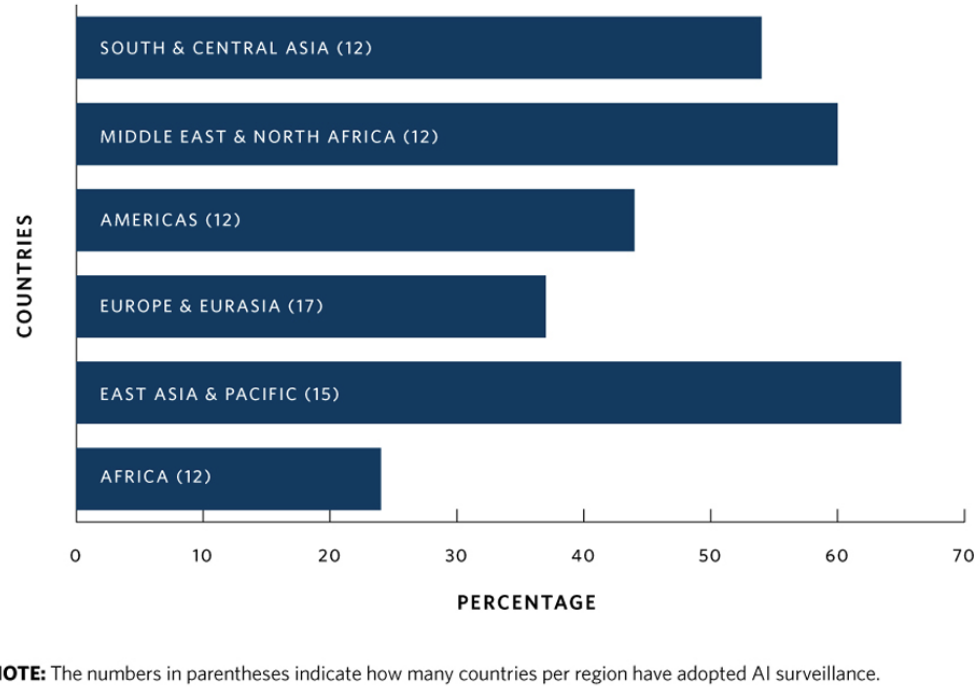


Figure 1.1: Percentage of countries by region adopting AI surveillance [4].

at a rapid pace around the world. As shown in Figure 1.1, Feldstein [4] reports that at least 75 out of 176 countries globally are actively using AI technologies for surveillance purposes including smart/safe city platform, facial recognition and smart policing.

With COVID-19, the use of contactless technology has also become more popular and it would further accelerate the growth of the new technology. Many applications in computer vision (CV) have been proposed in video analytic such as face detection, human detection, mask detection, object detection, human behavior classification, and so on. Using video analytics as part of an integrated security system can assist to enhance security and safety. In video analytics, the surveillance videos are continuously analysed in real-time to alert users to things that need attention, which adding an extra layer of protection by providing security risks before they occur.

Anomaly detection is one of useful applications in CV to detect unusual events in surveillance video due to the increasing demand in broad domains such as security and risk management. Using anomaly detection algorithm in video could help to send authorities to the incident quickly enough to disrupt ongoing events, potentially saving

lives or allowing fleeing criminals to be apprehended quickly. The term "anomaly" refers to an event that is different from normality. Real-world anomaly events are complicated because of the diversity of possible events. In most surveillance system centers, monitoring video feeds from hundred CCTV cameras are fairly manual activity using human labor. It requires operations center engineers to determine if an on-going situation is an anomaly, which makes this activity error-prone and affected to delays. In additions, anomaly events generally occur with a low probability, making the effort for people to track videos wasted.

To overcome this problem, developing intelligent computer vision algorithms to automatically detect events in a video scene is a viable solution. 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. 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. Todays anomaly detection approaches have reached the accuracy level under certain condition. However, 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. It should be noted that anomaly events in surveillance video is diverse, it is therefore impossible to learn a model of all that normal and anomaly events. This dissertation defines anomaly events as those events rarely occur in the entire video. Video datasets using in this dissertation contain videos of both normal and anomaly events, whereas the training set contains only the normal event.

In presents, existing anomaly detection approaches learn to model the normal pattern or normal feature representation from training videos and classify events as abnormal if they differ from the model. The idea of feature reconstruction from normal training videos is a commonly used strategy. Further, almost of anomaly detection approaches can be classified into two categories: handcrafted features [5–7] and deep learning-based methods [8–13]. Traditional methods based on handcrafted features focused on understanding, characterizing, and improving features that can be extracted from images. These handcrafted features usually extract texture, color, or edge information. However, the complex or crowded scene in the surveillance camera may contain various objects with occlusions that are difficult to deal with. Recently, video anomaly

detections based on deep learning have attracted significant attention from the research community and are able to improve the accuracy of the detection and localization in complex scenarios by taking advantage of the learnable model of nonlinear transformation robustness.

Even though the performance of video anomaly detection methods using deep learning algorithm is improved, it should satisfy the requirements of real-world scenarios detection in terms of high detection accuracy and low processing time. The system should detect an anomaly immediately to warn the surveillance system officer and the accurate detecion system can be crucial. Since it is impossible to predict all types of anomalies in advance that may occur in real-world applications, the system should be able to work without data labeling. Therefore, this dissertation presents frameworks to address these criteria. The proposed frameworks based on unsupervised machine learning algorithm, which is able to detect anomaly in videos without data labeling. Firstly, the framework focuses on reducing the network parameters and the processing time, while attain the detection accuracy by employing the Shortcut Inception Modules (SIM), residual skip connection and asymmetric convolution layers. Secondly, the two encoders structure is proposed to improve the detection accuracy. The RGB image sequence and the optical flow are used as inputs of the network, which can be trained to capture both appearance and motion information.

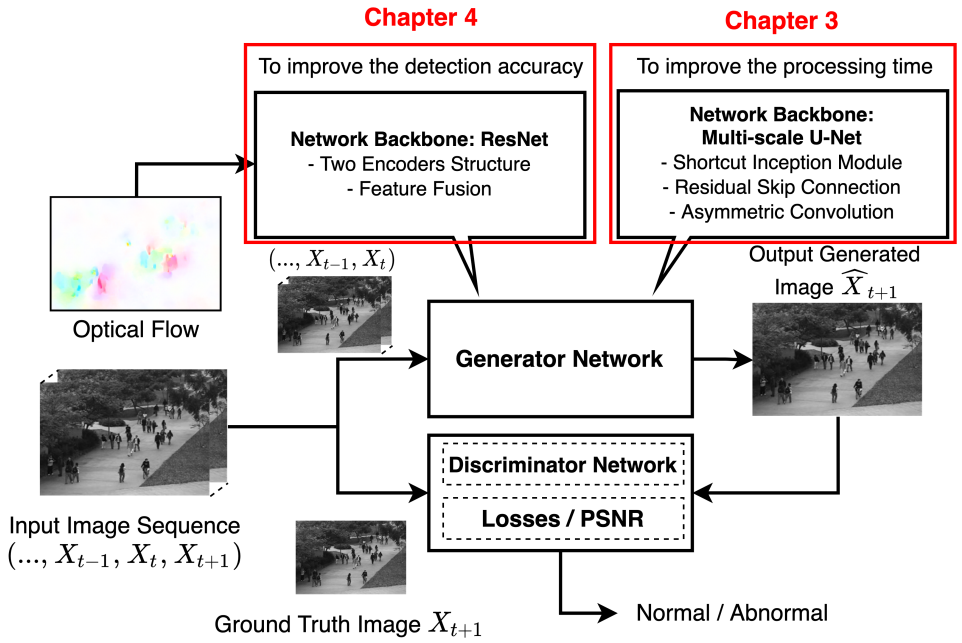


Figure 1.2: Overview of this dissertation.

1.2 Research Objectives and Contributions

This section presents the objectives and contributions of this research. The goal of this research is to create an anomaly detection framework for surveillance video using unsupervised machine learning, satisfying the requirements on detection accuracy and the processing time. Figure 1.2 illustrates the overview of this dissertation. This research improves the performance of video anomaly detecion by focusing on the generator network of the genrative adversarial network (GAN) in video anomaly detection framework. Objectives of this dissertation are as follows:

- To develop a framework based on unsupervised machine learning that can automatically detect the anomaly in surveillance videos.
- To reduce the parameters in the neural network, in order to reduce the processing time of video anomaly detection framework.
- To improve the detection accuracy by investigating the joint representation learning structure for video anomaly detection.
- To evaluate the performance of video anomaly detection approaches in terms of the processing time and the detection accuracy.

Contributions of this dissertation are described as follows.

Firstly, this dissertation proposes a video anomaly detection framework using a GAN structure. Chapter 3 proposes a framework that focuses on the design of the network architecture in order to improve the trade-off between detection accuracy and processing time. Reducing learning parameters of the network can lead to reduce the processing time, making the video anomaly detection meet the real-time requirements. In this research, a "multi-scale U-Net" network architecture has been employed which acts as a generator network. Shortcut Inception Modules (SIMs) and residual skip connection are employed to increase the ability of the training and testing of the neural network. In the stage of feature extractions, spatial and temporal features are extracted through the multi-scale U-Net architecture. Further, PatchGAN [14] has been utilized as a discriminator network to distinguish the ground truth image and the output image from the generator. This researches also uses the optical flow in training to optimize the training parameters. Instead of using traditional convolutional neuarl network (CNN) layers, the proposed framework 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. This research evaluates the performance of the framework with three benchmarks datasets including UCSD Pedestrian, CUHK Avenue, and ShanghaiTech datasets. Experiments on the benchmark datasets show the effectiveness of the proposed framework for video anomaly detection. In additions, the results show that the proposed module acheive better performance in term of detection accuracy compared to the traditional U-Net architecture.

Secondly, some of the anomaly detection approaches use only appearance information as the input, which lacks of motion information. Since anomalies in real-world scenarios are diverse, using only appearance information could not be enough to capture the anomaly object in video scenes. Therefore, this dissertation designs a new scheme of convolutional neural network architecture for video anomaly detection. Chapter 4 aims to improves the detection accuracy of video anomaly detection by proposing the two encoders network structure. The first encoder uses the RGB image sequence as an input to extract the appearance feature and the second encoder used the optical flow as an input to extract the motion feature. The proposed network architecture combines a representation of the appearance and the motion features. Two encoders in the proposed

network are based on ResNet architecture [15] as its capability of capturing appearance features. The RGB image sequence and the optical flow are fed to different encoders in order to improve the feature extraction of the object appearance and the motion. Finally, features from two encoders are fused to combine useful features obtained by encoding of two inputs. The network contains 18 layers, where the first 13 layers are encoder layers consisting of convolutional layers and stacking of residual blocks followed by the fusion layer, and the remaining is the decoder layers. This research uses a filter size of 7x7 in the first layer to extract the feature on a large object size in the image. The experimental results of the proposed network architecture demonstrate the remarkable detection accuracy with respect to existing state-of-the-art methods.

1.3 Overview of the Dissertation

This dissertation is organized into five chapters including this chapter. Contents of the remaining chapters are described as follows.

Chapter 2 provides some background and related works of the anomaly detection in a video. This chapter also provides the knowledge required for understanding the algorithms used in this dissertation such as U-Net and ResNet architecture. Traditional video anomaly detection and learning-based methods are described in this chapter, including histogram of oriented gradient (HOG), histogram of optical flow (HOF), autoencoder (AE), recurrent neural network (RNN), long short term memory (LSTM), generative adversarial network (GAN). The details of benchmark datasets are described in the remaining of this chapter.

Chapter 3 presents the proposed video anomaly detection framework using deep generative network. A multi-scale U-Net containing SIMs, residual skip connection and an asymmetric convolution layer are explained in details. The experimental setup and results of the proposed framework are described and discussed in this chapter.

Chapter 4 explains the details of the proposed network architecture of a joint representation learning for anomaly detection in surveillance videos. A feature fusion layer and a two-encoders structure based on ResNet backbone are described in this chapter.

Chapter 5 summarizes the main results and directions of the future work based on the research provided in this dissertation.

Chapter 2

Background and Literature Review

In this chapter, the background knowledge, concepts and methods that used for the main contributions are provided. This chapter also provided literature reviews of video anomaly detection. Machine learning algorithms based on unsupervised learning are explained in Section 2.1 such as autoencoder (AE), recurrent neural network (RNN), long short term memory (LSTM) and generative adversarial network (GAN). Network architectures using in this dissertation including U-Net and ResNet are explained in detail in Section 2.2. Next, Section 2.3 explain existing video anomaly detection algorithms including traditional handcrafted features and learning-based methods. Finally, the details of benchmark datasets are described in the remaining of this chapter.

2.1 Machine Learning Algorithms

Machine learning (ML) is a subfield of AI in which computers learn from data to improve their performance on a certain task without being explicitly programmed. Over the last few years, ML algorithms are widely used in many research fields and have demonstrated effectiveness in addressing challenging problems in vision, speech, natural language processing and understanding, and the other areas with a lot of input data. There are two main major branches of ML algorithms: supervised learning and unsupervised learning. In supervised learning, labels of the input data are given and it can be used to improve the model performance, whereas labels are not available in unsupervised learning. Since anomalies are rarely occurred in the video scenes, making it

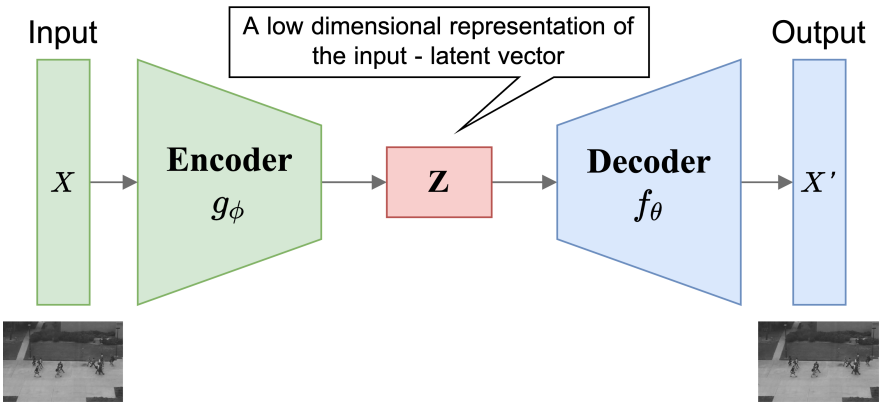


Figure 2.1: A strucutre of Autoencoder.

difficult to label any possible anomaly events. Therefore, this dissertation focuses on solving the problems using unsupervised learning where labels are not available in the training dataset. In the following subsections, this chapter present unsupervised learning algorithms using in video anomaly detections including AE, RNN, LSTM, and GAN.

