



HAL
open science

Enhancing Video Anomaly Detection by Leveraging Advanced Deep Learning Techniques

Wenhao Shao

► **To cite this version:**

Wenhao Shao. Enhancing Video Anomaly Detection by Leveraging Advanced Deep Learning Techniques. Cryptography and Security [cs.CR]. Institut Polytechnique de Paris, 2023. English. NNT : 2023IPPAS012 . tel-04328688

HAL Id: tel-04328688

<https://theses.hal.science/tel-04328688>

Submitted on 7 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2023IPPAS012

Thèse de doctorat



Enhancing Video Anomaly Detection by Leveraging Advanced Deep Learning Techniques

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626 Ecole Doctorale de l'Institut Polytechnique de Paris (ED IP
Paris)

Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Evry, le 21/11/2023, par

WENHAO SHAO

Composition du Jury :

Patricia Desgreys Professor, IP Paris, Telecom Paris - France	Président
Shiping Wang Professor, Fuzhou University - China	Rapporteur
Ioan Marius Bilasco Professor, University of Lille - France	Rapporteur
Bin Gu Assistant Professor, Mohamed bin Zayed University of AI - UAE	Examineur
Noel Crespi Professeur, IP Paris, Telecom SudParis - France	Directeur de thèse
Praboda Rajapksha Researcher, IP Paris, Telecom SudParis - France	Co-directeur de thèse

Title : Enhancing Video Anomaly Detection by Leveraging Advanced Deep Learning Techniques

Keywords : Video Surveillance, Space Security, Video Analysis, Video Classification, Anomaly Detection, Deep Learning, Object Detection, Action Recognition, Target Tracking, Spatio-Temporal Features, Optical Flow, Frame Feature, Temporal Convolutional Network, U-Net, Convolutional-3D(C3D), Inflated-3D(I3D).

Abstract : Security in public spaces is a primary concern across different domains and the deployment of real-time monitoring systems addresses this challenge. Video surveillance systems employing deep learning techniques allows for the effective recognition of anomaly events. However, even with the current advances in anomaly detection methods, distinguishing abnormal events from normal events in real-world scenarios remains a challenge because they often involve rare, visually diverse, and unrecognizable abnormal events. This is particularly true when relying on supervised methods, where the lack of sufficient labeled anomaly data poses a significant challenge for distinguishing between normal and abnormal videos. As a result, state-of-the-art anomaly detection approaches utilize existing datasets to design or learn a model that captures normal patterns, which is then helpful in identifying unknown abnormal patterns. During the model design stage, it is crucial to label videos with attributes such as abnormal appearance, behavior, or target categories that deviate significantly from normal data, marking them as anomalies. In addition to the lack of labeled data, we identified three challenges from the literature : 1) insufficient representation of temporal feature, 2) lack of precise positioning of abnormal events and 3) lack the consistency research of temporal feature and appearance feature. The objective of my thesis is to propose and investigate advanced video anomaly detection methods by addressing the aforementioned challenges using novel concepts and utilizing weak supervision and unsupervised models rather than relying on supervised models.

We actively explored the applications of new video processing technologies, including action recognition, target detection, optical flow feature extraction, re-

presentation learning, and contrastive learning in order to utilize them in video anomaly detection models. Our proposed models comparatively analysed with baseline models. This comparative analysis are conducted using prevalent public datasets, including UCSD(Ped2), Avenue, UCF-Crime, and Shanghai-tech.

The first contribution addresses the first challenge outlined above by introducing an enhanced Temporal Convolutional Network (TCN). This novel TCN model learns dynamic video features and optimizes features to mitigate errors due to contrastive learned initial weights. This method enhances the overall capability of weakly supervised models by reducing the loss caused by initial parameters in contrastive learning. Nevertheless, weakly supervised learning only reduces the reliance on labeled data but does not eliminate the dependence on such data. Hence, our subsequent two contributions rely on unsupervised learning to addressing the other two challenges mentioned above. The second contribution combines the self-attention mechanism to prioritize the weights of areas with obvious dynamic fluctuations in frames. And, during the testing, abnormal areas are located through comparison of object detection and loss functions. The combination of self-attention mechanism and object detection significantly improves the detection accuracy and expands the functionality. The third contribution explores the integration of collaborative teaching network models, which bridges consistency between optical flow information and appearance information. This integration aims to enhance the spatio-temporal capture capabilities of unsupervised models. The overall performance and capabilities of the unsupervised model are significantly enhanced compared to the other baseline models.

Titre : Amélioration de la Détection d'Anomalies Vidéo en Exploitant des Techniques Avancées d'Apprentissage Profond

Mots clés : Surveillance vidéo, sécurité spatiale, analyse vidéo, classification vidéo, détection d'anomalies, apprentissage en profondeur, détection d'objets, reconnaissance d'action, suivi de cible, caractéristiques spatio-temporelles, flux optique, fonction de trame, réseau convolutif temporel, U-Net, convolutionnel-3D (C3D), Gonflé-3D(I3D).

Résumé : La sécurité est une préoccupation majeure dans différents domaines, et le déploiement de systèmes de surveillance en temps réel permet de relever ce défi. En utilisant des techniques d'apprentissage profond, il permet de reconnaître efficacement les événements anormaux. Cependant, même avec les avancées actuelles des méthodes de détection des anomalies, distinguer les événements anormaux des événements normaux dans les scénarios du monde réel reste un défi en raison d'événements anormaux rares, visuellement diversifiés et non reconnaissables de façon prévisible. Cela est particulièrement vrai lorsque l'on s'appuie sur des méthodes supervisées, où le manque de données d'anomalies labellisées pose un problème important pour distinguer les vidéos normales des vidéos anormales. Par conséquent, les approches de détection d'anomalies les plus récentes utilisent des ensembles de données existants pour concevoir ou apprendre un modèle qui capture les modèles normaux, ce qui permet ensuite d'identifier les modèles anormaux inconnus. Au cours de la phase de conception du modèle, il est essentiel de labelliser les vidéos avec des attributs tels qu'une apparence anormale, un comportement ou des catégories cibles qui s'écartent de manière significative des données normales, en les marquant comme des anomalies. Outre le manque de données labellisées, trois autres défis principaux ont été identifiés dans la littérature : 1) la représentation insuffisante des caractéristiques temporelles, 2) le manque de précision dans le positionnement des événements anormaux et 3) l'absence d'informations sur le comportement.

Nous avons exploré les applications des nouvelles technologies de traitement vidéo, notamment la reconnaissance des actions, la détection des cibles, l'extraction des caractéristiques du flux optique, l'apprentissage de la représentation et l'apprentissage

contrastif, afin de les utiliser dans les modèles de détection des anomalies vidéo. Les modèles que nous proposons sont analysés de manière comparative avec les modèles de référence. Cette analyse comparative est réalisée à l'aide de jeux de données publics courants, notamment UCSD(Ped2), Avenue, UCF-Crime et Shanghaitech.

La première contribution relève le premier point décrit ci-dessus en introduisant un réseau convolutionnel temporel (TCN) amélioré. Ce nouveau modèle de réseau convolutionnel temporel apprend les caractéristiques dynamiques de la vidéo et les optimise afin d'atténuer les erreurs dues aux poids initiaux appris de manière contrastive. Cette méthode améliore la capacité globale des modèles faiblement supervisés en réduisant la perte causée par les paramètres initiaux dans l'apprentissage contrastif. Néanmoins, l'apprentissage faiblement supervisé ne fait que réduire la dépendance à l'égard des données labellisées, mais ne l'élimine pas complètement. C'est pourquoi nos deux contributions suivantes s'appuient sur l'apprentissage non supervisé pour relever les deux autres défis mentionnés ci-dessus. La deuxième contribution combine le mécanisme d'auto-attention pour donner la priorité aux poids des zones présentant des fluctuations dynamiques évidentes dans les images. Lors des tests, les zones anormales sont localisées en comparant les fonctions de détection et de perte d'objets. La troisième contribution explore l'intégration de modèles de réseaux d'apprentissage collaboratifs, qui assurent la cohérence entre les informations sur le flux optique et les informations sur l'apparence. Cette intégration vise à améliorer les capacités de capture spatio-temporelle des modèles non supervisés. Les performances et les capacités globales du modèle non supervisé sont considérablement améliorées par rapport aux autres modèles de base.

**Doctor of Philosophy (PhD) Thesis
Institut-Mines Télécom, Télécom SudParis
& Institut Polytechnique de Paris (IP Paris)**

Specialization

COMPUTING, DATA AND ARTIFICIAL INTELLIGENCE

presented by

Wenhao Shao

**Enhancing Video Anomaly Detection by Leveraging Advanced
Deep Learning Techniques**

Committee:

Shiping Wang	Reviewer	Professor, Fuzhou University - China
Ioan Marius Bilasco	Reviewer	Professor, University of Lille - France
Patricia Desgreys	Examiner	Professor, IP Paris, Telecom Paris - France
Bin Gu	Examiner	Assistant Professor, Mohamed bin Zayed University of AI - UAE
Noel Crespi	Advisor	Professor, IP Paris, Telecom SudParis - France
Praboda Rajapksha	Co-advisor	Researcher, IP Paris, Telecom SudParis - France

**Thèse de Doctorat (PhD) de
Institut-Mines Télécom, Télécom SudParis
et l'Institut Polytechnique de Paris (IP Paris)**

Spécialité

INFORMATIQUE, DONNÉES ET INTELLIGENCE ARTIFICIELLE

présentée par

Wenhao Shao

**Amélioration de la Détection d'Anomalies Vidéo en Exploitant
des Techniques Avancées d'Apprentissage Profond**

Jury composé de :

Shiping Wang	Rapporteur	Professeur, Fuzhou University - China
Ioan-Marius Bilasco	Rapporteur	Professeur, CRISAL - Université de Lille - France
Patricia Desgreys	Examiner	Professeur, IP Paris, Télécom Paris - France
Bin Gu	Examiner	Professeur Adjoint, Mohamed bin Zayed University of AI - UAE
Noel Crespi	Directeur de thèse	Professeur, IP Paris, Telecom SudParis - France
Praboda Rajapksha	Co-encadrant	Chercheuse, IP Paris, Telecom SudParis - France

Dedication

To My Family

Acknowledgements

First and foremost, I would like to express my gratitude for Prof. Noel Crespi who gave me the great opportunity to do this research work and provided me all the guidance and support. I really admire him for the given freedom to conduct independent research. I always enjoyed working with him and it increased my exposure and knowledge.

I would like to thank my co-supervisor Dr. Praboda Rajapksha who has supported me throughout my research. She continuously provided encouragement and was always willing and enthusiastic to assist in any way at anytime. Her help enabled me to integrate into the laboratory research quickly. Most importantly, she advised on every entangled situation and motivated whenever I lacked inspiration.

I wish to express my sincere gratitude to my thesis reviewers, Dr Mengzhu wang and Dr. Ruliang Xiao for their useful reviews and suggestions, which helped me to improve the quality of my thesis. A special thank to Prof. Ioan Marius Bilasco, Prof. Wang Shiping, Patricia Desgreys, Dr. Gareth Tyson and Dr. Bin Gu for being the part of my jury as examiners for my thesis defense.

My special thanks to all the lovely team members of Samovar, Telecom SudParis, Institut Polytechnique de Paris, especially Guanlin Li and Dun Li, Zhenjiao Liu, Amir Jafari and Dr Reza Farahbakhsh. Thanks to Prof. Roberto who always gave useful suggestions and comments while doing projects and appreciate your dedication on those.

Special thank goes to Valerie Mateus, the secretary of RS2M Department. She was always very kind and generous in solving tedious administrative tasks. A deep thanks to Veronique Guy, the administrative responsible of PhD program, who always helped me a lot in dealing with PhD administrative tasks.

My profound love, respect and thank goes to my family members : my father, my mother, my brother. They always prayed for me, supported me and encouraged me to achieve this new and hard milestone of my life. My gratitude goes to My friends Xuechen Zhao and Mengzhu Wang for help me prepare the documents for travelling and pursuing my PhD aboard. Thank you for believing in me and supporting me, and without you I would not have made it through my PhD.

Last but not least, my biggest thanks to my wife Yanyan Wei and my child Yiting Shao, who always care for me and tolerating all the hardships and providing continuous support during my PhD. Thank you for the countless times you went out of your way to make sure I was comfortable and for all the compromises. You are truly a wonderful wife and thank you is a small word for all that you have done for me. I hope they find here the expression of my deep gratitude and appreciation.

Wenhao Shao, 22th November 2023

Abstract

Security is a primary concern across different domains, mainly because of the escalating crime rates in both public spaces and isolated areas [1]. The significant increase in anomaly detection and real-time monitoring systems addresses this challenge to provide safety measurements, security, and protection of personal property. The deployment of video surveillance systems employing deep learning techniques allows for the effective recognition and interpretation of scenes and anomaly events. However, even with the current advancements in anomaly detection methods, distinguishing abnormal events from normal events in real-world scenarios remains a challenge because they often involve rare, visually diverse, and unpredictably unrecognizable abnormal events. This is particularly true when relying on supervised methods, where the lack of sufficient labeled anomaly data poses a significant challenge for distinguishing between normal and abnormal videos. As a result, state-of-the-art anomaly detection approaches utilize existing datasets to design or learn a model that captures normal patterns, which is then helpful in identifying unknown abnormal patterns. During the model design stage, it is crucial to label videos with attributes such as abnormal appearance, behavior, or target categories that deviate significantly from normal data, marking them as anomalies [2–4]. In addition to the lack of labeled data, two other main considerations and challenges we identified from the literature: 1) insufficient representation of temporal feature, 2) lack of precise positioning of abnormal events and 3) lack the consistency research of temporal feature and appearance feature. The objective of my thesis is to propose and investigate advanced video anomaly detection methods by addressing the aforementioned challenges using novel concepts and utilizing weak supervision and unsupervised models rather than relying on supervised models.

We conducted comprehensive experiments on different methods for detecting abnormal events in videos. We approached the task from three different perspectives, each contributing uniquely to improve existing techniques. In addition, we actively explored the applications of new video processing technologies, including action recognition, target detection, optical flow feature extraction, representation learning, and contrastive learning in order to utilize them in video anomaly detection models. Our proposed models comparatively analysed with baseline models through different evaluation metrics such as accuracy, precision, loss and AUC. This comparative analysis are conducted using prevalent public datasets employed in video anomaly detection, including Ped2 [5], Avenue [6], UCF-Crime [7], and Shanghaitech [8], to validate our assumptions.

Our first contribution addresses the first challenge outlined above by introducing an enhanced Temporal Convolutional Network (TCN). This novel TCN model aims to learn dynamic video features, utilizing the TCN output to optimize input features for multi-instance learning and mitigate errors due to contrastive learned initial weights. The purpose of this ensemble is to calculate the weights of temporal features in weakly supervised schemes. This method enhances the overall capability of weakly supervised models by reducing the loss caused by initial parameters in contrastive learning. Nevertheless, weakly supervised

learning only reduces the reliance on labeled data but does not completely eliminate the dependence on such data. Hence, our subsequent two contributions rely on unsupervised learning models to mitigate the challenge associated with labeled data, addressing the other two challenges. The second contribution combines the self-attention mechanism with an unsupervised video anomaly detection algorithm to prioritize the weights of areas with obvious dynamic fluctuations in video frames. At the same time, during the testing phase, abnormal areas are located through comparison of target detection and loss functions. The combination of self-attention mechanism and target detection technology significantly improves the detection accuracy of the anomaly detection model and expands the functionality of the model. The third contribution explores the integration of collaborative teaching network models, which helps establish consistency between optical flow information and appearance information. This integration aims to enhance the spatio-temporal capture capabilities of unsupervised models. By leveraging the collaborative teaching network model, the overall performance and capabilities of the unsupervised model are significantly enhanced compared to the other baseline models. These important contributions have been published and submitted to reputable journals and conferences in the field.

The video anomaly detection models proposed in this thesis not only promotes the development of academic research in this field, but also provides technical support for public safety system applications.

Keywords

Video Surveillance, Space Security, Video Analysis, Video classification, Anomaly Detection, Deep Learning, Object Detection, Action Recognition, Target Tracking, Spatio-Temporal Features, Optical Flow, Frame Feature, Temporal Convolutional Network, U-Net, Convolutional-3D(C3D), Inflated-3D(I3D)

Résumé

La sécurité est une préoccupation majeure dans différents domaines, principalement en raison de l'augmentation des taux de criminalité dans les espaces publics et les zones isolées [1]. L'augmentation significative des systèmes de détection d'anomalies et de surveillance en temps réel répond à ce défi en fournissant des mesures de sécurité, et la protection des biens personnels. Le déploiement de systèmes de vidéosurveillance employant des techniques d'apprentissage profond permet la reconnaissance et l'interprétation efficaces des scènes et des événements anormaux. Cependant, même avec les progrès actuels dans les méthodes de détection d'anomalies, la distinction entre les événements anormaux et les événements normaux en réalité. Les scénarios multi-mondes restent un défi car ils impliquent souvent des événements anormaux rares, visuellement divers et imprévisibles, méconnaissables. Cela est particulièrement vrai lorsque l'on s'appuie sur des méthodes supervisées, où le manque de données d'anomalies étiquetées suffisantes pose un défi important pour faire la distinction entre les vidéos normales et anormales. En conséquence, les approches de détection d'anomalies de pointe utilisent des ensembles de données existants pour concevoir ou apprendre un modèle qui capture des modèles normaux, ce qui est ensuite utile pour identifier des modèles anormaux inconnus. Au cours de la phase de conception du modèle, il est crucial d'étiqueter des vidéos avec des attributs tels qu'une apparence anormale, un comportement ou des catégories cibles qui s'écartent considérablement des données normales, les marquant comme anomalies [2–4]. En plus du manque de données étiquetées, deux autres considérations et défis principaux que nous identifions dans la littérature : 1) représentation insuffisante des caractéristiques temporelles, 2) manque de positionnement précis des événements anormaux et 3) manque de recherche de cohérence entre les caractéristiques temporelles et les caractéristiques d'apparence. L'objectif de ma thèse est de proposer et d'étudier la détection avancée d'anomalies vidéo. méthodes en relevant les défis susmentionnés en utilisant de nouveaux concepts et en utilisant une supervision faible et des modèles non supervisés plutôt que de s'appuyer sur des modèles supervisés.

Nous avons mené des expériences approfondies sur différentes méthodes de détection d'événements anormaux dans des vidéos. Nous avons abordé la tâche sous trois perspectives différentes, chacune contribuant à améliorer les techniques existantes. De plus, nous avons activement exploré les applications des nouvelles technologies de traitement vidéo, notamment la reconnaissance d'action, la détection de cibles. détection, extraction de caractéristiques de flux optique, apprentissage de représentation et apprentissage contrastif afin de les utiliser dans des modèles de détection d'anomalies vidéo. Nos modèles proposés ont été analysés de manière comparative avec les modèles de base à travers différentes mesures d'évaluation telles que l'exactitude, la précision, la perte et l'AUC. Cette analyse comparative est Utilisation des ensembles de données publiques les plus répandues réalisées dans le cadre de la détection d'anomalies vidéo, notamment Ped2 [5], Avenue [6], UCF-Crime [7] et Shanghaitech [8], pour valider nos hypothèses.

