Anomalous Event Detection and Localization Based on Deep Generative Adversarial Networks for Surveillance Videos

Miss Thittaporn Ganokratanaa

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Electrical Engineering
Department of Electrical Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2020

การตรวจจับและการระบุตำแหน่งเหตุการณ์ผิดปกติบนพื้นฐานการสร้างเครือข่ายปรปักษ์เชิงลึกสำหรับวิดีโอเฝ้าระวัง

น.ส.ฐิตาภรณ์ กนกรัตน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	Anomalous Event Detection and Localization Based on Deep Generative Adversarial Networks for Surveillance Videos
By	Miss Thittaporn Ganokratanaa
Field of Study	Electrical Engineering
Thesis Advisor	Associate Professor Dr. Supavadee Aramvith
Thesis Co Advisor	Professor Dr. Nicu Sebe

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the FACULTY OF ENGINEERING
(Associate Professor Dr. Supot Teachavorasinskun)

DISSERTATION COMMITTEE

..... Chairman
(Assistant Professor Dr. SUREE PUMRIN)
..... Thesis Advisor
(Associate Professor Dr. Supavadee Aramvith)
..... Thesis Co-Advisor
(Professor Dr. Nicu Sebe)
..... Examiner
(Associate Professor Dr. CHARNCHAI PLUEMPITIWIRIYAWEJ)
..... Examiner
(Associate Professor Dr. NISACHON TANGSANGIUMVISAI)
..... External Examiner
(Dr. Nattachai Watcharapinchai)

ฐิตาภรณ์ กนกรัตน : การตรวจจับและการระบุตำแหน่งเหตุการณ์ผิดปกติบนพื้นฐานการสร้างเครือข่ายปรปักษ์เชิงลึกสำหรับวิดีโอเฝ้าระวัง. (Anomalous Event Detection and Localization Based on Deep Generative Adversarial Networks for Surveillance Videos) อ.ที่ปรึกษาหลัก : รศ. ดร. สุภาวดี อร่ามวิทย์, อ.ที่ปรึกษาร่วม : ศ. ดร.นิคู เซเบ

การตรวจจับความผิดปกติมีความสำคัญอย่างยิ่งสำหรับวิดีโอเฝ้าระวังอัจฉริยะ งานวิจัยในปัจจุบันมักเจอกับปัญหาการตรวจจับและการระบุตำแหน่งวัตถุ เนื่องจากฉากที่แออัดและการไม่มีข้อมูลเบื้องต้นของวัตถุที่สนใจอย่างเพียงพอในระหว่างการเรียนรู้ของโมเดลซึ่งส่งผลให้ผลการตรวจจับเป็นเท็จ ดังนั้นในวิทยานิพนธ์นี้ จึงเสนอกรอบใหม่สองแบบสำหรับการตรวจจับและการระบุตำแหน่งความผิดปกติในวิดีโออันดับแรกเสนอเครือข่ายการแปลเวลาและพื้นที่เชิงลึก ซึ่งเป็นวิธีการตรวจจับและระบุตำแหน่งความผิดปกติแบบใหม่ที่ใช้การเรียนรู้แบบไม่มีผู้สอนบนพื้นฐานการสร้างเครือข่ายปรปักษ์เชิงลึกและการจับขอบ ในงานนี้ได้นำเสนอการรวมกันแบบใหม่ของภาพการลบพื้นหลังและภาพการเคลื่อนที่จริง โดยมีการต่อกันของภาพต้นฉบับและภาพการลบพื้นหลัง และปรับปรุงประสิทธิภาพของการระบุตำแหน่งความผิดปกติในการประเมินระดับพิกเซลโดยเสนอวิธีการจับขอบเพื่อลดสัญญาณรบกวนและลดขอบที่ไม่เกี่ยวข้องกับวัตถุที่ผิดปกติ เครือข่ายการแปลเวลาและพื้นที่เชิงลึกเป็นวิธีที่ประสบความสำเร็จโดยให้ประสิทธิภาพที่ดีเกี่ยวกับความแม่นยำในการตรวจจับความผิดปกติและความซับซ้อนของเวลาสำหรับวิดีโอเฝ้าระวัง อย่างไรก็ตามปัญหาการตรวจจับเป็นเท็จยังคงเกิดขึ้นในฉาก ดังนั้นจึงนำเสนอเครือข่ายการแปลเวลาและพื้นที่ที่เหลือเชิงลึก ซึ่งเป็นโมเดลเครือข่ายปรปักษ์เชิงลึกแบบมีเงื่อนไขที่เหลือแบบใหม่ที่ใช้การเรียนรู้แบบไม่มีผู้สอนด้วยวิธีการทำเหมืองเชิงลบอย่างมากแบบออนไลน์ เพื่อลบผลการตรวจจับที่เป็นเท็จโดยเฉพาะ เครือข่ายการแปลเวลาและพื้นที่ที่เหลือเชิงลึกนำเสนอเครือข่ายที่กว้างขึ้นเพื่อเรียนรู้การทำแผนที่จากการแสดงเชิงพื้นที่ไปจนถึงการแสดงเชิงเวลา และเพิ่มคุณภาพการรับรู้ของภาพที่สังเคราะห์จากเจเนอร์เรเตอร์ วิธีที่นำเสนอทั้งสองวิธีได้รับการทดสอบกับชุดข้อมูลความผิดปกติที่เปิดเผยต่อสาธารณะ ได้แก่ ชุดข้อมูลคนเดินเท้ายูซีเอสดี ชุดข้อมูลยูเอ็มเอ็น และ ชุดข้อมูลซียูเอชเค อเวนิว ซึ่งแสดงให้เห็นถึงผลลัพธ์ที่เหนือกว่าวิธีการแบบใหม่อื่น ๆ ทั้งในการประเมินระดับเฟรมและระดับพิกเซล

สาขาวิชา	วิศวกรรมไฟฟ้า	ลายมือชื่อนิสิต .....
ปีการศึกษา	2563	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6071415121 : MAJOR ELECTRICAL ENGINEERING
KEYWORD: Anomaly detection, anomaly localization, unsupervised learning, spatiotemporal translation, surveillance video

Thittaporn Ganokratanaa : Anomalous Event Detection and Localization Based on Deep Generative Adversarial Networks for Surveillance Videos. Advisor: Assoc. Prof. Dr. Supavadee Aramvith Co-advisor: Prof. Dr. Nicu Sebe

Anomaly detection is of great significance for intelligent surveillance videos. Current works typically struggle with object detection and localization problems due to crowded scenes and lack of sufficient prior information of the objects of interest during training, resulting in false-positive detection results. Thus, in this thesis, we propose two novel frameworks for video anomaly detection and localization. We first propose a Deep Spatiotemporal Translation Network (DSTN), a novel unsupervised anomaly detection and localization method based on Generative Adversarial Network (GAN) and Edge Wrapping (*EW*). In this work, we introduce (i) a novel fusion of background removal and real optical flow frames with (ii) a concatenation of the original and background removal frames. We improve the performance of anomaly localization in the pixel-level evaluation by proposing (iii) the Edge Wrapping to reduce the noise and suppress non-related edges of abnormal objects. DSTN is a successful approach, providing good performance regarding anomaly detection accuracy and time complexity for surveillance videos. However, the false-positive problem has still occurred in the scene. Thus, we continue to propose Deep Residual Spatiotemporal Translation Network (DR-STN), a novel unsupervised Deep Residual conditional Generative Adversarial Network (DR-cGAN) model with an Online Hard Negative Mining (OHNM) approach to specifically remove the false-positives. The proposed DR-cGAN provides a wider network to learn a mapping from spatial to temporal representations and enhance the perceptual quality of synthesized images from a generator. Our proposed methods have been tested on publicly available anomaly datasets, including UCSD pedestrian, UMN, and CUHK Avenue, demonstrating superior results over other state-of-the-art methods both in frame-level and pixel-level evaluations.

Field of Study:	Electrical Engineering	Student's Signature .....
Academic Year:	2020	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would first like to acknowledge my funding, Chulalongkorn University Dutsadi Phiphat Scholarship, for financially supporting me throughout my Ph.D. study.

I would like to express my sincere gratitude to my supervisor, Associate Professor Dr. Supavadee Aramvith, for the continuous support of my Ph.D. study and research and all of the opportunities I was given to further my career. I would like to thank for her motivation, invaluable guidance, immense knowledge, and for being a role model for me. Her advice and support are beyond words. I am genuinely grateful that I am among those under her supervision.

My sincere thanks also go to my co-supervisor, Professor. Dr. Nicu Sebe for his kind support, encouragement, enthusiasm, and valuable and insightful comments. His guidance pushed me to sharpen my thinking and brought my work to a higher level. I could not have imagined having a better co-supervisor for my Ph.D. study.

Besides, I would like to thank the rest of my thesis committee: Assistance Professor Dr. Suree Pumrin, Associate Professor Dr. Nisachon Tangsangiumvisai, Associate Professor Dr. Charnchai Pluempitiwiriyawej, and Dr.Nattachai Watcharapinchai, for their encouragement, guidance, and insightful comments.

I thank my colleagues in the VTRG group: Sovann Chen, Eiei Tun, and Watchara Ruangsang, for their help and the wonderful friendship I have had in these years. I would like to acknowledge my mentors from my research study at the University of Trento: Enver Sangineto and Paolo Rota, for their kindness and valuable guidance for my research. I also thank my fellow lab mates at the University of Trento for their friendliness and our beautiful moments in Italy.

I would also like to thank my family: Nawarat Ganokratanaa, Kanokluk Ganokratanaa, and Pichid Poomvichit for their mental and physical support. You are always there for me, and the reason why I am who I am today.

Thittaporn Ganokratanaa

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1. Overview of the Proposed Methods

This thesis presents two manuscripts for video anomaly detection in crowded scenes as follows:

i) Unsupervised anomaly detection and localization based on deep spatiotemporal translation network (DSTN);
ii) Deep residual spatiotemporal translation network for video anomaly detection and localization (DR-STN).

Both manuscripts are submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Electrical Engineering, Faculty of Engineering, Chulalongkorn University.

First, we proposed DSTN to solve the anomaly detection problems, e.g., complex scenes, time consumption, small anomaly datasets, and object localization. DSTN focuses on translating spatial to temporal information to obtain comprehensive information for both the appearance and motion features of the objects based on the image-to-image translation framework. DSTN is designed with the pre- and post-processing procedures to enhance its detection and localization performance and to eliminate non-object and redundant features based on Generative Adversarial Network (GAN) in an unsupervised manner. The pre-processing procedures include a background removal method, a novel fused optical flow, a patch extraction, and a concatenated spatiotemporal features. The post-processing procedure is an edge wrapping method. The proposed DSTN can handle any possible anomalous event in the complex scenes without tuning parameters during testing, making it particularly robust while achieving good running time performance due to the less complexity of the model. However, since DSTN works on the patch that does not specify only the objects but also the background, it faces problems in generating the motion information of the objects. Furthermore, using GAN may incorrectly generate the synthesized output (i.e., shapes of objects) from input pattern as the discriminator learns only the temporal representation (optical flow) without the appearance information or spatial representation. Additionally, the background removal method is quite sensitive to illumination changes and the patch extraction does not always capture the full appearance of the objects. These issues lead to object localization, misdetection, and false-positive detection problems, in which the normal event is incorrectly detected as the abnormal event.

To solve these problems, we continue to propose a novel Deep Residual Spatiotemporal Translation Network (DR-STN) framework for video anomaly detection and localization in crowds with a novel Deep Residual convolutional GAN (DR-cGAN) model. In this work, we did not use any traditional approaches, i.e., the background subtraction method, to extract features as the first work. Instead, we focus more on how to comprehensively obtain the full appearance of the objects of interest (the moving foreground objects). Thus, we apply the powerful object detector, YOLOv4, to extract features of individual objects in the frame to feed into our model.

Additionally, we specifically improve the performance of the translation model to extensively learn the translation of objects of interest from appearance (spatial) to motion (temporal) representations. Unlike DSTN, we built a novel DR-cGAN as a learning model by adding the residual units and the residual connections between layers in the encoder and decoder of the generator. The architecture of DR-cGAN in the generator creates a wider model, enhancing the accuracy and quality of the synthesized image. Besides, an online hard negative mining (OHNM) and a semantic region merging are proposed to remove the false-positive results and combine the synthetic results of objects of interest into a full output frame, respectively. With the significance of the proposed DR-STN, we can overcome the misdetection and false-positive detection problems and achieve competitive performance over other state-of-the-art methods, including the proposed DSTN.

1.2. Motivation and Problem Statement

Currently, the surveillance system is rapidly increased popularity as modern technology, which can be used to ensure life safety and break the wall of security mistrust. This modern technology can be installed in any environment (indoor and outdoor perimeter security) for various applications such as health monitoring, facility protection, vandalism deterrence, parking lots, traffic monitoring, and public safety. Those CCTV cameras have been widely used to record real-time situations and store in the system to help reduce crimes, monitor the activities, and collect the evidence. However, CCTV is only performed as a post video forensic process by investigating previous events. Human resources are required to manually screen all scenes monitored by the CCTV, leading to the difficulty for the monitors to find an abnormal event in the scene, even working as a team as they might lose some vital information when taking a break or switching the viewpoint from the screen. In addition, this behavior may cause computer vision syndrome to the monitors and affect their concentration [32]. Video anomaly detection [15] is a complex and challenging task for use in surveillance systems. An abnormal event refers to an activity that raises suspicions by differing from the majority of activities in the scene (e.g., a person driving a car while others are walking on the street). It can occur in any realistic scenario, e.g., indoor, outdoor, crowded, and uncrowded scenes. It may lead to significant problems, such as a robbery, an area invasion, and a terrorist attack, causing a lot of damage, injury, or death [16]. Since CCTV cameras can only monitor these events, there is a need to build an intelligent CCTV analysis system to precisely detect and localize abnormal events in realistic scenarios for surveillance videos. The main challenge of building an intelligent CCTV system is how to precisely detect and locate all possible abnormal events in crowded and complex scenes.

There are several issues to be considered for designing an effective anomaly detection system. The first issue is about the complex scene, which is considered a challenging factor for VAD since anomalies can present in various environments but mostly in crowded scenes where there might be more than one anomaly at a time. Most works focus on the crowded scenes due to its high complexity of the multiple objects with occlusion and clutter rather than the uncrowded scene which is much simpler. This scene complexity challenge has drawn interest from researchers in the computer vision research area. To handle this significant issue, two main approaches have been implemented for anomaly detection in crowds: (i) a traditional-based

approach and (ii) a deep learning-based approach. With (i) the traditional-based approaches [3; 4; 6; 9; 10; 23; 25; 26; 47], appearance and motion (e.g., trajectories) are employed to detect the anomaly events based on hand-crafted features. Their accuracy depends on object appearances and motion cues which can be found by extracting features and tracking the objects [26]. Even though the traditional-based approaches can detect multiple objects in crowded scenes, they are more difficult to generalize to complex scenarios than deep learning-based approaches. Hence, deep learning-based approaches [11; 14; 19; 27; 35-37; 50; 52-54], have been considered as being more appropriate for handling complex scenes as they can improve the performance of anomaly detection and localization with the use of a learnable model of nonlinear transformation [11; 19; 46; 52].

Following the complex scene issue, time-consumption is one of the challenging issues for using an anomaly detection system in real-world applications. If high accuracy is required, the detection of multiple objects in crowded scenes is very time-consuming, asking for an inherent speed-accuracy tradeoff [2; 8; 26; 29; 30; 39; 51]. Recently, the deep learning-based approaches are considered for reducing the time complexity while retaining good detection performance due to the importance of low computational complexity and high detection accuracy for the surveillance videos [12; 18; 28; 31; 40-43; 48]. The recent advanced techniques for speeding up CNNs are parameter pruning and sharing and transferred convolutional filters [5]. Many works [12; 18; 31; 40; 43; 48] try to optimize the computational time of CNN-based algorithms by focusing on convolutional architectures. They try to reduce convolutional layers and redundant parameters that are not drastically impacting the model performance, resulting in a smaller and faster network than traditional CNN [24]. Several works [28; 42] use pre-trained fully convolutional networks (FCNs) as a regional feature extractor for semantic segmentation to reduce the computational complexity of the traditional CNN.

Another significant issue is the lack of abnormal training samples in the datasets, leading to insufficient training information and the difficulty of designing good classifiers for indicating abnormal events. Besides, there is no chance to train for all possible abnormal events since they can occur unpredictably in real-world environments. An abnormal event in one dataset may be considered as a regular event in another dataset. Therefore, recent works focus on unsupervised deep learning-based approaches to overcome this problem, e.g., generative approaches [27; 36; 37].

Besides, the low performance of object localization in pixel-level anomaly detection is also addressed in the literature. Most works achieve high accuracy (measured by Area Under the Curve (AUC)) only on anomaly detection in a frame-level evaluation. In contrast, the AUC of object localization in the pixel-level evaluation is much lower. This issue occurs because of the lack of sufficient features of the objects of interest (e.g., appearance and motion patterns of foreground objects) for model training. These features should be extracted during training to learn the model. Precisely, the reasons lie in this problem are as follows:

i) A full-frame is fed into the model without prior knowledge of the objects, making it difficult for the model to correctly learn the mapping

from the appearance to the temporal information of objects and resulting in misdetection of abnormal objects [27; 35-37];

ii) Patch extraction is not effective enough to collect comprehensive features of the normal object for the model to learn its characteristics [11; 52]. Fig. 1.1 shows two examples of the missed detection of abnormal objects, including cycling in Fig. 1.1 (a) and a vehicle in Fig. 1.1 (b).

(a) Misdetection of cycling (b) Misdetection of vehicle

Fig. 1.1 The misdetection results on the UCSD Ped1 [3] and Ped2 dataset [35].

Finally, all these issues mentioned above lead to false-positive anomaly detection that decreases the accuracy and reliability of the system, e.g., AUC and pixel accuracy. Lots of works in anomaly detection research [27; 35-37] have faced false-positive results in their final output in which the system incorrectly detects regular events as abnormal ones. This problem is significant to be handled to enhance the overall performance of the system. Current works [34; 45] aim to enhance the accuracy with the use of a supervised learning method. However, even a supervised learning method provides high accuracy; it needs data labeling for all samples which is not suitable for video anomaly detection since anomalies are varied and unpredictable. The examples of false-positive detection results on pedestrians are illustrated in Fig. 1.2 [15].

Fig. 1.2 The false-positive detection results on UCSD Ped1 and Ped2 dataset [15].

According to these considerations, unsupervised deep learning-based approaches are considered the most suitable solution for handling the complex anomaly scenarios without defining any data labeling for anomaly samples. Unsupervised learning aims to learn only regular events since they are the majority of patterns in the scene. Any unknown patterns will be considered as anomalies by the significant difference in distance from the normal patterns. With the significance of unsupervised learning, Generative Adversarial Network (GAN) has gained more attention in anomaly detection research due to its outstanding performance in constructing images, affording data augmentation, and dealing with implicit data in complex scenarios. GAN consists of two competing networks: generator *G* and discriminator *D*. In common GAN [17], during training, *G* aims to generate the synthetic data that looks real and fools *D* that its generated data is real, while *D* tries to distingue whether its input is real or fake. *G* is the only network used to reconstruct new data at testing time. In addition, GAN allows the convolutional networks in the generator to generate data on different representations, e.g., sketch image to realistic image and vice versa. With the use of convolutional networks, many works have tried to achieve better performance on image reconstruction and to overcome vanishing gradients. He, *et al*., [20] proposed skip connection with identity mapping [21] instead of using it in FCNs [28]. U-Net is proposed in [38] to enhance the accuracy of image segmentation by concatenating feature maps from low- and high-level semantic information, achieving good segmentation results on the biomedical image. Isola *et al*. [22] proposed the translation of a sketch image to a realistic image based on conditional GANs (cGANs) using U-Net architecture as the generator and a patch-based discriminator.

This thesis presents two novel frameworks for anomaly detection and localization for surveillance videos. Both of the proposed methods have been tested on three publicly available benchmarks, including UCSD pedestrian, CUHK Avenue, and UMN datasets. We first proposed a novel Deep Spatiotemporal Translation Network (DSTN) with the main contributions listed below:

(i) We propose DSTN, a novel unsupervised deep learning architecture based on GAN, to transform information from the spatial to the temporal domain for addressing the anomaly detection and localization tasks in crowded scenes for surveillance videos. Our DSTN automatically learns the normal samples without varying any parameters, presenting remarkable advantages over previous traditional methods;

(ii) We propose a novel fusion of a background removal frame and a real dense optical flow frame to eliminate noise from appearance and motion representations and acquire explicit boundaries of foreground objects;

(iii) We propose concatenated spatiotemporal features to combine important feature information obtained from the new design of patch extraction requiring extensive low-level appearance and motion features;

(iv) This paper presents the first attempt to improve anomaly object localization at the pixel level by introducing an Edge Wrapping technique at the final stage of the framework.

This proposed DSTN is different from the early works [3; 4; 6; 9; 10; 23; 25; 26; 47] that focus on hand-crafted features since we can handle any possible anomalous event in the complex scenes without tuning parameters during testing, making the proposed DSTN particularly robust, while achieving good running time performance. Additionally, the proposed DSTN is different from [27; 35-37] that rely on deep learning-based approaches because DSTN is additionally equipped with pre- and post-processing procedures to enhance its detection and localization performance and to eliminate non-object and redundant features.

However, according to the experimental results of DSTN, there is still room for improvement in object localization and false-positive detection problems. Thus, to solve these problems, we continue to propose a novel Deep Residual Spatiotemporal Translation Network (DR-STN) framework for video anomaly detection and localization in crowds with a novel Deep Residual cGAN (DR-cGAN) model inspired by the deep residual learning [20] and image-to-image translation [22]. Different from DSTN and the previous works [27; 36; 37] which are based on the framework in [22], the DR-cGAN is built by adding the residual units and the residual connections in between layers in the encoder and decoder to learn the translation of objects of interest from appearance (spatial) to motion (temporal) representations. Our architecture in the generator creates a wider model, enhancing the accuracy and quality of the synthesized image. A powerful object detector [1] is applied to extract the objects of interest in the frame to feed into our DR-cGAN. Besides, an online hard negative mining (OHNM) and a semantic region merging are proposed to remove the false-positive results and combine the synthetic results of objects of interest into a full output frame, respectively. We compare our proposed DR-STN method with other state-of-the-art works, showing the superior performance of the proposed method in both frame-level and pixel-level evaluations. The DR-STN contribution can be concluded as four-fold:

(i) Our unsupervised DR-STN learns only normal events without using any hand-crafted features and effectively translate comprehensive information of the objects of interest from appearance to motion representations in crowded scenes;

(ii) We propose DR-cGAN, a novel end-to-end unsupervised deep residual connection network, to improve perceptual information of reconstructed images from the generator. DR-cGAN provides a wider network that extensively passes information from the previous to the next layer of encoder and decoder. To the best of our knowledge, this is the first attempt to build the deep residual connections (projection and identity shortcuts) on the U-Net architecture of cGAN for VAD;

(iii) We introduce the object detector as the pre-processing process to extract only the objects of interest to feed into the DR-cGAN model to help learn the pattern of normal objects. This approach provides better object localization for the pixel level. To our knowledge, we are the first to integrate the object detector with GAN;

(iv) We introduce OHNM and a semantic region merging as the post-processing processes to eliminate the false-positives without retraining the model and integrate the intensity of objects for the final anomaly output, providing more reliable and remarkable results than the state-of-the-art.

The outlines of this thesis consist of five chapters. Chapter two describes the background and literature reviews. Chapter three presents the methodology, which is divided into two main frameworks (DSTN and DR-STN). Chapter four shows the implementation details and the experimental results for each framework compared with other state-of-the-art works along with the discussions. Chapter five provides a conclusion and directions for future work.

1.3. Objectives

1. To investigate recent trends and datasets of anomaly detection in realistic scenarios.
2. To develop anomaly detection and localization for surveillance videos.
3. To enhance the ability of a CCTV surveillance system for security.

1.4. Scope of Work

1. Apply machine learning algorithms and determine the suitable algorithm for implementing in appearance and motion features
2. Use video from a stationary camera in the static environment.
3. Propose a framework on spatiotemporal anomaly detection that can detect and localize anomalous events for surveillance videos.
4. Evaluate the performance of the proposed algorithm on the ROC curve, with AUC and EER indicated, and compare with other state-of-the-art methods.

1.5. Research Benefits

1. The anomaly detection and localization system can be applied with realistic and crowded scenarios.
2. The accuracy of anomaly detection and localization is improved over other state-of-the-art methods while maintaining good running time, making the system more reliable and useful for real-world applications.
3. The system can be applied for security in smart city applications using surveillance videos.

CHAPTER 2
PROPOSED METHODS

This thesis presents two novel proposed methods: i) Deep Spatiotemporal Translation Network (DSTN) and ii) Deep Residual Spatiotemporal Translation Network (DR-STN). In this chapter, we introduce our proposed methods as the original manuscript in Section 2.1 and Section 2.2 for DSTN and DR-STN, respectively. Section 2.3 explains the training procedures in detail. Section 2.4 provides the evaluation criteria and Section 2.5 shows the discussion for both of our proposed methods.

Journal Information

Our first proposed method, DSTN, is published in IEEE Access. The detail of this work is presented as follows.

Topic: Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network
Authors: Thittaporn Ganokratanaa[1], (Graduate Student Member, IEEE), Supavadee Aramvith[1], (Senior Member, IEEE), and Nicu Sebe[2], (Senior Member, IEEE)
Address: [1]Department of Electrical Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Address: [2]Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy
Corresponding author: Supavadee Aramvith (supavadee.a@chula.ac.th)
Journal name: IEEE Access
Received: February 11, 2020
Accepted: February 29, 2020
Date of publication: March 10, 2020
Date of current version: March 20, 2020
Digital Object Identifier: 10.1109/ACCESS.2020.2979869

2.1. Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network

Abstract Anomaly detection is of great significance for intelligent surveillance videos. Current works typically struggle with object detection and localization problems due to crowded and complex scenes. Hence, we propose a Deep Spatiotemporal Translation Network (DSTN), a novel unsupervised anomaly detection and localization method based on Generative Adversarial Network (GAN) and Edge Wrapping [29]. In training, we use only the frames of normal events in order to generate their corresponding dense optical flow as temporal features. During testing, since all the video sequences are input into the system, unknown events are considered as anomalous events due to the fact that the model knows only the normal patterns. To benefit from the information provided by both appearance and motion features, we introduce (i) a novel fusion of background removal and real optical flow frames with (ii) a concatenation of the original and background removal frames. We improve the performance of anomaly localization in the pixel-level evaluation by proposing (iii) the Edge Wrapping to reduce the noise and suppress non-related edges of abnormal objects. Our DSTN has been tested on publicly available anomaly datasets, including UCSD pedestrian, UMN, and CUHK Avenue. The results show that it outperforms other state-of-the-art algorithms with respect to the frame-level evaluation, the pixel-level evaluation, and the time complexity for abnormal object detection and localization tasks.