2.1.1 Autoencoder (AE)

In the case of supervised learning problems, the input layer represents the features that fed into the neural network, and the label assigned to each observation is represented by the output layer. During the training, the neural network determines which parameters help minimize the error between the input and its predicted output for each observation and the true label. In unsupervised learning, the neural network uses multiple hidden layers to learn representations of the input layer, but the label is not given in the training. Autoencoder (AE) [16] is one of unsupervised machine learning algorithm, which comprises two parts, an encoder and a decoder, as illustrated in Figure 2.1. The encoder g_ϕ captures the context of the input X by extracting the feature to a vector Z, called latent vector or representation learning, and the decoder f_θ converts this newly learned representation and output the reconstructed image X' to be consistent to its input.

The aims of AE is similar to the concept of dimensionality reduction [17–19]. AE forced to learn the most useful properties of the input data, capturing the underlying structure or pattern of the data, which chooses important information to capture and

discards irrelavant or less important information. Specifically, the latent vector Z is extracted by the encoder function $g(\cdot)$ parameterized by ϕ and the output is reconstructed by a decoder function $f(\cdot)$ parameterized by θ. The parameters (ϕ, θ) are learned together to reconstruct the output. Mean Square Error (MSE) can be used to as the quality metric indicating the difference between input and output. MSE loss of AE ($L_{AE}(\phi, \theta)$) is calculated as follows.

$$L_{AE}(\phi, \theta) = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - f_\theta(g_\phi(X^{(i)})))^2, \tag{2.1}$$

where n is the total number of data elements and $X^{(i)}$ is the input data at the element i.

Variational AE (VAE) [20] is an alternative AE in which the encoder outputs a distribution instead of a fixed vector in AE. A distribution consists of two vectors, a vector of means μ and a vector of standard diviations σ. These two vectors form random variables. The VAE is able to sample across a continous space based on what it has learned from the input data [21].

2.1.2 Recurrent Neural Network (RNN)

RNN is a type of neural network that is suited for sequential data. The data can be sentences (sequence of words) or videos (sequence of images). In a traditional neural network, all inputs (and outputs) are independent of each other, where RNN qualifies a recurrent data at each time step generally related to the previous one. RNN uses the "memory," also called "state," to capture the information in the past and used it to predict the future. In the video task, the RNN uses a convolution neural network (CNN) to generate feature representations from the input image and passes it to the state. The state starts as a zero matrix and it is upated with each frame of the video. At the end of process, the final state is used to generate the output of the RNN.

The main component of RNN is the RNN cell, whose inputs are pass through these cells. Given a video consists of N frames, the structure of the RNN is depicted in Figure 2.2. Firstly, the cell combines the current state (h_0) with the current frame $(frame_1)$ to generate a new state (h_1). Then, the same process is applied to the next frames. At the end of this process, we end up with the final state (h_n). To generate a new state, the following equation is calculated.

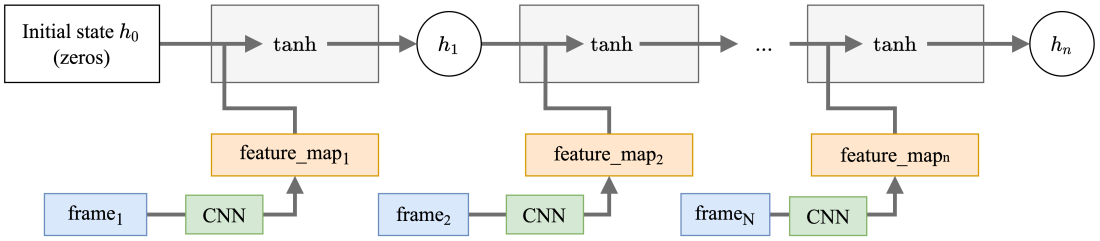


Figure 2.2: A basic strucutre of RNN cell for a video.

$$h_t = \tanh(W_{rec}h_{t-1} + W_{input}x_t + b) \tag{2.2}$$

Variables in Eq. 2.2 are denoted as follows.

- b is the bias.
- W_{rec} is the recurrent weight matrix, W_{input} is the weight matrix of the input.
- x_t is the input video at frame t.
- h_{t-1} is the current state, and h_t is the new state.

2.1.3 Long Short Term Memory (LSTM)

LSTM [22] is one of ML algorithms that are applied with a sequence data. It is a variant of the basic RNN. LSTM presented a new idea by incorporating the recurrent gate called "forget gate," which prevents backpropagated errors from vanishing or exploding occurred in RNN. In other words, LSTM worked well with long sequences data and can be stacked together to capture higher level information. The structure of LSTM is illustrated in Figure 2.3 and summarized with these equations as follows.

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{2.3}$$

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{2.4}$$

$$\hat{C}_t = \tanh(W_C \times [h_{t-1}, x_t] + b_C) \tag{2.5}$$

$$C_t = f_t \times C_{t-1} + i_t \times \hat{C}_t) \tag{2.6}$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{2.7}$$

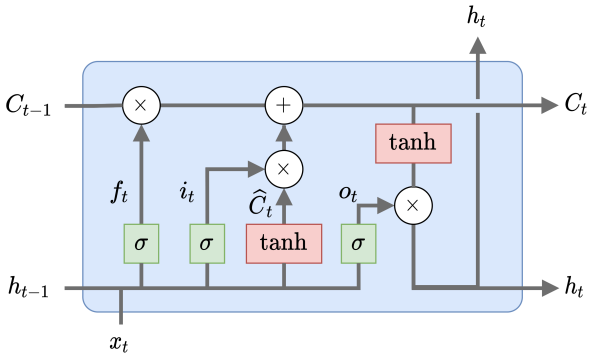


Figure 2.3: The structure of typical LSTM unit.

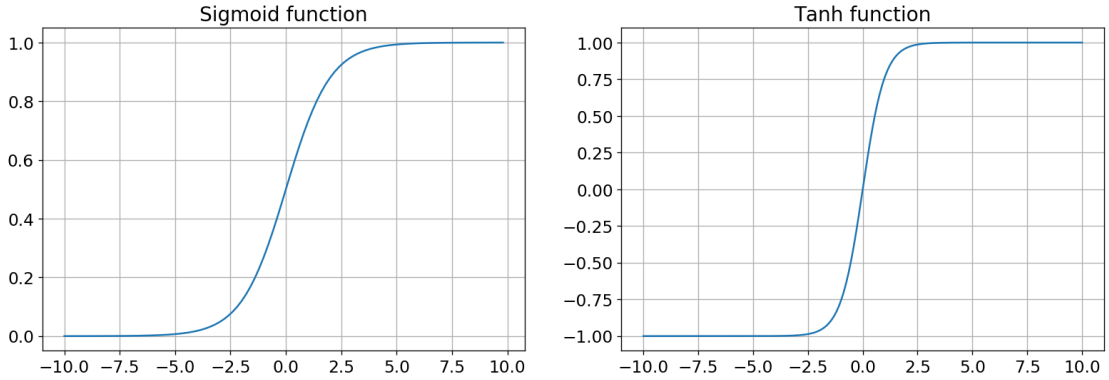


Figure 2.4: Graph of Sigmoid and Tanh functions.

$$h_t = o_t \times \tanh(C_t) \tag{2.8}$$

Let $(W_f, b_f), (W_i, b_i), (W_C, b_C)$ and (W_o, b_o) are weight matrixs and biases of the forget gate f_t, the input gate i_t, the cell state gate C_t and the output gate o_t, respectively. Equation 2.3 represents forget gate layer f_t at time t, which decides what information to throw a way from cell state using sigmoid function $\sigma(\cdot)$. The sigmoid function outputs values between 0 and 1, as demonstrated in Figure 2.4. Equations 2.4 and 2.5 are used to determine new information is to be added to. Equation 2.6 combines old and new information. The hidden state h_t is computed based on the cell state C_t and the output gate o_t using tanh (tanh outputs value between -1 to 1, as shown in Figure 2.4), which determines how much the cell state affects the output.

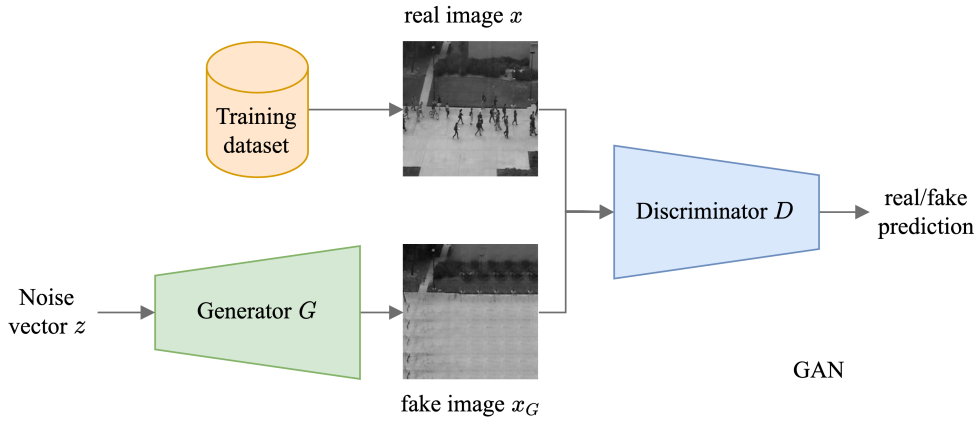


Figure 2.5: GAN representation.

2.1.4 Generative Adversarial Network (GAN)

GAN [23] is a generative model in neural network that uses an adversarial scheme, which can be trained in an unsupervised manner. Typically, GANs consist of two networks, the first network is a generator and another network is a discriminator. As shown in Figure 2.5, given an input noise vector z, the generator G learns to generate a synthesized data (fake image x_G), that becomes negative training examples for the discriminator D. The discriminator D learns to distinguish the output from the generator network whether it is a fake or real image x. During the training, the generator network uses the gradients to update its parameters and improve its ability to synthesize data. Therefore, at each iteration, the discriminator D (parameterized by P_D) tries to maximize the value function $V(G,D)$, where the generator G (parameterized by P_G) tries to minimize it, as shown in Equation 2.9.

$$\min_G \max_D V(G,D) = \min_G \max_D \mathbb{E}_x[\log D(x,P_D)] + \mathbb{E}_z[1 - \log D(G(z,P_G),P_D)]. \quad (2.9)$$

We assume that the label of real image is 1 and the fake image is 0. The first term of $V(G,D)$ represents the averaged log probability estimated by D that x is real. The second term represents the averaged log probability estimated by D that the output of the G is fake.

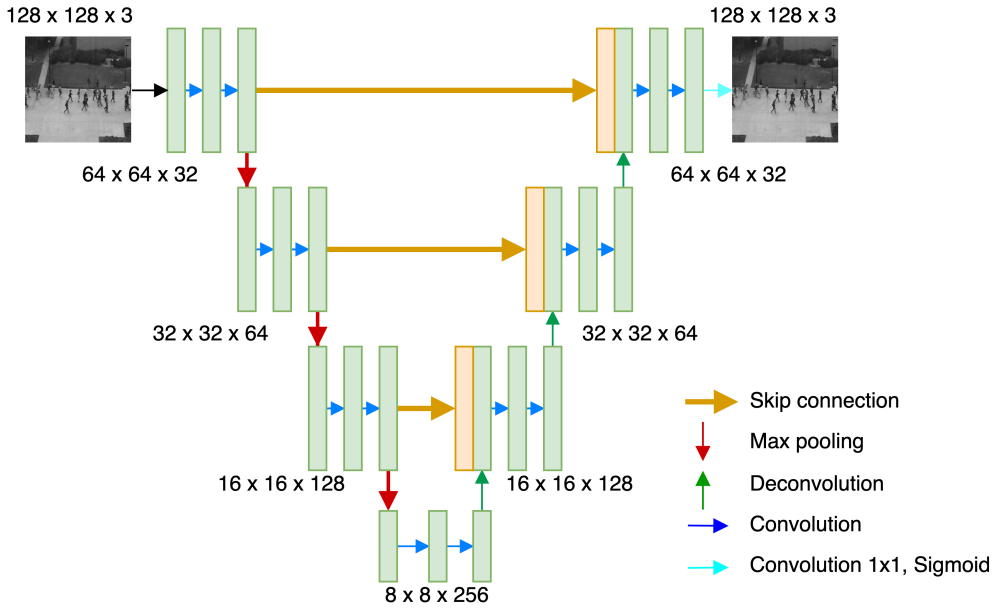


Figure 2.6: Example of U-Net architecture.

2.2 Network Architectures

Neural networks are complex structures made of artificial neurons that can take in multiple inputs to produce a single output. The advantage of neural network is that the network has a great ability to learn different levels of representation from both inputs and feature maps. Currently, there are many neural networks proposed in AI community to solve problems in classification, detection and other CV problems, in order to improve the performance and reducing the computation time. U-Net [24] and ResNet [15] are neural networks that are commonly used in CV researches, especially in video anomaly detection. This chapter presents the details of these two network architectures in the following subsections.