Notre première contribution répond au premier défi décrit ci-dessus en introduisant un

réseau convolutif temporel (TCN) amélioré. Ce nouveau modèle TCN vise à apprendre les fonctionnalités vidéo dynamiques, en utilisant la sortie TCN pour optimiser les fonctionnalités d'entrée pour l'apprentissage multi-instance et atténuer les erreurs dues aux contrastes. poids initiaux appris. Le but de cet ensemble est de calculer les poids des caractéristiques temporelles dans des schémas faiblement supervisés. Cette méthode améliore la capacité globale des modèles faiblement supervisés en réduisant la perte causée par les paramètres initiaux dans l'apprentissage contrastif. Néanmoins, l'apprentissage faiblement supervisé uniquement réduit la dépendance à l'égard des données étiquetées mais n'élimine pas complètement la dépendance à l'égard de ces données. Par conséquent, nos deux contributions suivantes s'appuient sur des modèles d'apprentissage non supervisés pour atténuer le défi associé aux données étiquetées, en relevant les deux autres défis. La deuxième contribution combine l'auto- Mécanisme d'attention avec un algorithme de détection d'anomalies vidéo non supervisé pour donner la priorité aux poids des zones présentant des fluctuations dynamiques évidentes dans les images vidéo. En même temps, pendant la phase de test, les zones anormales sont localisées grâce à la comparaison des fonctions de détection et de perte de cible. -Le mécanisme d'attention et la technologie de détection de cible améliorent considérablement la précision de détection du modèle de détection d'anomalies et étendent les fonctionnalités du modèle. La troisième contribution explore l'intégration de modèles de réseau d'enseignement collaboratif, qui permet d'établir une cohérence entre les informations de flux optique et les informations d'apparence. Cette intégration vise à améliorer les capacités de capture spatio-temporelle des modèles non supervisés. En tirant parti du modèle de réseau d'enseignement collaboratif, la performance globale et les capacités du modèle non supervisé sont considérablement améliorées par rapport aux autres modèles de base. Ces contributions importantes ont été publiées et soumises à des revues et conférences réputées dans le domaine.

Les modèles de détection vidéo d'anomalies proposés dans cette thèse favorisent non seulement le développement de la recherche académique dans ce domaine, mais fournissent également un support technique pour les applications des systèmes de sécurité publique.

Mots-clés

Vidéosurveillance, Sécurité de l'espace, Analyse vidéo, Classification vidéo, Détection d'anomalies, Apprentissage en profondeur, Détection d'objets, Reconnaissance d'action, Suivi de cible, Caractéristiques spatio-temporelles, Flux optique, Fonction de trame, Réseau convolutif temporel, U-Net, Convolutional-3D (C3D), Gonflé-3D (I3D)

Table of contents

1	Introduction	17
1.1	Motivation	18
1.2	Objectives of the Thesis	20
1.3	Contributions of the Thesis	21
1.4	Publications List	22
1.5	Relationship of Publications with Contributions	23
1.6	Outline of the Thesis	24
2	Background and Related Technologies	25
2.1	Overview	26
2.1.1	Definition and types of abnormal	26
2.2	Classification	28
2.2.1	Unsupervised Learning Video Anomaly Detection	29
2.2.2	Weakly Supervised Learning Video Anomaly Detection models	32
2.3	Technical Overview	32
2.3.1	Video Feature Extraction Techniques	36
2.3.2	Representation learning and multi-instance learning frameworks	37
2.3.3	Supplementary research	39
2.4	Datasets	40
2.5	Evaluation Metrics	41
3	Video Anomaly Detection with NTCN-ML: a Novel TCN for Multi-Instance Learning	43
3.1	Overview	44
3.2	Introduction	44
3.3	Methodology	46
3.3.1	Temporal Convolutional Networks	47
3.3.2	Extraction of Temporal Features of Video Sequences	49
3.3.3	The NTCN-ML Based on Temporal Convolutional Network Guidance	50
3.3.3.1	The proposed NTCN-ML model	50
3.3.3.2	The Training Phase	51

3.3.3.3	Loss function	53
3.3.4	The Anomaly Detection Phase	54
3.3.4.1	Steps of detection	54
3.3.4.2	Algorithm complexity analysis in the detection process	55
3.4	Experiments	55
3.4.1	Datasets	55
3.4.2	Experiment Details	56
3.4.3	Experimental results	57
3.4.3.1	Experiment 2: AUC comparison with state-of-the-art models	57
3.4.3.2	Experiment 3: Ablation Study	59
3.4.3.3	Experiment 4: Visual display during anomaly detection.	60
3.5	Conclusion and Future Work	60
4	COVAD: Content-Oriented Video Anomaly Detection using a Self-Attention based Deep Learning Model	63
4.1	Overview	64
4.2	Introduction	64
4.3	Methodology	66
4.3.1	Encoders and Decoders	67
4.3.2	Memory module	70
4.3.3	Coordination Attention	73
4.3.4	Loss Function	75
4.3.4.1	Loss function	77
4.3.4.2	A Visual Evaluation	78
4.4	Experimental Results	78
4.4.1	The proposed COVAD approach	78
4.4.2	Dataset Description	79
4.4.3	Hyperparameter selection process	80
4.4.4	Effectiveness of the attention mechanism in video anomaly detection	80
4.4.5	A Visual Test	82
4.4.6	Video anomaly detection using COVAD	82
4.5	Conclusion and Future work	84
5	Consistency-constrained unsupervised video anomaly detection framework based on Co-teaching	87
5.1	Overview	88
5.2	Introduction	88
5.3	Duel Channel model	90
5.4	Methodology	91
5.4.1	Preliminary	92
5.4.2	Consistency-constrained Framework Based on Co-teaching	93
5.4.3	Co-teaching within Memory Module	95
5.4.3.1	Reading and Updating Mechanisms of the Memory Module	96

5.4.3.2	Strong consistency constraints	99
5.4.3.3	Loss function Memorize Module	100
5.4.4	Anomaly detection stage	100
5.4.5	Experiment 1	101
5.4.6	Experiment 2	103
5.4.6.1	The Performance comparison of single-channel and various dual-channel models	103
5.4.6.2	The impact of skip connections and co-teaching on the per- formance of dual-channel models	105
5.4.7	Experiment 3	106
5.5	Conclusion and Future Work	106
6	Discussion	109
6.1	Implications of Findings	110
6.2	Challenges and Limitations	111
6.3	Future Research Directions	113
6.4	Conclusion	115
	References	116
	List of figures	126
	List of tables	129

Chapter **1**

Introduction

Contents

1.1	Motivation	18
1.2	Objectives of the Thesis	20
1.3	Contributions of the Thesis	21
1.4	Publications List	22
1.5	Relationship of Publications with Contributions	23
1.6	Outline of the Thesis	24

1.1 Motivation

Security concerns have become important across various domains due to the rising crime rates, necessitating sophisticated measures for anomaly detection and real-time monitoring. The continuous development of communication technology has led to the widespread deployment of monitoring equipment in various public areas, including traffic routes, schools, hospitals, shopping malls, supermarkets, and residential buildings. These surveillance devices not only provide covert security assurance to individuals but also produce a significant volume of surveillance videos [9]. Therefore, video surveillance systems, powered by deep learning, have emerged as crucial tools for scene interpretation and event detection targeting anomaly detection. The applications of video surveillance anomaly detection are extensive, covering areas such as intelligent transportation [10], smart home systems, patient monitoring, criminal investigation, campus security, and Internet of Things (IoT) applications [11]. In recent years, there has been active research on intelligent video anomaly detection technology utilizing surveillance video [12]. This has emerged as a research focus in many fields such as image processing, computer vision, deep learning, and various related domains.

However, majority of real-time video anomaly detection systems involve human-in-the-loop, either partially or entirely, for inspection and monitoring purposes. Unfortunately, this approach not only incurs high costs for prevention and control but also leads to human and property losses due to limited attention span, scalability issues and personal subjectivity [13–15]. One solution that can address this issue is by introducing efficient Intelligent video surveillance, which can automatically detect events that violate some operations in public scenes, ensuring personal safety and preventing emergencies in real-time. In general, anomalies refer to activities that deviate from normal patterns and are also known as novelties, outliers, and other similar terms. Video anomalies could include things like cars speeding, exceeding highway limits, flashing ambulances waiting at traffic lights, or passengers passing through ticket gates in an unusual manner. These anomalies may be unusual appearances or irregular movements of targets in specific locations. Therefore, video anomalies are scene-dependent and whether an event is anomalies or not depends entirely on the agreed meaning of exception within a particular context. Hence, the primary challenge lies in constructing a context-dependent deep model, which requires the collection of sufficient data to effectively train the model for a specific context. This challenge significantly impacts anomaly detection models that require labelled data.

The rise of deep learning has accelerated the swift advancement of anomaly detection research. In the literature, there exists many research works on video anomaly detection using supervised learning based models [16, 17]. However, video anomaly detection based on supervised learning requires sufficient labeled data, and most scenarios cannot generate

sufficient labeled data for training. Therefore, supervised learning model cannot be applied in most scenarios. On the other hand, unsupervised learning and weakly supervised learning rely far less labeled data than supervised learning and therefore both methods have begun to attract the attention of more researchers [18] [19]. In general, weakly supervised anomaly detection methods bases comparative learning between normal and abnormal data, utilizing a ranking loss function to define anomalies. It relies only on part of the anomaly dataset with video-level labels. There is no need to set supervision signals on the type of anomaly and the specific location where it occurs. Unsupervised method, without any label data, training a specialized model exclusively with normal data, subsequently calculating the error between the model's output data and the actual data to ascertain anomalies. Essentially, any patterns in the data that deviate from the norms are identified as exceptional occurrences.

Unsupervised video anomaly detection and weakly supervised video anomaly detection models have their own advantages and application scenarios. The unsupervised method provides adaptability to identify new anomalies that the model has not encountered before. In addition, these models can detect a wider range of anomaly types and exhibits strong versatility and generalization. On the other hand, the advantages of the weakly supervised method mainly include robustness and the ability to train models for specific application scenarios. Both unsupervised and weakly supervised models achieved excellent performance in many different public datasets, including Ped2 [5], Avenue [6], UCF-Crime [7], and Shanghaitech [8]. These datasets are used in our analysis to evaluate our newly proposed models comparatively. However, the existing models still have some shortcomings [15] such as i) challenge of mining temporal feature information from the video, ii) The second challenge is inaccurate abnormal positioning, unable to display specific abnormal areas, and iii) explore consistency of time features and appearance features in videos. In addition to these limitation, as we stated above, the other obvious challenge is to collect and label training datasets. My motivation is to address aforementioned challenges and existing limitations by introducing several innovative approaches for video anomaly detection. Hence, we proposed weakly supervised models and unsupervised models to address these issues rather than relying on supervised models. In my thesis, we introduce three innovative approaches for video anomaly detection as three novel contributions as explain below.

Weakly supervised models can be adapted to detect specific anomalies, but they rely on a multi-instance learning framework. Due to the influence of random initial weights, it is prone to optimizing towards the wrong target. Moreover, there is an excessive reliance on existing short-term motion capture methods(such as C3D and I3D) [20] [21] during the feature extraction process, preventing the extraction of comprehensive temporal features from video frames. This implies that the current weak supervision scheme fails to fully exploit

temporal features. And there is an important obstacle to the weakly supervised methods: it only reduces the dependence on data labels, but its learning model is still driven by labeled data. To solve this challenge, it presents a framework which integrates a weakly supervised model with multi-instance learning built with temporal convolutional networks. However, the weakly supervised scheme still needs enough abnormal data to achieve the robustness of anomaly detection. Therefore, in recent years researchers have turned more attention to unsupervised methods [22]. Unsupervised anomaly detection methods mainly centred around representation learning to reconstruct and predict target frames, still depend on appearance features to model normal patterns. However, the major concern is that anomalies are often appearing in small areas of video frames. Compared with weakly supervised models, unsupervised models have poorer ability to locate abnormal areas. While current unsupervised algorithms try to supplement the frame prediction/reconstruction framework using optical flow features, they still do not prioritize dynamic features as the core identification mode of the model. This deficiency leads to inadequate robustness and accuracy of the existing unsupervised model. Therefore, to address this challenge, my thesis introduces novel concept centered around the self-attention mechanism and it exhibits higher performance than our weakly supervised model. Lastly, we propose an unsupervised dual-channel consistency constraint prediction framework employing co-teaching networks and the experiment results indicates that this framework improves anomaly detection performances compared to baseline models. We evaluated the three models proposed in this thesis in public data sets (ped2 [5], Avenue [6], UCF-Crime [7], and Shanghaitech [8]) through accuracy, ablation experiments, and AUC. We comparatively evaluate our model performances identifying several baseline models and conducted ablation to verify the role and effect of each module proposed and each loss function in model. The experimental results prove that the three models proposed in this thesis have achieved state-of-the-art performance.

In conclusion, the development of intelligent video anomaly detection technology is a key step towards optimizing security and safety measures in various domains. It empowers surveillance systems to proactively respond to abnormal events, minimize potential risks and losses, and facilitate more efficient and effective monitoring in diverse scenarios.

1.2 Objectives of the Thesis

In this section, we present the main objectives of this thesis. We address each objective with one contribution. This thesis aims to implementing an efficient and stable video anomaly detection system. The main objectives to achieve this aim are as follows:

- Designs a novel weakly supervised video anomaly detection approach, which uses a

temporal convolutional network to generate pseudo labels for video clips to reduce the initial error of multi-instance learning,

- Improves current unsupervised algorithms based on video frame reconstruction/prediction [23], adds the self-attention mechanism and object detection technique to optimize the weight distribution of the model and locate anomalies
- Designs a novel unsupervised video anomaly detection scheme, synergizing the relationship between optical flow features and appearance features and improving the accuracy of feature extraction.

1.3 Contributions of the Thesis

Our approach to achieve the above research objectives is organized into three parts as three contributions, each corresponding to each research objective. We discuss them as follows:

C.1 The first contribution proposes to use TCN network to calculate the correlation between positive and negative instances, so as to enhance the temporal characteristics of the input. This contribution introduced an effective combination of temporal convolution networks and graph neural networks inspired by the literature [24]. More specifically, the first contribution provides two sub-contributions as follows (Chapter 3).

C.1.1 Firstly, we successfully introduce a novel temporal convolutional network in a weakly supervised learning for video anomaly detection and propose a novel video anomaly detection model NTCN-ML which has optimized the temporal feature extraction.

C.1.2 Secondly, we show that the NTCN-ML model proposed in this thesis can effectively enhance discriminative features between abnormal events and normal events. The experimental results on two widely-used benchmark datasets; 1) UCF-Crime dataset - 95.3% accuracy and 2) ShanghaiTech dataset - 85.1% accuracy, show that the performance of NTCN-ML reached state-of-the-art.

C.2 The second contribution proposed a COVAD [25](Content-Oriented Video Anomaly Detection) method which is based on an auto-encoded convolutional neural network and coordinated attention mechanism in order to effectively capture meaningful objects in the video and dependencies between different objects. Relying on the existing memory-guided video frame prediction network, our algorithm can more effectively

predict the future motion and appearance of objects in the video. Our proposed algorithm obtained better experimental results on multiple data sets and outperformed the baseline models considered in our analysis (Chapter 4).

C.3 The third contribution introduces a novel anomaly detection framework that balances dynamic and static information and builds the relationship between appearance features and corresponding optical flow features, where we set strong consistency constraints, which reduce the loss between dynamic information and corresponding static information. We utilize a collaborative teaching network to ensure consistent representation of static and dynamic information for prediction. The proposed framework consists of two sets of encoder-decoder pairs, supplemented by memory storage modules. Running in parallel with the dual encoder network is the collaborative teaching network, with shared memory modules serving as the cornerstone of collaborative training. Consistency constraints ensure strong consistency between dynamic and static information in the learned representation. During our experimental phase, we present convincing results demonstrating the superior performance of our algorithm on three publicly available datasets (Chapter 5).

C.3.1 Firstly, we propose an advanced approach for video anomaly detection by combining the power of the FlowNet2 for optical flow extraction with a co-teaching network structure. Our model seamlessly fuses optical flow information and the memory module of representation features to predict accurate representation features. Additionally, the integration of representation features and the memory module responsible for optical flow enables the prediction of light stream information.

C.3.2 Secondly, skipping connections are employed to convey background information to the decoder, aiding in the accurate prediction of background and color information.

1.4 Publications List

Accept Papers

- W. Shao, Y. Wei, P. Rajapaksha, D. Li, Z. Luo and N. Crespi, "Low-latency Dimensional Expansion and Anomaly Detection empowered Secure IoT Network," *IEEE Transactions on Network and Service Management*, doi: 10.1109/TNSM.2023.3246798.
- W. Shao, R. Xiao, P. Rajapaksha, M. Wang, N. Crespi, Z. Luo and R. Minerva.

"Video anomaly detection with NTCN-ML: A novel TCN for multi-instance learning.", *Pattern Recognition* (2023): 109765.

- W. Shao, P. Rajapaksha, Y. Wei, D. Li, N. Crespi and Z. Luo, "COVAD: Content-oriented video anomaly detection using a self-attention based deep learning model." 5.1 (2023): 24-41., *Virtual Reality & Intelligent Hardware*, 5.1 (2023): 24-41.
- C. Xiao, W. Shao and R. Xiao, "Toward More Efficient WMSN Data Search Combined FJLT Dimension Expansion With PCA Dimension Reduction," , *IEEE Access*, vol. 8, pp. 104139-104147, 2020, doi: 10.1109/ACCESS.2020.2999484.
- W. Shao, R. Xiao, J. Huang, H. Liu, and X. Du. "FJLT-FLSH: More efficient fly locality-sensitive hashing algorithm via FJLT for WMSN IoT search." , *IEEE Internet of Things Journal*, 6.4 (2019): 7122-7136..

Submitted Papers

- X. Zhao, B. Zhou, W. Shao, and H. Wu. "MSFR: Stance Detection based on Multi- aspect Semantic Feature Representation via Hierarchical Contrastive Learning". *ICASSP 2024. Conference*,
- W. Shao, P. Rajapaksha, N. Crespi, X. Zhao, M. Wang, N. Yin, X. Liu and Z. Luo, "Consistency-constrained unsupervised video anomaly detection framework based on Co-teaching". *IEEE Transactions on Circuits and Systems for Video Technology*

Ongoing Papers

- W. Shao, P. Rajapaksha, Y. Wei, N. Crespi and Z. Luo, "Video Abnormal Content Detection Based on Video Scene Understanding". *Target for TNNLS*
- W. Shao, P. Rajapaksha, N. Crespi, X. Zhao, M. Wang, N. Yin, and Z. Luo, "A Survey of Video Anomaly Detection Techniques". *Target for PROCEEDINGS OF THE IEEE*

1.5 Relationship of Publications with Contributions

In this section, we provide the relationships of publications with contributions.