Keywords anomaly detection, anomaly localization, spatiotemporal, unsupervised learning, video surveillance.

2.1.1. Introduction

Surveillance has rapidly gained increasing popularity as a modern technology, which can be used to ensure life safety and break the wall of security mistrust. Closed-Circuit Television (CCTV) cameras have been widely used for monitoring and recording situations, providing evidence to the surveillance system. According to [1], the growth of surveillance videos has increased by 9.3 percent in 2019. However, the CCTV cameras are mostly used for the post-video forensic process by allowing the investigation of previous events [2]. This means that the CCTV camera feed still needs to be manually monitored by a human operator for any abnormal events which can unpredictably occur in the scene. An abnormal or anomalous event refers to an activity that raises suspicions by differing from the majority of the activities. It can possibly occur in any realistic scenario (e.g., indoor, outdoor, crowded, and uncrowded scenes) and may lead to major problems, such as an area invasion, a robbery, and a terrorist attack, causing a lot of damage, injury, or death [3]. According to the performance of CCTV cameras [2], there is a need to build intelligent systems to analyze abnormal events in realistic scenes for surveillance videos. The main challenge of building an intelligent CCTV system is how to precisely detect and locate all possible abnormal events in crowded and complex scenes. To design an effective anomaly detection and localization system [7], [10], [13], [44], there are four main issues to be considered: the complex scene, time-consumption, dataset, and object localization.

The complex or crowded scene may contain multiple objects with clutter and occlusions which are difficult to deal with. Besides, it is more challenging than the uncrowded scene as it has higher complexity. This scene complexity challenge has drawn interest from researchers in computer vision research area [4]-[14], [19], [20], [45]-[48]. To handle this significant issue, two main approaches have been implemented for anomaly detection in crowds: (i) a traditional-based approach and (ii) a deep learning-based approach. With (i) the traditional-based approaches [20]-[28], [35], appearance and motion (e.g., trajectories) are employed to detect the anomaly events based on hand-crafted features. Their accuracy depends on object appearances and motion cues which can be found by extracting features and tracking the objects [20]. Even though the traditional-based approaches are able to detect multiple objects in crowded scenes, they are more difficult to generalize to complex scenarios than deep learning-based approaches. Hence, deep learning-based approaches [4], [7], [10]-[17], [29], have been considered as being more appropriate for handling complex scenes as they are able to improve the performance of anomaly detection and localization with the use of a learnable model of nonlinear transformation [7], [8], [13], [15].

Following the complex scene issue, time-consumption is one of the challenging issues for the use of an anomaly detection system in real-world applications. If high accuracy is required, the detection of multiple objects in crowded scenes is very time-consuming, asking for an inherent speed-accuracy tradeoff [5], [18], [20], [38]-[41]. Recently, the deep learning-based approaches were considered for reducing the time complexity while retaining good detection performance due to the importance of low computational complexity and high detection accuracy for the surveillance videos [44], [46], [52]-[57], [59]. The recent advanced techniques for speeding up CNNs are parameter pruning and sharing and transferred convolutional filters [30]. Many works [52]-[57] try to optimize the computational time of CNN-based algorithms, focusing on convolutional architectures by reducing convolutional layers and redundant parameters that are not drastically impacting the model performance, resulting in a smaller and faster network compared to the traditional CNN [58]. Several works [46], [59] use pre-trained fully convolutional networks (FCNs) as a regional feature extractor for semantic segmentation to help to reduce the computational complexity of the traditional CNN.

Another significant issue is the lack of abnormal training samples in the datasets, leading to insufficient training information and the difficulty of designing good classifiers for indicating abnormal events. In addition, there is no chance to train for all possible abnormal events since they can occur unpredictably in real-world environments. Therefore, recent works focus on unsupervised deep learning-based approaches, such as generative approaches [11], [14], [16], to overcome this problem.

Finally, the low performance of object localization in pixel-level anomaly detection is also addressed in the literature. Most works achieve high accuracy (measured by Area Under the Curve (AUC)) only on anomaly detection in a frame-level evaluation, while the AUC of object localization in the pixel-level evaluation is much lower. This occurs because of the lack of sufficient features of the objects of interest (e.g., appearance and motion patterns of foreground objects) for model training. These features should be extracted during training in order to learn the

model. Specifically, the full input frame is fed into the model without prior knowledge of the objects in the scene, making it difficult for the model to correctly learn the mapping from the appearance to the temporal information of objects and resulting in misdetection and false detection of abnormal objects [10], [11], [14], [16]. Current works try to improve the performance of object localization by isolating patches for deeper feature extraction [7], [13].

Following these considerations, we propose a novel unsupervised spatiotemporal translation based on Generative Adversarial Network (GAN) for anomaly detection and localization in crowded scenes. Our proposed framework, named Deep Spatiotemporal Translation Network (DSTN), is different from the early works [20]-[28], [35] that focus on hand-crafted features since we can handle any possible anomalous event in the complex scenes without tuning parameters during testing, making the proposed DSTN particularly robust, while achieving good running time performance. Additionally, the proposed DSTN is different from [10], [11], [14], [16] that rely on deep learning-based approaches because DSTN is additionally equipped with pre- and post-processing procedures to enhance its detection and localization performance and to eliminate non-object and redundant features. The proposed DSTN has been tested on three challenging anomaly benchmark datasets and compared with other state-of-the-art methods, showing the effectiveness of our proposed framework in terms of both accuracy and time complexity.

To conclude, our main contributions are four-fold:

(i) We propose DSTN, a novel unsupervised deep learning architecture based on GAN, to transform information from the spatial to the temporal domain for addressing the anomaly detection and localization tasks in crowded scenes for surveillance videos. Our DSTN automatically learns the normal samples without varying any parameters, presenting remarkable advantages over previous traditional methods;

(ii) We propose a novel fusion of a background removal frame and a real dense optical flow frame in order to eliminate noise from appearance and motion representations and acquire explicit boundaries of foreground objects;

(iii) We propose concatenated spatiotemporal features to combine important feature information obtained from the new design of patch extraction requiring extensive low-level appearance and motion features;

(iv) This paper presents the first attempt to improve anomaly object localization at the pixel level by introducing an Edge Wrapping technique at the final stage of the framework.

This paper consists of five sections. We review related works in Section 2.1.2 and present our proposed method, DSTN, in Section 2.1.3. Section 2.1.4 shows experimental results compared with several state-of-the-art algorithms and analysis of DSTN. Section 2.1.5 provides a conclusion and directions for future work.

2.1.2. Related Works

The related works of video anomaly detection can be grouped into two main categories: traditional-based and deep learning-based approaches.

A. Traditional-based Approaches

In this section, we focus on the frameworks that rely on hand-crafted features. These can be divided into two types: temporal (motion) approach and spatiotemporal (appearance and motion) approach. For the temporal approach, X. Tang, *et al.*, [21] proposed abnormal event detection based on motion attention using sparse coding by comparing current regions with neighboring regions to generate a motion attention map. Sparse reconstruction is proposed in [22] by extracting the optical flow and applying the Histogram of Maximal Optical Flow Projections with a sparse representation to generate the dictionary of the normal event. Recently, motion energy [23] and motion entropy [24] were proposed to characterize the abnormal event based on its temporal information only. Overall, the temporal approach is suitable only when dealing with scenes that have a simple background and a low number of foreground objects.

The spatiotemporal approach combines information from both appearance and motion features, making it more robust to complex scenes than the temporal approach. This approach has been addressed by using various local feature descriptors, including Gaussian Mixture Model (GMM) [25], Histogram of Oriented Gradients (HOG) [42], Histogram of Optical Flow (HOF) [42], Histogram of Optical Flow Orientation (HOFO) [43] and Magnitude (HOFM) [26], Gaussian regression [27], and Optical Flow (OF) [35] with Principal Component Analysis (PCA) [60], which can be grouped by applying classifier methods such as K-Means [28] and Bags of Visual Words (BoVW) [27]. However, the problem with the traditional-based approaches is that they rely on hand-crafted features that limit their generalization to other anomalous events.

B. Deep Learning-based Approaches

Deep learning-based approaches have gained wide popularity as they consistently achieve higher performance than the traditional state-of-the-art approaches [20]-[28], [35] in learning high-level features from a large amount of data and dealing with complex problems such as object detection and recognition and image classification. These approaches can be categorized based on the level of supervision involved. The supervised learning requires labeled data, causing difficulty in detecting unpredictable anomalous events in real-world use cases. Similarly to supervised learning, semi-supervised learning still needs some labeled samples to train the model [15], [29]. In contrast, unsupervised learning is able to handle various anomalous events without any labeling requirement, making it the most suitable approach for anomaly detection in real-world applications. Most frameworks of anomaly detection are based on unsupervised learning because of its high performance in terms of flexibility and reliability of anomaly detection and localization.

Unsupervised learning has been investigated for training in recognition tasks by using CNNs [10], [30]. Ravanbakhsh, *et al*., [10] proposed a Binary Quantization Layer as a final layer to plug into the top of the network for gathering motion information of abnormality. Xu, *et al*., [7] proposed an Appearance and Motion DeepNet (AMDN) for detecting anomalous events in the videos. The discriminative feature is used instead of hand-crafted features by applying Stacked Denoising AutoEncoders (SDAE) [61]. Fan, *et al*., [13] proposed a two-stream variational autoencoder by using Gaussian Mixture Model (GMM) with a Fully Convolutional Network (FCN) [46] at the bottleneck between encoder and decoder to compute the spatial and temporal score. In [17], the authors proposed a neural network for anomaly detection in video surveillance by using three processing blocks; feature learning, sparse representation, and dictionary learning, and also proposed and reformulate an adaptive iterative hard-thresholding algorithm as a new long short-term memory (LSTM). Liu, *et al.*, [16] introduced a video prediction framework for anomaly detection using GANs for training normal events, where the abnormal event is detected by leveraging the difference between a predicted future frame and its ground truth. A future frame is predicted based on appearance and motion feature information. Hasan, *et al*., [15] proposed an end-to-end deep learning framework for abnormal detection using a Convolutional Autoencoder for learning the normal events in crowds and generating the appearance of the normal pattern at testing time, where the abnormality score is measured by the reconstruction error. Similarly to [15], the authors in [14] recently proposed Generative Adversarial Nets (GANs) for an abnormal cross-channel event in which the discriminator is directly used as the final classifiers as an end-to-end anomaly detector. The difference between [15] and [14] is that the latter is based on the interplay between generator and discriminator networks. Another study [11] is dealing with the abnormal event detection in videos using GANs to train only normal events with the use of two networks, (i) generating the optical flow from the frame and (ii) generating the frame from the optical flow.

Following related works, GANs are an outstanding approach that achieves high performance in the anomaly detection task. GANs are a great solution to overcome classification problems as they are able to find the significant features in the frames without any predefined anomaly types. The fundamental architecture of GANs [31]-[33] comprises two networks, the generator G for generating synthetic data z that are likely to come from the same data-generating distribution as the real samples and discriminator D for discriminating whether the input data are real or fake data generated by G. More specifically, G generates a new image e from random noise z, while D tries to distinguish a real image x from e. In addition, D does its best to classify the synthetic image generated from G as the fake image, while G tries to fool D by producing the synthetic image which looks real, making it challenging to be differentiated. The parameters of G are optimized by updating only with gradients flowing through D in order to maximize the probability of $D(G(z))$ so that D makes a mistake by classifying the synthetic image as the real image, making G efficient in generating images [31]. With enough training time and capacity, G and D are incapable to improve because the probability distributions of the generator and the real data are equal, meaning that D can no longer distinguish between the two

distributions. GANs also afford data augmentation and implicit data management due to *D*, which benefits the deeper training of *G* on the same small anomaly dataset without training additional classifiers.

Even though GANs outperform several state-of-the-art works, there is still room for improvement of the object localization at the pixel-level evaluation as most of the current works [11], [14], [16] can significantly improve only the performance of frame-level evaluation for the object detection. Thus, apart from the anomaly detection in the frame-level evaluation, our DSTN specifically focuses on improving the performance of anomaly localization at the pixel level. Our model is implemented based on the image-to-image translation framework using the U-Net architecture with skip connections proposed in [34], using the generator with a patch-based discriminator and allowing transforming images to other representations. We take this ability to generate optical flow from raw pixel images by using GANs, so our *G* is used for spatiotemporal transformation. The difference between our DSTN and [34] is that we use *G* to learn the normal event to understand its pattern instead of using *G* to generate a realistic image. At testing time, *G* is only used for generating appearance (spatial) and motion (temporal) features of the normal event from the input image. With this generated frame, we can simply detect the anomalous areas by comparing the generated frame with the real frame.

2.1.3. DSTN for Anomaly Detection and Localization

A. Overview

Our DSTN consists of four main phases, including feature collection, spatiotemporal translation, differentiation, and edge wrapping for the object localization. Fig. 2.1 shows the overview of DSTN that can translate information from the spatial or appearance to the temporal or motion representations.

In the feature collection, we introduce a background removal method, a novel fusion between the background removal frame f_{BR} and the dense inverse search optical flow frame OF_{dis}, a patch extraction, and a concatenation between the original frame f and the background removal frame f_{BR}. Specifically, the novel fusion of f_{BR} and OF_{dis} is proposed to obtain the prior knowledge of the foreground objects in the scene for the model training. To our knowledge, this is the first attempt to fuse f_{BR} with OF_{dis} to enhance the performance of feature extraction of both appearance and motion patterns for anomaly detection and localization tasks. The background removal method provides the complete shape appearance for each moving foreground object while the dense inverse search optical flow method provides the temporal information corresponding to its input. However, OF_{dis} contains noises that may affect the quality of image generation during GAN training. Thus, due to the performance of the background removal method, we manage to fuse it with OF_{dis} to get rid of noises and make the edges of each foreground object sharper and more precise. The output of this fusion, represented as OF_{fus}, is considered as the real dense optical flow. The fusion of these simple but yet effective techniques provides remarkably good results in noise reduction in OF_{dis} which facilitates *G* to generate the desired temporal output.

Apart from the fusion, patch extraction is also applied to each frame before input it into a spatiotemporal deep GAN model, consisting of competing *G* and *D*

networks. Additionally, the concatenation between patches of f and f_{BR} is introduced to capture more information on the moving foreground objects in the scene. This concatenation is specifically designed for delivering the low-level appearance of the moving objects along with their temporal information within the scenes, assisting the model to learn to map the appearance information to temporal information in a more comprehensive way. To conclude, these feature collection methods are introduced in order to obtain better input data to feed into the spatiotemporal deep GAN model. In this way, the model is able to translate the information from the spatial or appearance to the temporal or motion representations efficiently.

In training, G learns only the normal events and translates the spatial to temporal image representations depending on the real dense optical flow. The output of G is a generated dense optical flow, represented as OF_{gen}. In D, it tries to discriminate the patches of real dense optical flow OF_{fus} from the patches of generated dense optical flow OF_{gen} while G tries to fool D by producing more OF_{gen} that is difficult to be discriminated. If D discriminates the patches of OF_{gen} as a fake or wrong image, G will regenerate OF_{gen} until the model reaches the target objective.

In testing, we input all video sequences so that G generates the generated dense optical flow of anomalous events based on the normal events. The anomalous events can be detected by differentiating the pixel intensity of the real optical flow OF_{fus} and the generated dense optical flow OF_{gen}. Finally, we analyze the final output with a novel edge wrapping technique to localize pixels that belong to the anomalous objects. The details of our DSTN are described in the following sections.

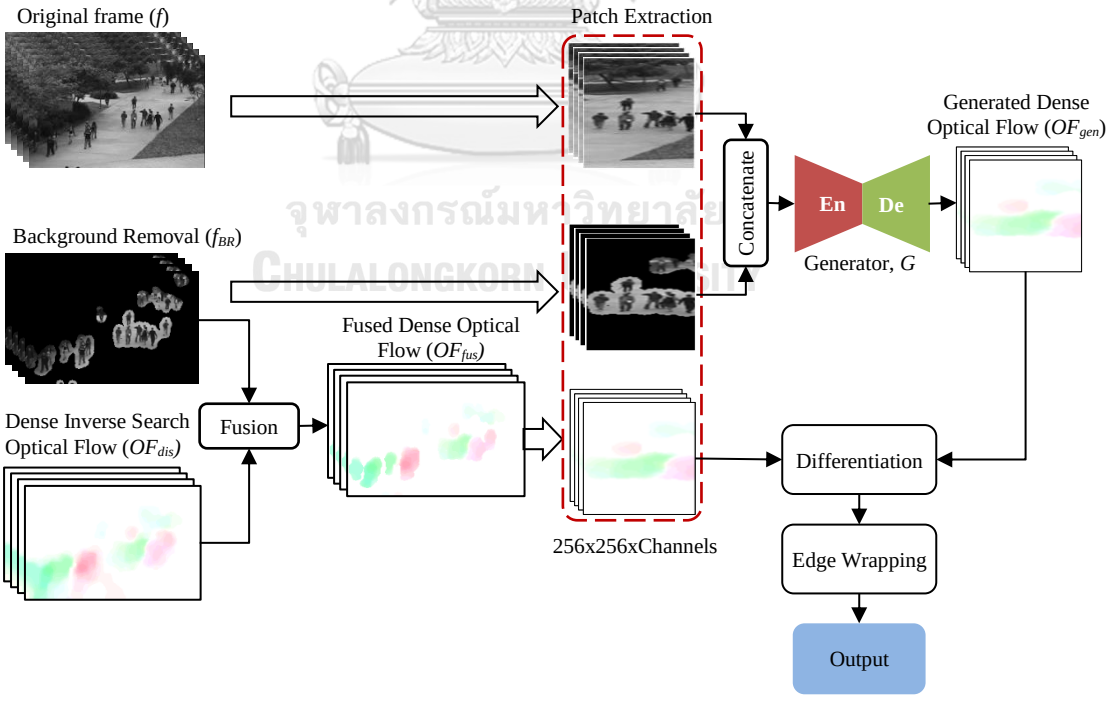


Fig. 2.1 The overview of our proposed framework.

B. Feature Collection

This is the most important initial task for obtaining the characteristics of objects in the scene. The details of the feature collection approaches are described in the following sections.

1) Background Removal

As we consider the real-world situations recorded from the static CCTV cameras, the objects of interest are only the moving foreground objects. In this case, where the background is stationary, we introduce a background removal method, represented as f_{BR}, to extract only the moving foreground object features and to remove unimportant pixels in the background. The f_{BR} image is the representation for appearance information which can be obtained by computing the frame absolute difference between two consecutive frames as shown in Eq. (2.1):

$$f_{BR} = |f_t - f_{t-1}|, \tag{2.1}$$

where f_t is the current frame and f_t-1 is the previous frame of a video sequence. In addition, to achieve more appearance features, we implement a binarization on f_{BR} and then concatenate the binarized f_{BR} with f. This concatenation provides more appearance information on the foreground objects of the binarized f_{BR} image, assisting in the learning of G.

In Section 2.1.4, we compare the background removal method with a popular technique for background subtraction, i.e., the GMM-based background subtraction [67]. The experimental results clearly show that the background removal method is more effective for anomaly detection in our experiments as it can preserve more appearance information of the moving foreground objects than the GMM-based background subtraction method.

2) Fused Optical Flow

Optical flow (*OF*) is a technique that is used to detect and track the motion of the object of interest obtained from two consecutive frames; f_t and f_{t-1}. Since we consider the motion of foreground objects in terms of running time and accuracy, we choose Dense Inverse Search (*DIS*), calculated by [37], to generate dense optical flow for our DSTN due to its high performance in real-world applications, including low complexity, less time-consumption, and accurate motion detection and tracking. Then we obtain the real dense optical flow generated from the *DIS* technique named OF_{dis}. However, OF_{dis} contains some noise dispersed in the scene apart from the objects. Hence, to eliminate it, we propose a novel fusion of f_{BR} and OF_{dis} for appearance and motion, respectively, by integrating these frames to acquire clear foreground object boundaries for the use of object detection, tracking, and localization. Equation (2.2) shows how to eliminate the noise in *DIS* optical flow by knowing the information of f_{BR} where its pixel values equal to 0 or 255. Then, the fusion OF_{fus} is defined by applying image masking of f_{BR} on OF_{dis} to change its pixel values. Thus, we obtain the new OF_{dis} represented as OF_{fus} that provides better boundaries of the foreground regions. The output of this fusion OF_{fus} is formulated as below:

$$OF_{fus} = OF_{dis} \left\lfloor \frac{f_{BR}}{f_{BR}+\zeta} \right\rfloor, \quad (2.2)$$

where ζ is a constant value.

3) Patch Extraction

Patch extraction is important for the feature collection process as it helps to obtain more appearance and motion features. Additionally, the patch extraction allows the model to learn the pattern of local pixels in the scene, resulting in achieving better feature collection performance than extracting the features directly from the full image frame. To extract the patch, we consider the magnitude and direction of the dense optical flow based on the moving objects in the scene. In addition, each moving object is needed to be detected in its full-size appearance at the current frame along with its motion and direction from the frame-by-frame image. The patch size can be determined by $\frac{w}{a} \times h \times c_p$, where w is the width of the frame, h is the height of the frame, a is a scale value, and c_p represents the number of channels. All patch elements are normalized into a range of [-1, 1]. In our DSTN, the patch is extracted by applying a sliding window approach with a stride d to feed into the spatiotemporal translation deep GAN model from its input frames, including f, f_{BR}, and OF_{fus}. While f and f_{BR} are the input for G, OF_{fus} is the input for D. This patch extraction provides the appearance of the moving foreground object along with its motion and direction in the scene, assisting in further processing of the concatenated spatiotemporal features.

4) Concatenated Spatiotemporal Features

In the learning of G, it is important to provide enough information on the appearance to make G understand the features of normal patterns in the scene extensively. The overview of data preparation of concatenated spatiotemporal features is shown in Fig. 2.2. To achieve more low-level information on the appearance, we propose the concatenation of f and f_{BR} patches for the input of G to learn the normal events. Specifically, the number of channels of the concatenated f and f_{BR} frames is 2 ($c_p = 2$). As a result, the G model obtains efficient information since f_{BR} gives the contour edge information of the foreground objects while f gives the overall information in the scene. After inputting the concatenated f and f_{BR} frames, the spatiotemporal translation deep GAN will learn this information until it reaches the desired temporal information as the target output.

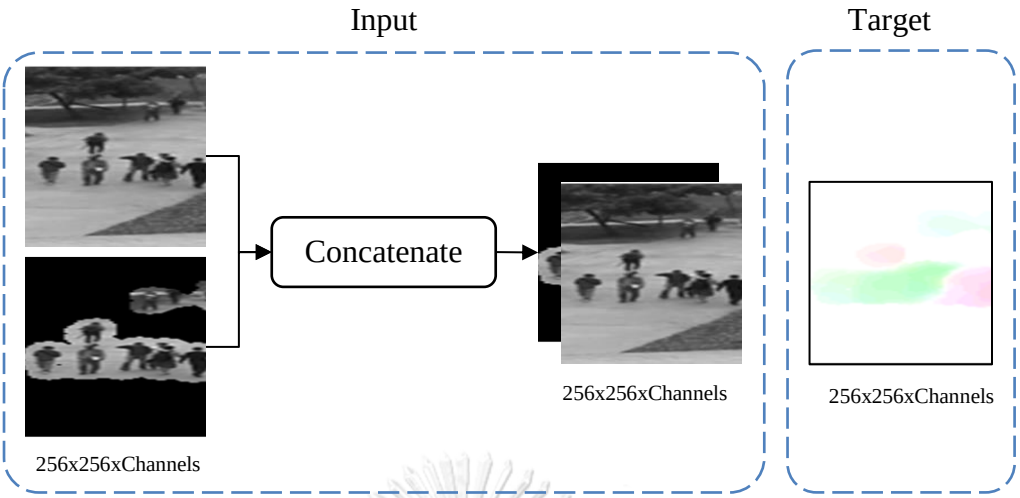


Fig. 2.2 The data preparation of concatenated spatiotemporal features for the temporal target output.

C. Spatiotemporal Translation Model

This work investigates the deep Generative Adversarial Network (GAN) as inspired by image-to-image translation [34] based on U-Net architecture [63]. The GAN network consists of two cores: generator *G* and discriminator *D*. It aims to learn a mapping from the inputs of spatial representation (*f* and f_{BR}) to the output of temporal representation (OF_{gen}).

1) Generator

The generator *G* model is the main model of the DSTN since it is applied in both training and testing. In the basic GAN [31]-[33], *G* takes an image *x* and a random noise *z* as the input. It generates the output image *e* with the same resolution as the input *x* but representing the different channel, using the random noise *z*, $e = G(x,z)$. In our DSTN, *G* tries to transform the spatial representation image of the concatenated *f* and f_{BR} frames to the temporal representation image of generated dense optical flow frame OF_{gen}. However, in this work, the random noise *z* is not effective to *G* because the input of *G* is the spatial representation data and *G* tries to generate the temporal representation data based on the input data. Hence, this model has been designed to include the drop-out instead of the additional Gaussian noise *z*. The Drop-Out algorithm [34] is applied within Batch Normalization [62] in the Decoder, resulting *e* to be reformulated as $e = G(x)$.