2.2.1 U-Net Architecture

U-Net architecture [24] is widely used in object segmentation tasks. The architecture was originally proposed for biomedical image segmentation. The network structure of U-Net is illustrated in Figure 2.6. U-Net consists of two parts: the downsampling part, where input images are compressed spatially but expanded the channel-wise, and

upsampling part, where representations are expanded spatially while the number of channls is reduced. U-Net architecture looks similar to the AE structure where the downsampling is the encoder and the downsampling act as the decoder. Unlike AE, U-Net proposed skip connections that are employed between the upsampling and downsampling parts of the network. Using skip connection allow information to shortcut parts of the network and flow through to later layers [25].

2.2.2 ResNet Architecture

One of the most powerful network architecture in ML called a residual network or ResNet [15]. Designing a deep neural network architecture by staking more layers with gradient-based learning methods and backpropagation may encounter the gradients of the loss function leads to zero, making the network hard to train. This problems is called a vanishing gradient problem [15,26]. Therefore, ResNet provides an efficient approach to create a deep neural network, which builds the stack of residual blocks on top of each other, where each block contains a skip connection that sums the input and output of the block, before passing to the next layer. Figure 2.7 shows a single residual block of ResNet architecture. The residual block can be defined with Equation 2.10 as follows.

$$y = F(x, \{W_i\}) + x, \tag{2.10}$$

where x and y are the input and output of the layer. The function $F(x, \{W_i\})$ is the residual mapping to be learned. The form of F is flexible, weight layers can increase more than two layers and it can be fully connected layers or convolutional layers. ResNet network uses a 34-layer plain network architecture inspired by VGG-19 [27], in which the shortcut connection is added. As a result, ResNet improves the efficiency of deep neural networks with more neural layers while minimizing errors of the network.

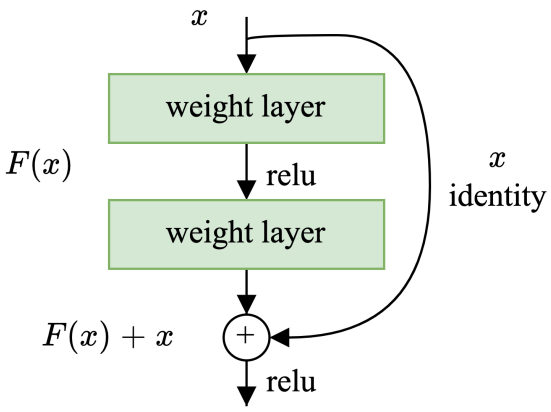


Figure 2.7: Residual block in ResNet architecture [15].

2.3 Anomaly Detection in Video

Anomaly detection approaches can be classified into two main categories including a hand-crafted feature and a learning-based method. Traditional methods based on hand-crafted features [5–7, 28–31] focused on understanding, characterizing, and improving features that can be extracted from images. These handcrafted features usually extract texture, color, or edge information. The most widely used handcrafted features for anomaly detection in the video are the histogram of oriented gradients (HOG) [32], and histogram of optical flow (HOF) [33, 34]. Lu et al. [6] proposed method to detect abnormal by using sparse combination learning. This proposed method directly finds the most suitable combination in the dictionary by evaluating the least square error to represent the input whether normal or abnormal events. Shangdong Wu et al. [35] proposed anomaly detection approach by utilizing low-level trajectory features to describe the abnormal event. Frequency-based analysis is performed in [36] to automatically discover rules of normal events, where anomalies could be classified from events that deviate from these rules. Gaussian mixture models (GMMs) also used in [37] to model the spatial scene structure from training trajectories. Zhao et al. [32] proposed a fully unsupervised dynamic sparse coding approach based on an online sparse reconstruction of query signals from learned event dictionaries. Yang et al. [5] proposed a dictionary selection model called the sparse reconstruction cost. Anomaly is determined by sparse reconstruction cost, through a weighted linear reconstruction. However, hand-crafted

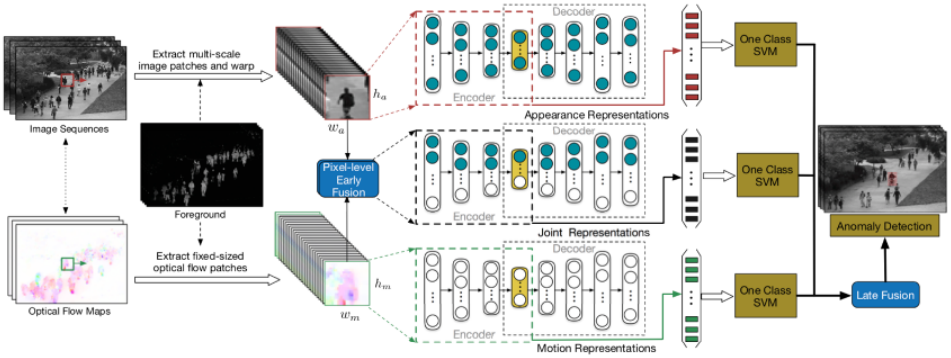


Figure 2.8: Appearance and Motion DeepNet (AMDN) framework for video anomaly detection proposed by [10].

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, especially in the presence of occlusion, noise, camera motion, and changes in illumination [38].

Recently, deep learning techniques have been attempted to overcome the limitations of the handcrafted feature by automatically extracting the high-level feature representations from the image. 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 [12, 13, 39–45]. Learning-based methods also applied in relative to video anomaly detection such as action recognitions [46–50]. 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 et al. [10] presented a novel Appearance and Motion DeepNet (AMDN) based on unsupervised deep learning for anomalous event detection, as illustrated in Figure 2.8. This method learns discriminative feature representation of both appearance and motion patterns, early and late fusion scheme are performed to combine appearance and motion features for discovering unusual activities as well as a joint representation are presented to capture the correlation between both features. Finally, to detect anomalous events, one-class SVMs has been proposed followed by late fusion in the last layer of the network.

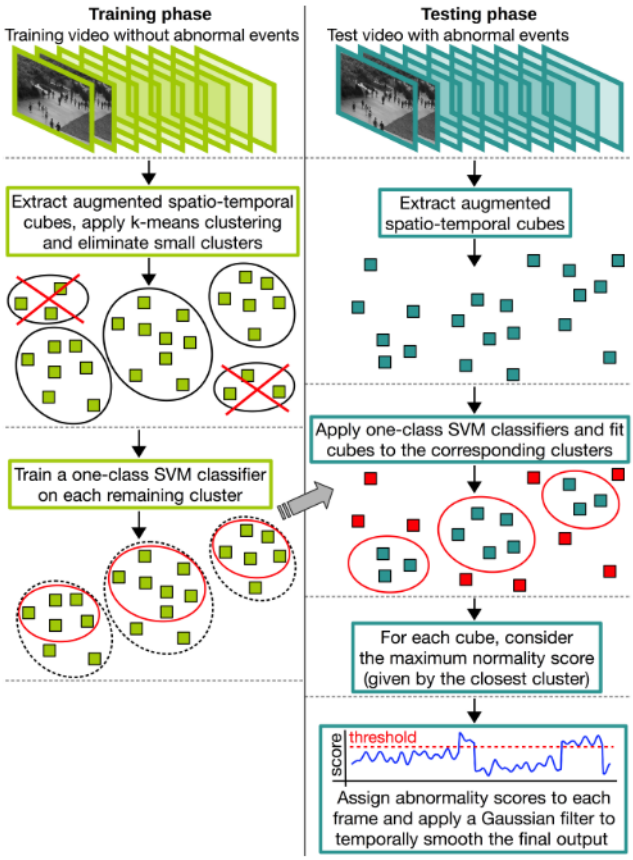


Figure 2.9: Narrowed normality clusters for video anomaly detection proposed by [11].

Ionescu et al. [11] presented abnormal event detection based on a two-stage outlier elimination algorithm. The method overview is illustrated in Figure 2.9. This work used appearance features extracted from the last convolutional layer of a pre-trained neural network to augment spatio-temporal cubes. The algorithm eliminates the outlier using k-mean clustering and classifies by training a one-class SVM. Each spatio-temporal cube is evaluated against each one-class SVM model to get a set of normality scores during the testing process. They obtain an anomaly prediction map for each frame by piecing together the cubes from all frames.

Pang et al. [51] applied self-trained deep ordinal regression to detect video anomaly events. The pretrained ResNet-50 [15] is used as a feature extractor. Self-training is a semi-supervised machine learning [52] that has an ability to cope with a small dataset in training. It creates more consistent labeled data in training a model with a small la-

beled dataset and then applying the trained model to unlabeled data. ResNet-50 [15] also employed in [41] framework called AnomalyNet. AnomalyNet consists of three blocks: the motion fusion block, the feature transfer block, and the coding block. The motion fusion block compresses video clips into a single image while removing unwanted background elements. Then compressed images are fed to the feature transfer block. Finally, features are extracted based on a transferable model.

Sultani et al. [53] 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 in order to consider as anomaly event. To avoid labor-intensive temporal annotations of anomalous segments in training videos, the model is learned using deep multiple instance ranking framework with weakly labeled data. This work also introduced a dataset of real-world surveillance videos with 13 realistic anomalies.

Fan et al. [43] proposed video anomaly detection method based on gaussian mixture variational autoencoder. The overview of this proposed method is shown in Figure 2.10. The assumption is that the normal samples can be associated with at least one Gaussian component of a GMM, while anomalies do not belong to any Gaussian component. The AE contains the encoder-decoder structure corresponding to feature extraction and model construction which permits to learn a mapping from high dimensional data to a low dimensional latent presentation while ensuring reconstruction accuracy. This method consists of two-stream framework, i.e. spatial and temporal streams. The spatial stream operates on RGB frames and captures the appearance anomalies while the temporal stream is generated to capture the long-term temporal information of motion anomalies. Finally, a sample based method is used to detect anomalies based on the joint probabilities of all the components in the GMM.

Luo et al. [54] utilizes convolutional LSTM based on AE (ConvLSTM-AE) to memorize all past frames in order to model both appearance and motion information. Deconvolution network was integrated to reconstruct past frames, and using reconstruction error to identify whether an anomaly event occurs. Another work proposed in [44] iteratively updates sparse coefficients via a stacked RNN to detect anomalies in videos. Chong et al. [8] proposed an end-to-end architecture for learning video representation, which included two main components, one for the spatial component and the other for

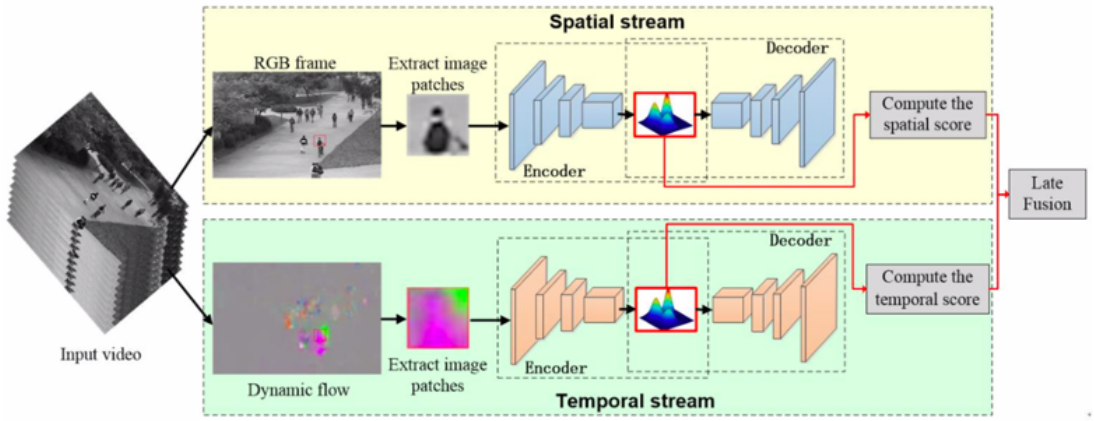


Figure 2.10: Gaussian mixture variational autoencoder for video anomaly detection [43].

the temporal component. The network architecture of [8] is shown in Figure 2.11. The structure of this network architecture is based on AE that aims to reconstruct the input image. The network of comprises LSTM layers in between encoders and decoders to learn the sequence spatial features from encoders. There is another approach that augment the AE algorithm. Gong at al. [55] proposed the improvement of AE with a memory module called memory-augmented autoencoder (MemAE). The proposed memory module consists of a memory to record the prototype encoded patterns of normal events from extracted features. The reconstruction error is also used as the anomaly detect criterion. The extended version of AE called variational AE (VAE) [20] is employed in video anomaly detection called Conv-VRNN [39], which considers temporal information in future frame prediction. An overview of Conv-VRNN is shown in Figure 2.12. Unlike the encoder in AE, which produces a vector called the latent variable, VAE produces a distribution on the latent variable that characterizes the pattern of normal events. This work also combines the ConvLSTM with VAE to better capture the temporal relationship among frames in a video.

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. Liu, et al. proposed a framework for anomaly detection based on GAN [12]. The framework overview is demonstrated in Figure 2.13. To generate a more realistic future

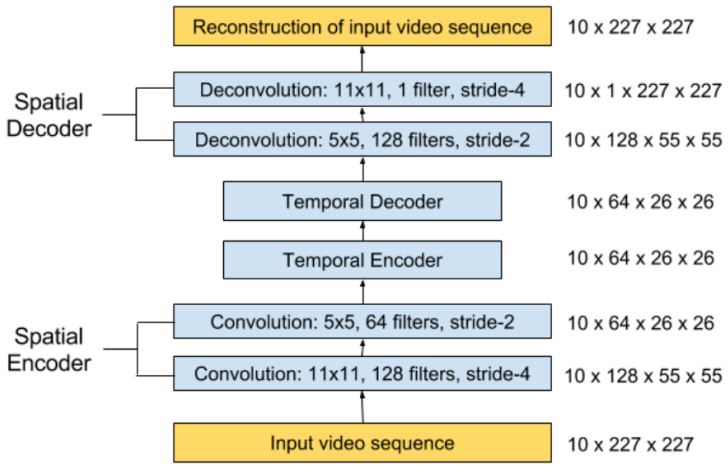


Figure 2.11: Convolution LSTM based on AE [8].