- The publication '*Video anomaly detection with NTCN-ML: A novel TCN for multi-instance learning*' corresponds to Contribution C.1 in Section 3 [26].

- The publication ‘*COVAD: Content-oriented video anomaly detection using a self-attention based deep learning model*’ corresponds to Contribution C.2 in Section 4 [25].
- The publication ‘*Unsupervised Anomaly Detection in Video Based on Consistency of Appearance and Optical Flow*’ corresponds to Contribution C.3 in Chapter 5.

1.6 Outline of the Thesis

The thesis is structured into six chapters.

- **Chapter 2** presents the background and related technologies relevant to the main topics of this thesis, i.e., deep learning, computer vision, video process, anomaly definition, and the existing unsupervised learning models and weakly supervised models and core technology classification.
- **Chapter 3** proposes the first innovation of this thesis, using temporal convolutional networks to extract temporal features, enhance the confidence of initial anomaly settings for multi-instance learning, and improve the algorithm performance of weakly supervised learning
- **Chapter 4** proposes the second innovation of this thesis, using self-attention to improve the traditional memory module-based unsupervised video anomaly detection
- **Chapter 5** describes the third innovation of this thesis, a weakly supervised video anomaly detection algorithm based on optical flow and representation.
- **Chapter 6** summarizes the thesis and discusses challenges, limitations and possible future directions for the advancements of this thesis.

Background and Related Technologies

Contents

2.1 Overview	26
2.1.1 Definition and types of abnormal	26
2.2 Classification	28
2.2.1 Unsupervised Learning Video Anomaly Detection	29
2.2.2 Weakly Supervised Learning Video Anomaly Detection models	32
2.3 Technical Overview	32
2.3.1 Video Feature Extraction Techniques	36
2.3.2 Representation learning and multi-instance learning frameworks	37
2.3.3 Supplementary research	39
2.4 Datasets	40
2.5 Evaluation Metrics	41

2.1 Overview

In this chapter, we introduce different categories of anomalies and the main definition of anomalies in video data, existing mainstream public data sets and evaluation indicators related to video anomaly detection. Finally, this chapter introduces related deep learning technology in detail, as well as the application and latest development of deep learning model in video anomaly detection.

2.1.1 Definition and types of abnormal

In general, the anomalies categories are point anomalies, contextual anomalies, and collective anomalies [3].

- Point anomalies: Point anomalies occur when only an entity’s data behaves somewhat irregularly compared to the rest of the data. Most anomaly recognition research focuses on this form of anomaly because it is the most fundamental. Cars staying in the middle of the road can be called outliers.
- Contextual anomalies: This occurs when a data value behaves erratically compared to the rest of the data in a specific context. Context includes the observer’s subjectivity and overall perception of the situation. Parking a coach in a bus parking lot can be considered a contextual anomaly.
- Collective anomalies: This occurs when a collection of data samples is considered anomalous compared to real data. A group of people gathering at the exit of a door can be called a collective anomaly.

These anomalies are also specifically demonstrated in video data, such as [27]. However, in the field of video data analysis, the distinction between point anomalies, collective anomalies and contextual anomalies is not clear. Video data has time attributes, and all events do not exist in isolation. It is impossible to train in isolation of events in a video. Compared with clear classification boundaries, video anomalies rely more on contextual information. Even a car staying in the middle of the road or a group of people gathering at the door can be identified based on contextual information. Therefore, combined with the continuous and uninterrupted nature of video data, there are other types of anomalies exist with video data [22]. Some of these video anomalies are explained below.

- Appearance only anomalies: These anomalies can be considered unusual appearances of objects in the video. For example, a cyclist on the sidewalk or a large rock on the road. Detecting these anomalies only requires examining local areas of a single frame of video.
- Short-term motion-only anomalies: These anomalies can be viewed as unusual object motion in the video. For example, a person is running in the library, or a car skids on the road. Detecting these anomalies often requires examining only localized areas of the

video over a short period of time. Appearance-only abnormalities and short-term motion abnormalities only can be further referred to as local abnormalities because they have this additional property.

– Long-term trajectory anomalies: These anomalies can be thought of as unusual object trajectories in the video. For example, people walk in zigzags on sidewalks, cars weave in and out of traffic, or linger around foreign embassy buildings. Detecting trajectory anomalies requires examining longer video clips.

– Group anomalies: Group anomalies can be thought of as unusual object interactions in video. An example would be a group of people walking in formation (such as a marching band, parade and group conflict). Detecting group anomalies requires analyzing the relationship between two or more video regions.

– Time of day anomalies: This type of anomaly is orthogonal to all other types. What is unusual about these activities is the timing of their occurrence. These anomalies are very similar in nature to the position-dependent anomalies discussed previously, with the "relevant contextual reference frame" being time rather than space. An example would be people entering a movie theater at dawn.

There are many different video anomaly detection models proposed in the literature, that are often applied to various intelligent monitoring systems, with early exploration dating back to the 1960s [28]. With the rise of deep learning, specifically deep anomaly detection, this field has shown remarkable progress due to the ability of deep learning to learn expressive representations of complex data, such as high-dimensional, temporal, spatial, and image data [15].

Video anomaly detection tasks are different from traditional machine learning tasks. The core problem is that there is no clear boundary between abnormal events and normal events. And it's hard to obtain sufficient amount of labeled abnormal data. Therefore, when designing anomaly detection algorithms, it is usually impossible to achieve a perfect anomaly detection algorithm by relying solely on data-driven methods.

- **Unknownness:** Anomalies often relate to unknown instances with sudden, unforeseen behavior, data structure, and distribution. They might not be known until they actually occur, like new types of terrorist attacks, fraud, or network intrusion.
- **Rarity:** Anomalies are usually rare instances, while normal instances significantly outnumber them. This scarcity makes it difficult to collect a large number of labeled anomalous instances for training.
- **Diversity:** Anomalous events in videos can be diverse, making it hard to cover all possible cases in the training data.

- **Dependence on Scene Definition:** The definition of abnormal events in videos often depends on specific scenes, making it challenging to generalize across different contexts.
- **Data Privacy:** There are limited public datasets available for video anomaly detection, which hinders the development of robust models.

To tackle these challenges, recent research has focused on deep learning using generative models, such as variational autoencoders, generative adversarial networks, and long short-term memory networks. These approaches have shown promise in addressing the unique complexities of video anomaly detection.

2.2 Classification

Existing video anomaly detection models based on learning can be mainly divided into three categories: supervised learning, unsupervised learning, weakly supervised learning as shown in Figure 2.1

Supervised learning algorithms offer several advantages, including the ability to accurately identify and classify anomalies using labeled data and the ability to identify specific types of anomalies [16, 17, 29]. However, large amounts of labeled data are required, and these techniques may be sensitive to environmental changes, affecting their accuracy. Supervised learning algorithms have low scalability [30]. Thus, unsupervised models and weakly supervised models have attracted more attention from researchers.

The unsupervised algorithm and the weakly supervised algorithm differ in how they define anomalies, leading to distinct data utilization during the training process. In the unsupervised algorithm, the characteristics of the data itself are used, requiring only normal data for training codes. However, when testing, since the model has not been exposed to abnormal data during training, the reconstruction or prediction errors can be substantial. On the other hand, the weakly supervised algorithm, specifically contrastive learning, relies on differentiating between normal and abnormal data. During training, both normal data and anomalous data with video-level labels are necessary. However, during testing, only individual data packets need to be input separately. The algorithm judges whether the maximum abnormal score of the instance in the packet exceeds the threshold, enabling the definition of the abnormal segment.

By understanding these distinctions in the training and testing processes, we can effectively utilize unsupervised and weakly supervised algorithms for anomaly detection. These methods cater to various scenarios where data labeling is limited or uncertain, making

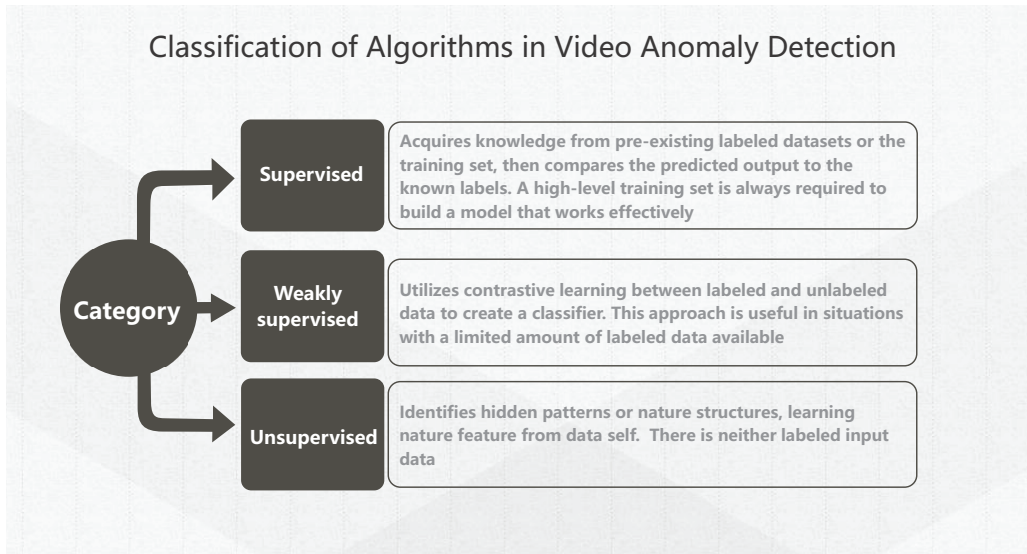


Figure 2.1: Classification of Algorithms in Video Anomaly Detection [22]

them valuable tools in anomaly detection tasks for video data, medical imaging, and other real-world applications.

2.2.1 Unsupervised Learning Video Anomaly Detection

Existing unsupervised algorithm framework can be divided into the following four categories: the first is the basic video frame reconstruction task model [23]; the second is the video frame prediction task model [31]; the third is a dual-channel model [32]; the fourth is a multi-task model [33].

Reconstruction-based methods operate on the assumption that normal data can be integrated into the low-dimensional domain, and that normal samples and anomalies are represented in different patterns in the low-dimensional space [34]. Autoencoder (AE) is mainly used, a feedforward neural network that includes an encoder and a decoder structure [35]. The goal is to capture important parts of the input data and learn low-dimensional representations of high-dimensional data; where the encoder network maps the input data to a low-dimensional latent space, and the decoder network maps the latent space back to the original data space. In the subsequent research process, variational auto-encoding, convolutional auto-encoding, graph convolution coding and adversarial auto-encoding were used to improve the quality of data compression and enhance the restoration ability of the model [23, 36–44]. In reconstruction-based methods, in 2020, Hyunjong Park et al. [23] optimized on the basis of anomaly detection tasks by combining with the U-Net network to further limit the expressive ability of the neural network, and proposed a

video anomaly detection algorithm based on future frame prediction and reconstruction. The turning point is that the method first proposed to limit the representation ability of the neural network. The main concern is that the powerful representation ability of the neural network will distort the description of the data in the feature space. As a result, some researchers began to work on interpretable detection models and video semantic analysis.

Prediction-based methods identify anomalies by evaluating the difference between expected and actual spatio-temporal properties of features [34]. These models assume that all normal activity is predictable after training, and any deviation from the prediction indicates an anomaly. During the training process, they used recurrent neural networks (RNN), long short-term memory networks (LSTM), convolutional neural networks (CNN), and convolutional long short-term memory networks (Conv LSTM) to take the previous consecutive frames as input and predict the results of the next frame as output. These techniques are commonly used to process sequence data. In the subsequent research process, local features, global features, and dynamic features are used as supplements to improve the accuracy of model prediction. Common prediction-based algorithms include [31, 34, 45–48], Vision Transformer (ViT) [49–51] and U-Net and C3D [23] can also be used as a prediction network. In prediction and reconstruction tasks, researchers often use dynamic information, such as optical flow information, to supplement input features and improve the prediction accuracy of the model. Therefore, a dual-channel model was proposed in the subsequent research process.

Dual-channel-based models setting on the assumption that the static feature and dynamic feature of video are related. They use multi-branch encoding and decoding structure to achieve multiple inputs to multiple outputs. In this process, auxiliary technologies are usually used, such as memory modules [38], graph neural networks [52], or joint loss functions [53] to build the connection between dynamic features and static features. These models regard the consistency of dynamic features and static features (appearance features-optical flow features, temporal features-spatial space) as an important discussion. Initially, optical flow features were used as a supplementary part of appearance features to enhance model accuracy [54], but now, researchers are more inclined to put optical flow features and appearance features in an equally important position [32, 44, 55, 56].

Multi tasking-based models assume that there is an essence to the data. Even if the video frames undergo various modifications, including rearrangement, scaling, rotation, or occlusion. During the learning process, the model can always discover internal patterns and recover video sequences. And multi-task models use multiple agent tasks to improve self-supervised learning capabilities. There are many kinds of proxy tasks for video frame sequences, such as reading video sequences out of order to predict intermediate frames, and reading sequences with occlusion in some video frames to restore the occlusion content.

Read the rotated video frames to identify or predict the rotation angle, read the scaled video frames to restore the normal size, and determine whether the video frame sequence is in forward or reverse order. The technologies used in this process include target detection, optical flow estimation, 3D convolution, U-net network, etc. General steps: The first step is to use object detection or feature extraction algorithms to extract the features of video frames. The second step is that modify these features and set up multiple agent tasks. The third step is that input the modified features individually or jointly to U-Net network or convolutional network, restore the original normal video frame sequence features. The current multi-task learning models mainly include [33, 57–59]

In this thesis, we list four representative methods of classification in Table 2.1, including the data sets used and core technologies, and we briefly described related algorithms in Table 2.2.

2.2.2 Weakly Supervised Learning Video Anomaly Detection models

Weakly supervised learning models, also known as semi-supervised models, use labeled and unlabeled data to create classifiers. This approach is particularly useful when the amount of available labeled data is limited. This approach mainly stems from the introduction of the UCF-Crime dataset [7]. This dataset has video-level labels in both training and test sets.

The core of this model is to enlarge the error between the segment with the highest abnormal confidence in the abnormal video and the segment with the highest abnormal confidence in the normal video. The weakly supervised model is based on the assumption that the abnormal confidence of all normal video clips is smaller than the maximum abnormal confidence clip in the abnormal video. During training, the predictor segments unlabeled samples and assigns confidence to each segmented sample segment.

Therefore, in weakly supervised models, the commonly framework Composed of contrastive learning model and multi-instance learning. Multi-instance learning provides convenience for assigning prediction confidence to each video instance and setting the contrastive loss function.

But usually when video features are input into the multi-instance learning network, some auxiliary work will be done, such as action feature network C3D, I3D, long short-term memory network (LSTM), VGG, Transformer, recurrent neural network (RNN) or automatic encoding and other technologies (AE) enhance the coherence of input features, or provide pseudo-labels to reduce errors in the initial random allocation process of multi-instance learning networks. In learning-based algorithms, Waqas Sultani et al. [7] in 2018, which first proposed to use C3D network to extract the video features after clips and input the features into a MIL to calculate anomaly scores for each instance. Now there are weakly supervised models [7, 19, 61–71] showing Table 2.3.

In addition, video anomaly detection technologies based on video understanding or natural language processing (NLP) triples have gradually emerged. These technologies are separated from the reconstruction or prediction of video representation information, but try to understand what happened in the video and based on these The semantics of content hiding are used to determine anomalies based on the context. At present, this classification is still immature, and the core obstacle is the lack of sufficient data sets with semantic labels to assist testing [76–78].

2.3 Technical Overview

In this study, several techniques are employed to aid video processing and analysis. These techniques are primarily categorized into three groups: video feature extraction techniques,

Table 2.1: Unsupervised Video anomaly detection classification

Type	Method	Techniques	Datasets
Reconstruction	LTR [36]	Autoencoder,temporal cuboid	Avenue, UCSD, subway
	ST-AE [37]	CNN,autoencoder	Traffic, UCSD, Avenue
	AMDN [38]	One-class SVM, Optical flow	Train, UCSD
	GMFC-VAE [39]	Convolutional autoencoder, Gaussian mixture model	Avenue, UCSD
	ConvAE-LSTM [40]	convolutional autoencoder, LSTM, optical flow	Avenue, UCSD
	Temporal cues [41]	GAN, LSTM, Optical flow	Avenue, ShanghaiTech
	MNAD [23]	Convolutional autoencoder, Memory module	UCSD, Avenue, Shanghaitec
	Ada-Net [42]	GAN, autoencoder	UCSD, Avenue, ShanghaiTech
	Adver-3D CAE [43]	Convolutional autoencoder	Subway, Avenue, UCSD, ShanghaiTech
	AE-U-Net [44]	Convolutional autoencoder	UCSD, Avenue, Subway,Traffic
Prediction	HF2-VAD [31]	Variational Autoencoder	UCSD, Avenue, Shanghaitec
	FFP [34]	Adversarial training, U-Net	UCSD, Avenue, Shanghaitec
	Residual LSTM [48]	Residual attention LSTM	Avenue, UMN, UCF-Crime,
	CT-D2GAN [47]	GANs, Transformer, CNN	Ped2, Avenue, ShanghaiTech
Dual channel	ITAE [60]	Convolutional autoencoder, Two path generative	Ped2, Avenue, ShanghaiTech
	AMSRC-Net [54]	Gated Fusion Module, Two path generative	Ped2, Avenue, ShanghaiTech
	Two-P [32]	Two convolution autoencoder	Ped2, Avenue, ShanghaiTech
Multi-tasking	Multi-T [57]	3D convolutional, YOLOv3,multi-task	Ped2, Avenue, ShanghaiTech
	Bi-d predict [59]	Bi-directional,Autoencoder	Ped2, Avenue, ShanghaiTech