Specifically, on the generator architecture, *G* consists of Encoder (*En*) and Decoder (*De*) [34]. Fig. 2.3 shows the Encoder and Decoder deep network architecture constructed by a residual connection. The Encoder network has been constructed from Convolution (Conv), Batch Normalization (BN), and the Activation Leaky-ReLU (L-ReLU). On the other hand, the Decoder network has been built from De-Convolution (De-Conv), Batch Normalization (BN) with Drop-Out, and the Activation ReLU that allows the model to speed up the learning to suffuse the color space of the training distribution [33]. This residual connection or a skip connection directly connects the encoder layers to the decoder layers based on the architecture of

U-Net [63]. The layers of the Encoder and the Decoder are indicated in Fig.2.4. In detail, the residual connection is inserted between each layer *l* at the Encoder and layer *t-l* at the Decoder, where *t* is the total number of layers. It allows the information to flow through the initial layer to the last layer by concatenating all channels at layer *l* with layer *t-l*. In other words, it often allows one to use smaller networks that are easier to optimize and provide higher quality results of image transformation with a lower complexity cost than the deep convolutional network such as VGG nets [64], [65]. The analysis of the residual connection is discussed in Section 2.1.4.

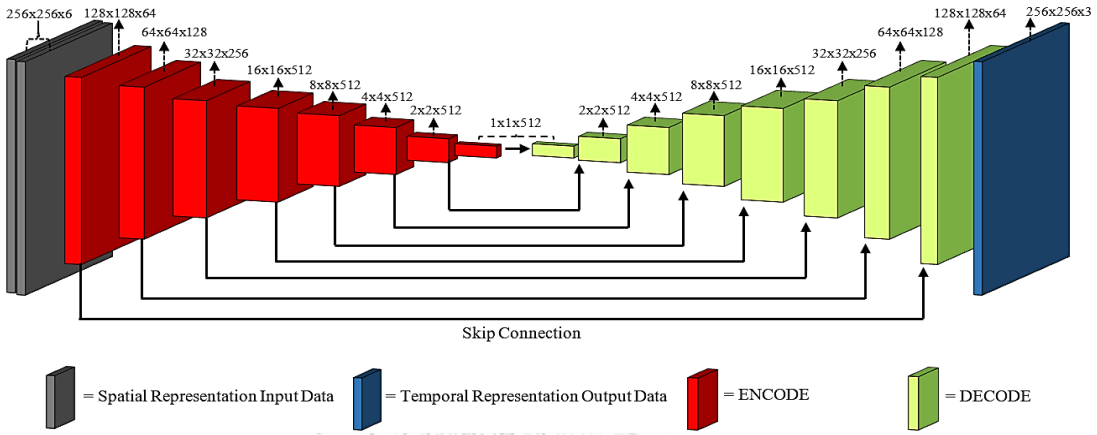


Fig. 2.3 An overview of our generator architecture in which its input is a spatial representation and its output is a temporal representation.

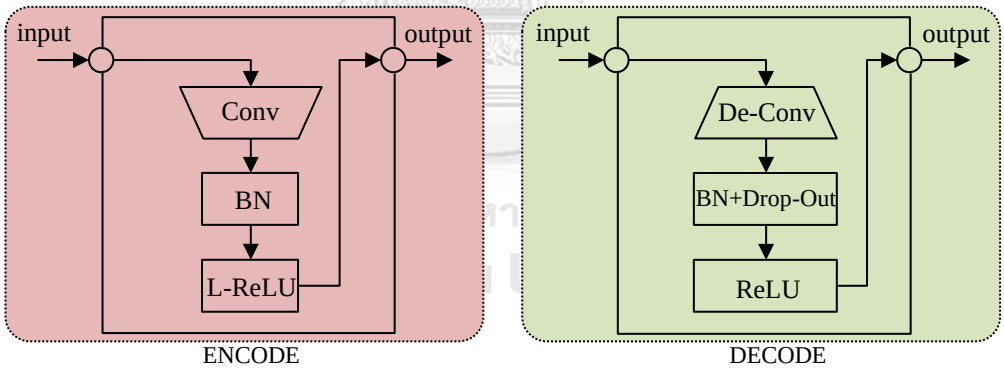


Fig.2.4 Encoder and Decoder Architectures

2) Discriminator

The discriminator (*D*) is used only at the training time. There are two inputs for *D* to discriminate: the fake patch of OF_{gen} ($OF_{gen} = e$) and the real patch of OF_{fus} ($y = OF_{fus}$) obtained from the fusion between f_{BR} and OF_{dis}. The job of *D* is to check whether *G* can produce OF_{gen} or not, and how it looks like comparing with OF_{fus}. *D* provides a scalar output denoting the probability of the inputs (OF_{fus}, OF_{gen}) for determining the real data.

In *D*, we use PatchGAN which is constructed as shown in Fig. 2.5. The PatchGAN can produce a faster training GAN than the full image discriminator net (e.g., 256×256) because it applies to each partial patch of the image. For the

implementation of D, the OF_{fus} image is subsampled from the resolution of 256×256 pixels to 64×64 pixels. Hence, the total patches of OF_{fus} image are 16 patches. These 16 patches are passed through the PatchGAN model to decide whether OF_{gen} from G is True or False. We analyze the impact of using 64×64 PatchGAN in Section 2.1.4, where we compare the performance of different sizes of PatchGAN in terms of FCN-scores and visual quality outputs.

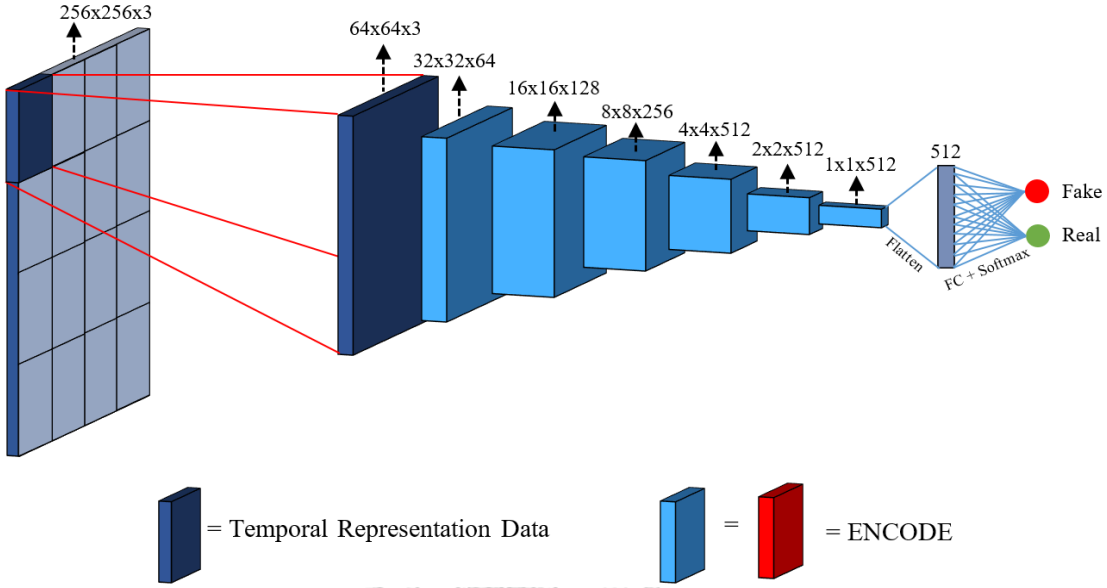


Fig. 2.5 The PatchGAN structure in the discriminator architecture.

Two objective functions including a Generator Loss or L1 Loss L_{L1} and a GAN Loss L_{GAN} are determined for training G and D. Our DSTN contains only one network consisting of the translation of spatial to temporal images where the dense optical flow is defined by three-channel components; horizontal, vertical, and magnitude. Suppose y is the target image, which is OF_{fus}, x is the input data of G, which is obtained by concatenating f and f_{BR} frames. Specifically, G learns the mapping from x to y without noise z, where the drop-out algorithm is used in the form of z in this work. The objective functions, L_{L1} and L_{GAN}, can be defined as below,

$$L_{L1}(G) = E_{x,y}[\|y - G(x)\|_1], \tag{2.3}$$

$$L_{GAN}(G, D) = E_y[log\, D\,(y)] + E_x\left[log\left(1 - D\big(G(x)\big)\right)\right]. \tag{2.4}$$

Finally, the network, G, is optimized as

$$G^* = arg \min_G \max_D L_{GAN}(G, D) + \lambda L_{L1}(G). \tag{2.5}$$

This one network of spatiotemporal translation deep GAN provides less complexity cost while contains enough important information for learning normal events. The reason that we do not train for anomalous events is that we need the model to know only normal patterns to be able to handle the possibility of occurrence

of various anomalous events without any descriptions for anomaly ground truth samples.

D. Anomaly Detection

After training the normal events by the spatiotemporal learning-based deep GAN, the model understands the translation from the spatial representation of the normal events (the concatenated frame of f and f_{BR}) to the temporal representation (OF_{fus}). Then, the model parameters of this training are used in the testing procedure.

During testing, all video sequences are used in the experiment. Each frame f and its previous frame t-1 from the test video sequences are input into DSTN. We use G in the spatiotemporal learning-based deep GAN as it corresponds to the trained model. In this case, if there are unknown events in the scenes, G will try to generate the dense optical flow based on the normal objects as it has been learned only with the normal events. Thus, it cannot reconstruct the anomalous event in the same way as normal events. This inaccuracy of G for anomalous event reconstruction leads to the detection of the possible occurrence of anomaly events.

To detect the anomalous events in the scene, we simply subtract the patches of OF_{fus} and OF_{gen} to find the pixel by pixel difference in the scene. In addition, the position of anomalous objects is required to be identified in the scene. Hence, we propose the edge wrapping for object localization in this work. The details of differentiation and edge wrapping are described as following.

1) Differentiation

After completing the model training, OF_{gen} can be observed by using the trained model parameters. To identify whether the scene contains the abnormal events or not, the pixel by pixel differentiation between OF_{fus} and OF_{gen} is simply defined by subtracting a patch of OF_{fus} and a patch of OF_{gen} as shown in Eq. (2.6) below,

$$\Delta_{OF} = OF_{fus} - OF_{gen} > 0, \tag{2.6}$$

where Δ_{OF} is the subtraction output after differentiating between OF_{fus} and OF_{gen} in which the output value is more than 0. This shows the possible abnormal events in the scene due to the fact that G was not able to reconstruct the anomalous events in OF_{gen} in the same way as the actual anomalous events in OF_{fus}.

After the subtraction, we consider the probability of pixels in Δ_{OF} as the score indicating whether the pixels in Δ_{OF} belong to normal or abnormal events. As each Δ_{OF} from different test video sequences needs to have the same range of pixel values where the lowest value is 0 and the highest value is 1, we consider the highest pixel value in Δ_{OF} as the abnormal pixel in the frame. We normalize Δ_{OF} by computing the maximum value M_{OF} of all components for each test video sequence, regarding its range of values. Then, the ROC curve is computed by gradually changing the

threshold of anomaly scores to determine the best decision threshold. The normalization of differentiation Δ_{OF} can be defined as N_{OF} as shown in Eq. (2.7):

$$N_{OF}(i,j) = 1/M_{OF}\Delta_{OF}(i,j) \tag{2.7}$$

where $N_{OF}(i,j)$ is the normalized differentiation of Δ_{OF} in the position of the pixel (i,j).

2) Edge Wrapping

After applying differentiation, the differences between OF_{fus} and OF_{gen} are revealed, showing the anomalous events in the scene. However, there are some problems with false anomaly detection on the normal events and over-detection on the abnormal object areas. Thus, to correctly localize the position of the anomalous objects and events in the scene, we propose the Edge Wrapping technique for specifically improving the object localization at the pixel level by preserving only the important edge information and suppressing the rest.

To suppress the unimportant edges along with the noise, we implement the Edge Wrapping based on the Canny edge detection [49]. This Edge Wrapping approach is a multistage procedure divided into three stages, including a noise reduction, a gradient intensity, and a non-maxima suppression, as described below.

- Noise Reduction

A Gaussian filter is used to smooth the normalized differentiation output image N_{OF} by removing noise from the background and removing pixels from non-related anomalous events. The size of the filter is $w_e \times h_e \times c_e$ where w_e and h_e represent the width and height of the Gaussian filter of the Edge Wrapping and c_e represents the number of channels such as $c_e = 1$ for the grayscale image and $c_e = 3$ for the color image. For our DSTN, we obtain the grayscale image after differentiation, then $c_e = 1$.

- Gradient Intensity

For the gradient intensity, the edge gradient (G_e) can be obtained by convolving the image with a gradient operator in horizontal (G_x) and vertical (G_y) directions. To find G_e, the image is filtered by a gradient operator, Sobel kernel, in G_x and G_y directions to obtain the gradient magnitude and its direction, which is perpendicular to the edges, for each pixel. The derivative filter size is the same as the Gaussian filter size in the noise reduction stage. G_e is computed at each pixel using the first derivative to obtain the edge gradient magnitude and the edge gradient direction, which is perpendicular to the edge direction, as shown in Eq. (2.8) and Eq. (2.9).

$$G_e = \sqrt{G_x^2 + G_y^2}, \tag{2.8}$$

$$\theta = tan^{-1}\left(\frac{G_y}{G_x}\right). \tag{2.9}$$

- Non-maxima Suppression

Finally, the non-maxima suppression is implemented by determining the threshold to preserve the ridge edges and suppress the noise. We check whether the magnitude at a pixel is greater than a threshold T (T=50). If it is greater than T, there is a point of the edge, representing a local maxima in the neighborhood. Thus, if it is the local maxima, preserve it. Otherwise, suppress it to 0. Therefore, we obtain the edges corresponding to the actual anomalous objects. The reason why we choose the threshold value of 50 is indicated in Section 2.1.4, where we consider different threshold values in our experiment.

In addition, the Gaussian filter with kernel size $w_e \times h_e \times c_e$ is applied to avoid the occurrence of spot noise in the image. The output of this procedure is represented as EW, which is defined for the final anomaly object localization OL as shown in Eq. (2.10):

$$OL = \Delta_{OF} \left\lfloor \frac{EW}{EW+\zeta} \right\rfloor \tag{2.10}$$

where ζ is a constant value.

2.1.4. Experimental Results

This section presents the evaluation of our DSTN on three challenging anomaly datasets, including UCSD pedestrian [5], UMN [6], and CUHK Avenue [18], with its implementation details. Our proposed method is analyzed to highlight the impact of residual connections, background removal, patch extraction, and edge wrapping with its base threshold value. The experiment results are comprehensively compared with other state-of-the-art methods in terms of the frame-level and pixel-level evaluations and the time complexity.

A. Dataset

1) UCSD Dataset

The UCSD pedestrian dataset [5] contains crowded scenes in outdoor environments with various anomalous events such as cycling, skateboard, vehicle, and wheelchair. It comprises two sub-sets, including Ped1 with 34 training and 16 test video sequences with around 5500 normal and 3400 anomalous frames and Ped2 with 16 training and 12 test video sequences with 3460 normal and 1652 anomalous frames. Ped1 has a resolution of 238×158 pixels, while Ped2 has a resolution of 360×240 pixels.

2) UMN Dataset

The UMN dataset [6] has been recorded for distinguishing the anomalous events in crowded scenes. It has 11 video sequences in three different scenes, containing both indoor and outdoor scenes with a total number of 7700 frames. The image resolution is 320×240 pixels. The main characteristics of this dataset are that the crowds walk normally and then suddenly run in different directions. The walking and running patterns are represented as normal and abnormal events, respectively.

3) CUHK Avenue Dataset

The CUHK Avenue dataset [18] has been recorded with a fixed camera installed in front of a school gate, containing frames with a total number of 30652 frames which are divided into 16 training and 21 test video sequences with 15328 and 15324 frames, respectively. The length of each video sequence is about 1-2 minutes (around 25 frames per second). The normal pattern includes pedestrians walking parallel to the camera, while the abnormal patterns contain different events (e.g., people throwing objects, jumping, running, and loitering). The ground truth of abnormal object that is labeled in the rectangular area is provided in this dataset.

B. Implementation

The proposed DSTN is implemented by using Python and Matlab based on Keras [50] backend TensorFlow [51]. At training time, we use a GPU with a high-performance graphics card, NVIDIA GeForce GTX 1080 Ti with NVIDIA CUDA Cores 3584, and a memory bandwidth of 484 GB/sec. The testing is implemented by using a 2.8 GHz CPU with the Intel Core i9-7960x processor. The reconstruction loss L_{L1} is optimized to 10^{-3} using Adam optimization.

C. Evaluation Criteria

We evaluate the quantitative performance of the proposed DSTN framework based on two criteria: frame level and pixel level. The frame-level evaluation checks whether there is at least one anomalous event that occurs in a test frame, and then the frame is defined as being abnormal. The pixel-level evaluation indicates the position of anomalous events, triggered if the detected abnormal area overlaps more than 40% with the ground truth [20]. The pixel-level evaluation is more challenging than the frame-level evaluation because of the complexity of anomaly localization.

D. Evaluation on UCSD Dataset

The first experiment is on the UCSD pedestrian dataset which contains 10 image sequences of the UCSD Ped1 and 12 image sequences of the UCSD Ped2 with the ground truth of pixel-level evaluation. In this dataset, both frame-level and pixel-level protocols are used to evaluate the UCSD Ped1 and the UCSD Ped2.

In the feature collection, we independently extract patches from each original image of the UCSD Ped1 with a size of 238×158 pixels and the UCSD Ped2 with a size of 360×240 pixels to multiple patches with a size of $\frac{w}{4} \times h \times c_p$. The total number of patches of the UCSD Ped1 and the UCSD Ped2 for training is about 22k and 13.6k image patches, respectively. The patches give information on the appearance of the foreground object along with its motion features due to the information of the changing vector within each patch in the frame. After collecting the appearance and motion features, all patches are resized to the resolution of 256×256 pixels to be fed into the model as the input for training and testing.

At the training time, the sizes of the input and target data are set to the resolution of 256×256 pixels as a default. The input of G has been defined by the concatenation of f and f_{BR} patches to provide the information on the appearance with

the foreground object boundaries. Since *G* comprises of Encoder and Decoder networks [34], there are different procedures implemented in each part. In the Encoder network, the image resolution of the first layer of the proposed framework is 256×256 pixels. Then it is encoded from 256→128→64→32→16→8→4→2→1 to get the variable vectors known as latent space that exploits data in one-dimensional space from the spatial representation of images. The downscale from the spatial representation image to latent space is implemented by using a CNN with a kernel size of 3×3 pixels and a stride of $s = 2$. In addition, the number of neurons in each layer of the Encoder network is set from 6→64→128→256→512→512→512→512→512, corresponding to its image resolution of each input layer.

After the encoding process, the Decoder network starts to generate the target data by performing a reverse process with the same structure. The Decoder decodes the latent space to the target image size of 256×256 pixels in order to reach the temporal representation of the optical flow output. The number of neurons in each layer of the Decoder is the same as the Encoder configuration with its image resolution of the input layer. Moreover, the drop-out is applied in the Decoder to be represented as the random noise *z* of GAN by removing connections of neurons with the default probability at $p = 0.5$. This drop-out helps to prevent over-fitting on the training dataset.

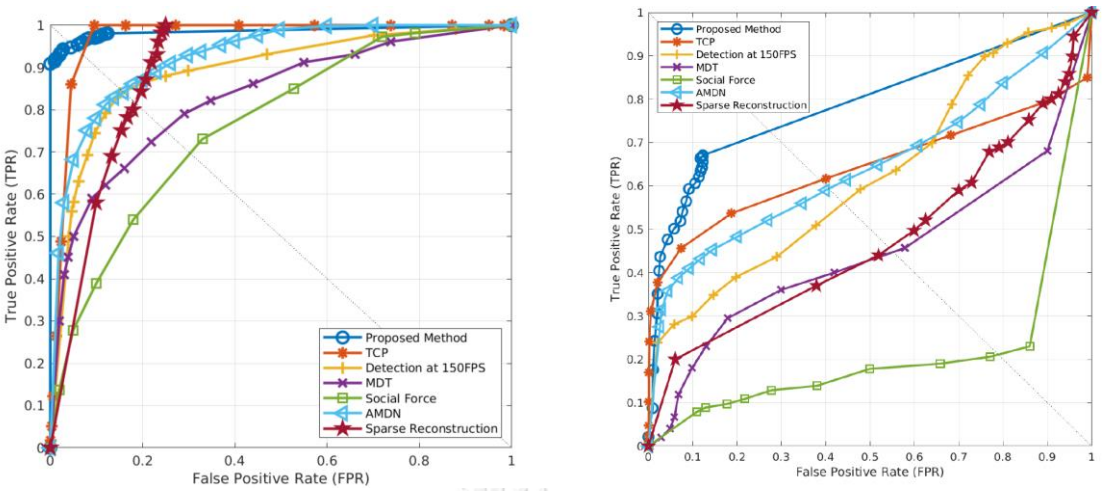
Furthermore, the training process requires *D* to vary *G* in order to optimize the distinction of a fake and a real image. *D* is represented by PatchGAN, having an input size of 64×64 pixels and output the probability showing whether the object belongs to a negative class (fake) or a positive class (real). The PatchGAN structure is defined as 64 →32→16→8→4→2→1, where it is flattened to 512 neurons which are then followed by a Fully Connection (FC) and a Softmax layer to link to the target output label. Since PatchGAN works on a partial image which has less learnable parameters, we observe that the training of the deep spatiotemporal translation GAN network is faster. For other parameter settings, the batch size is set to 1 and the reconstruction loss (norm L1) is optimized to be lower than 0.001. Adam optimization is used with a learning rate of 0.0002 and a momentum of 0.9.

At the testing time, *G* is the only model used to generate OF_{gen} to compare with the original temporal representation OF_{fus}. The resolution of the test images is the same as the training images for all datasets. Various state-of-the-art methods [4]-[8], [10], [11], [13]-[19], [35], [36] are compared with our DSTN. According to the quantitative comparison of different methods in terms of Equal Error Rate [49] and Area Under Curve (AUC) in Table 2.1, it is clearly shown that our proposed method outperforms all the methods as we achieve the highest AUC value in both frame-level and pixel-level evaluations of the UCSD pedestrian dataset. We also reach the lowest EER value compared to the other methods except only for the pixel-level evaluation on the UCSD Ped2 in [13].

Table 2.1 Performance comparison with state-of-the-art methods on UCSD dataset.

Method	Ped1 (frame level)		Ped1 (pixel level)		Ped2 (frame level)		Ped2 (pixel level)	
	EER	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**
MPPCA [70]	40%	59.0%	81%	20.5%	30%	69.3%	-	-
Social force (SF) [71]	31%	67.5%	79%	19.7%	42%	55.6%	80%	-
SF+MPPCA [72]	32%	68.8%	71%	21.3%	36%	61.3%	72%	-
SR [73]	19%	-	54%	45.3%	-	-	-	-
MDT [72]	25%	81.8%	58%	44.1%	25%	82.9%	54%	-
Detection at 150fps [74]	15%	91.8%	43%	63.8%	-	-	-	-
SR+VAE [75]	16%	90.2%	41.6%	64.1%	18%	89.1%	-	-
AMDN (double fusion) [59]	16%	92.1%	40.1%	67.2%	17%	90.8%	-	-
GMM [60]	15.1%	92.5%	-	69.9%	-	-	-	-
Plug-and-Play CNN [62]	8%	95.7%	40.8%	64.5%	18%	88.4%	-	-
GANs [63]	8%	97.4%	35%	70.3%	14%	93.5%	-	-
GMM-FCN [65]	11.3%	94.9%	36.3%	71.4%	12.6%	92.2%	**19.2%**	78.2%
Convolutional AE [15]	27.9%	81%	-	-	21.7%	90%	-	-
Liu *et al.* [16]	23.5%	83.1%	-	33.4%	12%	95.4%	-	40.6%
Adversarial discriminator [14]	7%	96.8%	34%	70.8%	11%	**95.5%**	-	-
AnomalyNet [17]	25.2%	83.5%	-	45.2%	10.3%	94.9%	-	52.8%
DSTN (proposed method)	**5.2%**	**98.5%**	**27.3%**	**77.4%**	**9.4%**	**95.5%**	21.8%	**83.1%**

The qualitative results of our proposed method can be visually illustrated in the standard protocol for abnormality detection as ROC curves, where the x-axis is the False Positive Rate (*FPR*) and the y-axis is the True Positive Rate (*TPR*). To produce the ROC curves, the threshold parameter has been varied from 0 to 1 to indicate the flow of *TPR* and *FPR*. We compare our performance with other state-of-the-art methods from their original papers (when available) as shown in Fig. 2.6 and Fig. 2.7, where Fig. 2.6 shows the ROC comparison on the UCSD Ped1 in both (a) frame-level evaluation and (b) pixel-level evaluation and Fig. 2.7 shows the ROC comparison on the UCSD Ped2 in the frame-level evaluation. According to Fig. 2.6 and Fig. 2.7, our proposed DSTN, represented as the dark blue curves, outperforms all the competing methods as our curves have the strongest growth on the *TPR*, meaning that the abnormal events in our proposed method are accurately detected and localized in both frame-level and pixel-level evaluations.



(a) (b)

Fig. 2.6 ROC comparison on UCSD Ped1 dataset: (a) frame-level evaluation and (b) pixel-level evaluation.