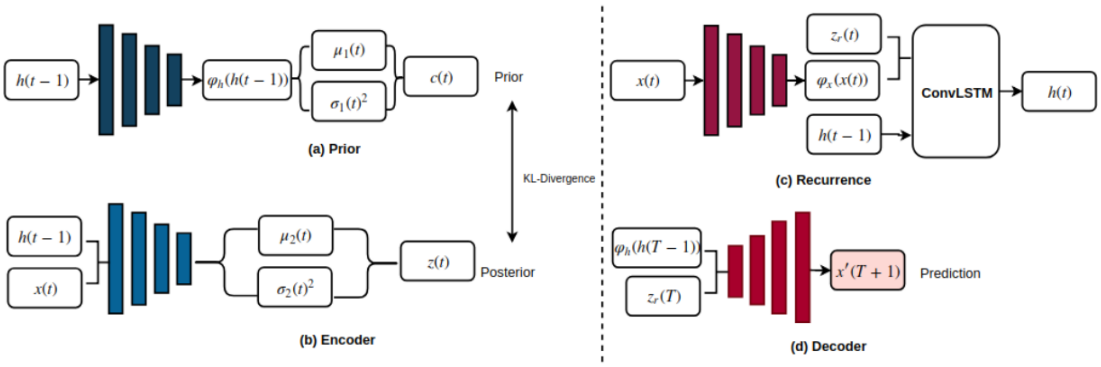


Figure 2.12: Variational Autoencoder (VAE) for video anomaly detection [39].

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 [24] 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. To improve the efficiency of training GAN for anomaly detection, Dong et al. [40] proposed a semi-supervised approach with a dual discriminator-based GAN, which considers more motion information in video clips compared with [12] work. However, some features are lost in these generator networks during training due to convolution operators of each layer. Network

architectures in these approaches are required extensive search for varying the depth of the network in order to balance the trade-off between the detection accuracy and the processing time. In addition, real-world video scenes are diverse, making anomaly detection more challenging. Motion information could be considered as input in the anomaly detecion method to capture the object characteristic in videos.

To resolve this issue, this dissertation presents a framework that consists of a multi-scale U-Net as a generator network. This research proposes the Shortcut Inception Module (SIM) and residual skip connection to make a network learn higher-level features of images. The proposed framework also employs the idea of asymmetric convolution instead of traditional convolution layer to reduce the number of parameters. Furthermore, this dissertation proposes a joint representation learning to improve the detection accuracy of video anomaly detection, the network architecture contains two separate encoders to first extract appearance and motion features. The feature maps from two encoders are then fused and fed to the same decoder. In practical, due to the ambiguous nature, rare occurrences, large variance within anomaly events, and the data imbalance problem of the video anomalies, it is almost impossible to accurately define the video anomalies with the associated labels in most cases. Hence, the recent trend demonstrates that unsupervised and semi-supervised training processes are far more effective than supervised video anomaly detection methods. This dissertation is based on fully unsupervised learning manner, which learns to detect anomaly in videos without data labeling. The details of two proposed frameworks are described in Chapter 3 and Chapter 4, respectively.

The summary of related works are shown in Table 2.1

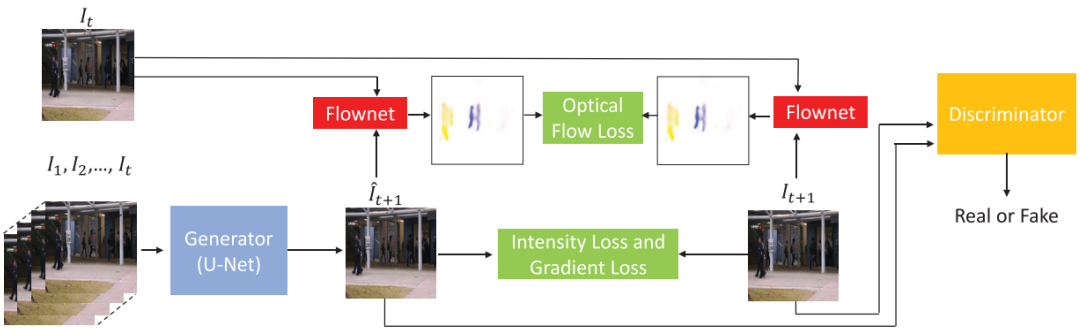


Figure 2.13: Future frame prediction for video anomaly detection based on GAN structure [12].

Table 2.1: Summary of related works for video anomaly detection based on machine learning algorithms.

Authors	Years	Network backbones	Learning algorithms	Datasets
Xu et al. [10]	2015	*	AE	UCSD, Train
Luo et al. [54]	2017	*	ConvLSTM-AE	Avenue, UCSD, Subway, Exit
Luo et al. [44]	2017	*	Stacked RNN	Avenue, UCSD, Subway, ShanghaiTech
Chong et al. [8]	2017	*	ConvLSTM-AE	Avenue, UCSD, Subway, Exit
Sultani et al. [53]	2018	Pretrained C3D	Multiple instance learning (MIL)	private dataset
Liu et al. [12]	2018	U-Net	GAN	UCSD, Avenue, ShanghaiTech
Lu et al. [39]	2019	VGG	VAE-LSTM	UCSD, Avenue
Gong at al. [55]	2019	*	AE	UCSD, Avenue, ShanghaiTech
Ionescu et al. [11]	2019	-	K-means, SVM	Avenue, Subway, UMN
Pang et al. [51]	2020	ResNet-50	Self-trained deep ordinal regression	UCSD, Subway, UMN
Fan et al. [43]	2020	*	VAE	UCSD, Avenue
Dong et al. [40]	2020	U-Net	GAN	UCSD, Avenue, ShanghaiTech

* These approaches uses their own designs of the network backbone.

2.4 Benchmark Datasets for Video Anomaly Detection

This section presents the benchmark datasets used in video anomaly detection. This dissertation evaluates the performance with three benchmark datasets including UCSD Pedestrian, CUHK Avenue and ShanghaiTech. These datasets contain only normal events in training set, whereas the testing set contains both normal and anomaly events. The input image resolution of these datasets are different, in this dissetation, each input is resized to 256x256 with three color channels. The annotation of anomaly is labled within the anomaly frame in each testing videos. The details of each datasets are described in following subsections and the summary of these three benchmark datasets are presented in Table 2.2.

2.4.1 UCSD Pedestrian Dataset

The UCSD Pedestrian dataset [56] provides 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. Aanomaly events in the UCSD Pedestrian dataset includes cars, scooters, wheelchairs, bicycles and people walking across a walkway or in the grass. The camera viewpoints of Ped1 and Ped2 subsets are different. Ped1 subset contains a group of people walking towards and away from the camera, where Ped2 subset contains pedestrian movement parallel to the camera plane, and hence anomaly objects of these video scenes become more diverse. The example of training and testing frames of Ped1 and Ped2 subsets are shown in Figure 2.14 and Figure 2.15, respectively.

2.4.2 CUHK Avenue Dataset

Another widely used dataset in video anomaly detection is CUHK Avenue dataset [6]. This dataset contains 16 training and 21 testing videos with a total of 47 abnormal events such as abnormal objects, throwing, and strange actions. This dataset is captured from single outdoor surveillance camera looking at a side of the building with a group of people walking surroundings. The size of people may vary due to the camera setting and angle of this dataset. The CUHK Avenue dataset contains some challenging such as cameras shaking and ourlier in training dataset. The example of this dataset is shown

Figure 2.14: Example frames of UCSD Pedestrian 1 subset (red bounding box indicates anomaly region and using only for visualisation purposes).

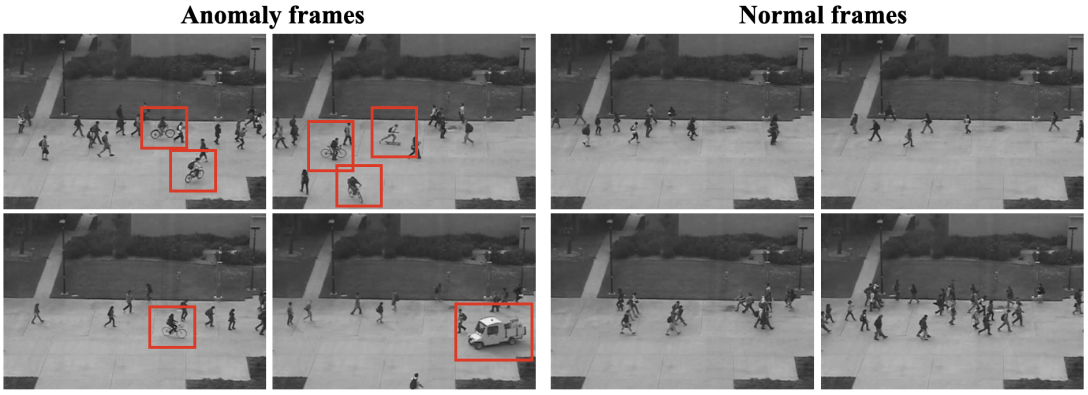


Figure 2.15: Example frames of UCSD Pedestrian 2 subset (red bounding box indicates anomaly region and using only for visualisation purposes).

in Figure 2.16.

2.4.3 ShanghaiTech Dataset

The ShanghaiTech dataset [12] is one of challenging datasets in video anomaly detection. This dataset covers challenging scenarios for video anomaly due to large variations in appearance and viewpoint. Almost all of anomaly datasets for video contain videos captured from fixed camera viewpoint, which lacks variations in camera angles and video scenes. The ShanghaiTech dataset consists of 13 scenes with complex light conditions and camera angles of 330 training and 107 testing videos. Moreover, several

Figure 2.16: Example frames of CUHK Avenue dataset (red bounding box indicates anomaly region and using only for visualisation purposes).

Figure 2.17: Example frames of ShanghaiTech dataset (red bounding box indicates anomaly region and using only for visualisation purposes).

scenes for a given scenario have substantial perspective changes, making this dataset more challenging. Example of training and testing frames of the ShanghaiTech dataset are illustrated in Figure 2.17.

Table 2.2: Summary of benchmark datasets for video anomaly detection using in this dissertation.

Datasets	Number of videos	Training/Testing videos	Example anomalies
UCSD Ped1 [56]	70	34/36	Cars, skateboards, scooters, wheelchairs, bicycles and people walking across a walkway
UCSD Ped2 [56]	28	16/12	Cars, skateboards, scooters and bicycles
CUHK Avenue [6]	37	16/21	Running, throwing and strange action
ShanghaiTech [12]	437	330/107	Cars, skateboards, scooters, bicycles motorbikes, strollers, and motion anomalies such as fighting, chasing, pushing and jumping

2.5 Summary

This chapter has provided the background knowledge of machine learning algorithms in video anomaly detection. Fundamentals of autoencoder (AE), recurrent neural network (RNN), long short term memory (LSTM) and generative adversarial network (GAN) are presented in Section 2.1. Network architectures are described in Section 2.2. Next, the literature reviews of related works in video anomaly detection are provided, where the approaches can be classified to two categories such as handcrafted features and learning-based approaches. Details of benchmark datasets are presented in Section 2.4. This dissertation used three benchmark datasets to evaluate the performance of the proposed frameworks, including UCSD Pedestrian, CUHK Avenue and ShanghaiTech datasets. The experimental results and discussions on these datasets describe in Chapter 3 and Chapter 4.

Chapter 3

Multi-scale U-Net for Video Anomaly Detection

This chapter presents the proposed framework for video anomaly detection using deep generative network structure. The proposed multi-scale U-Net, Shortcut Inception Module (SIM) and residual skip connection are described in details. Improvement of the traditional U-Net network architecture and the design idea of the shortcut inception module are also described. Experimental results are reported and evaluation in several aspects is coordinated in the remaining of this chapter.

3.1 Introduction

U-Net network architecture [24] is widely used as feature extraction method in CV researches especially in object segmentation tasks. Problems of using U-Net as a network backbone are that the network requires extensive architecture search for varying the depth of the network and its skip connection from encoder to decoder losts some informative features affecting the detection accuracy due to convolution operations. To solve these problems, this chapter proposes a framework of the convolutional neural network architecture for video anomaly detection. The proposed network uses the U-Net as the baseline architecture. This research modifies and replaces 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. This research also applies the idea of asymmetric convolution

layers so that its representation capability is equivalent to the traditional convolution layer, while it reduces the number of network parameters, requiring less computation in comparison with the traditional U-Net architecture.

The rest of this chapter is organized as follows. Section 3.2 describes multi-scale U-Net and explains the ability of the SIM and residual skip connection designs to preserve the information during training the neural network. Section 3.3 discusses the results of the evaluation, and Section 3.4 summarizes this chapter.

3.2 Multi-scale U-Net for Video Anomaly Detection

In this section, details of the proposed framework are described, as illustrated in Figure 3.1. Firstly, given an input image sequence, a multi-scale U-Net is utilized as a generator network G to extract spatial features. Figure 3.2 shows the structure of the multi-scale U-Net. This research employs 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 the architecture for the purpose of propagating spatial information which was lost during the convolutional operation from encoder to decoder. Using these residual skip connections is beneficial for learning higher feature of an image. In the training phase, the optical flow of the generated image is enforced to be close to that of the ground truth image in order to optimize the network parameters. Further, PatchGAN [14] uses as a discriminator network D into the framework to distinguish between the generated image and the ground truth image. Details of each part describes in subsections below.