Table 2.2: Unsupervised learning-based video anomaly detection models

Method	Description
LTR [36]	Use feedforward neural networks and encoder networks to construct local action features based on hand-crafted features to achieve end-to-end learning under restricted supervision
ST-AE [37]	The spatiotemporal AE comprises one encoder and two decoders, employs parallel training of decoders with monochrome frames, which is noteworthy compared to the distillation process
AMDN [38]	The appearance and motion DeepNet model employs AEs and a modified two-stream network with an additional third stream to improve detection performance.
GMFC-VAE [39]	The Gaussian mixture fully convolutional-variational AE uses the conventional two-stream network technique and uses a variational AE to enhance its feature extraction capability.
ConvAE-LSTM [40]	This method uses the convolutional AE and long short-term memory to detect anomalies. The framework produces the error function and reconstructed dense optical flow maps.
Temporal cues [41]	A conditional GAN is trained to learn two renderers that map pixel data to motion and vice versa. Normal frames will have little reconstruction loss, while anomalous frames is significant loss.
MNAD [23]	This algorithm proposes to store the behavioral patterns of normal videos through the memory module and limits the expressive ability of the convolutional neural network.
Ada-Net [42]	An attention-based autoencoder using contentious learning is proposed to detect video anomalies.
Adver-3D CAE [43]	A 3D CAE-based competitor anomalous event detection method is proposed to obtain the maximum accuracy by simultaneously learning motion and appearance features. It was developed to explore spatiotemporal features that help detect anomalous events in video frames.
AE-U-Net [44]	A two-stream model is created that learns the connection between common item appearances and related motions. A single encoder is paired with a U-net decoder to predict motion and a deconvolution decoder that reconstructs the input frame under the control of the reconstruction error
FFP [34]	Spatial and motion constraints are used to estimate the future frame for normal events in addition to density and gradient losses
HF2-VAD [31]	Extract spatial CNN features from a series of video frames and feed them to the proposed residual attention-based LSTM network, which can precisely recognize anomalous activity
Residual LSTM [48]	Using a light-weight CNN and an attention-based LSTM for anomaly detection reduces the time complexity with competitive accuracy.
CT-D2GAN [47]	A Conv-transformer is used to perform future frame prediction. Dual-discriminator adversarial training maintains local consistency and global coherence for future frame prediction.
ITAE [60]	Proposed a structure with two encoders and a single decoder, in which the two encoders capture static and dynamic features, the decoder learns to combine and reconstruct them together as original inputs.
AMSC-Net [54]	Use optical flow features as a complement to appearance features to enhance the accuracy of prediction or reconstruction
Two-P [32]	Design two proxy tasks to train the two-stream structure to extract appearance and motion features in isolation, the prototypical features are recorded in the corresponding spatial and temporal memory pools
Multi-T [57]	Design multiple proxy tasks: three self-supervised and one based on knowledge distillation. The self-supervised tasks are: (i) arrow of time, (ii) motion irregularity and (iii) reconstruction. The knowledge distillation task takes into account both classification and detection information,
Bi-d predict [59]	Propose a novel bi-directional architecture with three consistency constraints to comprehensively regularize the prediction task from pixel-wise, cross-modal, and temporal-sequence levels

Table 2.3: Weakly supervised algorithm classification

Methods	Techniques	Datasets	Codes
F-MIL [7]	MIL, C3D, TCNN	CCTV	https://github.com/WaqasSultani/AnomalyDetectionCVPR2018
GCN [64]	GCN,C3D,TSN	UCF-Crime, ShanghaiTech, UCSD ped2	https://github.com/jx-zhong-for-academic-purpose/GCN-Anomaly-Detection
MLEP [72]	ConvLSTM, Encoder-decoder	Avenue, Shanghaitech	https://github.com/svip-lab/MLEP
Motion-Aware [67]	Temporal augmented, VGG,C3D,I3D	UCF Crime	
Siamese [66]	Siamese network, CNN	UCSD, Avenue	
AR-Net [65]	Regression net, Dynamic loss	Shanghaitech	https://github.com/wanboyang/Anomaly_AR_Net_ICME_2020
XD-Violence [73]	Multimodal information, C3D,I3D	XD-Violence	https://roc-ng.github.io/XD-Violence/
CLAWS [68]	Clustering assisted, random selector,C3D	UCF-Crime, Shanghaitech	
MIST [62]	Self-guided attention, Sparse sampling	UCF-Crime, Shanghaitech	https://kiwi-fung.win/2021/04/28/MIST/
RTFM [61]	Top-k MIL, C3D,I3D	UCF-Crime, Shanghaitech, XD-Violence	https://github.com/tianyu0207/RTFM
STAD [69]	Spatio-temporal tube, Relationship reason	UCF-Crime, Shanghaitech	
WSAL [63]	High-order Context AE	TAD, UCF-Crime	https://github.com/ktr-hubrt/WSAL
CRFD [70]	Causal temporal Relation	UCSD, UCF-Crime, Shanghaitech	
MSL [71]	MSL, transformer	Shanghaitech, UCF-Crime, XD-Violence	
UR-DMU [74]	I3D, self attention	UCF-Crime, XD-Violence	
CMRL [75]	Context-Motion Interrelation	Avenue, XD-Violence, Shanghaitech, UCF-Crime	

representation learning, and supplementary techniques. This section will provide an introduction to each of these techniques.

2.3.1 Video Feature Extraction Techniques

This category encompasses a diverse set of methods aimed at extracting meaningful and informative features from raw video data. Including motion capture technology, optical flow extraction technology, and sequence feature extraction technology

- Motion capture technology: I3D (Inflated 3D) [21] and C3D (Convolutional 3D) [20] are commonly used motion capture technologies in the field of video anomaly detection. During the application process, the feature map of the last layer is usually intercepted as the feature of the current video frame. I3D takes a 2D image classification network and inflates all filters and pooling kernels-giving them an extra temporal dimension, converting it into a 3D convolutional network. Since the I3D network takes into account the temporal characteristics of image data, it is often applied to video sequence feature extraction tasks. The model undergoes pre-training on an extensive video dataset, equipping it with the capability to grasp intricate spatio-temporal patterns and representations from videos. This capacity enables I3D to efficiently capture both appearance and motion cues present in consecutive frames, making it especially well-suited for various video analysis tasks, including action recognition, video classification, and temporal localization. C3D is achieved by convolving a 3D kernel into a cube formed by stacking multiple consecutive frames together. With this structure, feature maps in convolutional layers are connected to multiple consecutive frames in the network. The "3D" in C3D refers to the convolutional layers being extended to operate in three dimensions (width, height, and time), which allows the model to consider the temporal dynamics of video data. The model was originally applied to action recognition. We utilize the powerful C3D feature extraction technique to enhance video analysis and comprehension.
- Optical flow extraction technology [79]: which evaluates the motion characteristics of the object by identifying the differences between consecutive video frames, that provides essential information about the dynamic aspects and movements within the video data. It is often used in the field of computer vision, such as object tracking, object recognition, and dynamic feature capture. Accurate estimation of optical flow is crucial for various computer vision tasks. Commonly used methods include IRR [80], GMA [81], Flownet [82]. Flownet2 utilizes a deep CNN architecture, enabling it to capture complex spatial and temporal features in video data. The network takes pairs of consecutive frames as input and processes them through multiple layers of

convolutions and non-linear activations. These layers empower the model to learn hierarchical representations efficiently encoding the patterns and motion information present in the video frames. Compared with [82], there are two important improvements in the Flownet2 network [83]. The first is the order of training data, from simple to complex, which improves the matching ability of the model; The second one is to fuse the output optical flow of the previous frame with the current frame and output the optical flow of the next frame. Therefore, Flownet2 has stronger performance. We used the Flownet2 in my thesis.

- Sequence feature extraction technology: used in natural language processing (NLP) contextual semantic analysis tasks. Sequence feature extraction technology refers to methods and technologies for extracting relevant features or information from data sequences. Sequences can come in many forms, including text, time series data, DNA sequences, audio signals, and more. Video is a type of time series data. The goal of sequence feature extraction is to convert raw, usually high-dimensional, unstructured sequences into a structured format. For video data, sequence feature extraction technology helps the model learn dynamic features to understand video content. This technology can be used for a variety of machine learning and data analysis tasks. Common techniques used for video sequence feature extraction include Transformer [49], recurrent neural network (RNN) [84], long short-term memory networks (LSTM) [85], convolutional neural network(CNN) [47], Temporal Convolution Network(TCN) [86], pyramid network [87]. autoencoders [72]. In this thesis, we used the TCN network. Temporal series learning networks usually need to follow two principles [88,89]: (1) The input and output structures of the network are the same; (2) The features of the current time node are not disturbed by the features of the next time node. Therefore The proposed TCN network consists of Dilated Causal Convolutions (DCC) [90] and residual networks [91]. The Dilated Causal Convolutions is used to pass the information of the previous nodes, and the residual network is used to supplement the information. In this thesis, TCN is used to extract temporal features to supplement the features extracted by C3D, and to optimize the multi-instance learning network

2.3.2 Representation learning and multi-instance learning frameworks

Unsupervised video anomaly detection and weakly supervised video anomaly detection differ in anomaly definition and data dependence. Therefore, there are two learning frameworks to handle these two tasks: representation learning and multi-instance learning. Self-supervised learning is a method of unsupervised learning. Its main purpose is to learn useful information by supervising itself from non-manually labeled data. The means to achieve

self-supervised learning is mainly to use auxiliary tasks to mine its own supervisory information from large-scale unsupervised data. The network is trained with constructed supervision information so that valuable representations for downstream tasks can be learned. We call these auxiliary tasks representation learning [52]. In application scenarios with limited data, self-supervised learning and representation learning have stronger application value. Encoding-Decoding is the most commonly used unsupervised video anomaly detection framework.

For weakly supervised video anomaly detection solutions, since video-level labels exist in some data, A combines the structures of contrastive learning and traditional multi-instance learning to design a multi-instance learning framework for video anomaly detection.

- **Encoding-Decoding:** The U-Net network is the most commonly used encoding-decoding structure, originally developed for biomedical image segmentation tasks. It was proposed by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015 [92]. The name "U-Net" is derived from the architecture's U-shaped design, resembling an inverted "U." In the U-Net architecture, the encoder conducts a series of convolution and pooling operations, effectively reducing the spatial dimension while increasing the number of channels. On the other hand, the decoder employs upsampling and transposed convolutions to restore the original spatial resolution. One distinctive feature of the U-Net is the incorporation of skip connections, enabling the network to utilize feature maps from the encoder during the restoration phase. This allows the decoder to combine both high-level contextual information and low-level spatial details, contributing to improved segmentation accuracy. Due to its unique design and skip connections, U-Net has found widespread application in representation learning [93,94].
- **MIL (Multiple Instance Learning) [95]** is a machine learning paradigm that addresses scenarios where the training data lacks explicit instance-level labeling and is instead organized into groups or "bags" of instances. Each bag can contain multiple instances, and the bag's label is determined by the presence or absence of certain patterns or properties within the instances. MIL finds particular significance in situations where video anomaly localization is unclear and labels are insufficient. In such cases, where only video-level labels are available without clear indications of the abnormality's specific location, multi-instance learning is employed. In 2018, Sultani et al [7] combined multi-instance learning and contrastive learning to build a weakly supervised video anomaly detection framework based on multi-instance models, where both normal and abnormal data are divided into the same number of instances and input into the multi-instance learning network simultaneously. The network calculates the

abnormal score for each instance within the two data packets. Notably, the maximum anomaly score in the abnormal data packet is always greater than any instance in the normal data packet. Based on this observation, a video anomaly definition is constructed using comparative learning [96]. MIL can effectively tackle challenges in anomaly detection where precise instance-level labeling is not feasible, and the focus lies on identifying abnormal patterns at the video level. This approach enables us to handle various real-world applications where anomaly localization may be ambiguous, such as video surveillance, medical imaging, and industrial monitoring, enhancing the accuracy and robustness of the anomaly detection process.

2.3.3 Supplementary research

In video anomaly detection, existing model [33] [62] [49] utilizes several supplementary technologies, including object detection [97], object tracking [98], and the self-attention mechanism [99]. Object detection enables the model to locate the area in the video frame where the anomaly occurs, providing a clearer representation of the abnormality. On the other hand, object tracking tracks video features and supplements the lack of temporal information. Lastly, the self-attention mechanism assists the model in better identifying the dynamic characteristics of targets when anomalies occur. Together, these additional techniques enhance the accuracy and effectiveness of video anomaly detection, enabling the model to detect and highlight anomalies more efficiently.

- **Object Detection:** Object detection is a fundamental computer vision task aimed at identifying and localizing multiple objects of interest in an image or video. The goal of object detection is to provide a class label and bounding box coordinates for each detected object present in the input data. Object detection can be roughly divided into two categories: Single Shot Detectors and Region-based Detectors. Single Shot Detectors, such as SSD (Single Shot Detectors) [100] and the YOLO series (you only look once) [101], predict object classes and bounding box coordinates directly from predefined anchor boxes placed at different positions and ratios within the image. These methods offer fast and efficient advantages. On the other hand, Region-based Detectors employ technologies like selective search or region proposal network (RPN) to generate region proposals (candidate object bounding boxes). The regions are then classified and refined to obtain the final object detection result. This approach provides high accuracy and scalability. Representative technologies in this category include Faster R-CNN [97], R-CNN [102], and Mask R-CNN [103]. Both types of object detection methods have their strengths and are widely used in various computer vision applications, such as autonomous driving, surveillance, robotics, and more.

- Object tracking [104]: Object tracking is a fundamental computer vision task that involves locating and tracking a specific target or object of interest over a series of consecutive frames in a video. The goal of object tracking is to maintain a consistent association between the target object in an initial frame and its corresponding instance in subsequent frames, even when the object undergoes changes in appearance, motion, or occlusion. Key components and techniques used in object tracking include [98]: 1.Object representation: Effectively representing each frame involves considering features such as color, texture, speed, and direction of the target object; 2.Object detection and initialization: Object tracking typically starts by detecting the target object in the first frame of the video or manually annotating its position; 3. Motion prediction: Object tracking algorithms often employ motion models to predict the position and movement of the target object in subsequent frames; 4.Data association: Data association methods are used to link the target object in the current frame with its corresponding instance in the previous frame, ensuring a smooth tracking transition; 5.Occlusion handling: Object tracking algorithms must handle situations where the target object may be partially or completely occluded by other objects or obstacles in the scene.

Common object tracking technologies include SiamRPN [105], DaSiamRPN [106] and DCF-Net [107] represented by deep learning approaches, which composed of Siamese sub-network for feature extraction and region proposal sub-network including classification branch and regression branch. These techniques have shown significant advancements in achieving robust and accurate object tracking in real-world scenarios. As technology continues to evolve, ongoing research and development efforts are likely to further enhance object tracking algorithms, leading to even more advanced and efficient tracking solutions in the future.

2.4 Datasets

There are four commonly used public data sets for video anomaly detection. According to different usage methods, there are different division methods in the specific implementation process. Since weak supervision requires partially labeled data, different segmentation schemes exist in the Shanghaitech dataset.

- UCSD: The UCSD dataset [5] consists of pedestrian data captured at two different pedestrian areas on the UCSD campus, with subsets named Ped1 and Ped2. The training set contains 34 clips from Ped1 and 16 clips from Ped2, all containing only normal frames. In contrast, the test set comprises 36 clips from Ped1 and 12 clips from Ped2, containing both normal and abnormal frames. Frame-level annotations are

available for all test clips, with pixel-level ground truth annotations provided for ten clips. In the UCSD dataset, pedestrian walking is considered a normal pattern, while non-pedestrian entities like bicycles and skaters are defined as anomalous instances.

- **UCF Crime:** The UCF Crime dataset [7] includes 1900 unedited videos capturing various real-world anomalous events, such as abuse, arrest, arson, assault, traffic accident, theft, explosion, fight, robbery, shooting, stealing, burglary, and vandalism. Out of these videos, 950 are normal, while the rest contain at least one anomalous event. The training set contains 800 normal videos and 810 abnormal videos. The remaining 150 normal and 140 abnormal videos are test set. Both the training and test sets encompass all 13 anomalous events. Some videos may contain multiple anomaly categories, such as robbery vs. fight or theft vs. vandalism. The videos in UCF Crime represent realistic surveillance applications and cover different lighting conditions, image resolutions, and camera poses, making it a challenging dataset for anomaly detection tasks.
- **Avenue:** The Avenue dataset [6] contains 15 videos, each lasting 2 minutes, resulting in a total of 35,240 frames. For training purposes, 8,478 frames from four videos are used. Typical unusual events captured in this dataset include running and throwing objects.
- **ShanghaiTech:** The dataset [8] consists of 13 scenes collected at ShanghaiTech University, featuring complex lighting conditions and camera viewpoints. It comprises 437 videos, with an average of 726 frames per video. The training set includes 330 normal videos, while the test set contains 107 abnormal videos and 130 abnormal videos. Unusual events in this dataset encompass uncommon patterns on campus, such as motorcycles or cars.

These datasets play a crucial role in advancing the research and development of anomaly detection algorithms, providing valuable benchmarks and metrics for evaluating the effectiveness and robustness of various anomaly detection methods.

2.5 Evaluation Metrics

The area under the ROC Curve (AUC) [108, 109] provides an aggregate measure of performance across all possible classification thresholds. The Area Under the PR curve (AUC-PR) is useful when true negatives are more common than true positives. The PR curve only focuses on the predictions of the positive (rare) class and therefore, this is a good metric for anomaly detection. The difficulty lies in predicting those rare truly positive events.

Accuracy is directly affected by the category or class imbalance effect and as a result, FP outcome of the model is also affected. Hence, the ROC curve does not capture this effect. For highly imbalanced datasets, the PR curves are more capable of highlighting differences between model outcomes. Therefore, for a highly unbalanced class setting, the AUC-PR score can be considered as the best metric to compare different models.

Video anomaly detection is a classification problem, and the ROC curve (Receiver Operating Characteristic curve) is one of the commonly used classification evaluation metrics [110]. The ROC curve plots two parameters: true positive rate (TPR) and false-positive rate (FPR). True positive rate is also called as Recall, which indicates the probability of an actual abnormal event will predict as a abnormal event. False positive rate indicates the probability of a true normal event will predicts as an abnormal event [111]. TPR and FPR can be expressed as mentioned in Equation 2.1 and Equation 2.2, which is the calculation methods of the ROC curve and the components of the ROC curve.

$$TPR(Recall) = \frac{TP}{TP + FN} \quad (2.1)$$

$$FPR = \frac{FP}{FP + TN} \quad (2.2)$$

where TP is the outcome when the model correctly predicts true abnormal event; FN is the outcome when the model incorrectly predicts true normal event and detected as an abnormal event; FP is an outcome when the model incorrectly predicts true normal and detected as an abnormal event, and TN is the model outcome when the model predicts true normal.

The area under the ROC Curve (AUC) [112] [113] provides an aggregate measure of performance across all possible classification thresholds. The Area Under the PR curve (AUC-PR) is useful when true negatives are more common than true positives. The PR curve only focuses on the predictions of the positive (rare) class and therefore, this is a good metric for anomaly detection. The difficulty lies in predicting those rare truly positive events.

Accuracy is directly affected by the category or class imbalance effect and as a result, FP outcome of the model is also affected. Hence, the ROC curve does not capture this effect. For highly imbalanced datasets, the PR curves are more capable of highlighting differences between model outcomes. Therefore, for a highly unbalanced class setting, the AUC-PR score can be considered as the best metric to compare different models.

Video Anomaly Detection with NTCN-ML: a Novel TCN for Multi-Instance Learning

Contents

3.1	Overview	44
3.2	Introduction	44
3.3	Methodology	46
3.3.1	Temporal Convolutional Networks	47
3.3.2	Extraction of Temporal Features of Video Sequences	49
3.3.3	The NTCN-ML Based on Temporal Convolutional Network Guidance	50
3.3.4	The Anomaly Detection Phase	54
3.4	Experiments	55
3.4.1	Datasets	55
3.4.2	Experiment Details	56
3.4.3	Experimental results	57
3.5	Conclusion and Future Work	60

3.1 Overview

This chapter proposes a new weakly supervised video anomaly detection model supplemented by temporal convolutional networks to address the insufficient temporal features in the weak supervision process and the error in initial weights in contrastive learning. This model takes into account the multi-instance learning process. The correlation between abnormal events and normal events in the system is trained on global temporal features and generated instance pseudo-labels to slow down the comparison error of multiple instances and improve the performance of the multi-instance learning framework. The experimental results suggest that the proposed temporal convolutional network shows a strong learning ability in enhancing temporal feature representation and reducing bias in training, with the detection performance of the proposed weakly supervised learning model outperforming the current mainstream models on two benchmark datasets (UCF-Crime and ShanghaiTech).