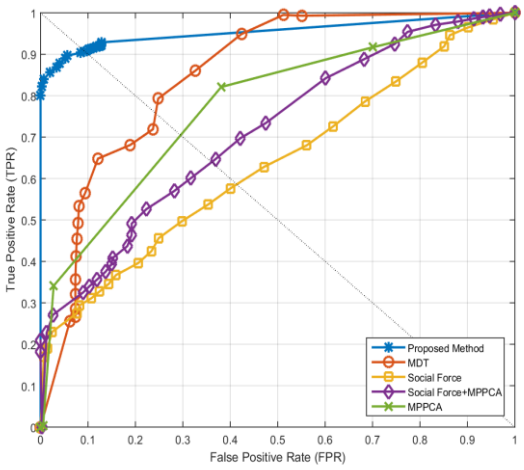
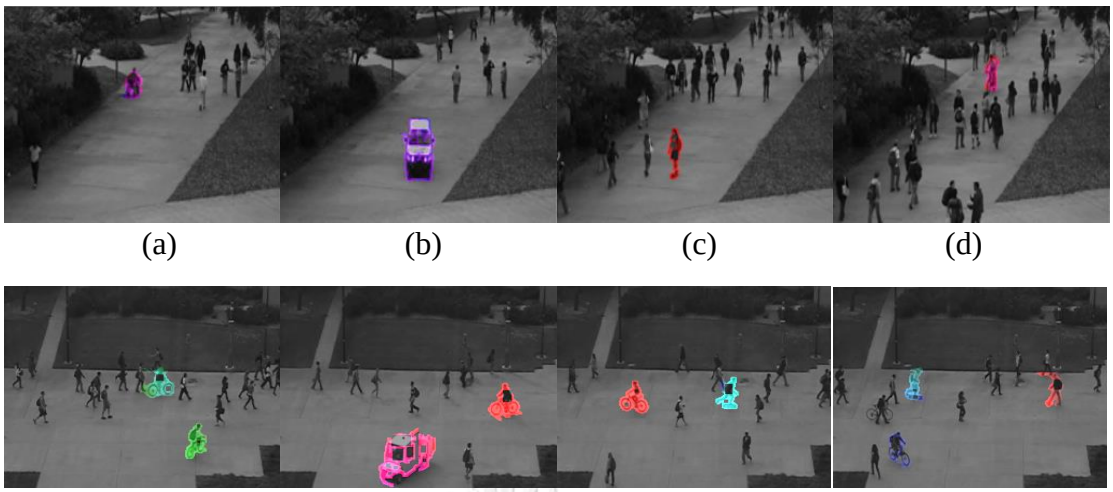


Fig. 2.7 ROC comparison on UCSD Ped2 dataset at frame level.

We also show some examples of the anomaly detection and localization on the UCSD dataset in Fig. 2.8. The results show that our proposed method can efficiently detect different anomalous events in the frame, including a single object (e.g., a wheelchair, a vehicle, a skateboard, and a bicycle) and multiple objects (e.g., bicycles, vehicle and bicycle, bicycle and skateboard). However, there is false anomaly detection in Fig. 2.8 (h), where the proposed method detects the normal event (walking pedestrians represented in red color) as an abnormal event. This is probably because the speed of walking pedestrians is the same as the cycling event in the scene.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2.8 Examples of anomaly detection and localization results on UCSD Ped1 and Ped2 dataset: (a) wheelchair, (b) vehicle, (c) skateboard, (d) bicycle, (e) bicycles, (f) vehicle and bicycle, (g) bicycle and skateboard, and (h) bicycle and skateboard.

E. Evaluation on UMN Dataset

We evaluate the performance on the UMN dataset using the same training parameter settings and network configuration as for the UCSD dataset. Table 2.2 shows the AUC comparison of our DSTN performance with other state-of-the-art methods [6], [10], [11], [14], [17], [19], [36].

Table 2.2 AUC comparison with state-of-the-art methods on UMN dataset.

Method	**AUC**
Optical-flow [6]	0.84
SFM [6]	0.96
Sparse Reconstruction [19]	0.976
Commotion [36]	0.988
Plug-and-Play CNN [10]	0.988
GANs[11]	0.99
Adversarial Discriminator [14]	0.99
AnomalyNet [17]	**0.996**
DSTN (proposed method)	**0.996**

From Table 2.2, it is clear that the proposed DSTN outperforms most of the baseline methods and its AUC performance is equal to the best method [17]. The examples of anomaly detection and localization on three different scenes of the UMN dataset are shown in Fig. 2.9.

(a) (b) (c) (d)

Fig. 2.9 Examples of anomaly detection and localization results on UMN dataset: (a), (b), and (c) show running activity in outdoor scenes, while (d) shows running activity in an indoor scene.

F. Evaluation on CUHK Dataset

In this section, we follow the previous training parameter settings and network configuration of the UCSD and UMN datasets for the evaluation on the CUHK Avenue dataset. Table 2.3 shows the comparison of our DSTN performance with other state-of-the-art methods [13], [15]-[18], in which the proposed DSTN outperforms all the competing methods for both AUC and EER.

Table 2.3. Performance comparison with state-of-the-art methods on CUHK Avenue dataset.

Method	**EER**	**AUC**
Convolutional AE [15]	25.1%	70.2%
Detection at 150 FPS [18]	-	80.9%
GMM-FCN [13]	22.7%	83.4%
Liu et al [16]	-	85.1%
AnomalyNet [17]	22%	86.1%
DSTN (proposed method)	**20.2%**	**87.9%**

Fig. 2.10 presents examples of anomaly detection and localization on the CUHK Avenue dataset, containing multiple abnormal activities, including (a) jumping, (b) throwing objects (papers), (c) falling objects (papers), and (d) grabbing a falling bag. From Fig. 2.10, it is clearly seen that our DSTN can detect and localize various anomalous events accurately, especially in Fig. 2.10(d) where the abnormal areas (e.g., a bag and a human head) are detected, even though they have only a slight difference in motion from the normal events.

(a) (b) (c) (d)

Fig. 2.10 Examples of anomaly detection and localization results on CUHK Avenue dataset: (a) jumping, (b) throwing objects, (c) falling objects, and (d) grabbing object.

G. Analysis of Residual Connection

As the residual connection or the skip connection in *G* is significant to our DSTN, we conduct additional experiments to indicate and analyze the performance of the residual connection compared to the autoencoder network which is created by removing the residual connections in the U-Net. First, we train on all training video sequences from the UCSD Ped2 dataset for 40 epochs on both networks to study their performance of minimizing the L1 loss on the training samples as shown in Fig. 2.11.

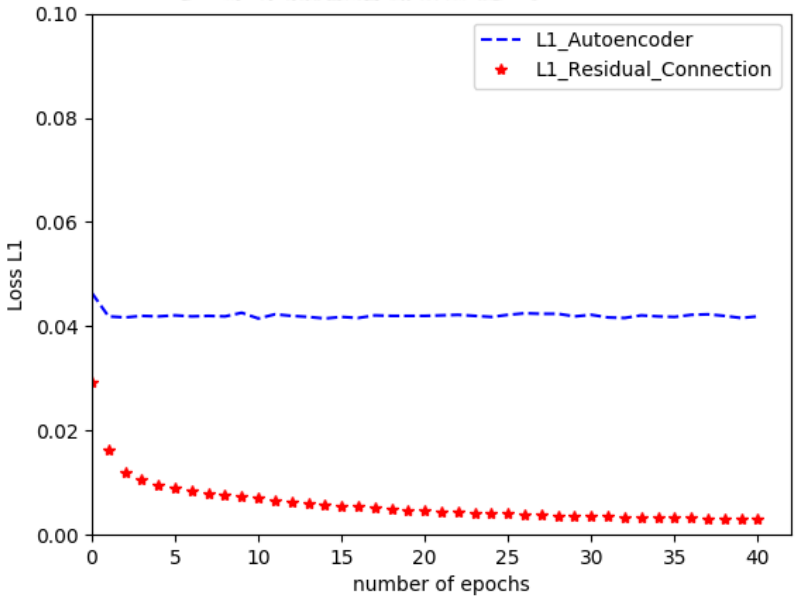
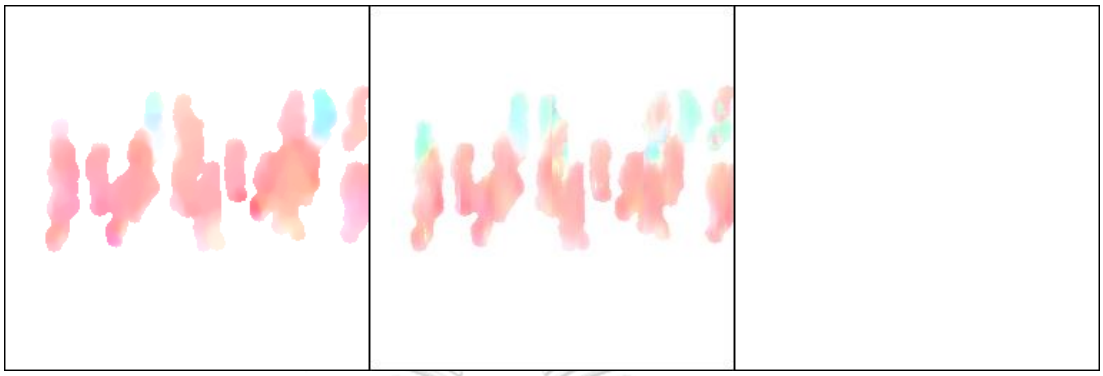


Fig. 2.11 Performance comparison between autoencoder and residual connection on UCSD Ped2 dataset.

The residual connection loss, represented as a red star curve, exhibits lower training error over training time compared to the autoencoder loss represented as a blue dash curve, meaning that the performance of the residual connection is remarkably higher than the one of the autoencoder.

In addition, we observe the ability of temporal information generation of the residual connection and the autoencoder from the test video sequences of the UCSD Ped2 dataset as shown in Fig. 2.12. Fig. 2.12(c) shows that the autoencoder is unable

to generate dense optical flow in our experiment. On the other hand, the residual connection in Fig. 2.12(b) can properly generate a new dense optical flow corresponding to the real dense optical flow in Fig. 2.12(a), providing a good quality result of the synthesized image.



(a) dense optical flow (b) residual connection (c) autoencoder

Fig. 2.12 Examples of dense optical flow generation results of residual connection and autoencoder on the UCSD Ped2 dataset.

Besides the above, we also compute FCN-scores on pixel accuracy [59] and Structural SIMilarity Index (SSIM) [66] metrics on the UCSD Ped2 dataset to compare the performance between the autoencoder and the residual connection as shown in Table 2.4. For both evaluations, a higher value means a better result. The pixel accuracy metric is a common semantic segmentation evaluation. In this work, there are two classes; a foreground region class and a background region class. Let n_{ij} be the number of wrong classified pixels of class i, and n_{ti} be the total number of pixels of class i. The pixel accuracy can be computed by $\sum_i n_{ii} / \sum_i n_{ti}$. For the SSIM index, we use it to measure the similarity between the original and the synthesized images. The more the synthesized image looks like the original image, the more efficient the model is. The results in Table 2.4 show that the residual connection clearly achieves superior results on the low-level information than the autoencoder for both pixel accuracy and SSIM evaluations.

Table 2.4. Performance comparison of the autoencoder and the residual connection in terms of FCN-scores on pixel accuracy and Structural SIMilarity Index (SSIM) on the UCSD Ped2 dataset.

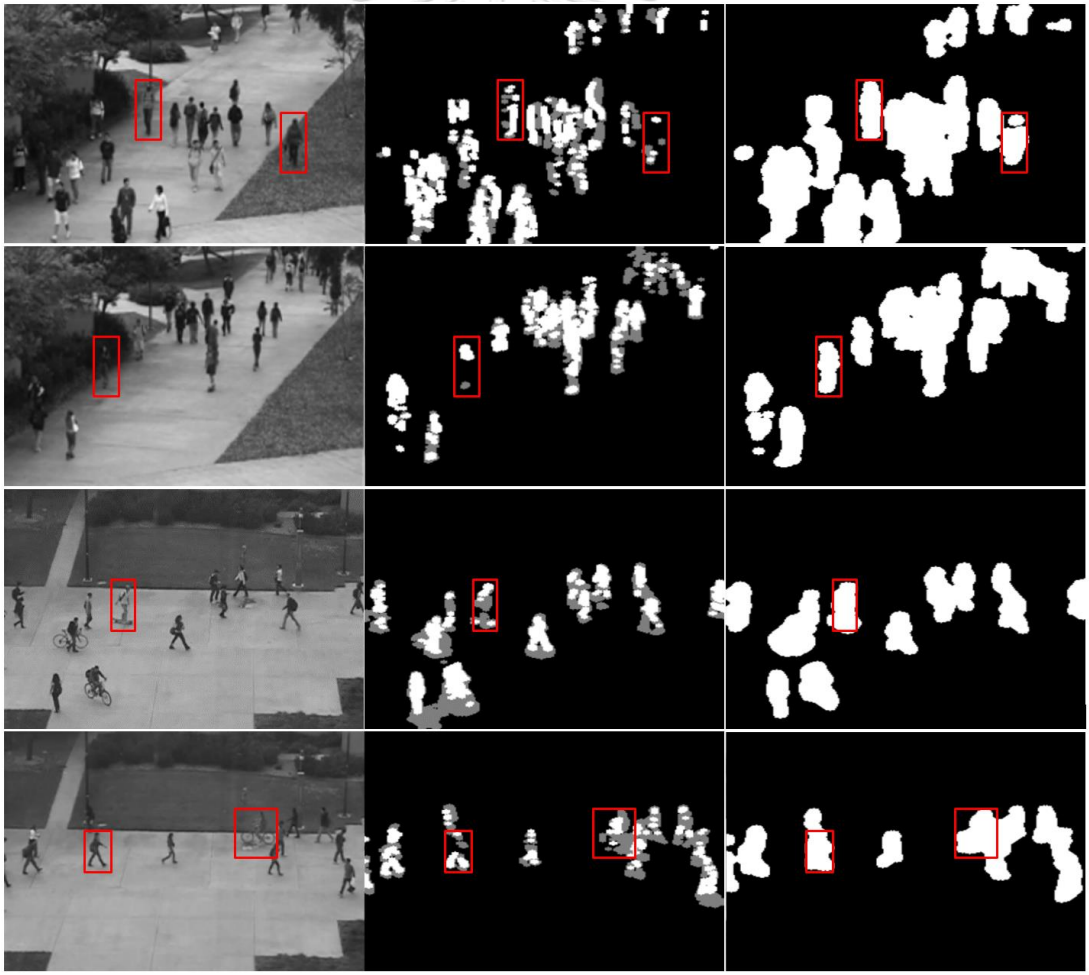
Network Architecture	**Pixel accuracy**	**SSIM**
Autoencoder	0.83	0.82
Residual connection	**0.9**	**0.96**

H. Analysis of DSTN

In this section, the proposed DSTN is analyzed to emphasize the significance of its main elements, including background removal, PatchGAN, patch extraction, and Edge Wrapping with its threshold value as follows.

- The Background Removal Method

First of all, to demonstrate the performance of the background removal method using the frame absolute difference on the proposed DSTN, we compare it with a popular technique for background subtraction, i.e., the Gaussian mixture model (GMM)-based background subtraction method [67], on the UCSD dataset as shown in Fig. 2.13. As we train only the normal event patterns in the scene, Fig. 2.13(c) shows that the background removal method can preserve more information on the normal events than the GMM-based background subtraction method, which loses some appearance information of the normal and abnormal events as shown in the red box in Fig. 2.13(b), providing incomplete and inaccurate information of the foreground objects. According to these experimental results, the background removal method is more suitable for our proposed method since it comprehensively preserves the appearance feature information of the moving foreground objects. Thus, we use it as the foreground feature extractor under the assumptions of static CCTV cameras.

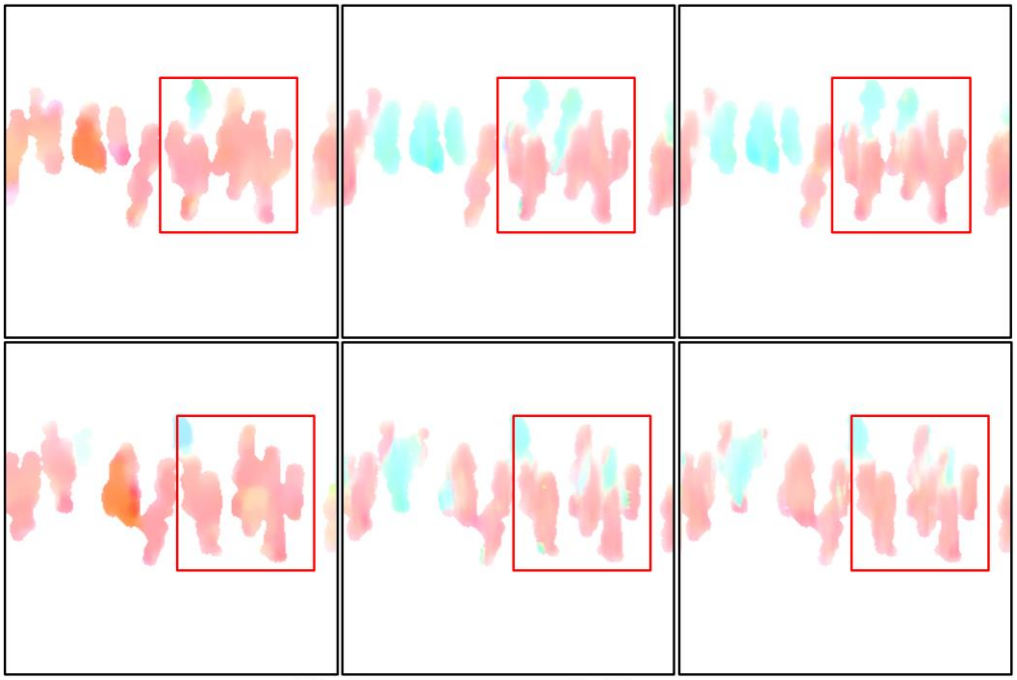


(a) original frame (b) GMM (c) background removal

Fig. 2.13 Performance comparison of background subtraction between (b) GMM-based background subtraction method and (c) background removal method on the UCSD dataset.

- The Impact of PatchGAN and Patch Extraction

Considering the impact of using the patch in our proposed method, we investigate different sizes of PatchGAN used in *D* to demonstrate its performance to DSTN. Based on [34], the full ImageGAN has greater depth and more parameters than PatchGAN, making it more difficult to train. Thus, we test additional PatchGAN with a patch size of 32×32 pixels and 64×64 pixels. The use of the 32×32 PatchGAN provides lower intensity on the appearance of objects than the 64×64 PatchGAN which is better in the visual quality of the synthesized images, meaning that the structure of synthesized images is more recognizable, as shown in Fig. 2.14.



(a) frame (b) 32×32 pixels (c) 64×64 pixels

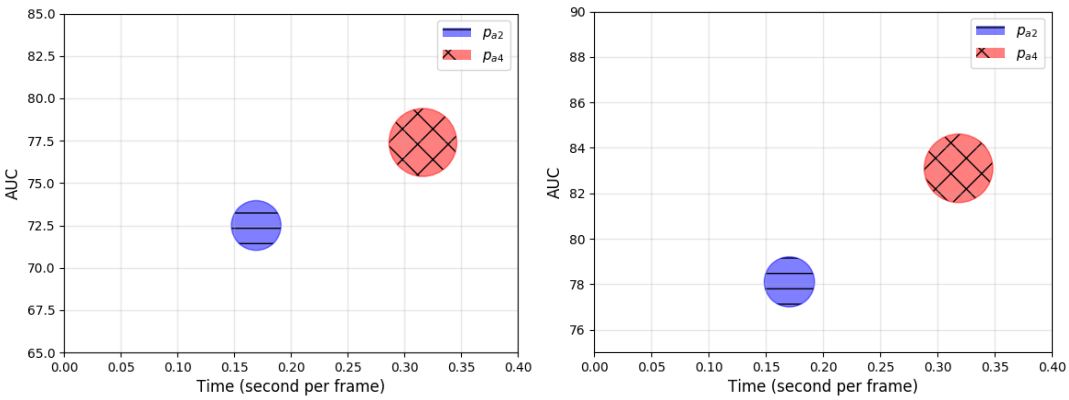
Fig. 2.14 Comparison of different sizes of PatchGAN: (a) frame, (b) 32×32 pixels, and (c) 64×64 pixels.

We also compute the FCN-scores on the pixel accuracy and the SSIM of the 32×32 PatchGAN and the 64×64 PatchGAN, as shown in Table 2.5. From Table 2.5, the 64×64 PatchGAN achieves slightly better pixel accuracy than the 32×32 PatchGAN. Thus, according to the performance of the 64×64 PatchGAN in Fig. 2.14 and Table 2.5, we decided to use it in all the experiments.

Table 2.5 Performance comparison of different sizes of PatchGAN in terms of FCN-scores on pixel accuracy and Structural SIMilarity Index (SSIM) on the UCSD Ped2 dataset.

PatchGAN Size	**Pixel accuracy**	**SSIM**
32×32	0.89	**0.96**
64×64	**0.9**	**0.96**

Furthermore, we also ran additional experiments to show the effect of the patch extraction from the feature collection process. We investigate two different patch sizes with the scale value a = 2 (p_{a2}) and a = 4 (p_{a4}) on the UCSD datasets. Fig. 2.15 shows the comparison of AUC and computational complexity of two different patch sizes, p_{a2} and p_{a4}, on the UCSD datasets.



(a) Ped1 (b) Ped2

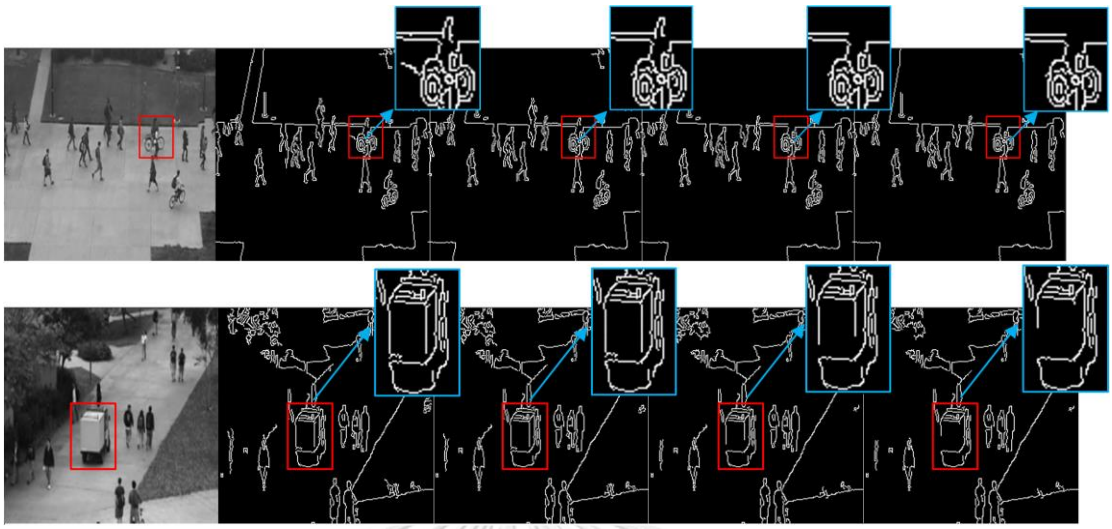
Fig. 2.15 Comparison of AUC and computational complexity of two different patch sizes, p_{a2} and p_{a4}, on the UCSD datasets.

p_{a2} provides low computational complexity as it achieves 50% faster processing than p_{a4} due to its bigger patch size. However, p_{a2} has a lower accuracy than p_{a4} on both frame-level and pixel-level evaluations. Specifically, the AUC values of p_{a2} on the UCSD Ped1 dataset are 96.9% for frame level and 72.5% for pixel level, while the AUC values of p_{a4} are 98.5% for frame level and 77.4% for pixel level. For the AUC values of p_{a2} on the UCSD Ped2 dataset, they are 95.4% for frame level and 78.1% for pixel level, while the AUC values of p_{a4} are 95.5% for frame level and 83.1% for pixel level. This remarkably shows that p_{a4} achieves more accurate results for both evaluations. Based on these experimental results, we can conclude that the patch size with a higher scale value provides better abnormal event localization. Since we aim to collect features from both appearance and motion information for enhancing the localization accuracy, we use p_{a4} for the training videos of all datasets. The stride d is assigned to $\frac{w}{a}$ for extracting the patches which are then resized to 256×256 pixels. Thus, the patch size is $\frac{w}{4} * h * c_p$.

- The Impact of Edge Wrapping and Threshold Value

As we aim to improve the performance of the anomaly localization in the pixel-level evaluation, we introduce the Edge Wrapping [29] at the final stage of our DSTN. To choose the threshold values in *EW*, Canny edge detection [49] recommends the ratio of the high to the low threshold in the range of two or three to one. In this work, the low threshold is observed from the high threshold divided by three. Since the pixels above the high threshold value considered as strong edges have the maximum value of 255, the lower threshold should be assigned as $\frac{255}{3}$ = 85. Then, we explore different threshold values, including the threshold value of 35, 50, 65, and

80. We conduct experiments on edge preservation of different threshold values as shown in Fig. 2.16.



(a) frame (b) T=35 (c) T=50 (d) T=65 (e) T=80

Fig. 2.16 Comparison of edge detection with different thresholds: 35, 50, 65, and 80.

The experimental results show that the threshold value of 50 (T = 50) can preserve better edges than other threshold values. Specifically, the threshold values of 35 (T = 35) and 50 (T = 50) are better than other threshold values (T = 65, T = 80) because they can preserve more soft edges of the objects in the scene, while the threshold values of 65 and 80 give incomplete edge results. However, the threshold value of 35 provides more edges (e.g., object shadows and background) which are not useful in our experiment. Thus, in this work, we select the threshold value of 50 as the base threshold.

Table 2.6 shows a comparison of the impact of *EW* on the proposed DSTN for the frame-level and pixel-level performances on the UCSD dataset. Using *EW*, we achieve a significant improvement in terms of the AUC and EER, especially in the pixel-level localization. To further demonstrate the importance of *EW*, we show a comparison of applying *EW* on examples from all datasets, the UCSD, UMN, and CUHK Avenue, in Fig. 2.17. From Fig. 2.17, it is clear that *EW* helps to locate the actual anomalous objects more precisely since all unrelated features (e.g., shadows, noises, and normal objects) are suppressed. These results prove the benefit of applying *EW* for anomaly detection and localization in combination with the proposed DSTN.

Table 2.6 Impact of Edge Wrapping [29] on UCSD frame-level and pixel-level performances.

Method	**Ped 1 (F)**		**Ped 1 (P)**		**Ped 2 (F)**	
	EER	**AUC**	**EER**	**AUC**	**EER**	**AUC**
DSTN without EW	9%	95.8%	35.6%	70.1%	9.8%	94.6%
DSTN with EW	**5.2%**	**98.5%**	**27.3%**	**77.4%**	**9.37%**	**95.54%**

(a) DSTN without Edge Wrapping

(b) DSTN with Edge Wrapping

Fig. 2.17 Examples of the impact of Edge Wrapping on all datasets: UCSD Ped1, UCSD Ped2, UMN, and CUHK Avenue.