3.2.1 Shortcut Inception Modules (SIM)

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 3×3 convolution layers followed by a max-pooling operation. As explained in [57], the

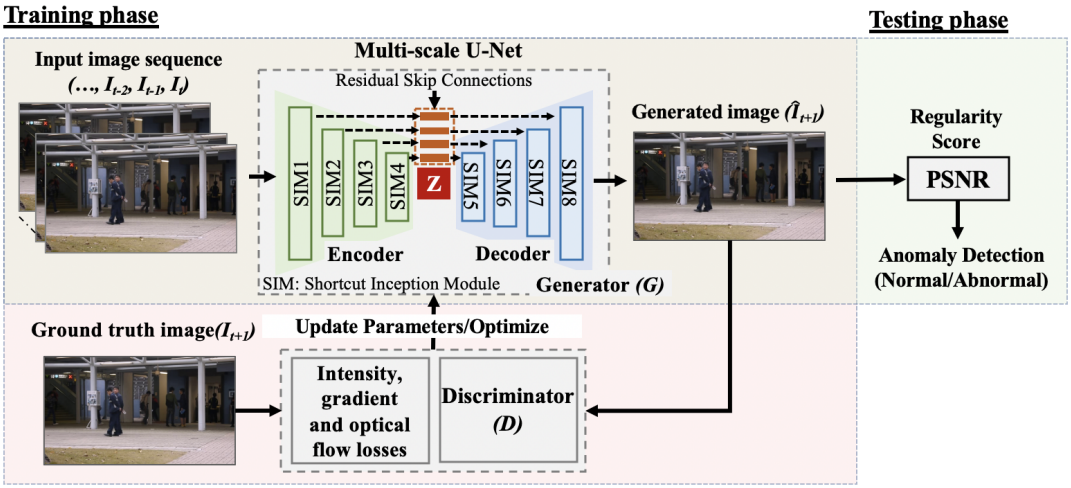


Figure 3.1: Overview of the proposed multi-scale U-Net network architecture.

sequence of two 3×3 convolutional layers actually resembles a 5×5 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 5×5 and 7×7 convolution operations in parallel to the 3×3 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 [58, 59].

This work uses the U-Net as the base network architecture for the generator network, however, original convolution layers are modified and replaced by SIM and increased the width of the network to make it possible to learn the input image in higher-level features. Figure 3.3 (c) illustrates a proposed SIM inspired by the idea of an inception module [57] (Figure 3.3 (a)). As described above, the convolutional layers can be replace with the inception blocks. Although the performance gain can be expected in introducing the larger size of convolution operations such as 5×5 and 7×7, the parallel network structure consumes high computational capacity. In the same manner as [57], this research factorizes and stacks the larger size of 5×5 and 7×7 convolution operators by a 3×3 convolution operation, the output of the last two 3×3 convolutions competent to the 5×5 and 7×7 convolutions approximately as shown in Figure 3.3 (b). This research also takes an advantage of a feature concatenation to extract the feature from different scales [58]. Then adding a shortcut connection with an additional 1×1 con-

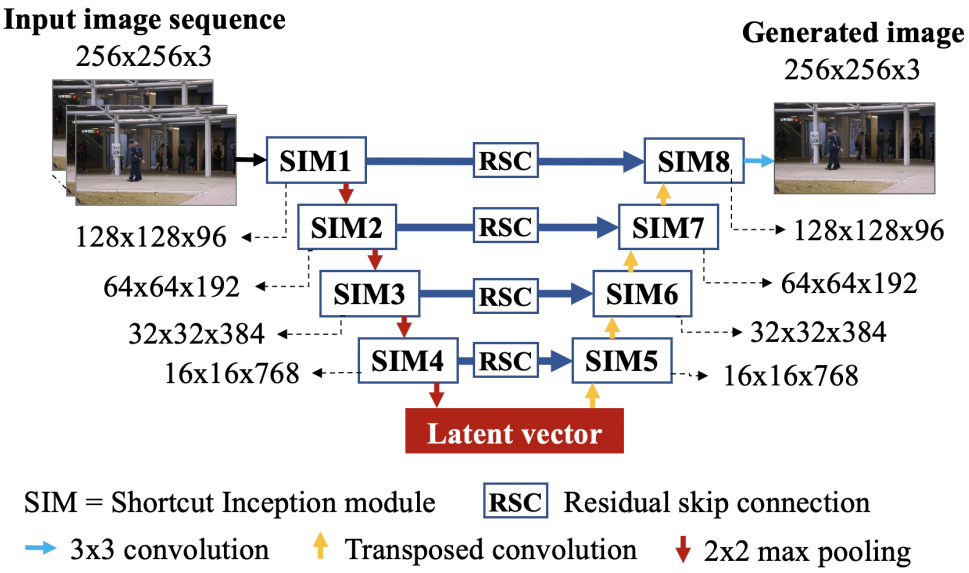


Figure 3.2: Proposed multi-scale U-Net architecture.

Table 3.1: The architecture details of proposed generator network.

Depth	SIM	Layer (filter size)	#filters	Output
1	SIM 1 & 8	AC,AC,AC,C1	16, 32, 48, 96	128×128×96
2	SIM 2 & 7	AC,AC,AC,C1	32, 64, 96, 192	64×64×192
3	SIM 3 & 6	AC,AC,AC,C1	64, 128, 192, 384	32×32×384
4	SIM 4 & 5	AC,AC,AC,C1	128, 256, 384, 768	16×16×768

AC= Asymmetric convolution layer, C1= Convolution layer of a 1x1 filter size.

volutional layer to add more non-linearity information to enhance the representation as well as reducing the network size without a performance penalty [60]. In [61], the traditional convolution layers are utilized in the inception module. Unlike [61], the 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 3×3 convolution is equivalent to a stacking of a 3×1 convolution followed by a 1×3 convolution, which reduces the size of the model and increasing the training efficiency [57, 62, 63].

Details of the proposed SIM are summarized in Table 3.1. This research assigns W_j to control the number of filters used in the proposed module in each depth j. Inside the SIM, this work assigns $\frac{W_j}{6}$, $\frac{W_j}{3}$, and $\frac{W_j}{2}$ to each of three convolution layers, respec-

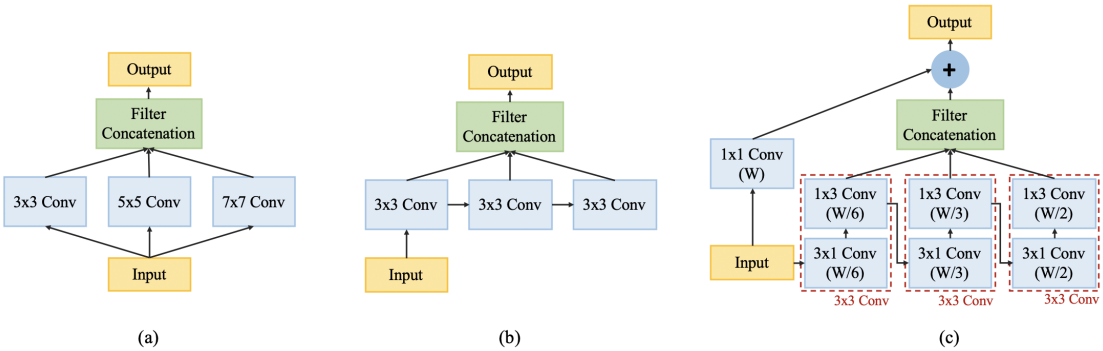


Figure 3.3: The comparison design of (a) the inception module [57] using 3×3, 5×5, and 7×7 convolutional filters in parallel. (b) Factorizing and stacking the larger size of 5×5 and 7×7 convolution operators in inception module by a 3×3 convolution operation, and (c) the proposed structure of a Shortcut Inception Module (SIM).

tively, as this combination achieved the best results in the preliminary experiment. W_j is computed as follows.

$$W_j = \alpha * N_j, \tag{3.1}$$

where N_j is the number of filters in the corresponding depth j of the multi-scale U-Net, and α 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}. This research selected $\alpha = 1.5$ as it keeps the number of parameters slightly below that of the original U-Net.

3.2.2 Residual Skip Connections

The U-Net architecture [24] 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,

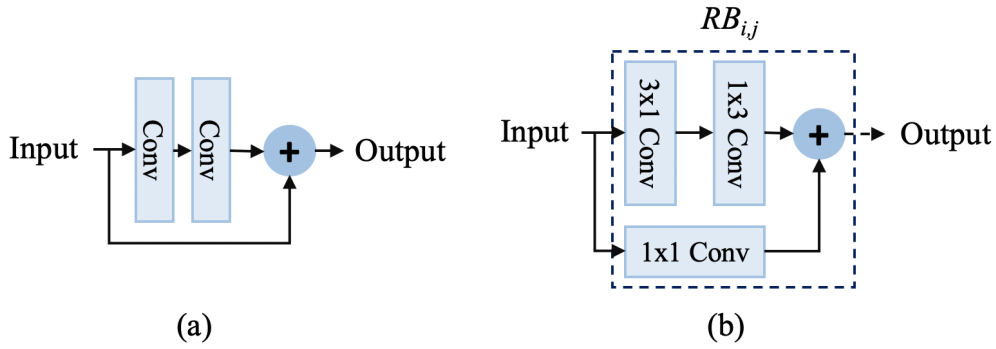


Figure 3.4: The comparison design of (a) the residual learning block [15] with (b) the proposed residual skip connection.

the fusion of these sets of features from encoder-decoder could cause feature learning which affects the reconstruction output.

Following the deep residual network [15] that proposed the idea of using residual learning block, as shown in Figure 3.4 (a), This research introduces residual skip connection blocks to the proposed generator network. As illustrated in Figure 3.4 (b), the proposed residual skip connection block consists of an asymmetric convolution layer of a 3×1 convolution layer followed by a 1×3 convolution layer and a shortcut connection of a 1×1 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, this work passes 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 [64]. 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. This work 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 the image down-sampling in every step using the max-pooling layer. Therefore, this work gradually decreases 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.

3.2.3 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_{gd} 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.

$$L_{int}(\hat{I}-I) = \|\hat{I}-I\|_2^2 \tag{3.2}$$

$$\begin{aligned} L_{gd}(\hat{I}-I) = &\||\hat{I}_{i,j}-\hat{I}_{i-1,j}|-|I_{i,j}-I_{i-1,j}|\|_1 \\ &+\||\hat{I}_{i,j}-\hat{I}_{i,j-1}|-|I_{i,j}-I_{i,j-1}|\|_1. \end{aligned} \tag{3.3}$$

The optical flow loss is applied to capture the motion information and to optimize training parameters. This work used a FlowNet [65] to estimate the optical flow. Following [12], this work applies the $L1$ distance loss to calculate the motion penalty:

$$L_{flow}(\hat{F}-F) = \|\hat{F}-F\|_2, \tag{3.4}$$

where F is the ground truth optical flow estimated from two consecutive frames 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, this research used an adversarial loss based on Generative Adversarial Network (GAN) [23] 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. This work utilized PatchGAN [14] as a discriminator network. The 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 classifies 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.

$$L_{adv}(\hat{I}) = \sum_{m,n} \frac{1}{2} L_{MSE}(D_{gen}(\hat{I})_{m,n}, 1) \tag{3.5}$$

$$L_{MSE}(\hat{I}, I) = (\hat{I} - I)^2, \tag{3.6}$$

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

$$L = \lambda_{int} L_{int} + \lambda_{gd} L_{gd} + \lambda_{adv} L_{adv} + \lambda_{flow} L_{flow}, \tag{3.7}$$

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

3.2.4 Anomaly Detection using Regularity Score

In testing, this work 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 the 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

$$PSNR(I_t, \hat{I}_t) = 10 \log_{10} \frac{[\max_{\hat{I}}]^2}{\frac{1}{N} \sum_{i=0}^{N} (p_t(i) - \hat{p}_t(i))^2}, \tag{3.8}$$

where $\max_{\hat{I}}$ represents the maximum intensity value in a generated image $\hat{I}$. $p_t(i)$ and $\hat{p}_t(i)$ are pixel intensity of index i in I and $\hat{I}$, respectively. N denotes the total number of pixels in the image. Then, the proposed framework 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:

$$R(t) = \frac{PSNR(I_t, \hat{I}_t) - \min_{PSNR}}{\max_{PSNR} - \min_{PSNR}}, \tag{3.9}$$

where the terms $\min_{PSNR}$ and $\max_{PSNR}$ are the minimum and maximum values of the PSNR in every frame of each test video. Finally, the proposed framework detects anomaly event in an image based on a threshold of regularity score $R(t)$.

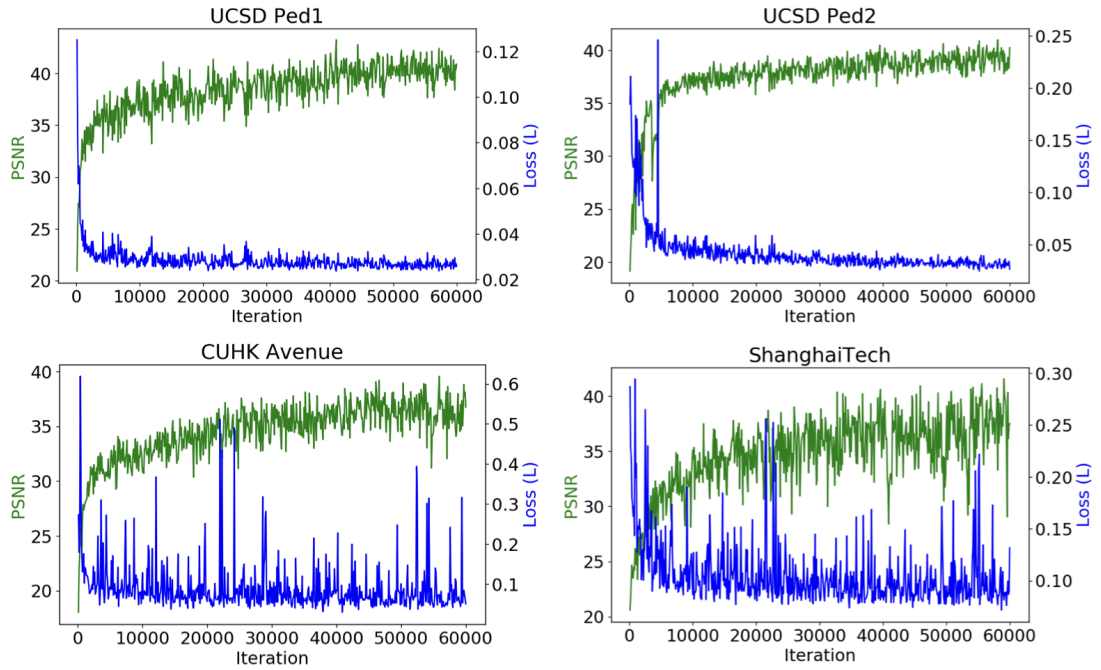


Figure 3.5: PSNRs and loss values of network trained on UCSD Pedestrian, CUHK Avenue and ShanghaiTech datasets.