3.2 Introduction

Video anomaly detection is a significant problem yet an active research area in which models observe patterns that deviate from normal behavior, which serves a crucial role in industrial production and transportation. There are still some challenges and problem complexities that require advanced approaches to model the patterns in complex video data to identify outliers. One main challenge is the recognition of positive instances or rare abnormal patterns as they manifest only small variations compared with normal events. In addition, rare positive instances are largely biased by the dominant negative instances.

In the literature, supervised learning strategies are mostly used for learning abnormal patterns and normal events, which require manually-annotated labels as learning signals [114]. However, it is challenging to acquire annotated data for all types of anomalous events, and therefore, supervised learning suffers from several disadvantages [22] such as, i) The boundary between normal and abnormal patterns is blurred in many video scenes, the same event can produce different consequences in different scenes resulting. ii) Anomalous video events are featured with global temporal properties, but deep learning usually ignores the global nature of temporal features and only extracts sequence temporal features through short-term motion capture. iii) Anomalous patterns cover a wide range of situations and it is unrealistic to define all patterns of anomalous events in a single scenario.

To this end, researchers have turned to explore unsupervised learning and weakly supervised learning models for video anomaly detection. Unsupervised methods solely rely on normal events for model training and anomalous events are identified by learning representation features and intrinsic patterns of normal events [34]. Compared to unsupervised algorithms, weakly supervised learning algorithms rely on training samples with both nor-

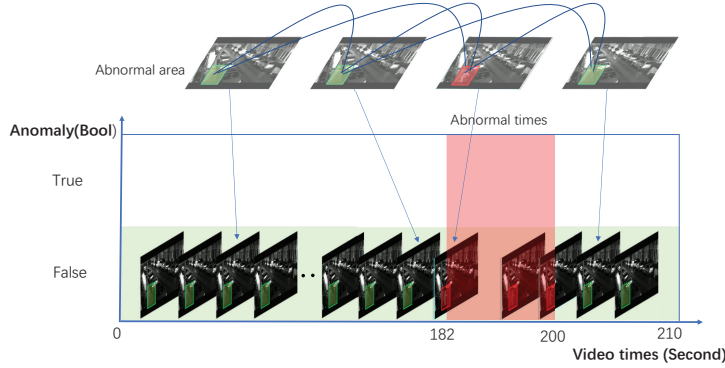


Figure 3.1: A representation of the spatiotemporal dimension of anomalous events.

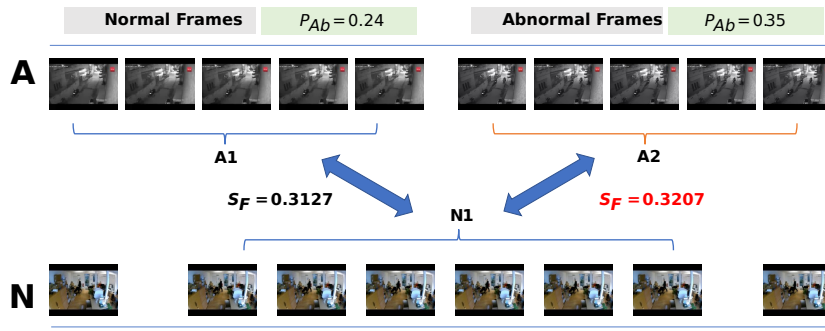


Figure 3.2: Feature similarity analysis of positive and negative instances: A means abnormal data, N means normal data. $A1$ represents normal instances in abnormal data, $A2$ represents abnormal instances, and $N1$ represents instances in normal data, S_F represents similarity of features between two instances and P_{Ab} is the probability of an anomalous instance.

mal and anomalous events. The core of weakly supervised algorithms is the Multi-Instance Learning (MIL) [115]. One assumption in MIL is that the optimization in each training process always targets the negative instance in the abnormal data. However, this assumption is unrealistic as it does not always learn the right patterns, because there is no guarantee that the ranking loss from different scenarios (pairs of normal data and abnormal data) always occurs on the negative instances of abnormal data.

As shown in Figure 3.2, the error between normal instances and normal instances in abnormal videos is larger than that with abnormal instances, which will cause the model

to learn in the wrong direction after training.

To mitigate the above issues, this chapter proposes to use TCN network to calculate the correlation between positive and negative instances, so as to enhance the temporal characteristics of the model. Inspired by the literature [86] [19], which introduced an effective combination of temporal convolution networks and graph neural networks. In this chapter, we consider the temporal and spatial features as equally important factors in video anomaly detection and propose a new weakly supervised video anomaly detection model, NTCN-ML (a **N**ew **T**emporal **C**onvolution **N**etwork for **M**ulti-**I**nstance **L**earning). The NTCN-ML model examines the correlation between positive and negative samples in the MIL process to enhance temporal patterns. Positive and negative correlation helps to balance the feature association between positive and negative instances, and then construct a novel temporal feature to optimize the MIL process.

Two main contributions of this part are: i) We successfully introduce a novel temporal convolutional network in a weakly supervised learning for video anomaly detection and propose a novel video anomaly detection model NTCN-ML which has optimized the temporal feature extraction and ii) We show that the NTCN-ML model proposed in this chapter can effectively learn the potential patterns between anomalous events and normal events. The experimental results on two widely-used benchmark datasets; 1) UCF-Crime dataset - 95.3% accuracy and 2) ShanghaiTech dataset - 85.1% accuracy, show that the performance of NTCN-ML reached state-of-the-art.

3.3 Methodology

This section explains a weakly supervised learning video anomaly detection model called NTCN-ML (a New Temporal Convolution Network for Multi-Instance Learning). In general, in the training process of paired data, when the model learns the features of sequential data, the positive (normal video frames) and negative instance (abnormal video frames) of the same video usually contain a large amount of similar content. There are a lot of similarities between positive examples and negative examples in the same video. In negative instances, the spatiotemporal region where the anomalous event occurs accounts for only a small portion of the entire video, as shown in video 23 in the vandalism subcategory in the UCF-Crime dataset [7]. As shown in Figure 3.1, measured in the time dimension (x-axis), the data unit Vandalism 23 for example, the video lasts about 210 seconds, but the time of the anomalous event occurs lasts only 18 seconds. It is about 8.6% of the whole video. Second, compared to the spatial dimension, the region where the anomaly occurs occupies only a very small number of pixels of the video frame. Figure 3.2 illustrates that the distinguishing features of abnormal instances in negative samples are not distinctly

prominent. Consequently, achieving accurate optimization of abnormal instances during training becomes challenging. As a result, the optimization of MIL may be steered in the wrong direction. Therefore, it is crucial to enhance the discriminative characteristics of positive and negative instances in the feature space by calculating the correlation between normal instances and abnormal instances in negative samples.

3.3.1 Temporal Convolutional Networks

1. Design principles Temporal Convolutional Networks are a type of neural network architecture designed for sequence modeling tasks, particularly those involving time-series data, which is derived from Time Series Networks [116]. Time series learning networks usually need to follow two principles [88]: (1) The input and output structures of the network should be the same; (2) The features of the current time node are not disturbed by the features of the next time node. The former is used to ensure that in the process of information mining, the sequence feature information will not be reduced and guarantee to extract high-quality representation features. The latter is to comply with objective facts. In the training process when using sequence data, since the complete sequence data has been obtained, the learning model can access the features after the current time node without obstacles. During the application process, the sequence data located after the current node cannot be accessed. Therefore, when designing the learning network structure, we should proceed from practical problems. That is, during the training phase, only the current node is provided with the features of its previous time nodes.

2. Feasibility of retrofitting traditional temporal convolutional networks In the traditional TCN(Temporal Convolutional Networks) structure [86], the convolutional network serves as the basic structural unit for extracting temporal features, and there is no aggregation mechanism or large memory module. The traditional TCN model has one-dimensional full convolutional structure [117], and the full convolutional structure ensures that the newly introduced network structure follows the first principle of temporal convolution, i.e., each hidden layer has the same length as the input layer and only the same input and output lengths are satisfied. However, this structure cannot store valid antecedent information and the posterior information may negatively affect the current features in the full convolutional network structure. Therefore, a novel TCN conforming to the second principle is proposed by Cheng et al [118], which consists of a fully convolutional network and a cascaded network. ($TCN = 1DFCN + CausalConvolutions$). The structure of this network implemented using cascading convolution, which uses the features of the same position of the previous layer and the features of its previous position to calculate the features of the current position. This temporal convolutional network conforms to the second

principle that the features of the current time node are not disturbed by the features of the next time node, which means that the features of this node are only affected by the features that precede this node in the video sequence. And the model provides stronger theoretical support when dealing with sequential data, such as text data, and video data.

However, this structure also has a major disadvantage, there is a possibility of parameter explosion, when we calculate the characteristics of all nodes in the previous layer before a certain node occurs, which also will require a large filter to process. And the information that is too old can also negatively affect the information of the current world nodes and reduce the quality of the extracted features. The existing video detection models usually use graph convolution or LSTM to store the sequence features (temporal features) of video data to complete the detection [85]. It mainly obtains indirect temporal features by LSTM and graph convolution. There is no strict definition and learning of temporal feature information of sequences. Since the convolutional network has a greater ability to scale [119], the performance of convolutional networks is improving in the learning task of sequence models. Based on this, this work introduces the dilated cascade technique into modern convolutional networks and implements a novel temporal convolutional network.

3 The Proposed Novel Temporal Convolutional Network The proposed TCN consists of Dilated Causal Convolutions (DCC) [90] and residual networks [91], and the cascaded dilated convolution layer is shown in Figure 3.4.a. Its core function can be defined as $X_{T+1} = X_T + F(X_T)$, where X_T denotes the current feature value and F denotes the cascaded convolution function.

Previous studies [120] have shown that TCN models outperform general-purpose recurrent architectures such as LSTM and GRU, and shown that the "infinite memory" advantage of RNN is basically non-existent in practice. Compared to recurrent architectures, TCN exhibits longer memory and wider convolutional horizons. In recurrent convolutional networks, many advanced schemes for regularizing and optimizing LSTMs have been proposed [121]. These schemes significantly improve the accuracy achieved by LSTM-based architectures on certain datasets. However, in the past two years, before the introduction of architectural elements such as dilated convolution and residual connections, the performance of convolutional architectures did not meet the needs of applications. Simple convolutional architectures are more effective than recurrent architectures such as LSTMs in various sequence modeling tasks. Due to the considerable clarity and simplicity of TCNs, convolutional networks should be seen as a natural starting point and a powerful toolkit for sequence modeling. Video data has sequence properties. In theory, any sequence data can be used to extract temporal features using the TCN model. The proposed TCN network [122] provides an important technique for mining the feature information of video sequences.

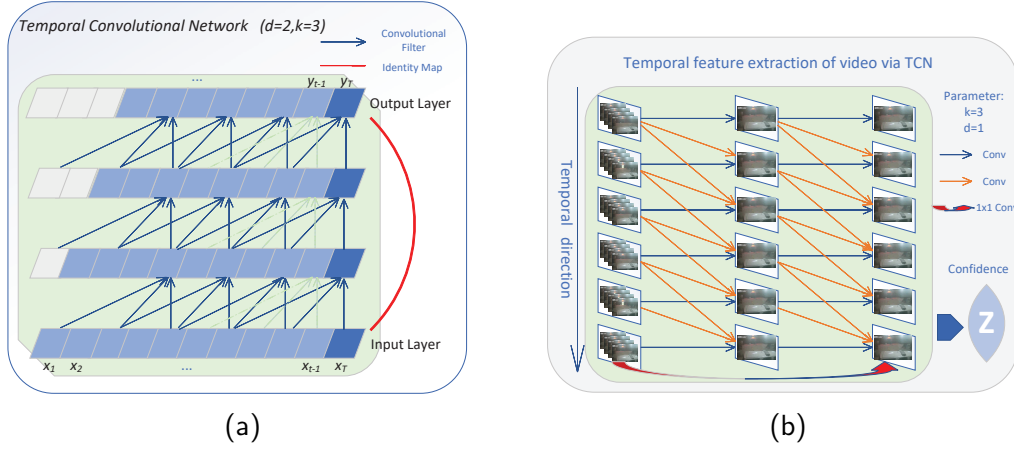


Figure 3.3: The structure of TCN and Application, (a) The proposed Temporal Convolutional Network structure, under $d = 2$, $k = 3$, the input is $X = x_1, x_2, x_3, \dots, x_T$, $k = 3$ is the number of kernels that three upper-layer neurons map a neuron of the current layer, $d = 2$ is stride representing the distance between two kernel units; (b) The novel TCN application in video processing. The red line represents the feature of the current node to be extracted, the blue line represents the feature of the previous node, and the red curve represents a 1×1 convolution unit that retains the most original features of the current node, and the output of the TCN is z , which is the probability value that the input node is an abnormal instance.

3.3.2 Extraction of Temporal Features of Video Sequences

The proposed TCN network is used to extract the features of the video sequence. This process is mainly divided into three steps: 1) Train vanilla discriminator C3D or I3D to extract the action features of the video data (the C3D is pre-training in Sports1M); 2) Input the features extracted by the vanilla discriminator into the new TCN network to extract high-quality temporal features, the steps of this process refer to Figure 3.3.b. The TCN network introduced in this chapter ensures the extraction of high-quality features through multi-layer concatenation and single-layer convolution; 3) According to the final temporal characteristics of the video, set the activation function to identify the video, and calculate the confidence of normal events and abnormal events.

The formalization process and the Qualitative Analysis of the NTCN-ML network:

Consider a video δ is divided into multiple segments δ_i^C , where $(i \in 0, 1, 2, 3 \dots I)$. The features extracted from the C3D network is represented by: $X_i = \phi_{vanilla}(\delta_i^C)$. Consider all video clip features belonging to the same data unit as a sequence of data $X_1, X_2, X_3, X_4, \dots, X_I$, where I represents the number of clips used, the first layer of

the hidden layer of the TCN is represented as X^1 , and the sequence is represented as $X^1 = X_i^1 | i = 1, 2, 3, \dots, I$, the calculation process:

$$X_i^1 = F(\prod_{t=0}^{k-1} (X_{i-td})) \quad (3.1)$$

As shown in Figure 3.3.b, when $k = 3, d = 1$; then $\delta_i^1 = F(\delta_i \cdot \delta_{i-1} \cdot \delta_{i-2})$, where F represents the convolution function, k represents the kernel during mapping. The number, d represents the stride, that is, the distance between two kernel units and so on for the rest of the nodes. The final output of the network structure of the output unit:

$$Output = Activate[(\delta_1, \delta_2, \dots, \delta_I) + F(\delta_1, \delta_2, \dots, \delta_I)] \quad (3.2)$$

Since the video is only divided into normal events and abnormal events, we set the output unit to two nodes, namely $Output = Z(z_1, z_2)$. z_1 represents the probability that the video segments belong to normal video, and z_2 represents the probability that the video segments belong to abnormal video. If the normal video contains elements in some abnormal events, the value of z_1 is more. On the contrary, if the abnormal events contain a large number of normal elements, the value of z_2 is low. Use the formula $\hat{X} = \max(z_1, z_2) \cdot X_i$ to construct a new video sequence feature. For normal segments δ_n , it belongs to the probability of a positive sample is $\max(z_1^n, z_2^n)$, and for abnormal segments δ_a its probability belonging to a negative is $\max(z_1^a, z_2^a)$, then the new feature of normal video $\hat{X}^n = \max(z_1^n, z_2^n) \cdot X^n$ the new feature of abnormal video is expressed as $\hat{X}^a = \max(z_1^a, z_2^a) \cdot X^a$. This work enhances the ability to determine abnormality by improving the separation characteristics between positive and negative samples. Therefore, this work proposes to use disentanglement to improve the performance of instance learning. The process of MIL is a paired training process in which a normal video sample and an abnormal video sample are included, and the probabilities of the normal video and abnormal video belonging to positive and negative samples are different.

3.3.3 The NTCN-ML Based on Temporal Convolutional Network Guidance

3.3.3.1 The proposed NTCN-ML model

A weakly supervised video anomaly detection model based on temporal convolutional network guidance is proposed in this chapter. The model uses a novel temporal convolutional network to extract the temporal features of video data and calculates the confidence of the samples. The overall framework is shown in Figure 3.4. The model also uses the classic vanilla discriminator (C3D - Convolutional 3D, I3D - Inflated 3D ConvNet) to extract the features of the video and combines the obtained confidence with C3D or I3D features to

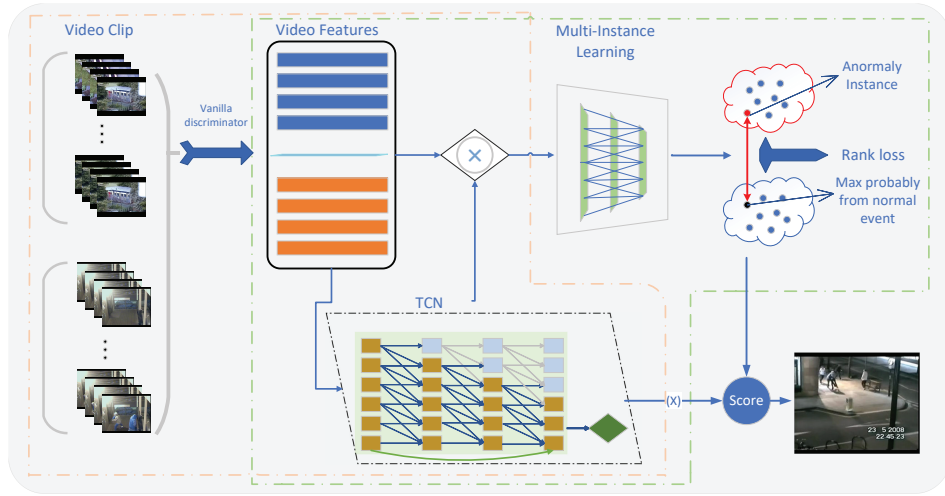


Figure 3.4: The NTCN-ML framework: The model training process is divided into two phases. The first phase is composed of a vanilla discriminator and novel TCN. The training purpose of this phase is to extract temporal features; the second phase is composed of a vanilla discriminator and TCN. The training model is composed of the MIL module, and the novel TCN module, this stage is to improve the classification ability of the MIL network.

form new input features. Then through the MIL network, the final abnormal probability of each instance is calculated; according to the abnormal probability, a loss function is constructed to train the parameters of the MIL. At the same time, the confidence of the video is also involved in the calculation of the abnormal score during the testing process. The NTCN-ML model proposed extracts temporal features through a novel TCN model and enhances the ability of MIL to learn instance labels. Compared with the mainstream algorithms, the NTCN-ML model has a more scientific and effective consideration of temporal features and has stronger robustness. Figure 3.4 shows the data processing flow of the proposed NTCN-ML model. We discuss model training, loss function, model testing, and algorithm complexity analysis during operation in the following sections.