I. Analysis on Time Complexity

We compare the computational time of the proposed DSTN with other state-of-the-art methods [5], [7], [18]-[20]. As these methods do not provide their original implementations, we follow the computational time and the environment from [7]. With regard to computational time in frame per second (fps), our DSTN achieves 3.17 fps, 3.15 fps, 3.15 fps, and 3 fps on the UCSD Ped1, UCSD Ped2, UMN, and CUHK Avenue datasets, respectively. We also compare our time complexity in seconds per frame with other baseline methods as shown in Table 2.7.

Table 2.7. Computational time comparison during testing (seconds per frame).

Method	**CPU**	**GPU**	**Memory**	**Running Time**			
				Ped1	**Ped2**	**UMN**	**Avenue**
Sparse Reconstruction [19]	2.6GHz	-	2.0GB	3.8	-	0.8	-
Detection at 150 fps [18]	3.4GHz	-	8.0GB	**0.007**	-	-	**0.007**
MDT [5]	3.9GHz	-	2.0GB	17	23	-	-
Li *et al*. [20]	2.8GHz	-	2.0GB	0.65	0.80	-	-
AMDN (double fusion) [7]	2.1GHz	Nvidia Quadro K4000	32GB	5.2	-	-	-
DSTN (proposed method)	2.8GHz	-	24GB	0.315	**0.319**	**0.318**	0.334

It is clear that our computational time is lower than most of the baseline methods except for [18]. This is because our architecture is based on a deep learning framework consisting of multiple convolutional layers while [18] is based on a sparse combination learning framework that has lower neuron connections. However, we obtain significantly higher AUC value and relatively much lower EER value in both

frame-level and pixel-level evaluations on the UCSD and the CUHK Avenue datasets than [18]. According to our experimental results, we can conclude that the proposed DSTN outperforms other competing methods by achieving the highest AUC value in both frame-level and pixel-level evaluations while providing a good running time for surveillance videos.

2.1.5. Conclusion

In this paper, we propose a novel unsupervised spatiotemporal anomaly detection and localization for surveillance videos. The proposed DSTN framework is embedded with concepts of deep convolution neural network of GAN based Edge Wrapping approach which brings advantages to anomaly localization. The deep spatiotemporal translation network is designed to learn the appearance and motion representations with the use of the fusion and the concatenation of patches for combining the learned features. Additionally, our proposed method does not rely on any prior knowledge in order to design features for the input (as we use raw pixels) and does not involve low-level object analysis, such as object detection and tracking. We provide extensive experimental results compared with other state-of-the-art methods and implemented on three publicly available datasets, including the UCSD pedestrian, UMN, and CUHK Avenue. We clearly show that our DSTN outperforms other state-of-the-art methods in terms of accuracy and time complexity as we obtain the highest AUC value in both frame-level and pixel-level evaluations for all datasets and achieve a good running time that outperforms most of the baseline methods. Our method is effective and robust for anomaly event detection and localization in the crowded scenes for surveillance videos. For future work, we will explore an object translation model with a clustering method to enhance the performance of the anomaly detection and localization from the complex scene. Other abnormalities will be observed for increasing the robustness of the model for real-world use.

Acknowledgment

This work was supported by the Chulalongkorn University Dutsadi Phiphat Scholarship.

References

[1] *Video Surveillance Intelligence Service-Annual-IHS Technology*. Accessed: Aug. 5, 2019. [Online]. Available: https://technology.ihs.com/Services/570988/video-surveillance-intelligence-service-annual

[2] T. Akinbinu and Y. Mashalla, “Impact of computer technology on health: Computer Vision Syndrome (CVS),” *Med. Pract. Rev.*, vol. 5, no. 3, pp. 20-30, 2014.

[3] K. Gates, “Professionalizing police media work: Surveillance video and the forensic sensibility,” in *Images, Ethics, Technology*. Evanston, IL, USA: Routledge, 2015, pp. 53-69.

[4] Y. Feng, Y. Yuan, and X. Lu, “Learning deep event models for crowd anomaly detection,” *Neurocomputing*, vol. 219, pp. 548-556, Jan. 2017.

[5] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975-1981.

[6] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935-942.
[7] D. Xu, Y. Yan, E. Ricci, and N. Sebe, “Detecting anomalous events in videos by learning deep representations of appearance and motion,” *Comput. Vis. Image Understand.*, vol. 156, pp. 117-127, Mar. 2017.
[8] J. Sun, X. Wang, N. Xiong, and J. Shao, “Learning sparse representation with variational auto-encoder for anomaly detection,” *IEEE Access*, vol. 6, pp. 33353-33361, 2018.
[9] S. Bouindour, M. M. Hittawe, S. Mahfouz, and H. Snoussi, “Abnormal event detection using convolutional neural networks and 1-class SVM classifier,” in *Proc. 8th Int. Conf. Imag. Crime Detection Prevention (ICDP)*, 2017, pp. 1-6.
[10] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, “Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1689-1698.
[11] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, “Abnormal event detection in videos using generative adversarial nets,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577-1581.
[12] H.Wei, Y. Xiao, R. Li, and X. Liu, “Crowd abnormal detection using two-stream fully convolutional neural networks,” in *Proc. 10th Int. Conf. Measuring Technol. Mechatronics Autom. (ICMTMA)*, Feb. 2018, pp. 332-336.
[13] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine, “Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder,” 2018, *arXiv:1805.11223*. [Online]. Available: http://arxiv.org/abs/1805.11223
[14] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, “Training adversarial discriminators for cross-channel abnormal event detection in crowds,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1896-1904.
[15] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733-742.
[16] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection_A new baseline,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536-6545.
[17] 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 Trans. Inf. Forensics Secur.*, vol. 14, no. 10, pp. 2537-2550, Oct. 2019.
[18] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720-2727.
[19] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3449-3456.
[20] W. Li, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18-32, Jan. 2014.
[21] X. Tang, S. Zhang, and H. Yao, “Sparse coding based motion attention for abnormal event detection,” in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3602-3606.

[22] A. Li, Z. Miao, Y. Cen, and Q. Liang, "Abnormal event detection based on sparse reconstruction in crowded scenes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1786-1790.
[23] T. Chen, C. Hou, Z. Wang, and H. Chen, "Anomaly detection in crowded scenes using motion energy model," *Multimedia Tools Appl.*, vol. 77, no. 11, pp. 14137-14152, Jun. 2018.
[24] Z. Wang, C. Hou, B. Li, T. Chen, L. Yao, and M. Song, "Global abnormal event detection in video via motion information entropy," in *Proc. 2nd URSI Atlantic Radio Sci. Meeting (AT-RASC)*, May 2018, pp. 1-4.
[25] D. Du, H. Qi, Q. Huang, W. Zeng, and C. Zhang, "Abnormal event detection in crowded scenes based on structural multi-scale motion interrelated patterns," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1-6.
[26] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical _ow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 27, no. 3, pp. 673-682, Mar. 2017.
[27] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2909-2917.
[28] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Bremond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683-695, Mar. 2017.
[29] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, "Combining motion and appearance cues for anomaly detection," *Pattern Recognit.*, vol. 51, pp. 443-452, Mar. 2016.
[30] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2017, *arXiv:1710.09282*. [Online]. Available: http://arxiv.org/abs/1710.09282
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672-2680.
[32] Tim Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234-2242.
[33] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125-1134.
[35] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2921-2928.
[36] H. Mousavi, M. Nabi, H. Kiani, A. Perina, and V. Murino, "Crowd motion monitoring using tracklet-based commotion measure," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2354-2358.

[37] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool, "Fast optical flow using dense inverse search," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 471-488.
[38] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17-31, Apr. 2007.
[39] M. J. Roshtkhari and M. D. Levine, "An online, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," *Comput. Vis. Image Understand.*, vol. 117, no. 10, pp. 1436-1452, Oct. 2013.
[40] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1477-1481, Sep. 2015.
[41] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1590-1599, Oct. 2013.
[42] Y. Yuan, Y. Feng, and X. Lu, "Statistical hypothesis detector for abnormal event detection in crowded scenes," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3597-3608, Nov. 2017.
[43] N. Patil and P. K. Biswas, "Global abnormal events detection in crowded scenes using context location and motion-rich spatio-temporal volumes," *IET Image Process.*, vol. 12, no. 4, pp. 596-604, Apr. 2018.
[44] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992-2004, Apr. 2017.
[45] Y. Yuan, Y. Feng, and X. Lu, "Structured dictionary learning for abnormal event detection in crowded scenes," *Pattern Recognit.*, vol. 73, pp. 99-110, Jan. 2018.
[46] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88-97, Jul. 2018.
[47] S. Wang, E. Zhu, J. Yin, and F. Porikli, "Video anomaly detection and localization by local motion based joint video representation and OCELM," *Neurocomputing*, vol. 277, pp. 161-175, Feb. 2018.
[48] X. Zhang, S.Yang, X. Zhang, W. Zhang, and J. Zhang, "Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning," 2018, *arXiv:1805.10620*. [Online]. Available: http://arxiv.org/abs/1805.10620
[49] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986.
[50] Keras-Team. (Nov. 6, 2019). *Keras*. GitHub. Accessed: Nov. 12, 2019. [Online]. Available: https://github.com/keras-team/keras
[51] M. Abadi, "TensorFlow: A system for large-scale machine learning," in *Proc. Symp. Operating Syst. Design Implement.*, 2016, pp. 265-283.
[52] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of autoencoder," *Electron. Lett.*, vol. 52, no. 13, pp. 1122-1124, Jun. 2016.
[53] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, and S. Chen, "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14617-14639, Nov. 2016.

[54] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for ef_cient neural network,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2015, pp. 1135-1143.
[55] K. Ullrich, E. Meeds, and M. Welling, “Soft weight-sharing for neural network compression,” 2017, *arXiv:1702.04008*. [Online]. Available: http://arxiv.org/abs/1702.04008
[56] T. N. Sainath, B. Kingsbury,V. Sindhwani, E. Arisoy, and B. Ramabhadran, “Low-rank matrix factorization for deep neural network training with high-dimensional output targets,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6655-6659.
[57] P. Maji and R. Mullins, “On the reduction of computational complexity of deep convolutional neural networks,” *Entropy*, vol. 20, no. 4, p. 305, 2018.
[58] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
[59] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431-3440.
[60] E. L. Andrade, S. Blunsden, and R. B. Fisher, “Modelling crowd scenes for event detection,” in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, 2006, pp. 175-178.
[61] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371-3408, Dec. 2010.
[62] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[63] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, 2015, pp. 234-241.
[64] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[65] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770-778.
[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004.
[67] Z. Zivkovic, “Improved adaptive Gaussian mixture model for background subtraction,” in *Proc. 17th IEEE Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2004, pp. 28-31.

Journal Information

Our second manuscript named a deep residual spatiotemporal translation network for video anomaly detection and localization (DR-STN) has been submitted to Pattern Recognition Letters on September 23, 2020. The detail of this work is presented as follows.

Topic: Deep Residual Spatiotemporal Translation Network for Video Anomaly Detection and Localization
Authors: Thittaporn Ganokratanaa[1], Supavadee Aramvith[2], and Nicu Sebe[3]
Address: [1]Department of Electrical Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Address: [2]Multimedia Data Analytics and Processing Research Unit, Department of Electrical Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Address: [3]Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy
Corresponding author: Supavadee Aramvith (supavadee.a@chula.ac.th)
Journal name: Pattern Recognition Letters
Initial Date Submitted: September 23, 2020

2.2. Deep Residual Spatiotemporal Translation Network for Video Anomaly Detection and Localization

Abstract Video anomaly detection has gained significant attention in the current intelligent surveillance systems. However, many existing works have difficulties in dealing with the anomaly localization in the crowded scenes due to the lack of sufficient prior information of the objects of interest during training, resulting in false-positive detection results. To cope with these issues, we propose Deep Residual Spatiotemporal Translation Network (DR-STN), a novel unsupervised Deep Residual conditional Generative Adversarial Network (DR-cGAN) model with an Online Hard Negative Mining (OHNM) approach. The proposed DR-cGAN provides a wider network to learn a mapping from spatial to temporal representations and enhance the perceptual quality of synthesized images from a generator. During DR-cGAN training, we take only the frames of normal events to produce their corresponding dense optical flow. At testing time, we compute the reconstruction error in local pixels between the synthesized and the real dense optical flow and then apply OHNM to remove false-positive detection results. Finally, a semantic region merging is introduced to integrate the intensities of all the individual abnormal objects into a full output frame. The proposed DR-STN has been extensively evaluated on three benchmarks, demonstrating superior results over other state-of-the-art methods both in frame-level and pixel-level evaluations.

Keywords anomaly detection, generative adversarial network, surveillance video, residual unit, hard negative mining

2.2.1. Introduction

Video anomaly detection [15] has recently become popular in computer vision research due to the growing demand for security aspects. An anomaly is a rare event occurring in crowded scenes and there might be more than one anomaly at a time. The challenges of VAD relate to complex and crowded scenes, anomaly localization, small anomaly datasets, and many false-positive detection results. The anomaly localization is required to indicate the position of the abnormalities in a scene and is more challenging than detecting an abnormal frame. Another challenge is the very small number of anomalies present in the available public datasets leading to the difficulty of learning a good classifier. Besides, these challenges result in false-positives in the final output through which the system incorrectly detects normal events as abnormal ones.

Many efforts in the community have been done to overcome these problems. Previous works used hand-crafted features (e.g., Gaussian regression with Bags of Visual Words [2], trajectories with K-means [3], and Histogram of Oriented Gradients [29]). However, it is difficult for these methods to precisely detect and localize different occluded and small objects in real crowded scenes even though they can detect multiple objects. In such a complex scene, deep learning methods [4, 5, 12, 13, 19, 22, 23, 25, 28, 30] are more suitable to generalize the representations of these objects due to the nonlinear transformation performance of learnable models. In addition, many of the deep learning methods [5, 12, 22, 23] are only able to obtain a high detection rate on the frame level while the detection rate at the pixel level is much lower. The reasons are as follows: i) a full-frame is fed into the model without

prior knowledge of the objects, resulting in insufficient features of objects of interest for performing deep data-hungry learning; ii) patch extraction is not effective in collecting comprehensive features of the object. Recent works [21, 26] aim to enhance the accuracy using supervised learning methods that need data labeling for all samples, making it not suitable for VAD as anomalies are varied and unpredictable. Hence, unsupervised deep learning methods are a more suitable solution as they aim to learn only normal events (the majority of patterns in the scene) without the need for labeling data. Any unknown patterns will be considered as anomalies by their large distance from the normal patterns. Following this consideration, Generative Adversarial Networks (GANs) have gained more attention in anomaly detection research due to their outstanding performance in constructing images, affording data augmentation, and dealing with implicit data in complex scenarios [6]. GANs consist of two competing networks: a generator *G* and a discriminator *D*. With the convolutional networks in *G*, many works have tried to achieve a high visual quality of image reconstruction and to overcome vanishing gradients. U-Net has been proposed in [24] based on the idea of skip connections [7] to enhance the accuracy of image segmentation for the biomedical image. Isola, *et al*., proposed [9] an effective translation of sketch images to realistic images based on conditional GANs (cGANs) with the use of U-Net.

In this work, we propose a novel Deep Residual Spatiotemporal Translation Network (DR-STN) approach for video anomaly detection and localization in crowds. Fig. 2.18 shows an overview of our proposed framework. Inspired by [7, 9], we propose a novel Deep Residual cGAN (DR-cGAN) to enhance the accuracy and quality of the synthesized image. A powerful object detector [1] is applied to extract the objects in the frame to be fed into our DR-cGAN. Different from previous works [5, 12, 22, 23] which are based on [9], our DR-cGAN is built by designing the residual units and the residual connections in *G* to learn the translation of objects of interest from appearance (spatial) to motion (temporal) representations.

Our contribution is four-fold:

(i) our unsupervised DR-STN learns only normal events without using any hand-crafted features and effectively translates comprehensive information of the objects of interest from appearance to motion representations in crowded scenes;

(ii) we propose DR-cGAN, a novel end-to-end unsupervised deep residual connection network, to improve perceptual information of reconstructed images from the generator. DR-cGAN provides a wider network that extensively passes information from the previous to the next layer of encoder and decoder. To the best of our knowledge, this is the first attempt to build deep residual connections (projection and identity shortcuts) on the U-Net architecture of cGAN for VAD;

(iii) we introduce the object detector as the pre-processing process to extract only the objects of interest to feed into the DR-cGAN model to help in learning the pattern of normal objects. This provides better object localization for the pixel level;

(iv) we introduce an Online Hard Negative Mining (OHNM) and a semantic region merging as the post-processing processes to eliminate the false-positives without retraining the model and integrate the intensity of objects for the final

anomaly output, providing more reliable and remarkable results than the state-of-the-art.

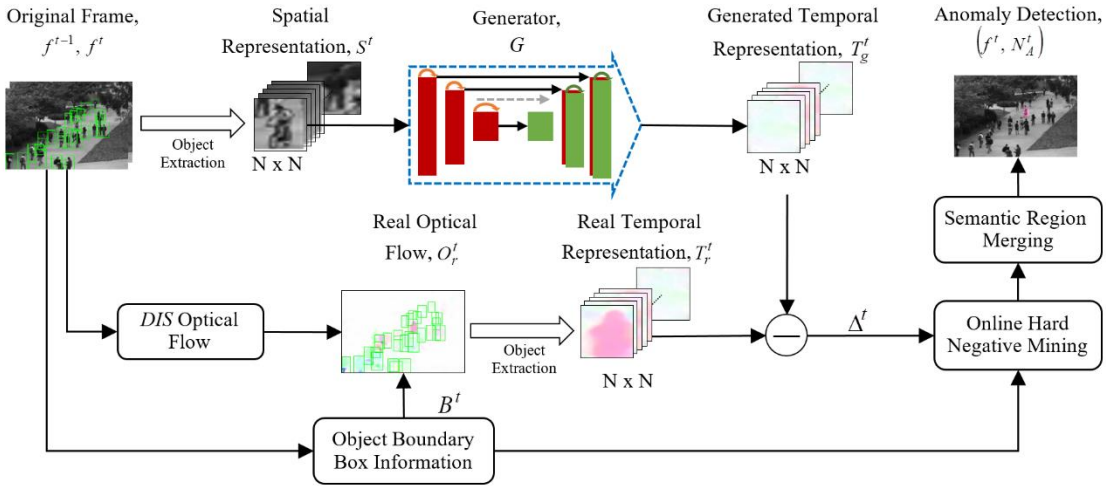


Fig. 2.18 Overview of proposed framework.

2.2.2. Related Works

Among existing works, the deep learning approaches are the most successful ones. The main approaches include supervised and unsupervised learning.

The supervised learning methods typically provide higher accuracy on classification problems. Ramachandra, *et al.*, [21] proposed anomaly localization in videos using Siamese CNN to compute a distance between the ground truth label on normal and abnormal video patches, causing over-fitting issues as the input of the network is limited to small patches of the abnormal event. Singh, *et al.*, [26] proposed Aggregation of Ensembles (AOE) of different fine-tuned CNNs with additional multiple SVM and Softmax classifiers to detect anomalies in crowds. This network is not end-to-end trainable and has a high cost of data annotation for obtaining a sufficient amount of data.

On the other hand, unsupervised learning is considered as being a more flexible approach for VAD. Xu, *et al.*, [28] proposed an appearance and motion anomaly detection network using Stacked Denoising AutoEncoders (SDAEs) as the feature extractor with the One-Class SVM classifier. Prawiro, *et al.*, [19] proposed a two-stream autoencoder where the decoder is used to learn the static background and the dynamic foreground objects. Ravanbakhsh, *et al.*, [22] proposed two cross-channel networks between appearance and motion and vice versa based on cGANs. This fusion strategy for the two networks makes it more complex to reconstruct images. Similarly, the adversarial discriminator based on cGANs is proposed in [23], where the discriminator is used as the classifier during testing, making it faster than [22] but yielding lower accuracy. Ganokratanaa, *et al.*, [5] proposed a deep spatiotemporal translation network (DSTN) based on GAN with pre- and post-processing procedures, resulting in good frame-level anomaly detection. However,

their background removal is quite sensitive to shadow and illumination changes and the patch extraction is not always able to obtain the full object appearance.

The proposed DR-cGAN is different from other previous works since we do not rely on hand-crafted features or require any labeled data as in the supervised-based approaches. Specifically, we are different from [5] as we build the deep residual cGAN architecture with the object detector without any pre-defined background subtraction model. Additionally, the OHNM method [10] has been implemented to explicitly address anomaly localization and false-positive detection problems, providing more robust and reliable results.

2.2.3. Methodology

A. Pre-processing DR-STN

The object detection is introduced at the first stage of DR-STN to locate and extract the objects of interest for the input of our DR-cGAN model, allowing us to gain more meaningful semantic information. We use You Only Look Once (YOLO) [1] to handle the challenges from the realistic scenes (e.g., noise, illumination changes, and object scaling and occlusions) due to its high robustness on images in different environments and its optimal speed-accuracy tradeoff. The pre-trained YOLO is applied on each frame f to predict a set of bounding boxes for the objects. These bounding boxes aim to extract spatial information of the objects from each frame f and temporal information of the objects from each dense optical flow O_r to pass into the DR-cGAN for model learning.

B. DR-cGAN in DR-STN

Our DR-cGAN is proposed for learning the translation from spatial to temporal information (dense optical flow). In training, we input only the objects of interest in the frames of normal events to G. G translates the spatial object f_{ob} to the synthesized dense optical flow object O_{obg} in such a way that it is challenging for D to differentiate it from the real dense optical flow object O_{obr}. Our G and D architectures are adopted from [8, 20]. The residual units in G are designed based on [7]. The details of our architecture are explained in the following sub-sections.

1) Generator with Residual Connections

The generator G is the core model used both in training and in testing in DR-cGAN. In the common GAN [6], G learns a random noise z as an input to construct an output image $\hat{y}$. Differently, cGAN [17] learns a conversion from an image x with a random noise z to output an image $\hat{y}$, $\hat{y} = G(x, z)$. However, the use of random noise z is not essential in G as G can still learn the mapping without the noise [9]. Following [9], we apply the noise in the form of dropout in the decoder, resulting in $\hat{y} = G(x)$.

A concerning issue of translating the spatial to temporal information is mapping the difference in surface appearance from a high-resolution input to a high-resolution output grid. Thus, we design the generator architecture to effectively align

the input structure to the output structure as shown in Fig. 2.19. This generator network consists of two models: encoder and decoder. The encoder functions as the data compressor, while the decoder reversely functions as the data decompressor. In the encoder, the spatial image is input to a series of down-sampling layers until reaching a bottleneck layer. Then, the decoder performs the reconstruction process to generate a semantic output image. Our structures of the encoding and decoding blocks are defined in [5].

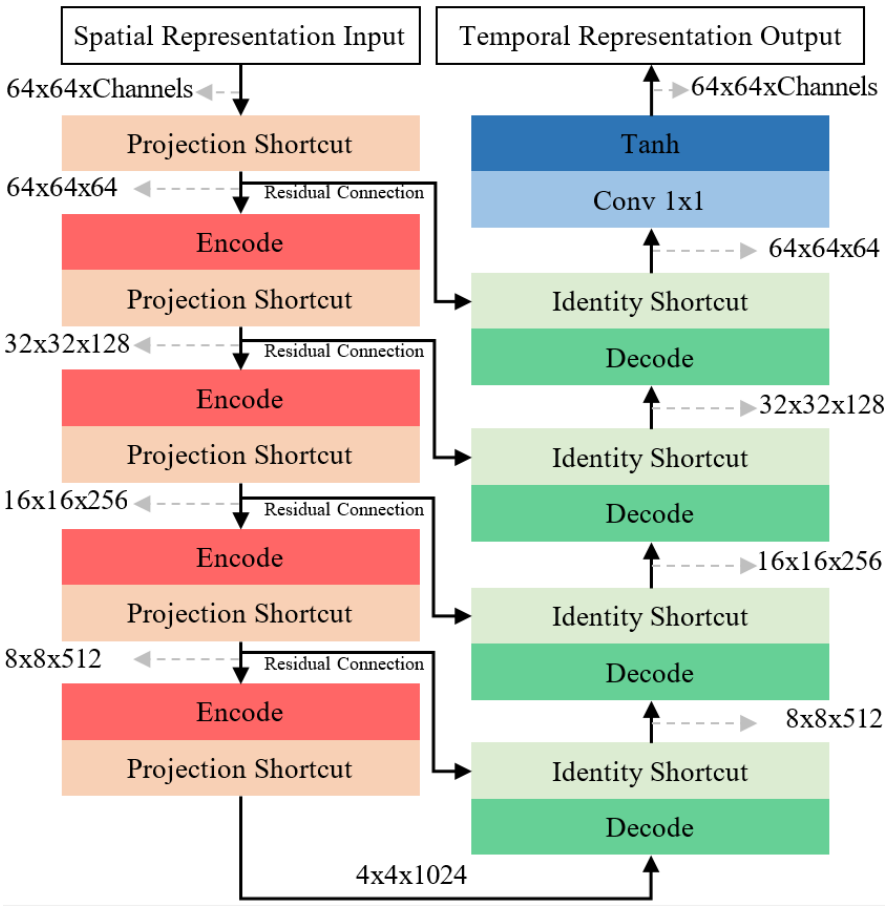


Fig. 2.19 The proposed generator architecture of DR-cGAN

To achieve finer semantic results, the low-level information is required to be shared between the input and the output in order to propagate the information through the network without degradation while maintaining the high-level information. Following this consideration, we introduce the novel generator architecture as i) we add the residual unit in each layer of the encoder and decoder to achieve a wider feature learning network; ii) we apply the residual connections from encoder layers to decoder layers to share the low-level information. Suppose n is the total number of layers. The residual unit is added after each encoder layer i and decoder layer n-i, while the residual connections are added from each encoder layer i to the decoder layer n-i. This implies better generalization and easier optimization for image translation as discussed in Section 2.2.4. Specifically, our residual units consist of projection and identity shortcuts as shown in Fig. 2.20. The projection shortcut is used to match the dimensions. Since the dimensions of our input and output in the encoder

are not the same, we define the projection shortcut to increase the dimensions of the input features to be able to add with the output features. For the decoder, its residual unit has two inputs: the output from the decoder layer and the residual connections from the encoder layer. The identity shortcut is then defined to add the concatenated inputs with the output using the same dimensions.