3.3 Experimental Results

3.3.1 Experimental Setup

The implementation is based on Tensorflow framework [66] using Python 3.7. The network architecture was trained and tested on NVIDIA Geforce RTX 2080 GPU. Training is based on Adam optimizer [67] and the batch-size is fixed to 4. In the training and testing, input images of the network are resized to 256 x 256 of 3 color channels and normalize it to the range of [-1, 1]. To be consistent with [12], this research has set 4 consecutive images as the input image sequence. Using large models and datasets, Adam optimizer can efficiently solve practical deep learning problems [67] and it performed better for training GAN as discussed in [68]. Figure 3.5 shows learning curves trained with different datasets. The final loss of all datasets are significantly decreased in high iterations. On the other hand, PSNRs of the generated image and the input are increased, showing the effectiveness of training the proposed network architecture.

Following the framework described in [12], this work uses a receiver operating char-

acteristic (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. TPR and FPR can be calculated as Eq.3.10 and 3.11, respectively. AUC used in most previous works [8, 12, 42–44, 54] 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.

$$TPR = \frac{\#TP}{\#TP + \#FN} \tag{3.10}$$

$$FPR = \frac{\#FP}{\#FP + \#TN} \tag{3.11}$$

where $\#TP$ denotes the number of true positive samples, $\#FP$ denotes the number of false positive samples, $\#TN$ denotes the number of true negative samples, and $\#FN$ denotes the number of false negative samples.

Table 3.2: Area under curve (AUC) comparison of proposed framework and [8, 12, 42–44, 54] works on UCSD and CUHK Avenue dataset

Methods	UCSD Ped1		UCSD Ped2		CUHK Avenue	
	AUC	EER	AUC	EER	AUC	EER
Chong et al. [8]	89.9%	12.5%	87.4%	12.0%	80.3%	20.7%
Conv-AE [42]	81.0%	27.9%	90.0%	21.7%	80.0%	25.1%
ConvLSTM-AE [54]	75.7%	-	88.1%	-	77.0%	-
StackRNN [44]	-	-	92.2%	-	81.7%	-
Fan et al. [43]	**94.9%**	**11.3%**	92.2%	12.6%	83.4%	22.7%
Liu et al. [12]	83.1%	23.5%	95.4%	12.0%	84.9%	20.6%
The Proposed	85.3%	23.6%	**95.7%**	12.0%	**86.9%**	**20.2%**

3.3.2 Results on the UCSD and the CUHK Avenue Dataset

This research has compared the proposed with several anomaly detection methods based on deep learning, including AE [8, 42, 43], LSTM [44, 54], and GAN [12] on UCSD Pedestrian dataset. As summarized in Table 3.2, the proposed framework surpasses the results reported in the previous works on UCSD Ped2 in terms of detection accuracy. The error rate is comparable to most recent works [8, 12, 42–44, 54]. The performance in UCSD Ped1 is 85.3% whereas the best result of 94.9% is achieved by Fan et al. [43], who employed a two-stream network that combines the appearance and motion of anomalies which may incur considerable computational cost. Another work that achieved better AUC is Chong et al. [8] that utilized the stacked convolution-deconvolution layer with ConvLSTM. Moreover, this work has compared the proposed framework with the baseline framework that utilized the U-Net architecture [12], the proposed framework outperformed their performances on both UCSD Ped1 and Ped2 subset with an AUC of 85.3% and 95.7%, respectively.

The comparative results of the CUHK Avenue dataset are also presented in Table 3.2. The proposed framework further surpass their existing learning-based methods, reaching the best AUC of 86.9% and EER of 20.2%, indicating that the proposed framework can precisely detect and reduce the false alarm rate of the anomaly detection in the CUHK Avenue dataset. The AUC is 2.0% above the baseline method reported in Liu et al. [12], and more than around 7% higher than other frameworks based on AE [8,42,54].

Table 3.3: Area under curve (AUC) comparison of proposed framework and [8, 12, 42–44, 54] works on ShanghaiTech dataset.

Methods	ShanghaiTech	
	AUC	EER
StackRNN [44]	68.0%	-
Liu et al. [12]	72.8%	32.6%
The Proposed	**73.0%**	**32.3%**

3.3.3 Results on the ShanghaiTech Dataset

ShanghaiTech is one of the most challenging dataset in video anomaly detection. It contains various type of anomalies, video scenes and complicated movement. Unlike training the network with only one video scene in UCSD Pedestrian and Avenue datasets, the network was trained in several scenes in the ShanghaiTech dataset. Experimental results presented in Table 3.3 demonstrated that the proposed framework achieved 73% of AUC, which is 5% better than the work that utilized the stacked of RNN layers. The performance is slightly improved compared to the baseline method [12].

3.3.4 Qualitative Result of the Proposed Framework

The qualitative results of the proposed framework on three testing videos in the CUHK Avenue, the UCSD Ped2 and the ShanghaiTech dataset are illustrated in Figure 3.6. 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). This research also shows the output of the proposed generator network in Figure 3.7. The result of the generated image and the image difference compared to the ground truth image indicates that the image quality of the anomaly area is blurred and distorted due to the fact that the generator network could not reconstruct the unseen object from the learned model (i.e. "throwing", "car and bicycle approaching", "strange action"), resulting in a lower regularity score in these video scenes. Although the proposed framework can detect anomalies in general scenes with cleared object appearances, the proposed framework still has limitations to detect occluded anomaly objects. For example, when a group of people occludes the anomaly object, the proposed framework cannot detect this kind of anomalies. In addition, camera viewpoints cause anomaly objects in different scenes to appear differently, which also affects the detection accuracy. For example, camera viewpoints of the UCSD Ped1 and UCSD Ped2 are different (Figure 3.7), the looks of cycling people in the UCSD Ped1 dataset is similar to walking people. However, cycling people can be clearly classified as an anomaly event in the UCSD Ped2 dataset.

Table 3.4: AUC for anomaly detection of networks with/wo SIM and residual skip connection. The results are evaluated on CUHK Avenue dataset.

Model	AUC
U-Net [24]	84.9%
Only residual skip connection	85.1%
Only SIM	86.6%
Multi-scale U-Net	**86.9%**

Table 3.5: Parameter number comparison of the proposed multi-scale U-Net with/wo and the original U-Net [24]

Model	Number of Parameters
U-Net (baseline) [24]	10,468,932
Proposed wo asymmetric convolution	9,748,210
Proposed with asymmetric convolution	**8,102,784**

3.3.5 Performance Comparison on Employing an Individual Module of the Proposed Framework

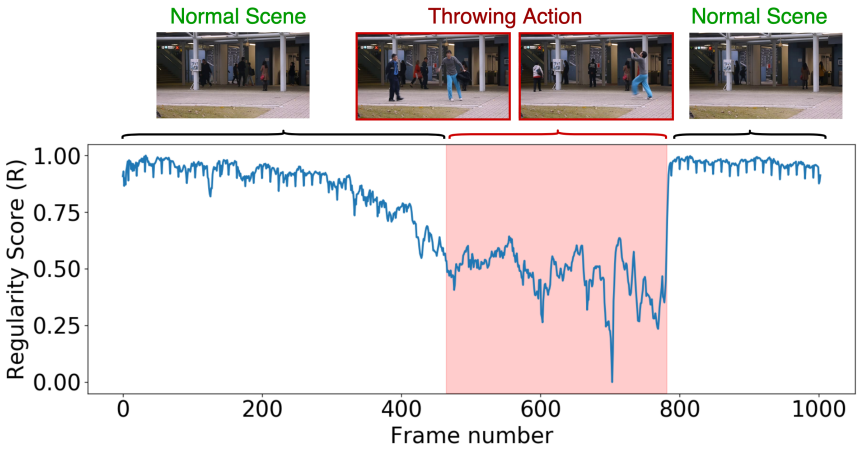
Table 3.4 summarizes the performance evaluation result of the proposed framework for confirming individual contribution. The experiments are performed on CUHK Avenue dataset using two parts based on the original U-Net architecture as a baseline: the inception module and the residual skip connection. Firstly, the traditional skip connections were replaced by the residual skip connection. Secondly, two consecutive 3x3 convolution layers were replaced by the inception module. The result shows that the residual skip connection slightly improves when included in the U-Net architecture, while employing the inception module is even more effective to the detection performance. However, combining the inception module and residual skip connection achieved the best detection performance compared to the original U-Net architecture.

3.3.6 Running Time Analysis

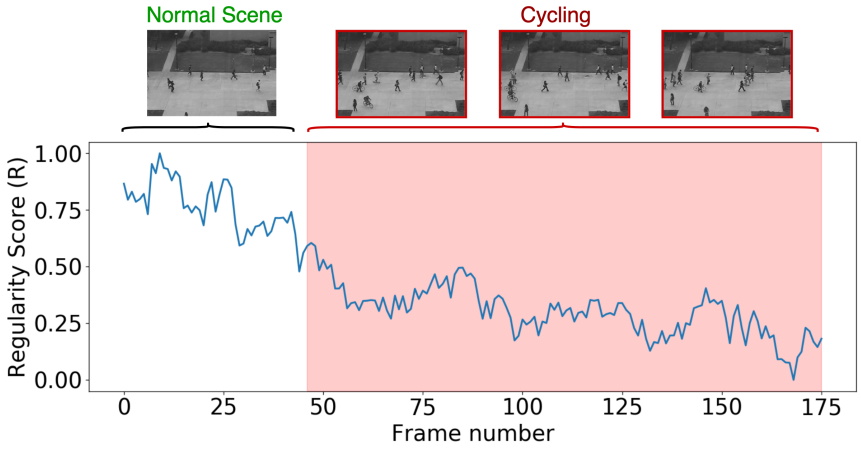
A comparison of the parameter number of the original U-Net architecture and proposed ones are presented in Table 3.5, demonstrating that the multi-scale U-Net reduces the number of parameters by 22.6% for training and testing an anomaly detection while improving the accuracy. This work also evaluated 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. The 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 [24] that is able to run at about 22 fps. Overall, the processing time of the proposed framework is speeded up by 9% on average.

3.4 Summary

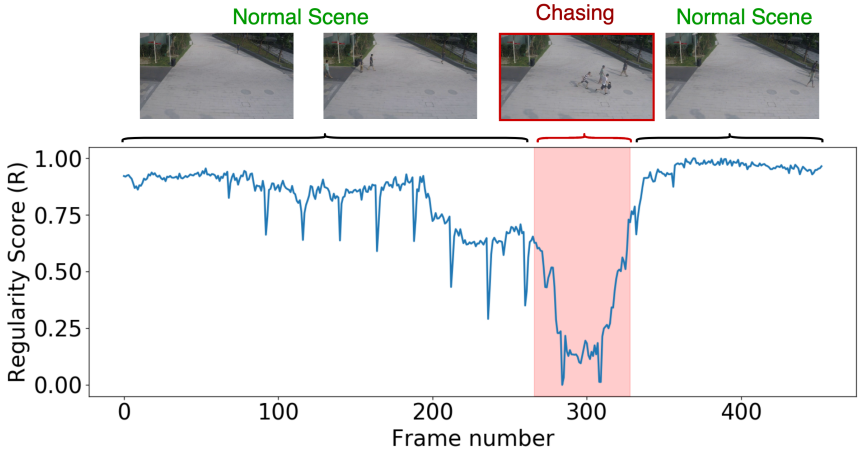
This chapter presented a framework based on multi-scale U-Net architecture for anomaly detection in video. The inception modules are employed instead of using the traditional convolution layers utilized in the original U-Net, making the multi-scale U-Net has an ability to learn image features in different scales. The skip connections were replaced by the proposed residual skip connections including shortcut connections, which 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 without degrading the detection accuracy. As a result in both qualitative and quantitative, the 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. The overall result illustrated that the design of the proposed framework is able to capture appearance and motion information to detect anomalies in real-world scene scenarios.



(a) CUHK Avenue



(b) UCSD Ped2



(c) ShanghaiTech

Figure 3.6: Qualitative results of the proposed framework on three testing videos in (a) CUHK Avenue, (b) UCSD Ped2 and (c) ShanghaiTech dataset. Colored window shows ground truth anomaly region.

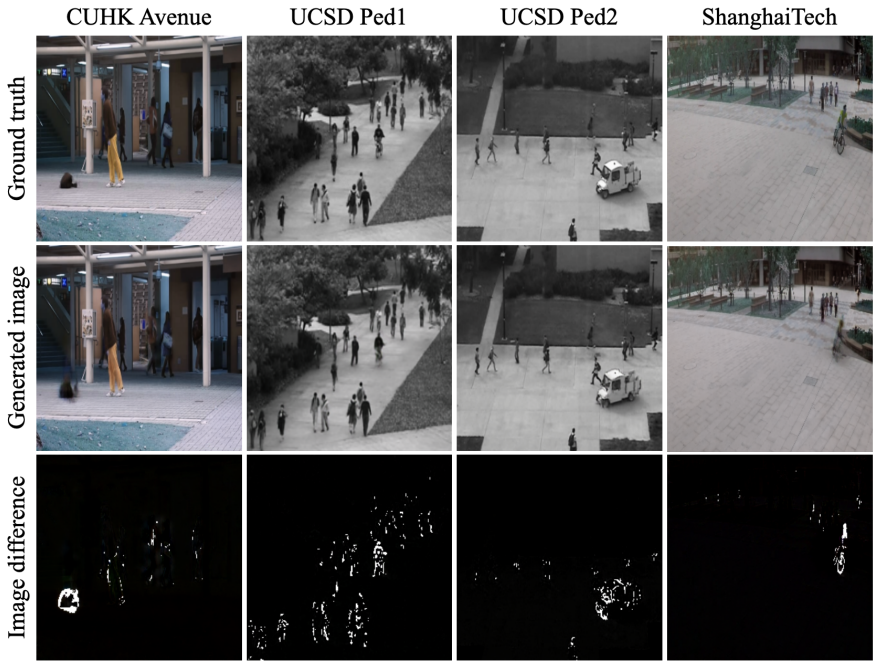


Figure 3.7: An example of the generated image of the proposed in all datasets: CUHK Avenue, UCSD Ped1, Ped2, and Shanghaitech datasets. The top row is ground truth images, the middle shows generated images, and the bottom row is an image difference between ground truth and generated images.