3.3.3.2 The Training Phase

The training process is divided into two parts. One is to train the temporal convolutional network. The second is to train a MIL network. The function of the testing phase is to calculate the anomaly score of each instance in the video and locate the time area where the anomaly occurs.

Training the temporal convolutional network is divided into three steps, 1. Input the video clips into the vanilla discriminator to extract features; 2. Input the extracted features into the designed temporal convolutional network ; 3. Output A 2D array pre-

dicting video instance classification. This two-dimensional array represents the probability that the video instance belongs to normal events and abnormal events.

The formalization process is as follows: X represents a video, which is divided into multiple segments $X_i, (i \in 0, 1, 2, 3 \dots I)$. Each video segment is called an instance, because sets 16 frames is a segment, so $I = F_n/16$, F_n is the total number of frames in the video. χ is extracts the features of the video segment X_i^C by vanilla discriminator $\phi_{vanilla}$ vanilla is belongs $I3C, C3D$. The TCN function is denoted by f_{TCN} . The final output is represented as:

$$z = f_{TCN} \sum_{i=0}^I \chi = f_{TCN} \sum_{i=0}^I \phi_{vanilla}(X_i) \quad (3.3)$$

$$loss_{TCN} = z - \bar{h}, \quad \begin{cases} \text{normal} & \bar{h} = (0, 1) \\ \text{abnormal} & \bar{h} = (1, 0) \end{cases} \quad (3.4)$$

z represents a two-dimensional vector, where z_1 represents the probability that the video belongs to a normal video, z_2 represents the probability that the video belongs to an abnormal video, the label of the video is \bar{h} , the label of normal video is 0, and the label of abnormal video is 1, If the video is a normal video, its label is $(0, 1)$, otherwise it is $(1, 0)$ at the phase of TCN training.

The training of the multi-instance anomaly detection algorithms is divided into four steps.

- Use the vanilla discriminator to extract video features;
- The extracted features are input into the pre-trained temporal convolutional network and output a two-dimensional vector for each instances;
- The inner product of the large value and the video feature matrix constructs new video features;
- The new video features are input into the MIL network, and the abnormal probability of each instance is calculated.

The formalization process is as follows: select the extracted C3D feature $\chi = \phi_{vanilla}(X_i)$ of a fixed length T (fixed number of instances) and the largest value dot product in the output z of the TCN trained in the first phase. Get new video features:

$$\hat{\chi} = z \cdot \chi = f_{TCN} \left(\sum_{i=0}^I \phi_{vanilla}(X_i) \right) \cdot \chi \quad (3.5)$$

During the learning process of the multi-instance algorithm, normal videos instance and abnormal videos instance are input to the neural network in pairs. We use χ^n to denote the features of normal videos, χ^a to denote the features of abnormal videos and the MIL is denoted as f_{MIL} .

$$Y = F_{MIL}(\hat{\chi}^n, \hat{\chi}^a) \quad (3.6)$$

Where $Y = (Y_a, Y_n)$, $Y_a = (y_1^a, y_2^a, y_3^a, \dots, y_T^a)$ represents the abnormal probability of all instances in the abnormal video package, $Y_n = (y_1^n, y_2^n, y_3^n, \dots, y_T^n)$ represents the abnormal probability of all instances in the normal video package.

3.3.3.3 Loss function

The loss function of MIL consists of four parts: ranking loss $L_{ranking}$, smooth loss L_{smooth} , sparse distribution loss $L_{sparsity}$, aggregation loss $L_{cluster}$. The ranking loss represents the difference between the highest abnormal probability in the normal video package and the highest abnormal probability in the abnormal video package at the training process. So the Ranking loss function is expressed as:

$$L_{Ranking} = ||\max(Y_a) - \max(Y_n)|| \quad (3.7)$$

The video is composed of multiple video clips and is sequence data. Therefore, the distribution of abnormal probability should be smooth, and the smooth loss indicates that the occurrence of abnormality in the video sequence is promoted by a process. The smooth loss function is expressed as:

$$L_{Smooth} = \lambda_1 \sum_{i=0}^{T-1} ||y_{i+1}^a - y_i^a||^2 \quad (3.8)$$

Loss of sparse distribution. In abnormal video, the time of abnormality only accounts for a very small part of the entire video data, so the average abnormal probability of the entire abnormal video is slightly higher than the average abnormal probability of normal video. The sparse loss function is expressed as:

$$L_{sparsity} = \lambda_2 \sum_{i=0}^T ||y_i^a - y_i^n||^2 \quad (3.9)$$

Aggregation loss: The difference between the maximum and minimum values of each instance in the video packets of normal events is not much different. On the contrary, the difference between the maximum value and the minimum value of each instance in the

video package of the abnormal event is relatively large. So the aggregation loss is expressed as:

$$L_{cluster} = \lambda_3(1 + \max(Y_n) - \min(Y_n) + \min(Y_a) - \max(Y_a)) \quad (3.10)$$

The total loss function is expressed as:

$$L = L_{ranking} + L_{smooth} + L_{sparsity} + L_{cluster} \quad (3.11)$$

3.3.4 The Anomaly Detection Phase

The anomaly detection phase is to describe the detection process of video data that cannot obtain any labels during the test process. The whole process is carried out unsupervisedly. In the detection stage, the algorithm complexity of video anomaly detection is also an important indicator for evaluating models.

3.3.4.1 Steps of detection

The detection steps in the anomaly detection phase are divided into four steps. The first step: preprocess the video data, divide the video into multiple video segments, and use the vanilla discriminator to obtain the feature (C3D, I3D) of these segments; The second step: input the feature into the trained TCN model, obtain the temporal feature and calculate pseudo labels for instance. The three step: combination to construct new instance features; The fourth step: the video features are input into the MIL, and the anomaly probability of each instance is calculated. The calculation of the anomaly score, the anomaly score is composed of the last instance anomaly probability, loss, and confidence.

$$Score = z \cdot y + \gamma_1(\delta_y) \quad (3.12)$$

The pseudocode of the anomaly detection phase is presented by Algorithmic 1:

Algorithm 1 Anomaly Detection

- 1: Initialization: f_{TCN} , f_{MIL} , $\phi_{vanilla}$, Pre-trained TCN network, MIL network and C3D vanilla discriminator;
- 2: $\chi = \phi_{vanilla}(X)$, Extract features from video clip X ;
- 3: $Z = f_{TCN}(\chi)$, Calculate the confidence of the video X Equation (3.3);
- 4: $Y = f_{MIL}(Z \cdot \chi)$, Calculate the anomaly probability of labels for each segment of the video Equation (3.6);
- 5: $Y_{var} = variance(Y)$, Calculate the volatility of video anomaly probability; Equation (3.12)

Output: Anomaly score= $\{\lambda_1 Y_{var} + Z \cdot Y\}$,
Calculate anomaly scores.

3.3.4.2 Algorithm complexity analysis in the detection process

The training phase only happens before the model is deployed, so only the algorithmic complexity of the inference process needs to be considered:

The complexity of the temporal convolutional network model: for a video sequence, extract T segments, input the TCN model to classify the video segments, the algorithm complexity depends on the number of input segments T , the dimension of the feature F of each segment, the number of hidden layer nodes, the number of hidden layers L , the number of kernels k in the TCN model, the stride d , and finally the category C . First, map the extracted features F to the first hidden layer, and each k feature is mapped to a unit.

$$O_{TCN} = (O(\phi_{vanilla}) \cdot k \cdot T)^L \cdot C \quad (3.13)$$

where C is a 2-category, normal or abnormal, and T is the number of segments. According to past experience, 32 are chosen, so the algorithm complexity mainly depends on the level of the network and the number of nodes.

The complexity of the MIL model: In the MIL process, the input unit usually consists of a feature sequence δ^n from normal videos and a feature sequence δ^a from abnormal videos. Each feature sequence contains T feature segments, and the MIL consists of three fully convolutional layers (l_1, l_2, l_3)

$$O_{MIL} = O(F_d(\delta^n) + F_d(\delta^a)) \cdot (l_1 + l_2 + l_3) = O(2F_d \cdot T) \sum_{i=0}^3 l_i \quad (3.14)$$

Therefore, in actual operation, the total algorithm complexity is:

$$O = (O(\phi_{vanilla}) \cdot k \cdot T)^L \cdot C + O(2F_d \cdot T) \sum_{i=0}^3 l_i \quad (3.15)$$

The above formula shows that the algorithm complexity mainly depends on the number of hidden layers of the neural network and the number of nodes in each layer.

3.4 Experiments

3.4.1 Datasets

There are two commonly used datasets for weakly supervised video anomaly detection algorithms, namely UCF-Crime [7] and ShanghaiTech datasets [34], which also is the benchmark datasets. So we validated the proposed model with these two datasets. Table 3.1 displays the data distribution of the two datasets, revealing that despite the UCF dataset containing an equal amount of normal and abnormal data, the distribution of training and testing sets

Table 3.1: Dataset Overview: Nor and Abnor are normal and abnormal videos; Atype is the number of abnormal types; N/A denotes the number of normal videos / abnormal videos

Numbers	Total	Nor	Abnor	ATypes	Train(N/A)	Test(N/A)
UCF_Crime	1700	950	950	13	810/800	140/150
ShanghaiTech	437	330	107	13	175/63	155/44

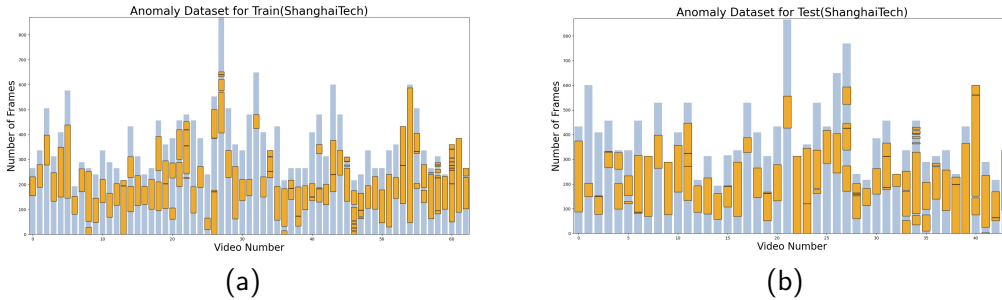


Figure 3.5: The distribution of the ShanghaiTech dataset, (a) denotes the abnormal distribution of abnormal data entries(63 videos) in training, x-axis represents the video number and y-axis represents the total number of frames. Yellow colour is the location for abnormal frames; (b) denotes the abnormal distribution of abnormal data entries in testing for 44 videos.

is imbalanced. Furthermore, during the data reading process, disrupting the arrangement of normal data and abnormal data. This is for when the number of normal data sets and abnormal data is not equal, taking one data from each at the same time will ensure that each normal data can be matched with different abnormal data during the multi-instance comparative learning process, thereby enhancing the robustness of the model. Zhong et al. [64] split the data into two subsets: a training set consisting of 175 normal videos and 63 abnormal videos. As shown in Figure 3.5:

3.4.2 Experiment Details

The core elements in our implementation process include the following steps: 1. Cut the video into 32 equidistant instances, each containing 16 video frames; 2. The extracted C3D and I3D features are stored in numpy format to speed up the training; 3. During the implementation of the TCN network, and the number of input channels is 32 for the number of instances of a video. In order to prevent overfitting, The hidden layer of the TCN network used in our work is $(32*8)$, a total of 8 layers, and each layer has 32 nodes; 4. The MIL consists of $(512, 32, 1)$ three-layer fully connected convolutional networks. The evaluation index refers to the literature of this series, with AUC as the main evaluation

Table 3.2: The TCN classification performance analysis under the UCF-Crime

UCF-Crime(C3D)	32	64	96
32*8	85.1	85.3	84.9
64*8	86.0	85.1	85.2
128*8	84.8	84.3	84.4

index [108, 109].

3.4.3 Experimental results

The experiments are set in three groups. The first group examines the classification performance of the TCN network and obtains the TCN network structure with the best performance. The second group is the AUC evaluation experiment for video anomaly detection. The purpose of this experiment is to measure the performance of the model proposed. The last group is the visualization experiment, the main purpose of which is to promote the important evaluation method for the model to transfer from the experimental scene to the application scene.

The experimental results indicate that with the increase in the number of divided segments, the classification accuracy of TCN does not show a linear increase. Through the analysis of the dataset, it is inferred that this is due to the "invalid filling" caused by the different time lengths of each video in the data. The reason for data padding is that the duration of some videos is too short to meet the number of divisions, and it is necessary to repeatedly borrow some video clips and video frames to construct a specified number of clips. As an example, 32 video clips each one is made up of 16 frames, and the total number of video frames of each video cannot be less than 512. Through the analysis of the data set, only a few videos have a total number of frames less than 512. If we divide each video into 64 segments, there are nearly 12% of the data does not have enough frames to split, and therefore overfitting occurs and the detection accuracy decreases. Hence, we decided to use 32 fragments as a reasonable number in our experiments.

3.4.3.1 Experiment 2: AUC comparison with state-of-the-art models

The purpose of this experiment 2 is to compare the AUC accuracy of the algorithm proposed with the current mainstream algorithms.

In this process, we first train a novel temporal convolutional network. The output value is the probability that the segments of normal video and abnormal video belongs to abnormal. After completing the TCN model training, input the feature to MIL model. First, divide a video into multiple segments and extract features; second, extract features of a fixed number of segments as input, and the number of segments is the number of

Table 3.3: Accuracy test of current mainstream algorithms on the UCF-Crime dataset

Method	Source	Technique	Performance (AUC)
Sultani et al [7]	CVPR18	C3D	75.4
TAEDM [123]	SCN20	ResNet	78.5
TCN-IBL [124]	ICIP19	TCN & IBL	78.7
Zaheer et al [125]	SPL21	Self-Reasoning	79.5
GCN-AD. [64]	CVPR19	GCN & Action Classifier	82.1
XD-Violence [126]	ECCV20	Holistic-Localized Networks	82.4
CLAWS [68]	ECCV20	Clustering	83.0
SACRF [127]	ICCV21	Relation-Aware	85.0
RTFM [61]	ICCV21	Feature Magnitude	84.0
STGCNs [19]	IPM22	Spatio-temporal GCN	84.2
BN-SVP [128]	CVPR22	Bayesian	83.4
Ours		Novel TCN	85.1

Table 3.4: Accuracy test of current mainstream algorithms on the ShanghaiTech dataset

Method	Source	Technique	Performance (AUC)
TCN-IBL [124]	ICIP19	TCN & IBL	83.5
Zaheer et al. [125]	SPL21	Self-Reasoning	84.2
GCN-AD [64]	CVPR19	GCN & Action Classifier	84.4
CLAWS [68]	ECCV20	Clustering-Based	89.7
AR-Net [65]	ICME20	AR Network	91.2
TAEDM [123]	SCN20	ResNet	94.2
MIST [62]	CVPR21	Self-Guided Attention	94.8
BN-SVP [128]	CVPR22	Bayesian	96.0
Ours		Novel TCN	95.3

input channels; third, input features to the TCN model, calculate the probability; The fourth step is to take the larger value in the two-dimensional array and perform the point multiplication operation with the extracted features, and then input it into the MIL model to calculate the abnormal probability of each segment.

Table 3.3 results shows that the algorithm proposed in this chapter has achieved an accuracy of 85.1% on the I3D features of the UCF-Crime dataset, which has reached the most advanced accuracy. In addition, in order to test the classification performance of the TCN network, C3D features were extracted from the original video to analyze the performance of TCN. For details, see Experiment 1.

Table 3.4 shows that the AUC accuracy of the model proposed has reached 95.3%. Compared with the current most mainstream algorithms, the algorithm proposed has surpassed the performance of most published mainstream algorithms. Through the experimental results of the two data sets, it is concluded that the correlation between normal data and abnormal data is also an important consideration in the process of abnormal detection. The model proposed overcomes the above two shortcomings.

Table 3.5: Ablation study: Divided two datasets into four groups: I3D+MIL, C3D+MIL, I3D+TCN+MIL, C3D+TCN+MIL, to evaluate the TCN module.

	I3D	C3D	TCN	AUC
ShanghaiTech	✓		✓	95.3
		✓	✓	88.3
	✓			86.1
		✓		85.3
UCF_Crime	✓		✓	85.1
		✓	✓	78.2
	✓			82.3
		✓		76.1

Table 3.6: The Study of Loss Function: Set different loss function combination modes to explore the impact of different loss functions

I3D+TCN+MIL	$L_{ranking}$	$L_{sparsity}$	L_{smooth}	$L_{cluster}$	AUC
	✓				83.6
UCF-Crime	✓	✓			83.7
	✓	✓	✓		84.7
	✓	✓	✓	✓	85.1
	✓				91.1
ShanghaiTech	✓	✓			91.3
	✓	✓	✓		92.7
	✓	✓	✓	✓	95.3
	✓	✓	✓	✓	95.3

3.4.3.2 Experiment 3: Ablation Study

To test the model’s capability, we conducted two sets of ablation experiments: An ablation study and a Loss Function study. The former involved training and testing different components of the TCN model independently to confirm their effectiveness. The latter involved combining various loss functions during training to examine their impact on performance. Our aim was to verify the impact of different loss functions on the model’s performance.

The Ablation study conducted in this chapter involves the verification of the model with two datasets (UCF-Crime and ShanghaiTech) using C3D and I3D to independently extract video features and input them into MIL training. Additionally, The model training is divided into four groups, namely I3D+MIL, C3D+MIL, I3D+TCN+MIL, and C3D+TCN+MIL, and the performance was calculated for each group as shown in Table 3.5.

Table 3.5 shows the results of the ablation experiments. The results show that the TCN module used in this article can effectively improve the accuracy of the model on the benchmark. On the ShanghaiTech dataset, the model proposed in this chapter improves by 9% compared with the baseline I3D+MIL. Compared with the baseline C3D+MIL, the model improves by 3%; for the UCF-Crime dataset, our model improves by 3% compared with the baseline I3D+MIL, and compared with the baseline C3D+MIL, the accuracy improves

by 2%. It shows that the TCN module proposed in this article is effective..

In the study of loss function, this chapter uses the ranking loss function as the benchmark, and cooperates with several other novel loss functions to test the performance of the model in two data.

Table 3.6 shows that when there are more types of loss functions combined, the performance tends to increase slowly. Among them, the $L_{sparsity}$ loss has a general effect on improving the model performance, and the loss L_{smooth} and $L_{cluster}$ have a greater impact on performance. The experimental results show that the $L_{cluster}$ loss function is helpful for performance improvement.