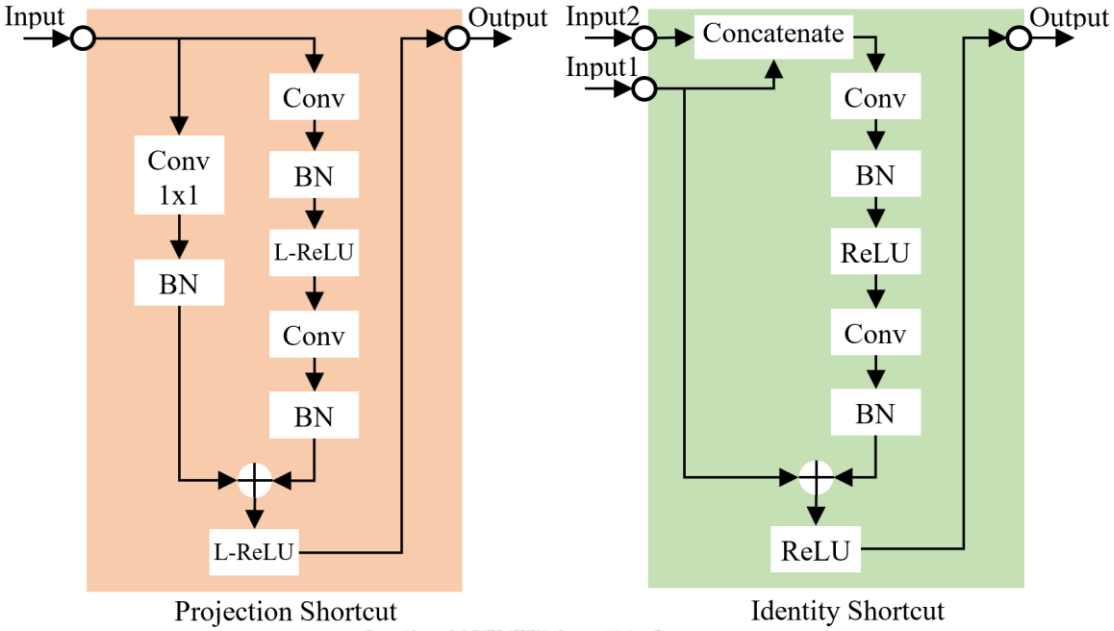


Fig. 2.20 Structure of the residual unit

2) Discriminator

We use the discriminator D only during the training process. D classifies two classes of spatiotemporal objects: a real class $\{x = f_{ob}, y = O_{obr}\}$ and a fake class $\{x = f_{ob}, O_{obg} = G(x)\}$. We train D to maximize the correct classification problem on both real and fake classes. A binary cross-entropy loss with logits loss is computed as the objective function of D. In contrast, G is trained to minimize the objective function of D with a reconstruction error between O_{obg} and O_{obr}. In other words, the adversarial D and G learn a two-player minimax game with value function $V(D, G)$:

$$\min_G \max_D V(D,G) = \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \tag{2.11}$$

where $\mathcal{L}_{cGAN}(G, D)$ presents as a cGAN loss, and $\mathcal{L}_{L1}(G)$ is a reconstruction loss in G. Both losses are determined as below,

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[\|y - G(x)\|_1], \tag{2.12}$$

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}\left[log\left[\sigma\left(D(x,y)\right)\right]\right] + E_x\left[log\left[1 - \sigma\left(D\left(x, G(x)\right)\right)\right]\right]. \tag{2.13}$$

where σ is a sigmoid function, $\sigma(D) = 1/(1 + e^{-D})$.

Our DR-cGAN provides good feature learning of the learned normal events while being less complex. Since we do not train with the abnormal event, the model understands only the normal patterns at the training time and then can observe the irregular objects following the reconstruction error at the testing time. The anomaly detection process is explained in detail in the following section.

C. Anomaly Detection

At testing time, only *G* is applied to translate f_{ob} of the test video frame to O_{obg} in order to compare with its corresponding O_{obr} for obtaining the irregular object. Specifically, the spatial objects $S^t = \{f_{ob_1}, f_{ob_2}, \ldots, f_{ob_K}\}_t$ and their corresponding bounding boxes $B^t = \{b_1, b_2, \ldots, b_K\}_t$ are extracted from each frame at time *t*, where *K* is the total number of the detected objects in a frame. To detect the irregular object, the reconstruction error $\Delta^t = \{\Delta_{ob_1}, \Delta_{ob_2}, \ldots, \Delta_{ob_K}\}_t$ is computed by differentiating between the real temporal objects $T_r^t = \{O_{obr_1}, O_{obr_2}, \ldots, O_{obr_K}\}_t$ and the synthesized temporal objects generated from *G*, $T_g^t = \{O_{obg_1}, O_{obg_2}, \ldots, O_{obg_K}\}_t$. The reconstruction error on k^{th} object is:

$$\Delta_{ob_k} = O_{obr_k} - O_{obg_k} > 0 \tag{2.14}$$

Δ_{ob_k} provides an irregular score representing the possible anomalous event in the scene when the value of Δ_{ob_k} is greater than 0. However, the output of Δ_{ob_k} may result in a false positive, meaning that the normal object (negative sample) is incorrectly detected as the abnormal object (positive sample). This false-positive object represents a hard negative example. To ensure that we obtain the actual abnormal object, we determine the high confidence score to decide whether Δ_{ob_k} belongs to the normal or abnormal object. Then OHNM is proposed to get rid of the negative example in the anomaly detection. The probability of anomaly score P_{a_k} on k^{th} object is computed as:

$$P_{a_k} = \frac{\sum_{(i,j)\in \Delta_{ob_k}} \Delta_{ob_k}(i,j)}{\sum_{(i,j)\in O_{obr_k}} O_{obr_k}(i,j)} \tag{2.15}$$

Since the model is trained only with the normal patterns, it performs a good reconstruction on the normal objects, causing a low value of Δ_{ob_k} and P_{a_k}. In contrast, the model is not able to correctly reconstruct the abnormal object, causing a high value of Δ_{ob_k} and P_{a_k}. Following these characteristics, the high confidence scores of the normal and abnormal objects are set based on two-interval thresholds: confident normal threshold C_n and confident abnormal threshold C_a. After this setting, we obtain a true detection of normal and abnormal objects. However, there are some objects which are not enrolled in these two criteria ($C_n < P_{a_k} < C_a$). Then, we take these objects into consideration of the OHNM examples to finalize the true detection of anomaly outputs.

To observe hard negative examples, the template matching is performed as a short tracklet to match each detected object in f^t to the search patch p in its adjacent frames within a window ±1 frame. The size of p is assigned to extensively cover the displacement of the object by enlarging the bounding box b_k of the k^{th} reference object f_{ob_k} to the size of 20×20 pixels. This size of p is defined due to the small movement of the object between frames. Specifically, the main idea of our OHNM is to move f_{ob_k} (template) at f^tover p in its adjacent frames (f^{t-1} and f^{t+1}) in order to measure the highest similarity patch and record the template as a normal object. The highest similarity of the pattern between $f_{ob_k}$and p is determined via block matching by shifting f_{ob_k} with the distance (u,v) in the horizontal and vertical directions within the corresponding sub-patch of p. To find the similarity score from the best-matching position between f_{ob_k} and p, we use the standard normalized cross-correlation (*NCC*) algorithm which is formulated as:

$$NCC(u,v) = \frac{\sum_{(i,j)\in f_{ob_k}} p(u+i,v+j)\cdot f_{ob_k}(i,j)}{\sqrt{\sum_{(i,j)\in f_{ob_k}} p^2(u+i,v+j)\cdot \sum_{(i,j)\in f_{ob_k}} f^2_{ob_k}(i,j)}}. \quad (2.16)$$

After acquiring the NCC similarity score, we are able to determine whether the object is abnormal or not based on the confident similarity score C_s. If there is a large appearance change between frames, we assign the object as being abnormal otherwise, we consider it to be a normal object or an isolated object yielded by flicker noise. Finally, the semantic region merging is implemented by combining all the detected abnormal objects into a full semantic frame A computed as follows,

$$A(i,j) = \begin{cases} \Delta_{ob_k}(i,j), & non-overlapping\ object \\ 1/K \cdot \sum_{k\in K} \Delta_{ob_k}(i,j), & otherwise \end{cases} \quad (2.17)$$

where K is the total number of the final abnormal objects and (i, j) are the pixel positions of A.

A is normalized to get the probability score N_A in a range of [0, 1] of the full semantic frame. The highest pixel intensity value of A, M_A, is considered as the abnormal pixel in the frame. The ROC curve is performed on N_A by slightly shifting the threshold of anomaly scores in a range of [0,1] to determine the best decision threshold. N_A can be defined as follows,

$$N_A(i,j) = 1/M_A \cdot A(i,j). \quad (2.18)$$

2.2.4. Experimental Results

In this section, we evaluate the performance of the proposed DR-STN on three anomaly benchmarks and compare it with state-of-the-art methods on both frame level and pixel level. The impact of our proposed DR-cGAN model and OHNM method is analyzed in detail.

A. Datasets

The UCSD dataset [15] includes two sub-folders: Ped1 and Ped2. There are 34 training and 16 test videos in Ped1 with around 5500 normal and 3400 abnormal frames. For Ped2, it has 16 training and 12 test videos with around 2500 normal and 1652 abnormal frames. The image sizes of Ped1 and Ped2 are 238×158 pixels and 360×240 pixels, respectively. The abnormal events in this dataset include cycling, skateboarding, vehicles, and wheelchairs.

The UMN dataset [16] has 11 videos recorded in crowded indoor and outdoor scenes with around 7700 frames and an image size of 320×240 pixels. The abnormal events refer to running, while the normal events refer to the normal walking.

The CUHK Avenue dataset [14] has 16 training and 21 test videos with 15328 and 15324 frames and an image size of 360×640 pixels. There are various anomalies in the scenes, e.g., jumping, loitering, running, and throwing objects, while the normal events are the walking crowds.

B. Implementation Details

Our proposed DR-STN is based on Python and Matlab with PyTorch [18]. The training and testing processes are implemented on NVIDIA GeForce GTX 1080 Ti. Adam optimization is used to optimize our reconstruction loss ($\lambda\mathcal{L}_{L1}$) that targets to 2E-1. The optimization parameters are defined as [9].

In our DR-cGAN, the sizes of the input and output of *G* for both training and testing processes are set to 64×64 pixels. With the encoder network in *G*, the input image is encoded by using a CNN with a kernel size of 3×3 pixels and a stride $s = 2$ to reach a bridge representing the spatial data. For the decoder network in *G*, each layer is built as the reverse of each encoder layer. To avoid the over-fitting problems on the training dataset, the random noise *z* is provided in the form of dropout in the decoder with the default probability value $p = 0.5$. In addition, the residual units for both encoder and decoder are designed by using 3×3 convolution and 1×1 convolution with $s = 1$, respectively. For *D*, it takes two input images with the resolution of 64×64 pixels to produce the 6×6 output feature.

C. Evaluation Criteria

We evaluate the quantitative performance of the proposed DR-STN considering two criteria: frame level (F) and pixel level (P). In F, the frame is considered as an anomaly if there is at least one abnormal event in a test frame. On the other hand, P specifies the location of the abnormal event. The frame is a true detection when the detected abnormal region overlaps with the ground truth region more than 40% [11].

D. Performance Evaluation

In this section, we compare Area Under the Curve (AUC) and Equal Error Rate [49] performance of DR-STN with other state-of-the-art methods as shown in Table 2.8.

We use the same network configuration and training parameter settings for all three datasets. The experiment on the UCSD dataset is implemented with 10 and 12 videos of the UCSD Ped1 and UCSD Ped2, respectively, along with their pixel-level ground truth. GANs [22] and DSTN [5] are set as the baseline methods due to their success in leveraging frame-level and pixel-level detection accuracy and achieving state-of-the-art performance in an unsupervised manner. Table 2.8 shows that our DRSTN surpasses not only the baseline methods but also most of the competing works in both F and P criteria in which we achieve higher AUC and lower EER than other works, except only for the AUC of the UCSD Ped1 dataset at P in [21]. This is probably due to their supervised learning on labeled abnormal data. However, our experimental results can significantly overcome other criteria in [21] and all criteria in [26] which also relies on a supervised-based method, showing the competitive performance of DR-STN in anomaly detection and localization tasks. In addition, the examples of our detection and localization results on three datasets are shown in Fig. 2.21 where we can detect and localize both single and multiple abnormal events in the crowded scenes even when they are occluded (e.g., a bicycle and a skateboard in Fig. 2.21(b)).

Table 2.8 AUC and EER Comparison with State-of-the-Art Methods on UCSD, CUHK Avenue, and UMN datasets

Method	**UCSD Ped1 (F) AUC/EER**	**UCSD Ped1 (P) AUC/EER**	**UCSD Ped2 (F) AUC/EER**	**UCSD Ped2 (P) AUC/EER**	**CUHK Avenue (F) AUC/EER**	**UMN (F) AUC/EER**
Social force (SF) [16]	67.5%/31.0%	19.7%/79.0%	55.6%/42.0%	-/80.0%	-/-	96.0%/-
Detection at 150fps [14]	91.8%/15.0%	63.8%/43.0%	-/-	-/-	80.9%/-	-/-
AMDN (double fusion) [28]	92.1%/16.0%	67.2%/40.1%	90.8%/17.0%	-/-	-/-	-/-
GANs [22]	97.4%/8.0%	70.3%/35.0%	93.5%/14.0%	-/-	-/-	99.0%/-
Liu, *et al*. [12]	83.1%/23.5%	33.4%/-	95.4%/12.0%	40.6%/-	85.1%/-	-/-
Adversarial Discriminator [23]	96.8%/7.0%	70.8%/34.0%	95.5%/11.0%	-/-	-/-	99.0%/-
AnomalyNet [30]	83.5%/25.2%	45.2%/-	94.9%/10.3%	52.8%/-	86.1%/22.0%	99.6%/-
DSTN [5]	98.5%/5.2%	77.4%/27.3%	95.5%/9.4%	83.1%/21.8%	87.9%/20.2%	99.6%/-
GMM-FCN [4]	94.9%/11.3%	71.4%/36.3%	92.2%/12.6%	78.2%/19.2%	83.4%/22.7%	-/-
Siamese [21]	86.0%/23.3%	80.4%/-	94.0%/14.1%	**93.0%**/-	-/-	-/-
AOE [26]	94.6%/-	-/-	95.9%/-	-/-	-/-	-/-
Two-stream decoder [19]	84.2%/-	-/-	96.1%/-	-/-	-/-	-/-
DR-STN (proposed method)	**98.8%/2.9%**	**82.5%/21.5%**	**97.6%/6.9%**	86.4%/**16.3%**	**90.8%/11.0%**	**99.7%/-**

(a) (b)

(c) (d)

Fig. 2.21 Examples of anomaly detection and localization results.

E. Analysis of DR-STN

To emphasize the importance of our DR-STN, we analyze two main components of the proposed framework:

i) The performance of DR-cGAN compared with the baseline methods including U-Net [9] and autoencoder which is simply built by removing the skip connections in U-Net;
ii) The impact of OHNM on DR-STN with regard to AUC.

First, we divide the training folder of the UCSD Ped1 dataset into two subsets: 70% for training samples and 30% for testing samples. We train the DR-cGAN model

and other baseline methods for 20 epochs to see their effectiveness in minimizing the $\lambda\mathcal{L}_{L1}$ loss as illustrated in Fig. 2.22, where our DR-cGAN (red square) reaches the lowest error over the training epochs, showing faster and superior performance in model learning than other baseline methods about 50%.

Fig. 2.22 Training loss comparison between Autoencoder, U-Net, and DR-cGAN on the UCSD Ped1 dataset.

To clarify the ability in generating the synthesized image on normal events during testing, we evaluate the proposed network using two common methods. First, FCN-scores for semantic segmentation on pixel accuracy [13] are computed to obtain the probability of correct pixels on a set of defined object classes (foreground and background region classes). The pixel accuracy is defined as $\sum_i n_{ii}/\sum_i n_{ti}$, where n_{ii} is the number of the correct classified pixels of class i, and n_{ti} is the total number of pixels of class i. Second, Structural SIMilarity Index (SSIM) metric [27] is used to evaluate the similarity between the synthesized and the real images. For both evaluations, a higher value indicates a better result of the synthesized image. Table 2.9 shows that our DR-cGAN significantly surpasses all baseline methods regarding both evaluations, providing a good synthesized image quality that is highly similar to the real image.

Table 2.9 Performance comparison of the Autoencoder, U-Net, and DR-cGAN in terms of FCN-scores on pixel accuracy and Structural SIMilarity Index (SSIM) on the UCSD Ped1 dataset.

Method	**Pixel accuracy**	**SSIM**
Autoencoder	0.81	0.78
U-Net	0.82	0.8
DR-cGAN	**0.87**	**0.85**

Apart from the above experiments, our OHNM relies on both temporal and spatial conditions. For the temporal condition, we can determine whether the object is

normal or abnormal based on P_{a_k} under the criteria of two-interval thresholds, C_n = 0.1 and C_a = 0.8. The object is classified as normality if its P_{a_k} is less than or equal to 0.1 ($P_{a_k} \leq 0.1$) and as an abnormality if its P_{a_k} is greater than or equal to 0.8 ($P_{a_k} \geq 0.8$). This is probably because the model has only the knowledge of the learned normal events at the training time. Hence, during testing, when we input all objects from each frame into the model, Δ_{ob_k} provides less difference in local pixels between the learned and the test samples in case the input is the normal object, resulting in a small value of P_{a_k} which falls into the criteria of C_n. On the other hand, there is a great difference of Δ_{ob_k} if the input is the abnormal object, resulting in a high value of P_{a_k} which is considered an abnormality following the criteria of C_a. For P_{a_k} value that does not belong to these two criteria ($0.1 < P_{a_k} < 0.8$), we apply the template matching to observe the *NCC* score of the objects between frames to indicate the appearance displacement whether the objects are the same. *NCC* results in a high similarity score if there is a small change in the appearance of the objects between frames, considering as the false-positive anomaly result. Based on the experiment, we set the confident similarity score on the normal object C_s = 0.8. We analyze the impact of OHNM on our DR-STN for reducing the false-positive detection results in terms of AUC on the UCSD dataset. With the use of OHNM, the model can remarkably improve the AUC values in both F and P as shown in Table 2.10.

Table 2.10 AUC Performance of OHNM on DR-STN

Method	**Ped 1 (F)**	**Ped 1 (P)**	**Ped 2 (F)**
DR-STN without OHNM	97.85*%*	72.65*%*	96.16*%*
DR-STN with OHNM	**98.83*%***	**82.50*%***	**97.62*%***

The AUC of P on the UCSD Ped1 dataset is increased up to about 10% compared to the plain DR-STN, providing a more precise location of the abnormal events in the scene. Following these experimental results, it is clear that applying OHNM with the proposed DR-STN benefits both anomaly detection and localization tasks.

2.2.5. Conclusion

This paper introduced a novel unsupervised deep residual spatiotemporal translation network for video anomaly detection and localization. The proposed DR-STN is embedded with a wider DR-cGAN and OHNM which benefits in reducing false-positive anomaly detection. The DR-cGAN is designed for the translation learning of appearance and motion representations by integrating the residual units, residual connections, and cGAN. Additionally, our DR-cGAN takes only raw pixels as the input from the object detector without relying on any prior knowledge of hand-crafted features. We conducted extensive experiments on three benchmarks and showed the robustness and effectiveness of the proposed framework which clearly outperforms other state-of-the-art methods.

Acknowledgment

This work is supported by Chulalongkorn University Dutsadi Phiphat Scholarship.

References

[1] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
[2] Cheng, K.W., Chen, Y.T., Fang, W.H., 2015. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2909–2917.
[3] Cosar, S., Donatiello, G., Bogorny, V., Garate, C., Alvares, L.O., Brmond, F., 2016. Toward abnormal trajectory and event detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology 27, 683–695.
[4] Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F., 2020. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding, 102920.
[5] Ganokratanaa, T., Aramvith, S., Sebe, N., 2020. Unsupervised anomaly detection and localization based on deep spatiotemporal translation network. IEEE Access 8, 50312–50329.
[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.
[7] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
[8] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
[9] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
[10] Jin, S., RoyChowdhury, A., Jiang, H., Singh, A., Prasad, A., Chakraborty, D., Learned-Miller, E., 2018. Unsupervised hard example mining from videos for improved object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 307–324.
[11] Li, W., Mahadevan, V., Vasconcelos, N., 2013. Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence 36, 18–32.
[12] Liu, W., Luo, W., Lian, D., Gao, S., 2018. Future frame prediction for anomaly detection-a new baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536–6545.
[13] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.
[14] Lu, C., Shi, J., Jia, J., 2013. Abnormal event detection at 150 fps in matlab, in: Proceedings of the IEEE international conference on computer vision, pp. 2720–2727.

[15] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N., 2010. Anomaly detection in crowded scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1975–1981.
[16] Mehran, R., Oyama, A., Shah, M., 2009. Abnormal crowd behavior detection using social force model, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 935–942.
[17] Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in neural information processing systems, pp. 8026–8037.
[19] Prawiro, H., Peng, J.W., Pan, T.Y., Hu, M.C., 2020. Abnormal event detection in surveillance videos using two-stream decoder, in: 2020 IEEE International Conference on Multimedia and ExpoWorkshops (ICMEW), IEEE. pp. 1–6.
[20] Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
[21] Ramachandra, B., Jones, M., Vatsavai, R., 2020. Learning a distance function with a siamese network to localize anomalies in videos, in: The IEEE Winter Conference on Applications of Computer Vision, pp. 2598–2607.
[22] Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N., 2017. Abnormal event detection in videos using generative adversarial nets, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 1577–1581.
[23] Ravanbakhsh, M., Sangineto, E., Nabi, M., Sebe, N., 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1896–1904.
[24] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.
[25] Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., Klette, R., 2018. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding 172, 88–97.
[26] Singh, K., Rajora, S., Vishwakarma, D.K., Tripathi, G., Kumar, S., Walia, G.S., 2020. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. Neurocomputing 371, 188–198.
[27] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612.
[28] Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N., 2015. Learning deep representations
of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553.
[29] Yuan, Y., Feng, Y., Lu, X., 2016. Statistical hypothesis detector for abnormal event detection in crowded scenes. IEEE transactions on cybernetics 47, 3597–3608.

[30] Zhou, J.T., Du, J., Zhu, H., Peng, X., Liu, Y., Goh, R.S.M., 2019. AnomalyNet: An anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security 14, 2537–2550.

2.3.Training Procedures

2.3.1. DSTN

As we described the proposed DSTN method in Section 2.1, we shall explain it more in detail on the feature extraction and training procedures. The training flow diagram of DSTN is shown in Fig. 2.23.

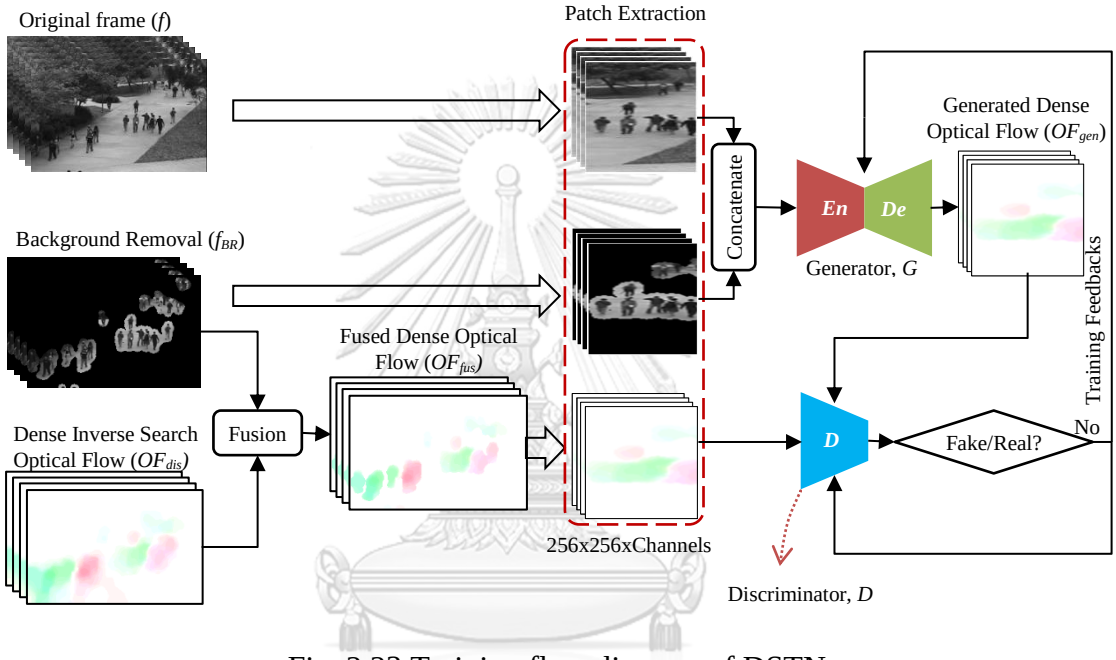


Fig. 2.23 Training flow diagram of DSTN

In the training framework, only normal event patches P_f of original frame f are input with their corresponding foreground patches P_{BR} of the background removal frame f_{BR} into the Generator G of the deep GAN to generate the synthesized patches $\hat{P}_{OF}$ of dense optical flow frame OF_{gen}, representing the motion information of the normal events. To obtain a good optical flow, f_{BR} is fused with the real dense inverse search optical flow frame OF_{dis} to eliminate noise in OF_{dis}. The fusion of f and f_{BR} frames called a fused dense optical flow frame OF_{fus} provides clear motion information for model learning.