Chapter 4

Joint Representation Learning for Video Anomaly Detection

This chapter presents a method to detect anomaly in surveillance video using GAN structure based on a joint representation learning. The network architecture consists of two encoders, one for appearance feature and another one for motion feature. The RGB image sequence and the optical flow are fed to different encoders in order to improve the feature extraction processes. The details of the proposed two encoders structure and network architecture are described in Subsection 4.2.1 and 4.2.2, respectively. In the remaining of this chapter, experimental results and discussions are provided.

4.1 Introduction

Recently, video anomaly detections based on deep learning have attracted significant attention from the research community and are able to improve the accuracy of the detection and localization in complex scenarios. Many researches in video anomaly detection attempted to model the normality pattern using unsupervised learning [12, 13, 39–44]. These approaches extract the feature from the input RGB image and reconstruct the output to be the same as inputs. Then the reconstruction error is performed to compute the anomaly score. However, using only the appearance information in these network structures is not capable of capturing enough object characteristics for video anomaly detection. Therefore, this research proposes a joint representation learning structure for

anomaly detection. The proposed architecture consists of two encoders that extract the feature from the appearance and the motion. The appearance and motion features are extracted using ResNet architecture [15]. The ResNet architecture is adopted as the generator network, as it allows information from previous layers in the network to skip ahead of one or more layers. Thus, the ResNet architecture can be trained in many layers without suffering from the vanishing gradient problem.

4.2 Joint Representation Learning for Video Anomaly Detection

An overview of the proposed method is visualized in Figure 4.1. The proposed network architecture consists of two separate encoders to extract appearance and motion features. The feature maps from two encoders are then fused and fed to the same decoder. This work utilized ResNet [15] as the network backbone due to its capability of capturing appearance features. The objective of the generator network is to minimize the reconstruction error between the input image and the generated image by the learned model. Then this work uses the reconstruction error to compute the regularity score. A normal scene is expected to have a high score in the testing phase, whereas an anomaly is expected to have a low score.

4.2.1 Two Encoders Structure

AE is a type of unsupervised learning, which consists of encoder and decoder parts. The encoder takes an input image to extract the feature into a small size vector called a latent representation. In contrast, the decoder aims to reconstruct the original image from the latent representation. Unlike the traditional AE, this research instead employed two encoders. The first encoder used the RGB image sequence as an input to extract the appearance feature and the second encoder used the optical flow as an input to extract the motion feature. In addition, this work uses ResNet [15] as a network backbone in the proposed architecture. ResNet was designed to allow information from previous layers in the network to skip ahead of one or more layers, which has shown impressive performance in anomaly detection and action recognition benchmarks [41, 69, 70].

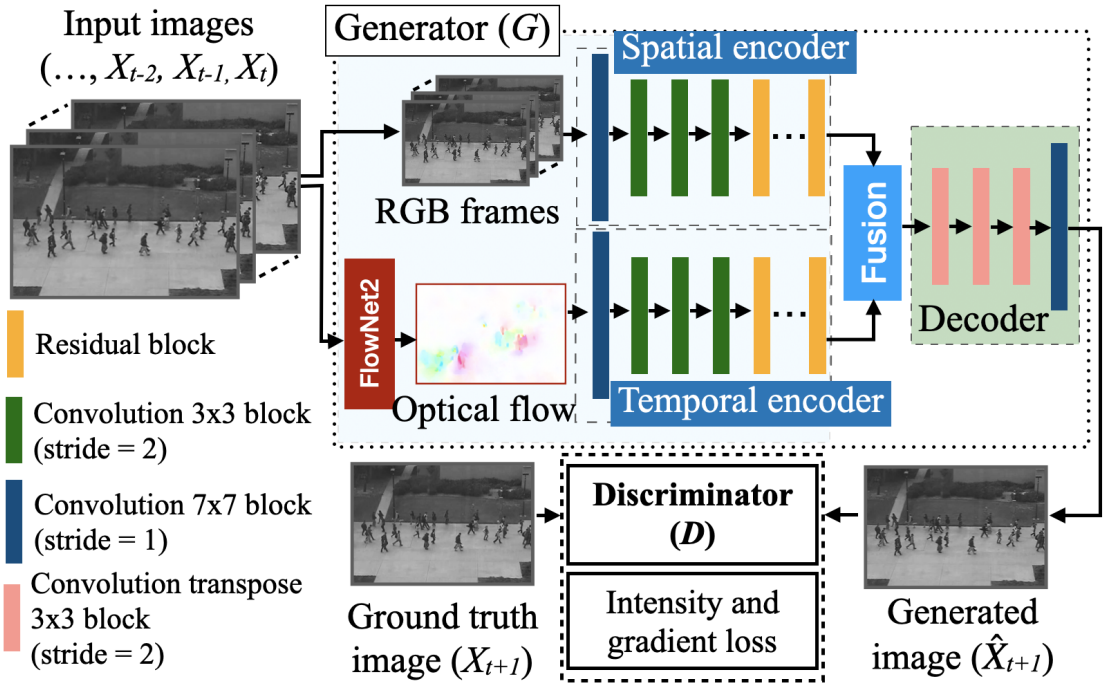


Figure 4.1: The overview of the proposed method based on two encoders network structure.

The generator network G (as shown in Figure 4.1) aims to generate the output image $\hat{X}_{t+1}$ that looks similar to the ground truth image X_{t+1}. Specifically, given input image sequence $(...,X_{t-2},X_{t-1},X_t)$, the spatial encoder takes the image sequence as input, whereas the temporal encoder uses the motion information by generating the optical flow F_t using Flownet [71]. The optical flow uses the change of the image intensity of two consecutive frames to calculate the motion information caused by the movement of the object itself or the camera movement. In the proposed method, this work computed the optical flow using X_{t-1} and X_t of the input image sequence. The feature maps of two encoders are then fused and fed to the same decoder.

4.2.2 Network Architecture

The detail of the proposed network architecture is described in Table 4.1. The network contains 18 layers; the first 13 layers are encoder layers consisting of convolutional layers and stacking of residual blocks. RGB images and optical flow are fed to a separate encoder. A fusion layer is applied in the middle of the network, followed by the decoder

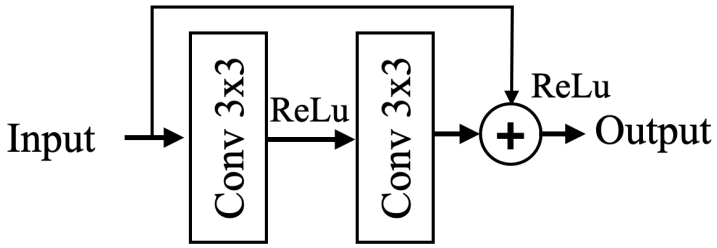


Figure 4.2: Residual block.

layer in the remaining. In detail, the input fed into the first layer with a filter size of 7×7 with a stride of 1 and 64 feature maps are generated so as to output a feature map of 256×256×64, this work uses the large filter size in the first layer of the network in order to extract the feature on a large object size in the image.

The structure of the residual block is illustrated in Figure 4.2. A residual block contains two 3×3 convolutional layers with 512 filters and the stride of 1, Rectified Linear Unit (ReLU) activation function [72] is applied after each convolution layer. Each residual block contains a skip connection that adds the input and the output of the convolutional layer before passing the output to the next layer. In the middle of the network, a fusion layer is utilized to combine a feature map from two encoders with the output of 32×32 and 1024 feature maps. Finally, the decoder comprises the reverse architecture of the encoder. Three 3×3 deconvolutional layers and a 7×7 convolutional layer are placed at the end of the network architecture to form a decoder part.

4.2.3 Model Learning

Generally, the AE approach in video anomaly detection aims to learn the useful features of normal events. As the structure of the proposed network consists of two encoders, this research considered the intensity loss L_{int}, and gradient loss L_{gd} of the generated image $\hat{X}$ and the ground truth X as the main objective functions to minimize the reconstruction error, which can be calculated as follows.

$$L_{int}(\hat{X} - X) = \|\hat{X} - X\|_2^2 \tag{4.1}$$

Table 4.1: The proposed network model structure and configuration.

Layer		Filter Size/Stride	#filters	Output
Encoders	S1, T1	Conv(7×7)/1	64	256×256×64
	S2, T2	Conv(3×3)/2	128	128×128×128
	S3, T3	Conv(3×3)/2	256	64×64×256
	S4, T4	Conv(3×3)/2	512	32×32×512
	S5-13, T5-13	Residual block	512	32×32×512
Fusion	14	Features fusion	1024	32×32×1024
Decoder	15	DeConv(3×3)/2	256	64×64×256
	16	DeConv(3×3)/2	128	128×128×64
	17	DeConv(3×3)/2	64	256×256×64
	18	Conv(7×7)/1	3	256×256×3

Conv = Convolutional layer, DeConv = Deconvolutional layer
S = the spatial encoder layer, T = the temporal encoder layer.

$$L_{gd}(\hat{X}-X) = \||\hat{X}_{i,j}-\hat{X}_{i-1,j}|-|X_{i,j}-X_{i-1,j}|\|_1 + \||\hat{X}_{i,j}-\hat{X}_{i,j-1}|-|X_{i,j}-X_{i,j-1}|\|_1, \tag{4.2}$$

where i, j is the spatial index of an image.

In addition, PatchGAN [14] is utilized as a discriminator network D in the proposed architecture to discriminate $\hat{X}$ from X, which maps $\hat{X}$ to small patches and takes each individual patch to predict whether a patch comes from X or $\hat{X}$. An output of D is a scalar that classifies the patch from X as class 1 and the patch from $\hat{X}$ as class 0. The goal of training G is to generate an image where D classifies it into class 1. The adversarial loss L_{adv} can be calculated as follows.

$$L_{adv}(\hat{X}) = \sum_{m,n} \frac{1}{2} L_{MSE}(D_{gen}(\hat{X})_{m,n}, 1) \tag{4.3}$$

$$L_{MSE}(\hat{X}, X) = (\hat{X}-X)^2, \tag{4.4}$$

where this work denoted $D_{gen}(.)$ as the discriminator network of the generated image $\hat{X}$ and m, n denote patch indexes. Finally, the objective function L can be computed as follows.

$$L = \lambda_{int} L_{int} + \lambda_{gd} L_{gd} + \lambda_{adv} L_{adv}, \tag{4.5}$$

where $\lambda_{int}, \lambda_{gd}, \lambda_{adv}$ are weights of each loss.

4.2.4 Anomaly Detection

This research quantifies the normalities or abnormalities by using only the generator network G in the testing phase. The work proposed by [12] and Mathieu et al. [73] proved that Peak Signal to Noise Ratio (PSNR) is a more optimal metric to evaluate the quality of the next frame prediction. Therefore, PSNR is utilized as the detection score, where high PSNR indicates that the image is more similar to the ground truth, which considers to be a normal event. PSNR can be calculated as follows.

$$PSNR(X_t, \hat{X}_t) = 10 \log_{10} \frac{[\max_{\hat{X}}]^2}{\frac{1}{N} \sum_{i=0}^{N} (p_t(i) - \hat{p}_t(i))^2}, \tag{4.6}$$

where $\max_{\hat{X}}$ represents the maximum intensity value in a generated image $\hat{X}$. $p_t(i)$ and $\hat{p}_t(i)$ are pixel intensities of index i in X and $\hat{X}$, respectively. N denotes the total number of pixels in the image. Finally, a regularity score $R(t)$ at frame t is obtained by normalizing the PSNR to the range of [0,1] as follows:

$$R(t) = \frac{PSNR(X_t, \hat{X}_t) - \min_{PSNR}}{\max_{PSNR} - \min_{PSNR}}, \tag{4.7}$$

this work denoted $\min_{PSNR}$ and $\max_{PSNR}$ as the minimum and the maximum among PSNRs of all frames in testing video.

4.3 Experimental Results

4.3.1 Implementation Details

This research performed the experiment on three benchmark anomaly datasets including the UCSD Ped2 [56], the CUHK Avenue [6], and the ShanghaiTech [12]. The training set of each dataset contains only normal events. Input images are resized to 256×256 with three color channels. This research has set 4 consecutive images as the input im-

age sequence. The network was trained on Tensorflow using NVIDIA RTX 2080 Ti GPU. In the preliminary experiment, this work achieved the best performance by setting $\lambda_{int}, \lambda_{gd}, \lambda_{adv}$ to 2, 1, and 0.05, respectively. In the same manner as [12, 13, 39, 40], this work plot the receiver operating characteristic (ROC) curve and compute the area under the curve (AUC), and use it as the evaluation metrics.

Table 4.2: Performance comparison with [12, 13, 39–44] on CUHK Avenue, UCSD Ped2, and ShanghaiTech dataset

Methods	CUHK Avenue	UCSD Ped2	ShanghaiTech
Conv-AE [42]	80.0%	85.0%	60.9%
Luo et al. [44]	81.7%	92.2%	68.0%
Fan et al. [43]	83.4%	92.2%	-
Liu et al. [12]	84.9%	95.4%	72.8%
Spatio-Temporal U-Net [13]	84.5%	**96.5%**	-
Dual Discriminator [40]	84.9%	95.6%	**73.7%**
AnomalyNet [41]	86.1%	94.9%	-
Conv-VRNN [39]	85.7%	96.1%	-
The Proposed	**86.9%**	**96.5%**	73.4%

4.3.2 Results of the Proposed Architecture Compared with Learning-based Methods

Table 4.2 shows the results of the proposed method compared with several learning-based methods, including two-stream network architectures, AE and GAN. The proposed method achieved an AUC of 86.9% and 96.5% in the CUHK Avenue and the UCSD Ped2, respectively, which is outperforming the existing methods. The result of the ShanghaiTech dataset is 73.4% and it was also improved compared to the work based on AE [42, 44] and GAN [12], but slightly lower than the work from Dong et al. [40] that utilizes both appearance and motion discriminator-based of the generative adversarial network.