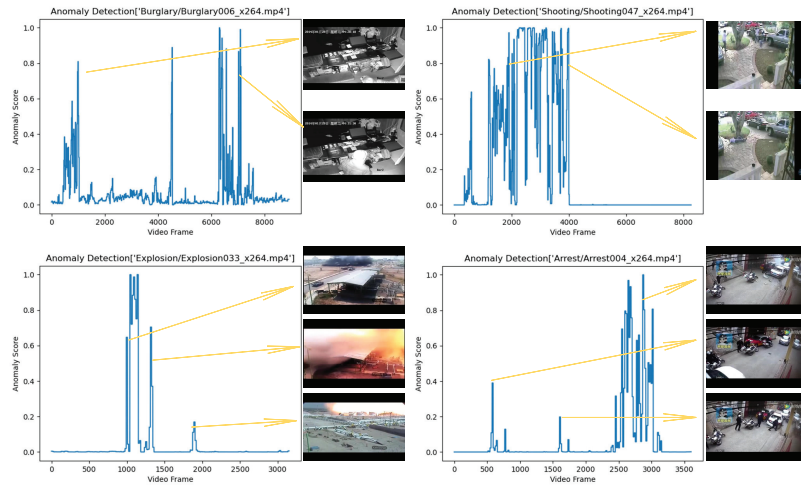
3.4.3.3 Experiment 4: Visual display during anomaly detection.

In the testing phase, outliers for anomaly detection are constructed from the output of the pretrained TCN network, the output of the MIL model, and the loss function.

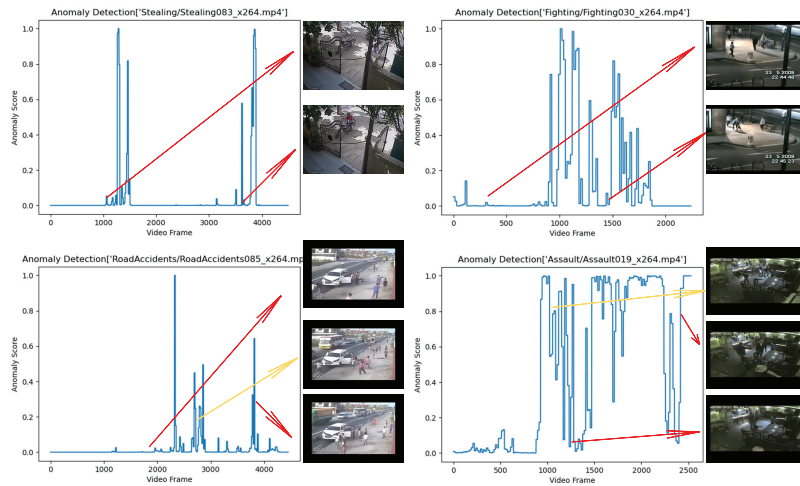
The results of experiment 4 show that when an abnormal event occurs in a video, the abnormal score will fluctuate violently (the abnormal score is generated after normalization), so the fluctuation of the abnormal value can be used as the identifier of an abnormal occurrence. Second, the results in Figure 3.6 (b) show that for long-lasting abnormal events, the fluctuations of outliers will be abnormal, resulting in inaccurate detection results. The reasons for this problem mainly come from two aspects: 1. C3D and I3D motion capturers tend to capture short-term actions, as show in Figure3.6 (a); the action extractor is training in Sports1M, and the action duration of this data set is relatively short. Therefore, the action capture used to preprocess the data set is more favorable for the short duration. 2. Video instance division and the generation of instance outliers do not meet the actual situation of long-term actions. During the experiment, 16 frames are usually delineated as an instance, and there are also cases where the duration is shorter. We hope that follow-up research in this chapter can optimize this problem. Figure 3.7 is a supplement to the visual experiment. In order to show the experimental effect more clearly, the video data is divided into 32 instances for calculation.

3.5 Conclusion and Future Work

This work proposes a novel weakly supervised anomaly detection model (NTCN-ML), a new Temporal Convolutional Network (TCN). The NTCN-ML model shows an excellent performance in temporal information mining and provides high-level temporal feature information for weakly supervised learning. The advantage of the NTCN-ML model is that it can enhance temporal features for the entire video sequence, which is different from other related works as they calculate temporal features in segments, and redefine the integrity



(a)



(b)

Figure 3.6: Visual effects of the anomaly detection phase, (a) the detection results of anomalous events with a short duration, (b) the detection results of anomalous events with a longer duration d, the yellow line indicates the correctly detected samples, and the red line indicates the detection results is wrong.

and coherence of temporal features of video data. Our experimental results show that the NTCN-ML model learns the potential patterns from both anomalous and normal events, and outperformed the baseline anomaly detection models considered in this work. The algorithm presented in this research chapter introduces a novel approach for video anomaly detection algorithms, delving into the distribution of data within the feature space in weakly

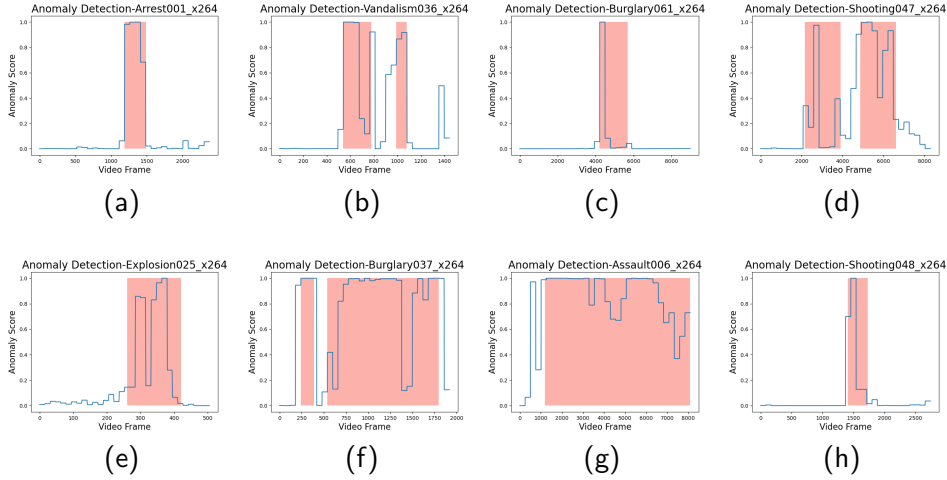


Figure 3.7: Visual effects of the anomaly detection phase. Red is the area where real anomalies occur, and the curve is the anomaly score.

supervised algorithms, and optimizing the process of weakly supervised learning. Furthermore, the proposed model can be seamlessly integrated into other systems, enhancing the algorithm’s robustness in real-world applications. However, this chapter is subject to certain interpretability limitations. It is expected that future research in the domain of video anomaly detection will primarily focus on improving interpretability.

In the future, we will assess whether the temporal features extracted from the video sequences align with real-world scenarios, and how the integrity and coherence of temporal features affect video data analysis. We will also evaluate whether the integrity of temporal signatures has positive implications with both unsupervised and supervised models. Based on the current work, we will further try to define a new anomaly definition in which anomalous events are deeply associated with global temporal signatures. This will certainly help to integrate temporal video analysis patterns in real traffic scenarios.

COVAD: Content-Oriented Video Anomaly Detection using a Self-Attention based Deep Learning Model

Contents

4.1	Overview	64
4.2	Introduction	64
4.3	Methodology	66
4.3.1	Encoders and Decoders	67
4.3.2	Memory module	70
4.3.3	Coordination Attention	73
4.3.4	Loss Function	75
4.4	Experimental Results	78
4.4.1	The proposed COVAD approach	78
4.4.2	Dataset Description	79
4.4.3	Hyperparameter selection process	80
4.4.4	Effectiveness of the attention mechanism in video anomaly detection	80
4.4.5	A Visual Test	82
4.4.6	Video anomaly detection using COVAD	82
4.5	Conclusion and Future work	84

4.1 Overview

This chapter proposes a novel video anomaly detection method named COVAD, which mainly focuses on the region of interest in the video instead of the entire video. Our proposed COVAD method is based on an auto-encoded convolutional neural network and coordinated attention mechanism, which can effectively capture dynamic objects in the video. Relying on the existing memory-guided video frame prediction network and object detection method, our algorithm can more effectively predict the future motion and appearance of objects in the video, and mark abnormal areas or objects. Our proposed algorithm obtained better experimental results on multiple data sets and outperformed the baseline models considered in our analysis. As visual result shows: the proposed model can provide pixel-level anomaly explanations.

4.2 Introduction

Many videos anomaly detection algorithm uses convolutional neural network (CNN) to learn video features, including temporal dimension features and spatial dimension features. Then, use decoding to reconstruct the video or combine with optical flow technology to predict the next frame. According to the definition of training loss, existing unsupervised and semi-supervised video anomaly detection algorithms are divided into two categories, one is reconstruction-based anomaly detection [36, 129, 130], and the other is prediction-based [45, 46] anomaly detection algorithms. The reconstruction-based anomaly detection algorithm defines the reconstruction loss as the training loss. Reconstruction-based method assumes that the detection model is trained by a large amount of normal data, the model can accurately describe normal events, extract video features, and restore video features to video frames with small reconstruction errors. If no data objects participate in the training especially for abnormal events, then the model will get a large loss when reconstructing abnormal videos. In the detection phase, error thresholds are set to detect abnormal events. For future frame prediction, the training loss is the prediction error, and the basic structure is to extract the video features of the previous frames and predict the features of the future frames. During the training phase, the loss between the predicted future frame and the real future frame is calculated, and the network parameters are updated.

This chapter proposes a video anomaly detection algorithm for future frame prediction and thus, it follows the assumptions that the models trained on normal data sets have small errors in predicting future frames of normal events, and abnormal events have higher prediction errors due to their uncertainty [131].

After the emergence of deep learning techniques, the use of CNN to extract video features instead of the original hand-made features greatly saves time and cost, and achieved

higher accuracy after training the models on specific scenario. The basic structure of current video anomaly detection algorithms is almost the same. It is mainly divided into the following steps: input the video frame into the encoder to extract the features using the training method of the adversarial CNN, and use the decoder to restore the features. Then, calculate the error between restored features and the original features, and adjust the network parameters to make the extracted features closer to the video frame. The neural network has strong representation ability, but in order to prevent unbounded expression, it is necessary to limit the representation ability of the neural network by adjusting the pooling part of the network structure. In addition, it is difficult to obtain an accurate model to discriminate anomalies with a single network structure parameter training and thus, it is necessary to record the extracted video frame features (all training sets are from normal events). One of the most typical solutions for this is to adapt memory-guided video anomaly detection algorithm as proposed in [23] in 2020. This method adopts the latest U-Net symmetric network, which has strong representation ability. In U-Net network, the back-sampling technology in the decoder can make up for the loss of spatial information in the pooling process and the memory module between the encoding and decoding further retains the features and feeds it back to the decoder for preserving spatial information.

In this chapter, a content-based video anomaly detection algorithm - COVAD, is proposed and its network structure is modified based on the original memory-based video anomaly detection algorithm. The main goal of optimization in the training network is to focus on the objects in the video frame. We use contentbased attention mechanism to optimize the structure of the encoding network and removed the last batch of normalization layer of the U-Net network. The former is used to focus on the target or content in the video and the latter is used to limit the powerful bias of the neural network as it is important to blur the boundary between normal data and abnormal data in powerful representations. Compared with the object detection algorithm, the attention mechanism is lightweight, does not take up a lot of time, and can effectively process video. The memory module stores more important content information, rather than the entire video frame pixels. Our experiments are deployed on the USCD [5] and Avenue datasets [6],and the experimental results show that the algorithm proposed in this contribution has better results compared to the bench mark models.

The main contributions of this chapter are 1) to propose a novel video anomaly detection method, called COVAD, for future frame prediction by combining the content-based attention mechanism, which can resist the interference of noise and focus on extracting the features of objects in the video, 2) to redefine memory module, which is used to classify and memorize various normal behavioral patterns available in video streams, and 3) to further improve the performances of video anomaly detection models focused on both normal and

exceptional events. The experimental results show that the performance of the proposed COVAD algorithm in this chapter is significantly higher than that of the baseline models considered in this work.

4.3 Methodology

This chapter focuses on combining memory module guidance and the content-based self-attention mechanism to propose a new video anomaly detection algorithm, which is mainly based on future frame prediction. The COVAD method proposed in this chapter, first, learns the temporal and spatial features of the video and then, uses cosine similarity function to maps its features to the memory module and updates the records of the memory module. Finally, the decoder network is used to restore the video features, calculate the difference between the predicted video frame and the real video frame, and evaluate the error. However, unlike previous methods, this chapter modifies the encoder and decoder networks and proposes a content-oriented self-attention mechanism by integrating the encoder/decoder network, which helps us focus features on video frames. The dynamic area helps the model better locate abnormal areas. This chapter further improves the data update method of the memory module in COVAD mode to support the memory module for various normal events. Figure 4.1 describes the COVAD system architecture, more details about the system are provided in the following sections

The area where abnormal events occur in a video only occupies a small part of the entire video frame, and therefore, most of the scenes in video frames are useless for detecting abnormal events, which we call in this research as the background. In video anomaly detection, it is generally accepted that stereoscopic, interdependent content, or objects in the video are more worthy of attention. However, most algorithms today are not designed with this argument in mind. Therefore, motivated by this, this chapter proposes a novel video anomaly detection algorithm that incorporates a state-of-the-art content-oriented self-attention mechanism to training on the important content of video frames, rather than the providing much attention to the background.

The algorithm proposed in this chapter is mainly divided into three parts: the encoder, the memory module, and the decoder.

- The encoder is used to extract the temporal and spatial features of the video,
- The memory module records the behavior patterns of normal events, and
- The decoder restores the extracted features as video frames.

Encoder and decoder: The most popular encoder and decoder used for video processing at present is the U-Net symmetric network [132]. The structure of the network is sym-

metrically distributed, which can effectively represent the process of feature extraction and feature restoration of video frames, as shown in Figure 4.1. Apart from that, due to the special aggregation mode of the U-Net network (meaningful data is appended when restoring features) and the up-sampling process, the motion and appearance information of the video can be preserved to the greatest extent.

Memory module: This module is a sparse binary matrix, which is updated during the training process, and constantly fits the behavioral patterns of normal events to realize the function of memorizing normal behavioral patterns. The basic principle is to use a sparse binary matrix to record the video features in each iteration. As the number of iterations increases, the sparse matrix of the memory module fits the normal behavior pattern during training.

In our proposed approach, model input which is the continuous video frame sequence $Seq = \{I_1, I_2, I_3, \dots, I_n\}$, $I_N \in \mathbb{R}^{W,H}$ of length N is divided into two parts, $I_{n-1} \in \mathbb{R}^{W,H}$ and $I_{nth} \in \mathbb{R}^{W,H}$. The first I_{n-1} frames are used as the input in the training process to extract features set $f_{I_{n-1}} \in \mathbb{R}^{W,H,C}$, where C is the final number of channels, and then read the memory $Mem \in \mathbb{R}^{M,C}$ to get the similarity index matrix $V \in \mathbb{R}^{M,W \times H}$. Then, update the memory module by V , and aggregate feature f_{n-1} and Mem to obtain $Agg_f \in \mathbb{R}^{2C,H \times W}$. Following that, model restores the features Agg_f to get the predicted \hat{I}_{nth} frame. Finally, calculate the loss between the predicted \hat{I}_{nth} frame and the real I_{nth} frame after retrieving the predicted value from the model. There are also some other additional loss functions applied during the training phase.

In the following sections, we explain each module presented in Figure 4.1 that are used in our COVAD framework.

4.3.1 Encoders and Decoders

U-Net was originally designed as a CNN for image segmentation and has achieved excellent results in many international competitions [132] [133]. Its unique structure and design philosophy inspired researchers in the field of computer vision, such as symmetrical ideas, up-sampling, and skip connections. The necessary functions of the CNN for video anomaly detection are to extract video feature frames and restore the feature to video frames through encoding/decoding process. The U-Net has a natural advantage that other network structures do not have, which is the symmetric structure of the network as shown in Figure 4.2. It consists of repeated applications of convolutions each followed by pooling at the extract feature phase and upsampling at the restore phase. For the upsampling, we all know that the max pooling is non-invertible, so we can add switch variables recording the information of max pooling, such as the position of the maximum value. In the decode, the upsampling uses these switches to reconstruct current layer above into appropriate locations of next

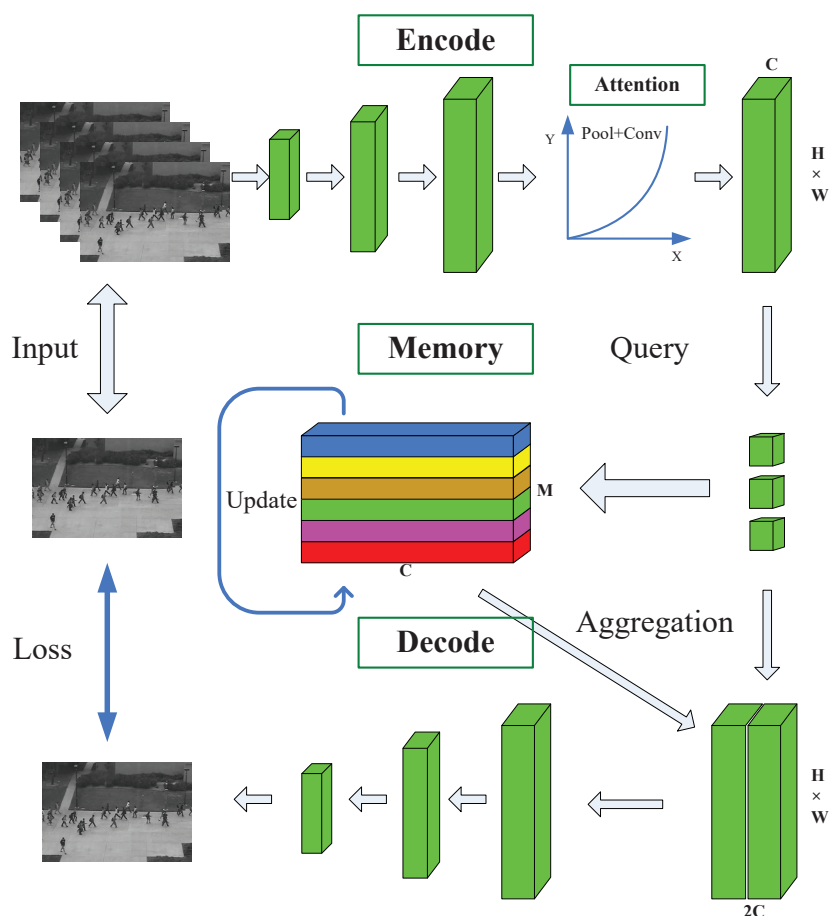


Figure 4.1: Algorithm Framework: 1. Extract video features through an encoder, 2. Then input collaborative attention mechanism to redistribute weights, 3. Read memory module and update, 4. Restore the aggregated query features and memory module features to video frames, 5. Calculate the loss, backpropagate, and update parameters

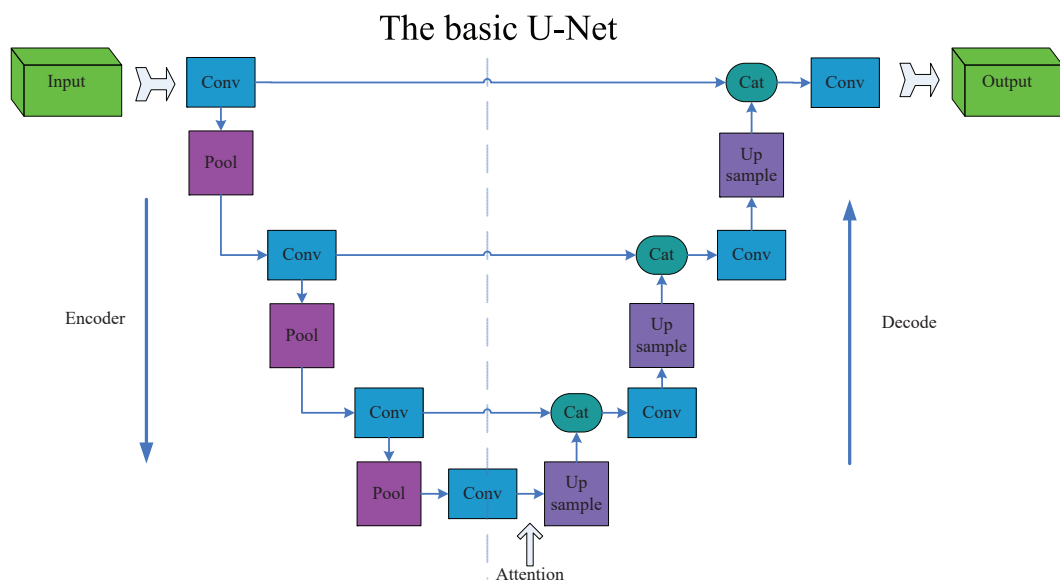


Figure 4.2: The basic U-Net: The U-Net network is composed of convolution, pooling, upsampling, and skip connections, where convolution and pooling are used to extract input features, upsampling is to restore the pooled and scaled features, and skip connections are feature splicing, trying to use a wider range of information to help restore video frames

layer, preserving the structure of the stimulus [134].