Let P_{OF} presents the patches of OF_{fus}. Thus, the full training set of N patch samples are defined as $T = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$, where $X_i = \left(P_{f_i}, P_{BR_i}\right)$, $Y_i = P_{OF_i}$, and $i \in N$. Specifically, before feeding the training set into the model, we upscale P_f, P_{BR}, and P_{OF} to a resolution of 256×256 pixels and normalize the intensity value in the range of [-1, 1]. Then, P_f and P_{BR} are concatenated and fed into G. The model is learned by optimizing the equation (2.5) to obtain the desired reconstruction loss $L1$ between the generated patch of dense optical flow $\hat{Y} = \hat{P}_{OF}$ and the corresponding patch of fused dense optical flow Y.

In the process of the Discriminator *D*, *D* learns *Y* and $\hat{Y}$ via a deep patch discriminate network to classify that *Y* is real and $\hat{Y}$ is fake, while *G* tries to fool *D* by producing more synthetic images $\hat{Y}$ that are difficult to be discriminated against from *Y*. If *D* classifies $\hat{Y}$as a fake image, *G* will generate the new $\hat{Y}$ until *D* classifies it as a real image (the same as *Y*).

The detail of the optimization of the training process of DSTN is explained in Algorithm 1 as follows;

Algorithm 1 Minibatch Adaptive moment estimation [33] in the training of deep GAN. The parameters of discriminator *D* are updated to its hyperparameter *k*, which we used $k = 1$ as the least expensive option in our experiments. The default values of Adam parameters: $\eta = 2\text{E} - 3, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\text{E} - 08$.

- Initial parameters of *G*, Θ_G
- Initial parameters of *D*, Θ_D

for number of training iterations **do**

for *k* steps **do**

- Sample minibatch of *m* examples $\{X^{(1)}, \ldots, X^{(m)}\}$ from the concatenated input patch of spatial representation (P_f, P_{BR}).
- Sample minibatch of *m* examples $\{Y^{(1)}, \ldots, Y^{(m)}\}$ from the original patch of temporal representation P_{OF}.
- Update the discriminator's parameters:

$$g_{\theta_D} \leftarrow \nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(Y^{(i)}\right) + \log\left(1 - D\left(G\left(X^{(i)}\right)\right)\right)\right].$$

$$\theta_D \leftarrow Adam\left(g_{\theta_D}, \theta_D, \eta, \beta_1, \beta_2, \epsilon\right)$$

end for

- Sample minibatch of *m* examples $\{X^{(1)}, \ldots, X^{(m)}\}$ from the concatenated input patch of spatial representation (P_f, P_{BR}).
- Sample minibatch of *m* examples $\{Y^{(1)}, \ldots, Y^{(m)}\}$ from the original patch of temporal representation P_{OF}.
- Update the generator's parameters:

$$g_{\theta_G} \leftarrow \nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(X^{(i)}\right)\right)\right) + \lambda \left\|Y^{(i)} - G\left(X^{(i)}\right)\right\|.$$

$$\theta_G \leftarrow Adam\left(g_{\theta_G}, \theta_G, \eta, \beta_1, \beta_2, \epsilon\right)$$

end for

Algorithm 1 stops updating the parameters after the reconstruction loss is less than 1E-3. The performance of *G* in generating synthesized temporal representation is evaluated by computing the pixel accuracy and SSIM as described in Section 2.1. Finally, *G* is the only network used in our testing experiment.

2.3.2. DR-STN

In this section, we explain more detail on the training procedures of the proposed DR-STN as presented in Section 2.2. The training flow diagram of DR-STN is shown in Fig. 2.24.

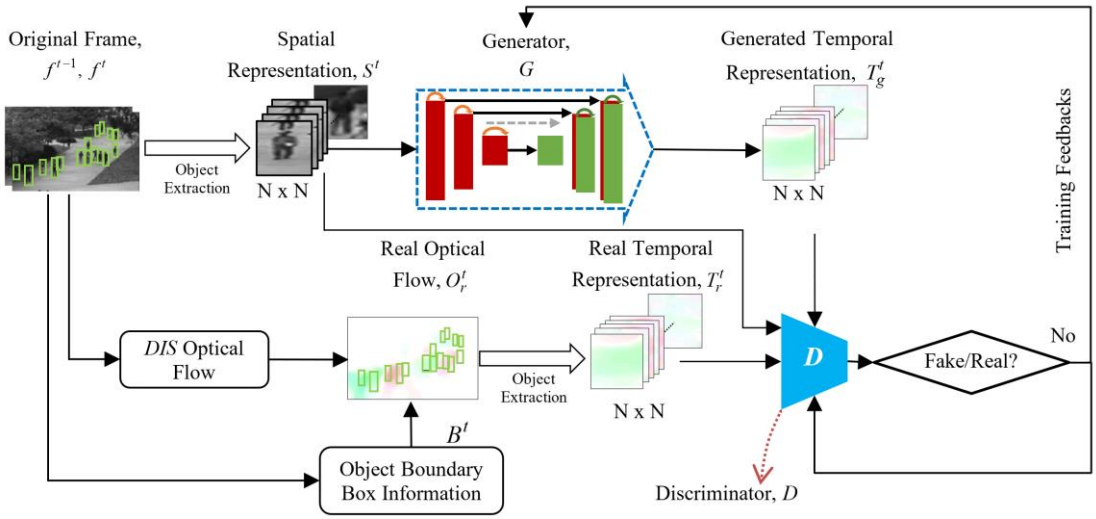


Fig. 2.24 Training flow diagram of DR-STN

In the training framework, only the spatial representation S of objects of interest in the frame f is input into the Generator G in our novel DR-cGAN to generate the synthesized dense optical flow objects T_g, representing the temporal information of the normal events. In this work, the pre-trained object detector is used to obtain individual objects in the scene. To the best of our knowledge, YOLOv4 is the latest and the most suitable object detection approach for our model due to its high detection rate and low complexity performances. We then apply it on each original frame f to acquire the object boundary box information B to extract spatial objects and dense inverse search (*DIS*) optical flow objects, representing both appearance and motion representations for the model learning.

Different from DSTN, the full training set of N patch samples are defined as $T = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$, where $X_i = f_{ob_i}$, $Y_i = O_{obr_i}$, and $i \in N$. We upscale f_{ob} and O_{obr} to a resolution of 64×64 pixels and normalize their intensity value in the range of [-1, 1] before feeding them into model learning. The model is learned by optimizing Eq. (2.11) to obtain the desired reconstruction loss $L1$ between the generated object of dense optical flow $\hat{Y} = O_{obg}$ with the corresponding object of real dense optical flow Y.

For Discriminator D, the pair of (X, Y) and the pair of $(X, \hat{Y})$ are learned via a deep discriminate network to classify that Y is real and $\hat{Y}$ is fake as shown in Fig. 2.25, while G tries to fool D by producing more synthetic images $\hat{Y}$ that are difficult to be discriminated against from Y. If D discriminates $\hat{Y}$ as a fake image, G will generate the new $\hat{Y}$ until D discriminates it as a real image (the same as Y).

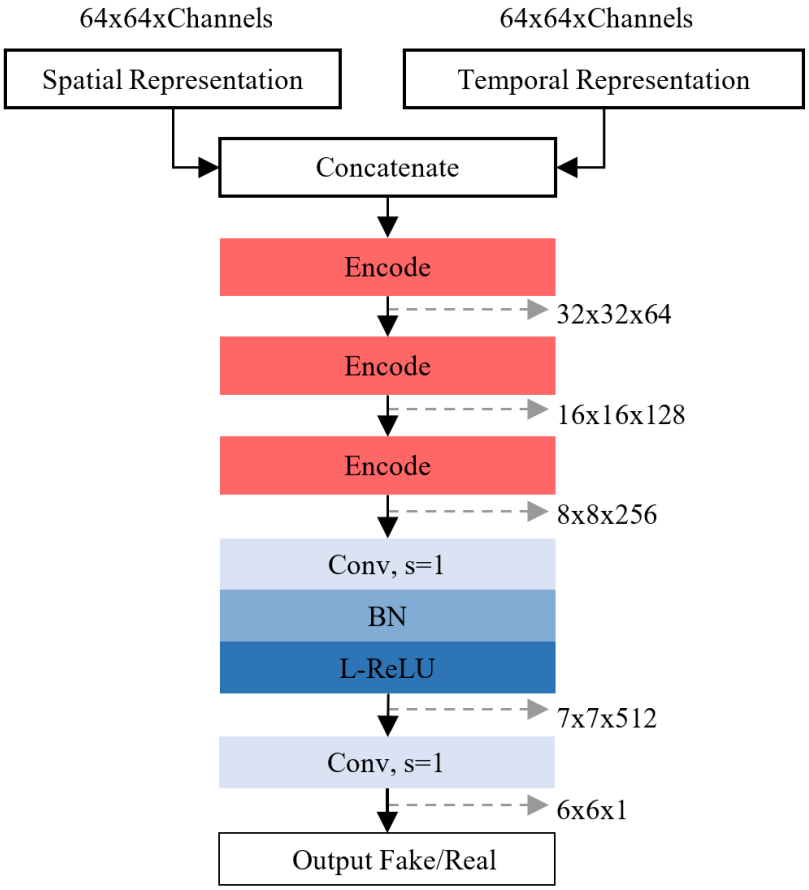


Fig. 2.25 A conditional discriminator architecture of DR-cGAN.

The detail of the optimization of the training process of DR-STN is explained in Algorithm 2 as follows;

Algorithm 2 Minibatch Adaptive moment estimation [33] in the training of DR-cGAN. The parameters of discriminator D are updated to its hyperparameter k, which we used $k = 1$ as the least expensive option in our experiments. The default values of Adam parameters: $\eta = 2\text{E} - 3, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\text{E} - 08$.

- Initial parameters of G, Θ_G
- Initial parameters of D, Θ_D

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m examples $\{X^{(1)}, \ldots, X^{(m)}\}$ from the input object of spatial representation f_{Ob}.
- Sample minibatch of m examples $\{Y^{(1)}, \ldots, Y^{(m)}\}$ from the original object of temporal representation O_{obr}.
- Update the discriminator's parameters:

$$g_{\theta_D} \leftarrow \nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \left[\log \sigma D\left(X^{(i)}, Y^{(i)}\right) \right.$$

$$+\log\left(1-\sigma D\left(X^{(i)},G\left(X^{(i)}\right)\right)\right)\Big]$$

$$\theta_D \leftarrow Adam\left(g_{\theta_D},\theta_D,\eta,\beta_1,\beta_2,\epsilon\right)$$

end for

- Sample minibatch of *m* examples $\{X^{(1)},\ldots,X^{(m)}\}$ from the input object of spatial representation f_{Ob}.
- Sample minibatch of *m* examples $\{Y^{(1)},\ldots,Y^{(m)}\}$ from the original object of temporal representation O_{obr}.
- Update the generator's parameters:

$$g_{\theta_G} \leftarrow \nabla_{\theta_G}\frac{1}{m}\sum_{i=1}^{m}\Big[\log\left(1-\sigma D\left(X^{(i)},G\left(X^{(i)}\right)\right)\right) +\lambda\left\|Y^{(i)}-G\left(X^{(i)}\right)\right\|\Big]$$

$$\theta_G \leftarrow Adam\left(g_{\theta_G},\theta_G,\eta,\beta_1,\beta_2,\epsilon\right)$$

end for

Algorithm 2 stops updating the parameters after the reconstruction loss is less than 2E-3. Similar to DSTN, we evaluate the performance of *G* in generating synthesized temporal representation by computing the pixel accuracy and SSIM metrics as explained in Section 2.2 and use it in our testing experiment.

2.4. Evaluation Criteria

- Receiver Operating Characteristic (ROC)

The Receiver Operation Characteristic curve (ROC) is a standard method used for evaluating the performance of an anomaly detection system. It is a plot that indicates a comparison between True Positive Rate (*TPR*) and False Positive Rate (*FPR*) at various threshold criteria and benefits the analysis of the decision-making process [13].

In the anomaly detection observation, the abnormal events that are correctly determined as the positive detections (abnormal event) from the entire positive ground truth data are represented as *TPR* known as the probability of detection. The more the curve of *TPR* goes up, the better the detection accuracy of abnormal events is. The normal event (negative data) that are incorrectly determined as the positive detections from the entire negative ground truth data are represented as *FPR*. The higher *FPR* means the higher rate of the misclassification of normal events. There are four types of binary predictions for *TPR* and *FPR* computation, as described below.

True Positive (*TP*) is the correct positive detection of an abnormal event when the prediction outcome and the ground truth data are positive (abnormal event).

False Positive (*FP*) is the false positive detection when the outcome is predicted as positive (abnormal event), but the ground truth data is negative (normal event), meaning that the normal event is incorrectly detected as an abnormal event. This problem often occurs in the video anomaly detection task (e.g., a walking person is detected as an anomaly).

True Negative (*TN*) is the correct detection of a normal event when the outcome is predicted as negative (normal event) and the ground truth data is also negative.

False Negative (*FN*) is the incorrect detection when the outcome is predicted as negative (normal event) and the ground truth data is positive (abnormal event).

Hence, *TPR* and *FPR* can be computed, as shown in Eq. (2.19) and Eq. (2.20), respectively:

$$TPR = \frac{TP}{TP+FN}, \tag{2.19}$$

$$FPR = \frac{FP}{FP+TN}. \tag{2.20}$$

Moreover, there are two related evaluation methods of the ROC as follows.

- Area Under Curve (AUC)

Area Under Curve, also known as AUC, is used in classification analysis problems to define the best prediction model. It is computed from all the areas under the ROC curve, where *TPR* is plotted against *FPR*. The higher value of AUC indicates the superior performance of the model. Ideally, the model is a perfect classifier when all positive data are ranked above all negative data (AUC = 1). In practice, most of the AUC results are required in the range between 0.5 and 1.0 (AUC = [0.5,1]), meaning that the random positive data are ranked higher than the random negative data (greater than 50%). Besides, the worst case is when all negative data are ranked above all positive data, leading the AUC to 0 (AUC = 0). Hence, AUC classifiers can be defined as AUC ∈ [0,1] where AUC values for real-world use are greater than 0.5. The AUC values that are less than 0.5 are not acceptable for the model. To conclude, we prefer higher AUC values than the lower ones.

- Equal Error Rate (EER)

Apart from the AUC, the performance of the model can be quantified by using an Equal Error Rate known as EER. It is the point that occurred on the ROC curves when there is an equal probability of misclassified positive or negative data where *FPR* equals 1-*TPR*. The EER comes from the intersection of the ROC curve and the unit square diagonal. The lower the EER values, the better the performance of the model. It is in contrast to the AUC values, which refer to the higher value.

- Frame-level Evaluation for Anomaly Detection

We evaluate the quantitative performance of the proposed method based on two criteria: frame level and pixel level. The frame-level evaluation is used for detecting the anomalous pixel in the frame at the time. If one or more anomalous pixels are detected, the frame will be labeled as the abnormal frame no matter what size and location of the objects are. In this case, the frame becomes the true positive (*TP*) when the ground truth is also abnormal as the test frame. However, if the ground truth is not abnormal, it will become a false positive (*FP*). In this work, the ROC curve is used to

illustrate the accuracy of frame-level anomaly detection on the UCSD dataset and compare the results with other state-of-the-art works. Also, the AUC and the EER are applied as the criteria for evaluating the results.

- Pixel-level Evaluation for Anomaly Localization

This aims to evaluate the accuracy of the anomalous events at the pixel level. It attempts to localize the anomaly pixel in the scenarios and enhances the precision of anomaly detection from the frame-level measurement. To indicate whether the frame is the true positive (*TP*) or not, the detected abnormal area is needed to be overlapped more than 40% with the ground truth [26]. In addition, if one pixel is detected as abnormal events, the frame will be distinguished as the false positive (*FP*). The pixel-level evaluation is more challenging and stricter than the frame-level evaluation because of the complexity of anomaly localization. For the accuracy measurement, the ROC curve is used to measure the accuracy of pixel-level anomaly localization.

- Pixel Accuracy

The pixel accuracy metric is a standard semantic segmentation evaluation [28]. In this work, there are two classes; a foreground region class and a background region class. Let n_{ij} be the number of wrong classified pixels of class i, and n_{ti} be the total number of pixels of class i. The pixel accuracy can be computed by $\sum_i n_{ii} / \sum_i n_{ti}$.

- Structural SIMilarity Index (SSIM)

SSIM index is used to measure the similarity between the original and the synthesized images [49]. The more the synthesized image looks like the original image, the more efficient the model is.

2.5. Discussion

In this section, we discuss our proposed methods, DSTN and DR-STN, and point out their advantages and limitations as follows.

- Advantages and Disadvantages

For our DSTN framework, we designed the video anomaly detection framework by embedding successful GAN with the additional pre- and post-processing approaches. Regarding the pre-processing approach, we proposed the simple but yet effective background removal method. The experiment of the background removal method in Section 2.1.4.H clearly shows that the background removal method can preserve the full appearance of foreground objects in the scene, making it suitable for model learning. However, after the differentiation, we noticed that the output of the irregular frame contains a large area of unnecessary pixels over the actual object in the scene. This area probably occurs because of the motion of the object. In other words, a fast-moving foreground object tends to provide a large reconstruction error. To solve this problem, we proposed edge wrapping as a post-processing approach to eliminate these unnecessary pixels from the reconstruction error of the GAN model. According to Section 2.1.4.H, edge wrapping also helps to remove the false-positive detection results. Thus, following the experimental results, we can conclude that using

pre- and post-processing approaches benefit GAN in terms of the anomaly localization. Besides, the ROC curves show that our proposed DSTN overcomes other state-of-the-art methods in both frame-level anomaly detection and pixel-level anomaly localization, especially at the frame level. This is because it is designed based on only one deep network with help from the edge wrapping.

However, our first proposed DSTN method has some limitations described as follows:

i) The background removal is not robust to illumination changes, noise, and occlusion, especially in complex and crowded scenes;
ii) The patch extraction is not good enough to provide a concise spatial object for GAN learning;
iii) The GAN model has a problem in learning low-level features in the scene (see Section 2.1.4.G);
iv) The edge wrapping does not guarantee that the false positive can be detected.

Due to the limitations of DSTN, we designed new algorithms to improve the performance of anomaly localization. We take advantage of the pre- and post-processing concepts from DSTN for our novel DR-STN.

For DR-STN, we first introduce the object detector to extract the spatial objects in the complex and crowded scenes. This object detector helps our model in learning the normal patterns of the objects more precisely. Second, we integrate the residual units, which are proposed to enhance the low-level feature information, in cGAN, resulting in a good performance on the pixel accuracy and SSIM, as indicated in Section 2.2.4. Moreover, the objective function in our model is smoother in model learning than the previous work due to the use of the binary cross-entropy loss with logits loss in the discriminator. Finally, the online hard negative mining (OHNM) method is presented to remove the false-positives in the final output without retraining the model, making the proposed DR-STN suitable for real use. Regarding the AUC and EER, DR-STN outperforms DSTN in all aspects, as shown in Table 2.11. Thus, the proposed DR-STN can better target the correct observation of normal and abnormal events in the crowded scene.

Table 2.11 AUC and EER Performance comparison between DSTN and DR-STN

Method	**UCSD Ped1 (F) AUC/EER**	**UCSD Ped1 (P) AUC/EER**	**UCSD Ped2 (F) AUC/EER**	**UCSD Ped2 (P) AUC/EER**	**CUHK Avenue (F) AUC/EER**	**UMN (F) AUC/EER**
DSTN	98.5%/5.2%	77.4%/27.3%	95.5%/9.4%	83.1%/21.8%	87.9%/20.2%	99.6%/-
DR-STN	**98.8%/2.9%**	**82.5%/21.5%**	**97.6%/6.9%**	86.4%/**16.3%**	**90.8%/11.0%**	**99.7%/-**

Examples of the comparison between DSTN and DR-STN methods on the UCSD, CUHK Avenue, and UMN datasets are shown in Fig. 2.26.

(a) DSTN on UCSD Ped1 (b) DR-STN on UCSD Ped1

(c) DSTN on UCSD Ped2 (d) DR-STN on UCSD Ped2

(e) DSTN on CUHK Avenue (f) DR-STN on CUHK Avenue

(g) DSTN on CUHK Avenue (h) DR-STN on CUHK Avenue

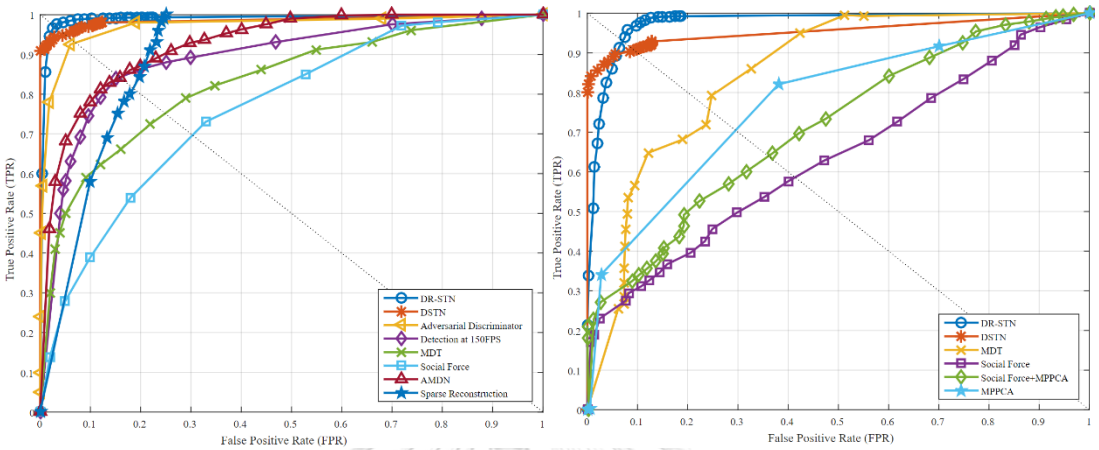
(i) DSTN on CUHK Avenue (j) DR-STN on CUHK Avenue

Fig. 2.26 Examples of the comparison between DSTN and DR-STN methods on UCSD, CUHK Avenue, and UMN datasets

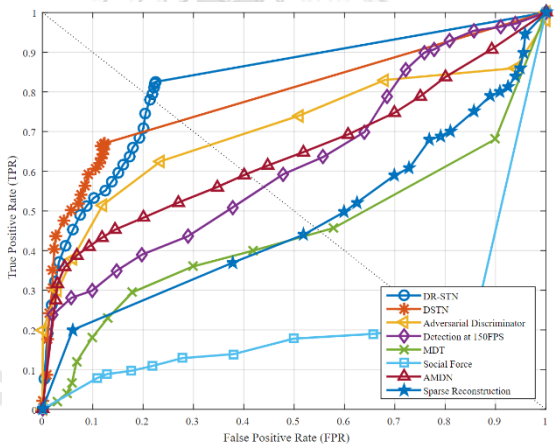
In Fig. 2.26, we demonstrate the same scene for each dataset to compare the performance of each proposed method. Fig. 2.26 shows that DR-STN provides a fuller anomalous mask on objects than DSTN in most scenes except in Fig. 2.26 (j) on the UMN dataset, in which some objects are missing. This misdetection problem might occur due to the low resolution of the input image that causes the errors in the object detector. However, DR-STN still can detect this frame as an abnormal frame based on the neighboring objects. Moreover, DR-STN can effectively remove the false-positives that occur on the objects, as shown in Fig. 2.26 (a), where there is some false detection on the normal events (e.g., the lower part of walking pedestrian) represented as red and green colors. The performance of removing false-positives of the proposed DR-STN significantly benefits the use of anomaly detection and localization system in real-world cases.

We also demonstrate the qualitative results of our proposed methods compared with other state-of-the-art works in the ROC curves, as shown in Fig. *2.27*, consisting

of (a) frame-level evaluation on the UCSD Ped1, (b) a frame-level evaluation on the UCSD Ped2, and (c) a pixel-level evaluation on the UCSD Ped1. Our proposed DR-STN (dark blue circle curves) outperforms all the competing methods, including the proposed DSTN (red star curves). The DR-STN's curves have the highest growth on the *TPR*, meaning that the abnormal events in DR-STN are precisely detected and localized in both frame-level and pixel-level evaluations.



(a) frame-level evaluation on UCSD Ped1 (b) frame-level evaluation on UCSD Ped2



(c) pixel-level evaluation on UCSD Ped1

Fig. 2.27 ROC comparison on UCSD dataset

- Model Parameters

We show our model parameters and sizes and the AUC performance in both frame-level and pixel-level evaluations to deliver the models' performance and characteristics. The comparison of AUC and model parameters and sizes between DSTN and DR-STN on the UCSD dataset is shown in Table 2.12.

Table 2.12 Comparison of AUC and model parameters and sizes between DSTN and DR-STN on the UCSD dataset

Model	Average AUC		Total parameters		Size (MB)
	Frame	Pixel	Generator	Discriminator	Generator
DSTN	97	80.25	30,631,299	6,279,620	122.6
DR-STN	98.2	84.45	43,230,400	2,768,705	173.0

Table 2.12 shows that DR-STN provides higher AUC in both frame-level and pixel-level evaluations than DSTN for the UCSD dataset. DR-STN has more parameters in the generator and is bigger than DSTN because we built a more comprehensive network (wider) to enhance the learning performance in the spatiotemporal translation of the generator instead of a deeper network. As the wider network is significantly more effective than the deeper one [20], we introduce the residual units inside the original GAN [17], resulting in an increasingly small number of total parameters but high accuracy.

Furthermore, we show the comparison between AUC in the frame-level and pixel-level evaluations and the models' parameters on the UCSD dataset, as shown in Fig. 2.28 and Fig. 2.29 as follows.