This research demonstrated the qualitative result of the proposed method on testing videos in the CUHK Avenue and the UCSD Ped2 datasets in Figure 4.3 and Figure 4.4, respectively. It can be seen that the proposed method can detect anomaly events, even

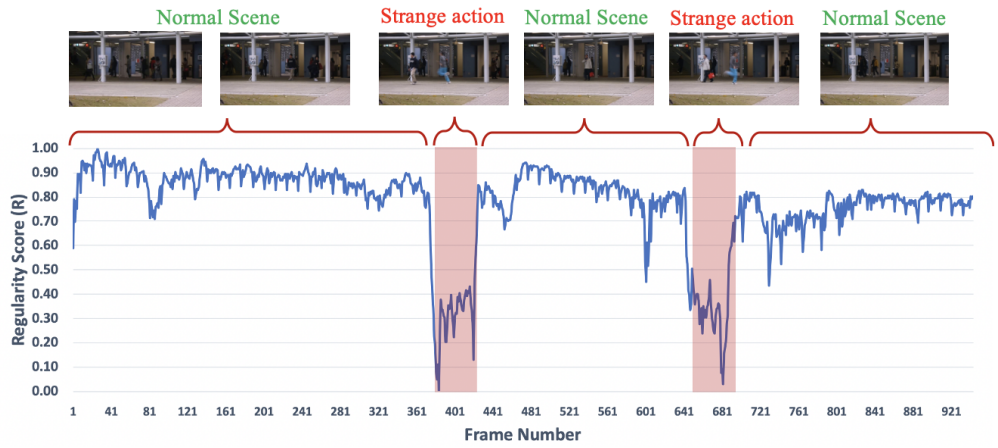


Figure 4.3: Qualitative results of the proposed method on Avenue dataset.

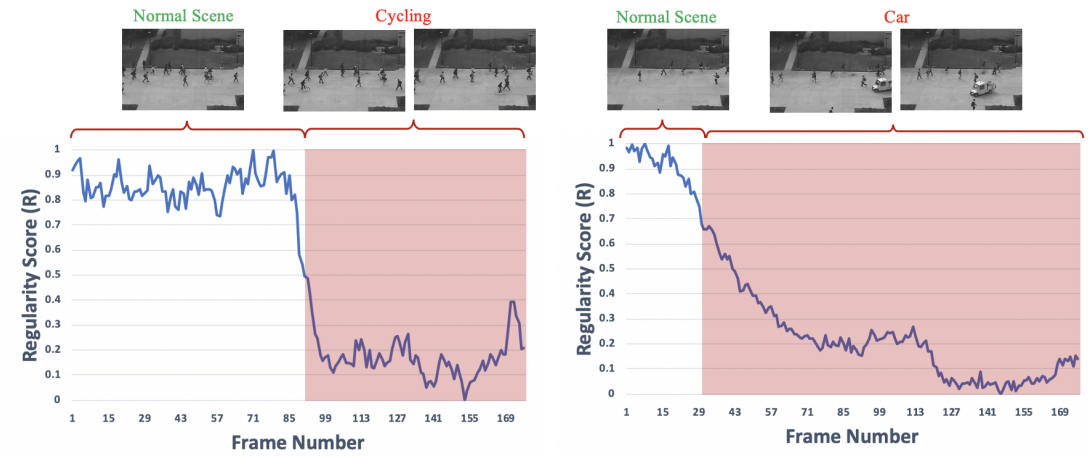


Figure 4.4: Qualitative results of the proposed method on UCSD Ped2 dataset.

in the case of the video scene of crowded people. However, for some of the anomaly events (i.e., strange action), such as people running across the video scene, the regularity score is rapidly changed to low due to the fact that the motion pattern of the people movement is unseen in the trained model. For example, in Figure 4.4, when the car was approaching the video scene, the regularity score trend slowly decreased comparing to Figure 4.3. This seems to be because only some part of the car has appeared in the scene, and the proposed method is not able to detect an anomaly object, but when the whole part of the car gets in the scene, the regularity score is completely low.

Table 4.3: AUC comparison of the network backbone in the proposed network architecture on UCSD Ped2, CUHK Avenue and ShanghaiTech dataset

Network Backbone	CUHK Avenue	UCSD Ped2	ShanghaiTech
U-Net [24]	85.3%	95.1%	72.6%
ResNet [15]	**86.9%**	**96.5%**	**73.4%**

4.3.3 Performance Comparison of using Different Network Backbones

The performance of the proposed network architecture was evaluated through a comparison with the conventional U-Net network backbone. All the setting is kept in the same condition including the latent vector size, filter size, number of filters and training hyperparameters. As shown in Table 4.3, the proposed network architecture utilizing the ResNet as a backbone network achieved better performance compared to the U-Net network in terms of detection accuracy. Therefore, this work adapts ResNet as the generator network in the proposed method. The effectiveness of using a different network backbone is also discussed in AnomalyNet [41]. The ResNet backbone achieved higher performance compared to some network backbones that are used in the classification and detection task. The result suggests that the network backbone is essential for the generator network in GAN to keep a useful feature during the feature extraction process.

4.4 Summary

This chapter proposed a joint representation learning for anomaly detection in surveillance video. The network architecture consists of two encoders. The first encoder used the RGB image sequence to extract the appearance feature, and the second encoder used the optical flow to extract the motion feature. Feature maps from two encoders are fused and fed to the same decoder. This research used a ResNet [15] as the network backbone. The experiment results on three benchmark datasets show the effectiveness of the proposed method in terms of detection accuracy compared to the existing learning-based video anomaly detection methods.

Chapter 5

Conclusion and Future Work

5.1 Summary of Methods and Contributions

In present, machine learning algorithms show promising performance in the context of AI in various area.Video analytic applications are widely developed in the past decade, making a surveillance system more autonomous. Anomaly detection in surveillance video is essential part in security system and attracts attention in computer vision research community. Since anomaly events in surveillance video is diverse, it is impossible to define all possible events of normal and anomaly video. Moreover, 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. This dissertation proposed the framework for video anomaly detection based on unsupervised machine learning structure to extract semantic representations from video data. These representations are then used to detect whether the video scene is normal or abnormal events. Methods proposed in this dissertation are based on a generative network, which shows the effectiveness of detection accuracy while reducing the processing time and the error of the detection in real-world scenarios. This dissertation consists of two contributions, where each contribution used different neural network architecture. The first contribution used the multi-scale U-Net that reduces the processing time, where the second contribution aims to improve the detection accuracy by utilizing the ResNet architecture. The ResNet consumes high computational resource due to the fact that the network architecture contains many layers compared to the multi-scale U-Net, making

these two contributions have different directions. In summary, this dissertation provides the following contributions to solve the problems in anomaly detection in surveillance video.

- **Multi-Scale U-Net:** Currently, most of the network architecture in video anomaly detection attempted to balance the detection accuracy and the processing time. In the generative network, some features are lost in the generator during training due to convolution operators of each layer, which affected to the accuracy. In Chapter 3, this dissertation proposed a video anomaly detection based on GAN structure. The network architecture is based on the U-Net but some modifications are applied. this work 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. Although neural networks acheived significant performance in terms of detection accuracy, speed and computational cost should be considered for real-world applications. Therefore, this research 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.
- **Joint representation learning:** Typically, the input of neural network in video anomaly detection approach contains only the appearance information and it is not capable of capturing enough object characteristics for video anomaly detection. In Chapter 4, this dissertation proposed a joint representation learning structure for anomaly detection. This research designed a convolutional neural network for video anomaly detection that combines a representation of the appearance and the motion features. The proposed network architecture consists of two encoders based on ResNet architecture. The RGB image sequence and the optical flow are fed to different encoders in order to improve the feature extraction of the object appearance and the motion. This research proposed features fusion of encoders to combine useful features obtained by encoding the appearance and the motion features.
- **Evaluations on several benchmark datasets:** To measure the performance, this dissertation evaluate the proposed framework and network architecture with three

benchmark datasets of different scene scenarios including the UCSD Pedestrian, the CUHK Avenue and the ShanghaiTech datasets. Experiments on the benchmark datasets show the effectiveness of the proposed framework and network architecture for video anomaly detection in terms of both detection accuracy and processing time.

5.2 Future Work

The proposed designs of network architecture based on generative network currently achieved a significant improvements in terms of detection accuracy and processing time. However, the generator network is unable to distinguish an ambiguous anomalous object in a scene such as some part of anomaly object is occluded by other objects, same motions but classifies as different events (e.g. the people running in the park is classified as normal event but the people running in the bank might be classified as anomaly event). For future work, this research will explores on experimenting on applying preprocessing techniques such as background segmentation, human detection before passing to the feature extraction processes to illuminate unwanted objects. Next, this research will redesigns the network architecture to used the motion features for training, in order to capture both the appearance and the motion of the object characteristic and to enhance the performance of the anomaly detection. As the improvement of AI technologies are attractive by the researchers and the CCTV camera are more automated, we will also explore the usage of computational resources such as memory size, power consumption, and the processing time in order to make the video anomaly detection framework able to work in the embedded system for real-world computer vision applications. Further, the results of testing anomaly detection including anomaly scenes and anomaly objects in the real-world scenario can be collected, these data can be used to create another dataset for anomaly detection and action recognition tasks.

Bibliography

[1] L. Lin and N. Purnell. (2019) A world with a billion cameras watching you is just around the corner. [Online]. Available: https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402.

[2] E. L. Piza, B. C. Welsh, D. P. Farrington, and A. L. Thomas, "CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis," *Criminology & Public Policy*, vol. 18, no. 1, pp. 135–159, 2019.

[3] R. Armitage, "To CCTV or not to CCTV?: A review of current research into the effectiveness of CCTV systems in reducing crime," *Nacro Briefing Note*, 2002.

[4] S. Feldstein, *The global expansion of AI surveillance*. Carnegie Endowment for International Peace Washington, DC, 2019, vol. 17.

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

[6] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.

[7] A. Li, Z. Miao, Y. Cen, and Q. Liang, "Abnormal event detection based on sparse reconstruction in crowded scenes," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1786–1790.

[8] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proceedings of the International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.

[9] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2019.

[10] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," *arXiv preprint arXiv:1510.01553*, 2015.

[11] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, "Detecting abnormal events in video using narrowed normality clusters," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1951–1960.

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

[13] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-temporal unity networking for video anomaly detection," *IEEE Access*, vol. 7, pp. 172 425–172 432, 2019.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proceedings of the 2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] A. A. Patel, *Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data*. O'Reilly Media, 2019.

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

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the Advances in neural information processing systems*, 2014, pp. 2672–2680.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[25] D. Foster, *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media, 2019.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[29] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 8, pp. 873–889, 2001.

[30] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Proceedings of the European Conference on Computer Vision*. Springer, 2010, pp. 563–576.

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

[32] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proceedings of the CVPR 2011*. IEEE, 2011, pp. 3313–3320.

[33] Y. Yuan, Y. Feng, and X. Lu, "Statistical hypothesis detector for abnormal event detection in crowded scenes," *IEEE transactions on cybernetics*, vol. 47, no. 11, pp. 3597–3608, 2016.

[34] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 2921–2928.

[35] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proceedings of the 2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2054–2060.

[36] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[37] 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.

[38] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, pp. 13–es, 2006.

[39] Y. Lu, K. M. Kumar, S. shahabeddin Nabavi, and Y. Wang, "Future frame prediction using convolutional vrnn for anomaly detection," in *Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[40] F. Dong, Y. Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," *IEEE Access*, vol. 8, pp. 88 170–88 176, 2020.

[41] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019.

[42] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.

[43] 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," *Computer Vision and Image Understanding*, p. 102920, 2020.

[44] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.

[45] T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised anomaly detection and localization based on deep spatiotemporal translation network," *IEEE Access*, vol. 8, pp. 50 312–50 329, 2020.

[46] Y.-L. Hsueh, W.-N. Lie, and G.-Y. Guo, "Human behavior recognition from multiview videos," *Information Sciences*, vol. 517, pp. 275–296, 2020.

[47] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[48] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[49] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4597–4605.

[50] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[51] G. Pang, C. Yan, C. Shen, A. v. d. Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 173–12 182.

[52] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

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

[54] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 439–444.

[55] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.

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

[57] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[59] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[60] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[61] S. Saypadith and T. Onoye, "Video anomaly detection based on deep generative network," in *Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[62] Y. Li, X. Li, C. Xiao, H. Li, and W. Zhang, "Eacnet: Enhanced asymmetric convolution for real-time semantic segmentation," *IEEE Signal Processing Letters*, vol. 28, pp. 234–238, 2021.

[63] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proceedings of the ACM Multimedia Asia*, 2019, pp. 1–6.

[64] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.

[65] 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 *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[66] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[68] I. Gemp and B. McWilliams, "The unreasonable effectiveness of adam on cycles," in *Proceedings of the NeurIPS Workshop on Bridging Game Theory and Deep Learning*, vol. 4, 2019.

[69] A. Sikdar and A. S. Chowdhury, "An adaptive training-less framework for anomaly detection in crowd scenes," *Neurocomputing*, vol. 415, pp. 317–331, 2020.

[70] H. Ergun, Y. C. Akyuz, M. Sert, and J. Liu, "Early and late level fusion of deep convolutional neural networks for visual concept recognition," *International Journal of Semantic Computing*, vol. 10, no. 03, pp. 379–397, 2016.

[71] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[72] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the ICML*, 2010, pp. 807–814. [Online]. Available: https://icml.cc/Conferences/2010/papers/432.pdf

[73] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.