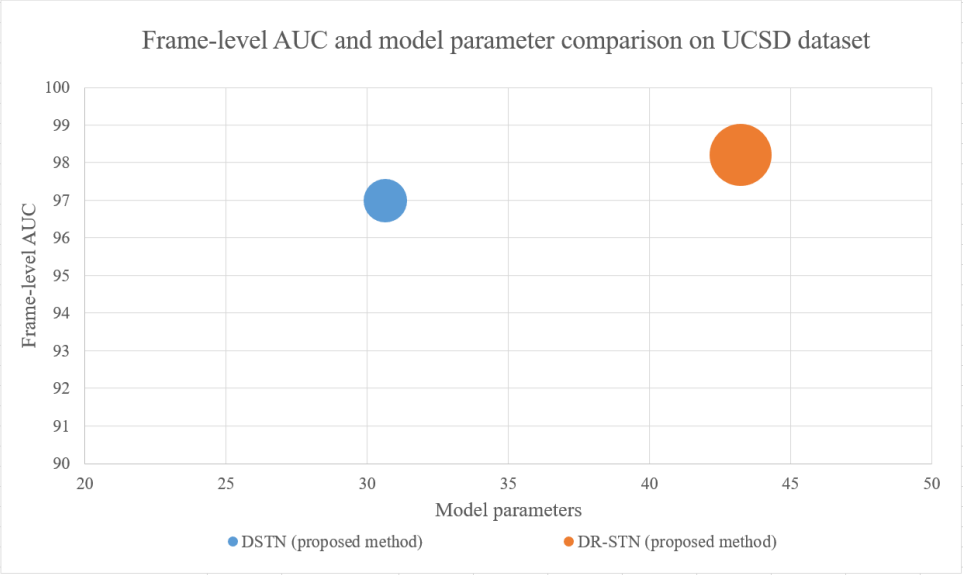


Fig. 2.28 Comparison of frame-level AUC and models' parameters on UCSD dataset

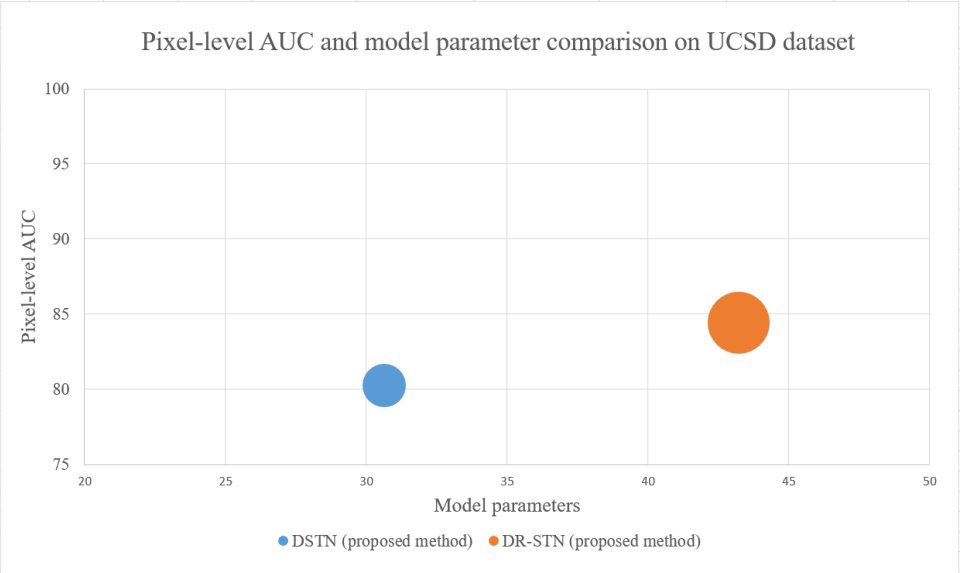


Fig. 2.29 Comparison of pixel-level AUC and models' parameters on UCSD dataset

In Fig. 2.28 and Fig. 2.29, the proposed DR-STN is presented in a circle orange mark while the proposed DSTN is a blue circle mark. The size of the mark in the plot represents the size of the model. The bigger size means that the model has more parameters. Thus, Fig. 2.28 and Fig. 2.29 show that DR-STN achieves higher AUC in both evaluations but has more parameters than DSTN.

- Computational Time

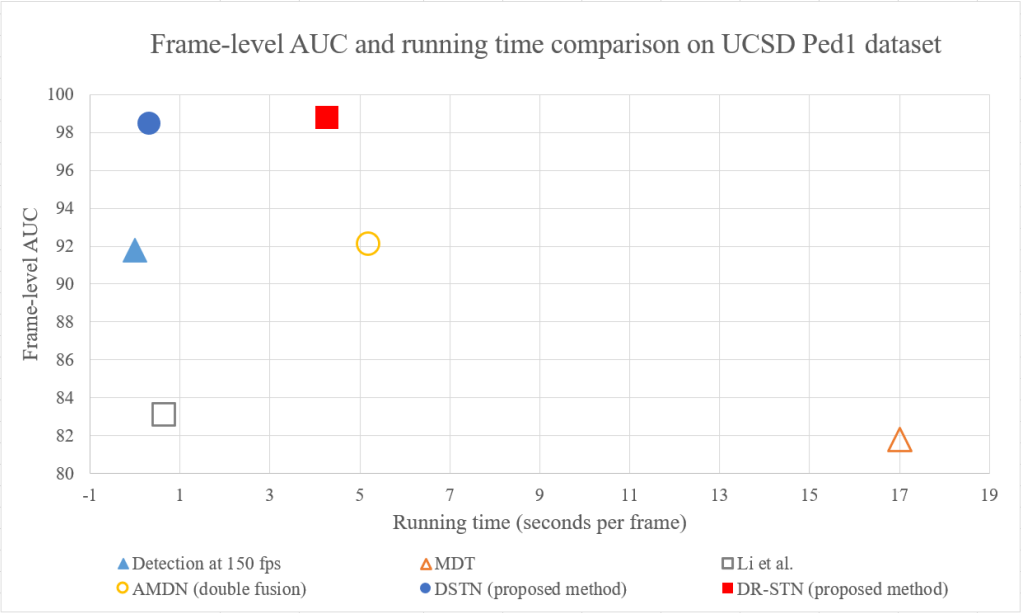
To comprehensively analyze our proposed methods, we compare the computational time in frames per second using CPU with other state-of-the-art methods on all three datasets: UCSD, UMN, and CUHK Avenue. The comparison of computational time during testing is shown in Table 2.13.

Table 2.13 Computational time comparison during testing (seconds per frame).

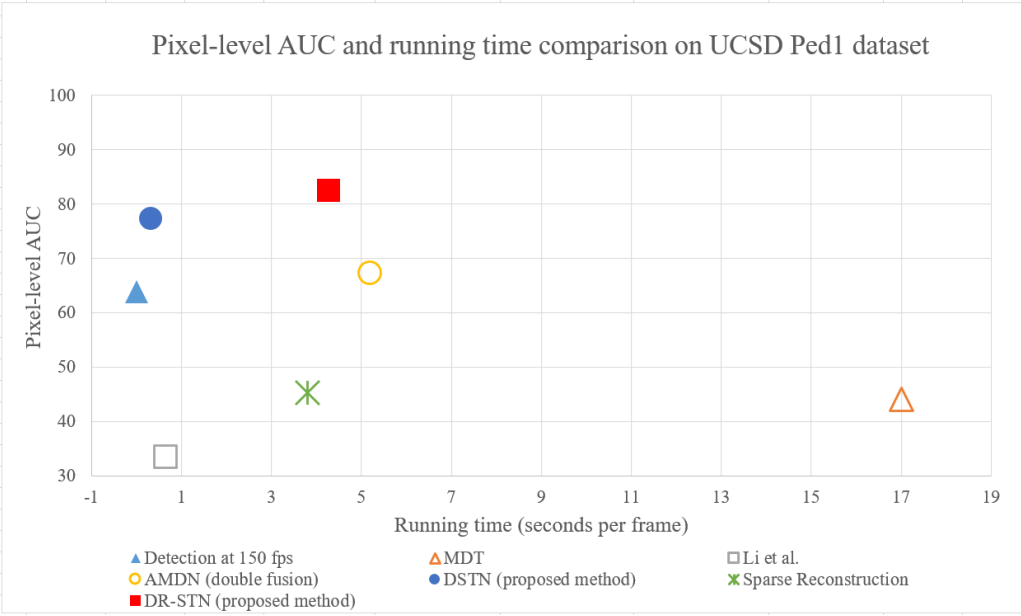
Method	**CPU**	**GPU**	**Memory**	**Running Time**			
				Ped1	**Ped2**	**UMN**	**Avenue**
Sparse Reconstruction	2.6GHz	-	2.0GB	3.8	-	0.8	-
Detection at 150 fps	3.4GHz	-	8.0GB	**0.007**	-	-	**0.007**
MDT	3.9GHz	-	2.0GB	17	23	-	-
Li *et al.*	2.8GHz	-	2.0GB	0.65	0.80	-	-
AMDN (double fusion)	2.1GHz	Nvidia Quadro K4000	32GB	5.2	-	-	-
DSTN	2.8GHz	-	24GB	0.315	**0.319**	**0.318**	0.334
DR-STN	3.4GHz	-	24GB	4.26	4.44	4.07	3.62

From Table 2.13, it shows that DSTN is faster than DR-STN for all datasets. The reason is that DSTN implements a small number of patches extracted from the scene, while DR-STN implements all individual objects in the scene. This issue refers to an inherent speed-accuracy tradeoff as DR-STN achieves higher accuracy than DSTN but requires more computational time for precisely detecting multiple objects in crowded scenes.

In particular, we also show our performance in terms of both accuracy and time complexity. We compare the AUC in the frame level and pixel level and the running time of our proposed methods with other state-of-the-art works (when available) on the UCSD Ped1 and Ped2 datasets, as shown in Fig. 2.30 and Fig. 2.31, respectively.

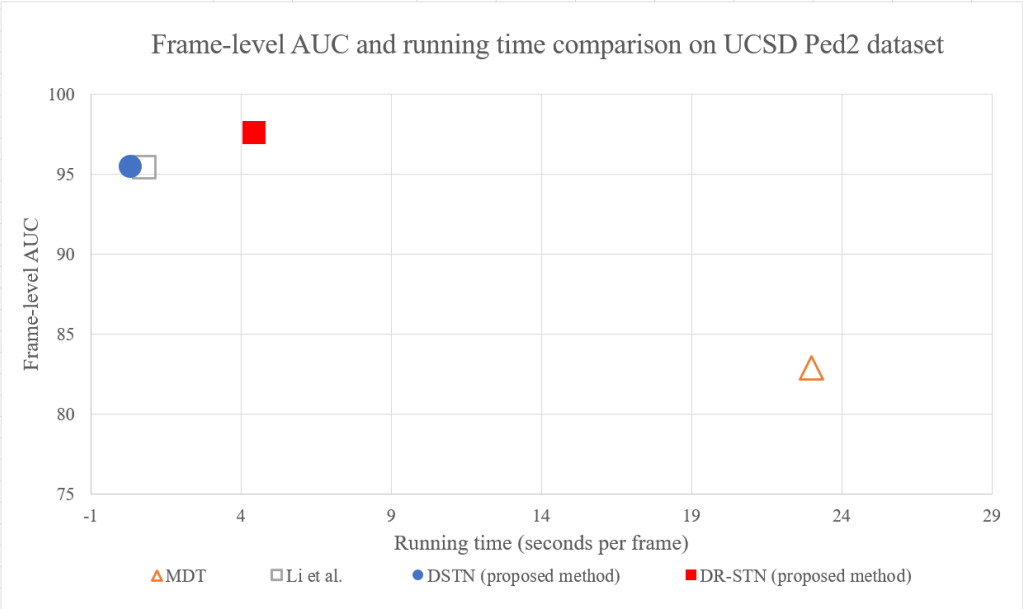


(a) Frame-level AUC and running time comparison on UCSD Ped1

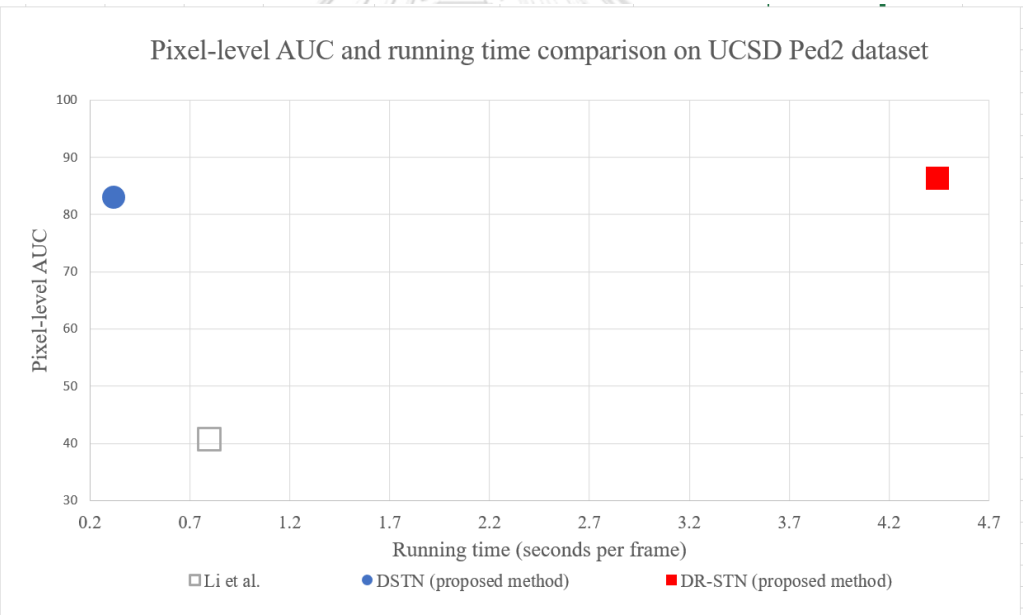


(b) Pixel-level AUC and running time comparison on UCSD Ped1

Fig. 2.30 Comparison of AUC and running time on UCSD Ped1



(a) Frame-level AUC and running time comparison on UCSD Ped2



(b) Pixel-level AUC and running time comparison on UCSD Ped2

Fig. 2.31 Comparison of AUC and running time on UCSD Ped2

In Fig. 2.30 and Fig. 2.31, we represent our proposed methods as a blue circle mark for DSTN and a red square mark for DR-STN. Fig. 2.30 shows that DR-STN outperforms other state-of-the-art works in terms of the frame-level AUC accuracy on the UCSD Ped1 for both frame-level and pixel-level evaluations. In the frame level, even DR-STN takes more time to implement than DSTN [15], Detection at 150 fps [29], and Li *et al.* [26], it can still surpass other competing works like AMDN [52]

and MDT [30]. Similar to the frame level, these works [15; 26; 29], and Sparse reconstruction [7] surpass DR-STN regarding the running time in the pixel level. However, DR-STN is faster than AMDN [52] and MDT [30]. Fig. 2.31 shows that DR-STN outperforms other competing works regarding the AUC in both frame-level and pixel-level evaluations, while DSTN still performs well in both AUC and running time aspects.

Overall, DSTN performs the best performance for surveillance videos with respect to both accuracy and running time concerns as it achieves high AUC and low time complexity. As requiring high accuracy leads to the speed-accuracy tradeoff, DR-STN is also considered as one of the best approaches as it achieves the highest AUC for both evaluations with an acceptable time consumption for real use. According to our experimental results, we can conclude that both of our proposed methods, DSTN and DR-STN, outperform other competing methods. They all achieve high AUC value in both frame-level and pixel-level evaluations and providing a good running time for surveillance videos.

CHAPTER 3
CONCLUSION

3.1. Conclusion

In this thesis, we propose two novel frameworks for anomaly detection in crowded scenes: i) unsupervised anomaly detection and localization based on deep spatiotemporal translation network (DSTN) and ii) deep residual spatiotemporal translation network for video anomaly detection and localization (DR-STN).

The proposed DSTN framework is embedded with a deep convolution neural network of GAN based Edge Wrapping approach, which brings advantages to anomaly localization. The deep spatiotemporal translation network is designed to learn the appearance and motion representations using the fusion and the concatenation of patches for combining the learned features. Additionally, our proposed DSTN does not rely on any prior knowledge in order to design features for the input (as we use raw pixels) and does not involve low-level object analysis, such as object detection and tracking.

The proposed DR-STN is embedded with a wider deep residual convolutional GAN (DR-cGAN) and online hard negative mining (OHNM) which benefits in reducing false-positive anomaly detection. The DR-cGAN is designed to translate appearance and motion representations by integrating the residual units, residual connections, and cGAN. Additionally, our DR-cGAN takes only raw pixels as the input from the object detector without relying on any prior knowledge of hand-crafted features.

For both of our proposed methods, we conducted extensive experiments and compared them with other state-of-the-art methods on three publicly available benchmarks, including the UCSD pedestrian, UMN, and CUHK Avenue. We clearly show that our DSTN outperforms other state-of-the-art methods of accuracy and time complexity that surpass most baseline methods. Additionally, our DR-STN performs best regarding the accuracy, especially at the pixel level. In DR-STN, we obtain the highest AUC value in both frame-level and pixel-level evaluations for all datasets and achieve a good running time suitable for surveillance videos. To conclude, our proposed methods are effective and robust for anomaly event detection and localization in the crowded scenes for surveillance videos.

3.2. Suggestion

A multimodal learning model might enhance the anomaly detection and localization performance and reduce time consumption for future work. Moreover, other regular events should be observed for improving the performance and robustness of the model in learning normal patterns for real-world use.

REFERENCES

[1]. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934.

[2]. Boiman, O., Irani, M., 2007. Detecting irregularities in images and in video. International journal of computer vision 74, 17-31.

[3]. Chen, T., Hou, C., Wang, Z., Chen, H., 2018. Anomaly detection in crowded scenes using motion energy model. Multimedia Tools and Applications 77, 14137-14152.

[4]. Cheng, K.-W., Chen, Y.-T., Fang, W.-H., 2015. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2909-2917.

[5]. Cheng, Y., Wang, D., Zhou, P., Zhang, T., 2017. A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282.

[6]. Colque, R.V.H.M., Caetano, C., de Andrade, M.T.L., Schwartz, W.R., 2016. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Transactions on Circuits and Systems for Video Technology 27, 673-682.

[7]. Cong, Y., Yuan, J., Liu, J., 2011. Sparse reconstruction cost for abnormal event detection, CVPR 2011. IEEE, pp. 3449-3456.

[8]. Cong, Y., Yuan, J., Tang, Y., 2013. Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE transactions on information forensics and security 8, 1590-1599.

[9]. Coşar, S., Donatiello, G., Bogorny, V., Garate, C., Alvares, L.O., Brémond, F., 2016. Toward abnormal trajectory and event detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology 27, 683-695.

[10]. Du, D., Qi, H., Huang, Q., Zeng, W., Zhang, C., 2013. Abnormal event detection in crowded scenes based on structural multi-scale motion interrelated patterns, 2013 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1-6.

[11]. Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F., 2020. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding, 102920.

[12]. Fang, Z., Fei, F., Fang, Y., Lee, C., Xiong, N., Shu, L., Chen, S., 2016. Abnormal event detection in crowded scenes based on deep learning. Multimedia Tools and Applications 75, 14617-14639.

[13]. Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters 27, 861-874.

[14]. Feng, Y., Yuan, Y., Lu, X., 2017. Learning deep event models for crowd anomaly detection. Neurocomputing 219, 548-556.

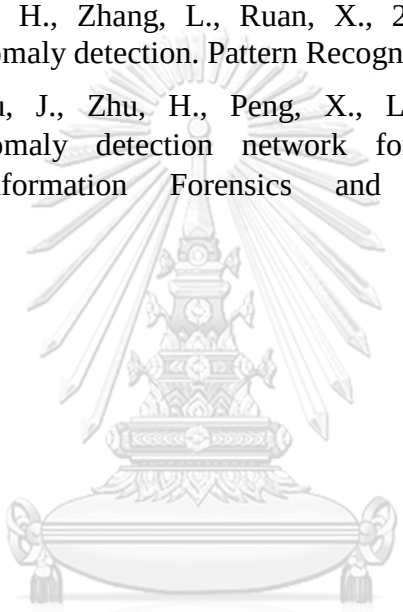
[15]. Ganokratanaa, T., Aramvith, S., Sebe, N., 2020. Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network. IEEE Access 8, 50312-50329.

[16]. Gates, K., 2015. Professionalizing police media work: Surveillance video and the forensic sensibility, Images, Ethics, Technology. Routledge, pp. 53-69.

[17]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, Advances in neural information processing systems, pp. 2672-2680.
[18]. Han, S., Pool, J., Tran, J., Dally, W., 2015. Learning both weights and connections for efficient neural network, Advances in neural information processing systems, pp. 1135-1143.
[19]. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S., 2016. Learning temporal regularity in video sequences, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 733-742.
[20]. He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.
[21]. He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, European conference on computer vision. Springer, pp. 630-645.
[22]. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125-1134.
[23]. Kim, J., Grauman, K., 2009. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates, 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2921-2928.
[24]. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278-2324.
[25]. Li, A., Miao, Z., Cen, Y., Liang, Q., 2016. Abnormal event detection based on sparse reconstruction in crowded scenes, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1786-1790.
[26]. Li, W., Mahadevan, V., Vasconcelos, N., 2013. Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence 36, 18-32.
[27]. Liu, W., Luo, W., Lian, D., Gao, S., 2018. Future frame prediction for anomaly detection–a new baseline, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536-6545.
[28]. Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.
[29]. Lu, C., Shi, J., Jia, J., 2013. Abnormal event detection at 150 fps in matlab, Proceedings of the IEEE international conference on computer vision, pp. 2720-2727.
[30]. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N., 2010. Anomaly detection in crowded scenes, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1975-1981.
[31]. Maji, P., Mullins, R., 2018. On the reduction of computational complexity of deep convolutional neural networks. Entropy 20, 305.
[32]. Mashalla, Y., 2014. Impact of computer technology on health: Computer Vision Syndrome (CVS). Medical Practice and Reviews 5, 20-30.
[33]. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems, pp. 8026-8037.

[34]. Ramachandra, B., Jones, M., Vatsavai, R., 2020. Learning a distance function with a Siamese network to localize anomalies in videos, The IEEE Winter Conference on Applications of Computer Vision, pp. 2598-2607.
[35]. Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., Sebe, N., 2018. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1689-1698.
[36]. Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N., 2017. Abnormal event detection in videos using generative adversarial nets, 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1577-1581.
[37]. Ravanbakhsh, M., Sangineto, E., Nabi, M., Sebe, N., 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds, 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1896-1904.
[38]. Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234-241.
[39]. Roshtkhari, M.J., Levine, M.D., 2013. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Computer vision and image understanding 117, 1436-1452.
[40]. Sabokrou, M., Fathy, M., Hoseini, M., 2016. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electronics Letters 52, 1122-1124.
[41]. Sabokrou, M., Fayyaz, M., Fathy, M., Klette, R., 2017. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing 26, 1992-2004.
[42]. Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., Klette, R., 2018. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding 172, 88-97.
[43]. Sainath, T.N., Kingsbury, B., Sindhwani, V., Arisoy, E., Ramabhadran, B., 2013. Low-rank matrix factorization for deep neural network training with high-dimensional output targets, 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp. 6655-6659.
[44]. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, Advances in neural information processing systems, pp. 2234-2242.
[45]. Singh, K., Rajora, S., Vishwakarma, D.K., Tripathi, G., Kumar, S., Walia, G.S., 2020. Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets. Neurocomputing 371, 188-198.
[46]. Sun, J., Wang, X., Xiong, N., Shao, J., 2018. Learning sparse representation with variational auto-encoder for anomaly detection. IEEE Access 6, 33353-33361.
[47]. Tang, X., Zhang, S., Yao, H., 2013. Sparse coding based motion attention for abnormal event detection, 2013 IEEE International Conference on Image Processing. IEEE, pp. 3602-3606.
[48]. Ullrich, K., Meeds, E., Welling, M., 2017. Soft weight-sharing for neural network compression. arXiv preprint arXiv:1702.04008.

[49]. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600-612.
[50]. Wei, H., Xiao, Y., Li, R., Liu, X., 2018. Crowd abnormal detection using two-stream fully convolutional neural networks, 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, pp. 332-336.
[51]. Xiao, T., Zhang, C., Zha, H., 2015. Learning to detect anomalies in surveillance video. IEEE Signal Processing Letters 22, 1477-1481.
[52]. Xu, D., Yan, Y., Ricci, E., Sebe, N., 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding 156, 117-127.
[53]. Zhang, Y., Lu, H., Zhang, L., Ruan, X., 2016. Combining motion and appearance cues for anomaly detection. Pattern Recognition 51, 443-452.
[54]. Zhou, J.T., Du, J., Zhu, H., Peng, X., Liu, Y., Goh, R.S.M., 2019. Anomalynet: An anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security 14, 2537-2550.



REFERENCES

VITA

NAME	Thittaporn Ganokratanaa
DATE OF BIRTH	5 July 1992
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Chulalongkorn University (2017-present) Chulalongkorn University (2015-2017) King Mongkut's University of Technology Thonburi (2011-2015)
HOME ADDRESS	55/5 The Place by Alicha, Phutthabucha 32 Alley, Phutthabucha Road, Bang Mod, Thung Kru, Bangkok 10140
PUBLICATION	THITTAPORN GANOKRATANAA received the B.Sc. (first-class honors) degree in Media Technology from King Mongkut's University of Technology Thonburi, Bangkok, Thailand, in 2015. She received the M.Eng. in Electrical Engineering (EE) from Chulalongkorn University (CU), Bangkok, Thailand, in 2017. She is pursuing a Ph.D. degree in Electrical Engineering at Chulalongkorn University, Bangkok, Thailand. She was a lecturer in Image Processing at King Mongkut's University of Technology Thonburi (KMUTT). Her research interest includes Computer Vision and Machine Learning, specifically Human-Computer Interaction for Surveillance Videos. She published about 20 papers in international conference proceedings and journals. She received an acceptance for publishing an international book chapter.
AWARD RECEIVED	Ms. Thittaporn's awards and honors include the trophy of appreciation for being a representative of Phra Maha Mongkut (Crown of King Rama IV), the gold medal for the excellent academic score from KMUTT, the certification of appreciation for volunteering at the EE department from CU, best paper awards (ICACME, SKIMA), best student paper awards (SKIMA, SNLP), the silver medal from Inventions Geneva, 1st runner-up of Three Minute Thesis® competition, 2nd runner-up of National Software Contest (NSC), and the Royal Monogram Brooch Investiture of His Majesty King Phra Pok Klao (King Rama VII) and Her Majesty Queen Rambai Barni of Siam from the Army Artillery Club, Pol. AAA (Antiaircraft Artillery).