At present, U-Net is widely used in video frame reconstruction and future frame prediction tasks. In addition, due to the skip connection of the U-Net network, fine-grained details can be recovered during prediction by extracting more video information during the decoding process. However, in the U-Net network structure, skip connections are not always useful specially for reconstruction tasks. This is mainly due to having noisy data in the previous feature set and not conducive to restore the most realistic features. Thus, skip connections are unrealistic to apply in this scenario. For the prediction-based video anomaly detection task in this chap focuses on the previous features that contains part of the information lost during the training process, and connecting the previous features to the current features can improve the accuracy of prediction [135]. The Attention in Figure 4.2 provides the interaction between video features and memory module.

Another issue with the strong representation ability of CNN is that the inability of defining the exact boundary between the normal event and abnormal event [23]. The final feature extracted from the encoding of the training phase, which obtained from normal data might deviate from the normal pattern, or out of its boundaries. In the testing phase, the features extracted from abnormal data may be regarded as normal features, resulting misclassification. Therefore, identifying and limiting the representation ability of neural network model is one of the most important aspects of network structure optimization. We removed the last batch of normalization [136] and ReLU layers [137] in the encoder, limiting different feature representations. We instead add an L2 normalization layer to make the features have a common scale.

4.3.2 Memory module

This module is composed of a randomly generated sparse matrix $M \times C$. The length and width of the matrix is M , depending on the actual application scenario, usually representing the number of normal behaviors in the training set, or the number of videos in the training set, or the number of different camera positions. The length of the feature extracted by the CNN is C , which is the same as the width of memory. Here, the operation of reading and updating the memory module in this chapter basically follows the processing in [23] [138]:

Read: The read operation is to calculate the similarity between the query point and all the entries in the memory module, and find the closest entry and the second entry from the query point. The former is used to fit the query-worthy behavior pattern, and the latter is used to expand the class spacing, where there are two components of the loss function. Second, in the update operation, the weighted average of the query points is accumulated to the nearest entry by the L2 norm.

In the process of reading the memory module, first calculate the similarity between the

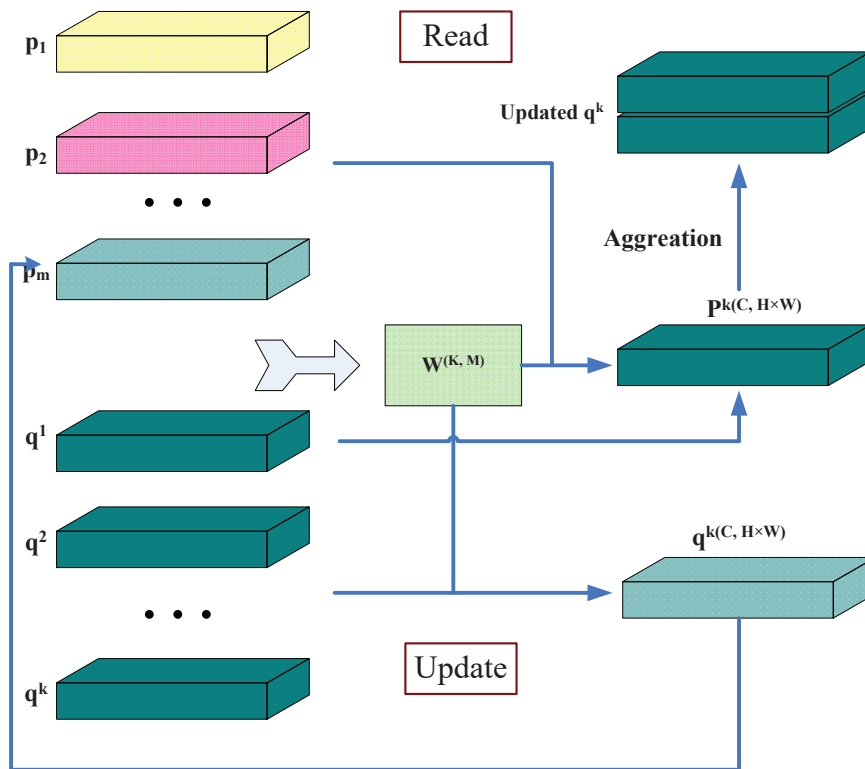


Figure 4.3: This is the algorithm flow of the memory module, including the flow chart of reading and updating memory

query feature value and all the entries in the memory module, that is, the cosine similarity, which is calculate by equation 4.3.2:

$$w^{k,m} = \frac{\exp(p_m^T q^k)}{\sum_m (\exp(p_m^T q^k))} \quad (4.1)$$

where p_m represents the entry in memory and q^k represents the query point, the encoded feature of the input video. So, we compute the similarity($w^{k,m}$) of the query point(q^k) to the memory module(p_m) as the weight of the memory module and read the memory module according to this weight($w^{k,m}$).

$$p^k = \sum^M w^{k,m'} p_{m'} \quad (4.2)$$

This chapter reads all memory entries instead of the closest entry, to consider the integrity of the normal pattern, which is beneficial to get a more accurate model. because anomaly detection is essentially a binary classification problem. In the real scene, different normal patterns may coexist at the same time, and there is an interdependence between the normal.

Update We use the probabilities in equation 1 to select all the nearest query points corresponding to each memory. U^m is defined as the index set of the m -th memory entry corresponding to the nearest query point, then the update mechanism is completed by the following equation.

$$p^m = \varphi(p^m + \sum_{k \in U^m} \hat{\nu}^{k,m} q^k) \quad (4.3)$$

The weighted average is used here instead of *sum*, so that the query points closer to m -th have a greater impact on the update of m -th. The way of calculating $\nu^{k,m}$ is the same as equation 1, but the normalization in the horizontal direction is used here.so $\nu^{k,m}$ has the following equation to calculate. However, since the value pit obtained by the weighted average can be too large or too small, it cannot have a sufficient impact on the data update, so after obtaining $\nu^{k,m}$, it should be normalized again following equation 4.4.

$$\nu^{k,m} = \frac{\exp((p_m)^T q^k)}{\sum_{k'=1}^K \exp((p_m)^T q^{k'})} \quad (4.4)$$

$$\hat{\nu}^{k,m} = \frac{\nu^{k,m}}{\max_{k \in U^m} \nu^{k,m}} \quad (4.5)$$

There is a problem here, since the initial memory modules are randomly generated, there is no guarantee of sufficient distance between memory entries. Therefore, this chapter adds a limit to the initial value of the randomly generated memory module R (Equation 4.6) to ensure that each entry is sufficiently independent.

$$R = \|CC^T - I\|_F^2 \quad (4.6)$$

where I is an identity matrix, and $\|\cdot\|$ is the Frobenius norm of the matrix. This function is used to limit the initially generated memory modules to ensure that there is enough distance between different memory entries to distinguish them and prevent confusion.

This chapter proposes another explanation scheme for the above memory module mechanism. In the process of multiple iterations, similar query points are continuously weighted and averaged to the nearest memory entry. This chapter proposes another way of thinking, that is, the memory entry corresponds to the clustering center of each normal event, the iterative process is a continuous clustering process, and its processing method is equivalent to k-means clustering. In the process of exploration, this chapter tried to add clustering loss to the iterative process of CNN, but did not achieve good results. We are still exploring this.

4.3.3 Coordination Attention

The attention mechanism emerged as an improvement over the encoder decoder-based neural machine translation system. Since video processing applications have no limitation on the length of the input and output sequences and need to allocate more computing resources, encoder decoder-based attention mechanisms are widely used [139] [140]. Traditional channel attention allows neural networks to learn what should be focused on during the learning by allowing the network to iteratively focus on the attention of its filters. These channel attentions generally transforms the feature tensor into a single feature vector through 2D global pooling. General self-attention based algorithms often use attention pooling to encode global spatial information, but compressing the spatial information into one channel interpreter loses many features and it is difficult to preserve the spatial information. As it is important to preserve video features during the long-term interactions, it is required to improve the accuracy of visual tasks. In addition, the attention module needs

to acquire more precise spatial information, and these precise spatial information can help to capture the target of long-term interactions.

As channel attention mechanisms neglect positional information that helps to generate spatial information, we can embed coordinated attention mechanism to aggregate features along the spatial directions [141]. The coordinated attention mechanism consists of two steps:

- coordinate information embedding
- coordinate attention generation.

Figure 4.4 depicts the coordinate attention block that will be used to integrate with two steps encode channel correlations and long-term dependencies using precise location information.

Coordinate information embedding: Channel attention is established as two 1D feature encoding that aggregate these features along with two spatial directions. Therefore, long-term dependent features can be captured along one spatial direction, and precise location information is preserved along with the other.

Given an input X , use two pooling kernels $(H, 1)$ and $(1, W)$ to encode all channels along with the horizontal and vertical directions. The c -th channel information in the horizontal and vertical directions can be expressed as shown in Equation 4.7 and Equation 4.8.

$$z_c^h(h) = \frac{1}{W} \sum_{0 < i < W} x_c(h, i) \quad (4.7)$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 < i < H} x_c(i, w) \quad (4.8)$$

Coordinated attention generation: Coordinated attention generation follows three important steps for computer vision tasks:

- When designing the network structure, it should be designed as simple as possible and need to make sure that it does not utilize additional memory;
- The network should be able to understand the relationship between different channels, which is the key to the attention mechanism;
- According to the analysis and findings in this chapter, the network should have the ability to capture the region of interest (the most important region) in the video with precise location information.

Once the coordinated attention has generated features of the embedded video, the connection information is sent to a shared convolutional transformation function F_1 in Equation 4.9.

$$f = \gamma(F_1([z^h, z^w])) \quad (4.9)$$

[.] represents a connection operation along the third spatial dimension. The operation of splicing the weights in the w direction and the weights in the h direction into a weight matrix. The third spatial dimension generally refers to the dimension occupied by the channel. γ is a activate function. $f \in R^{C/r \times (H+w)}$, r is used to control the block size reduction ratio. Then, we divide f into $f^h \in R^{C/r \times H}$, $f^w \in R^{C/r \times w}$. There are additional convolutional transforms F_h and F_w that transform f^h and f^w respectively into tensors with the same number of channels as the input X , yielding.

$$g^h = \sigma(F_h(f^h)) \quad (4.10)$$

$$g^w = \sigma(F_w(f^w)) \quad (4.11)$$

g^h and g^w are expanded as weights after feature value update, The final eigenvector Y is represented as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (4.12)$$

This encoding process allows our coordinated attention to more accurately locate the location of the object of interest, thus contributing to the overall model for better recognition. The specific flow chart of the coordinated attention is shown in Figure 4.4 indicating how to integrate pooling, convolutions, and other required methods together.

4.3.4 Loss Function

We introduce the loss function of the model proposed in this chapter, and the evaluation algorithm for detection accuracy. At the same time, we propose a visual explanation scheme to advance the detection accuracy from the frame level to the pixel level. The interpretation scheme is more intuitive and appreciative.

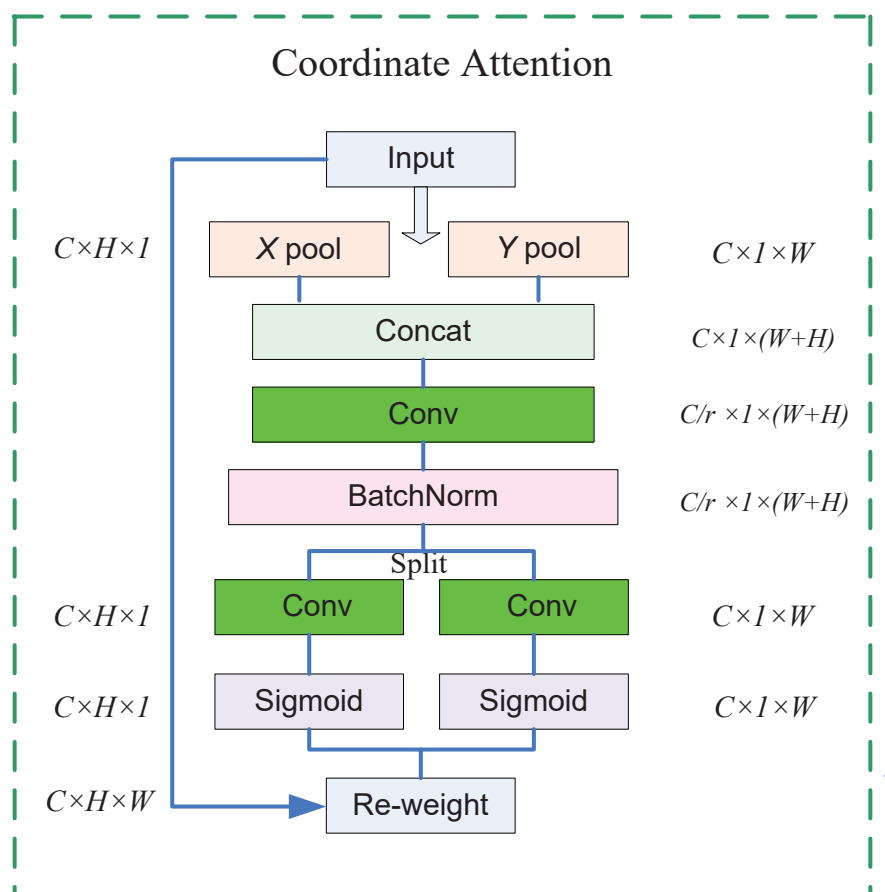


Figure 4.4: Coordinate Attention C is the number of channels; H, W represent the length and width of the current feature, respectively

4.3.4.1 Loss function

Following three main loss functions are used in this work:

- the prediction error(ς_{pred}),
- the L2 norm loss between the query point and its nearest memory entry(ς_{fit}), and
- the segmentation loss between the query point and the next closest memory entry(ς_{sp})

We can also use the similarity loss for randomly generated memory entries as shown Equation 4.6, but this is not the training loss as the training loss consists of ς_{pred} , ς_{fit} and ς_{sp} and evaluated based on the Equation 4.13. However, we are not using the similarity loss to evaluate our model performances in this work.

$$\varsigma_{Train} = \varsigma_{pred} + \lambda_f \varsigma_{fit} + \lambda_s \varsigma_{sp} \quad (4.13)$$

In prediction loss(ς_{pred}), we minimize the L2 distance between the future frames \hat{I} generated by the decoder and the true future frames I as shown in Equation 4.14.

$$\varsigma_{pred} = \sum_{w=1}^W \sum_{h=1}^H \|\hat{I}^{w \times h} - I^{w \times h}\|^2 \quad (4.14)$$

The feature fit loss(ς_{fit}) encourages queries to be closer to the nearest item in the memory, which is computed by the L2 norm between them. Following Equation 4.15 shows the feature fit loss(ς_{fit}), where $p_{q_t^k}$ is the memory entry closest to the query point q_t^k . This loss can also be considered as the clustering error.

$$\varsigma_{fit} = \sum_{k=1}^K \sum_{t=1}^T \|q_t^k - p_{q_t^k}\|^2 \quad (4.15)$$

To ensure that different memory entries still maintain a certain distance during the updating and training process, we introduce the term ς_{sp} to prevent different memory entries from being confused during training by penalizing the distance between the query feature and the next closest memory entry.

$$\varsigma_{sp} = - \sum_{k=1}^K \sum_{t=1}^T \|q_t^k - p_{se}\|^2 \quad (4.16)$$

where p_{se} is the second closest entry to the query point.

The three loss functions mentioned in Equation 4.14, 4.15 and 4.16 are considered together during the training phase and we use these metrics to evaluate the performance of the model. In comparison to the other two loss functions, the prediction loss is the most important loss function as the other two loss functions play a relatively small role and can be considered as secondary training loss functions. The λ_f and λ_s values used in Equation 4.13 usually ranges in between 0.1 and 0.01. The best fit values to our proposed models will be explored using different experiments and explain in the Section 4.4.

4.3.4.2 A Visual Evaluation

Since the detection of this model is only set at the frame level, it cannot label abnormal modules in video frames. This chapter proposes a visual anomaly interpretation module. The anomaly interpretation module consists of two parts, one is the feature error map, and the other is the object detector. The feature error map is matched to the output of the object detector, the largest one has the largest error and is specially labeled. When greater than the specified threshold, set as an exception.

4.4 Experimental Results

4.4.1 The proposed COVAD approach

The following step-by-step process provides detailed information on how our proposed anomaly detection approach, called COVAD, is implemented and evaluated.

1. As the first step, our algorithm randomly generates memory modules, and build $M \in R^{m \times c}$ matrix according to the number of videos and behavioral patterns, where R represents the Equation 4.6, m is the number of normal behavior patterns and c represents the number of features per channel. Initially, the value of m is set to 10.
2. Next, read the dataset and divide it into multiple consecutive T frames. The first $t - 1$ frame assigns as the input to the encoder network, and use convolution and pooling to scale down the extracted features to make a $32 * 32 * 512$ feature space.
3. Following that, input the extracted features into the collaborative attention mechanism and re-allocate weights to obtain new video features.
4. Randomly generated memory module calculates its similarity, and updates the memory module according to the method explained in Section 3.2. It also aggregates the

		λ_s				
λ_f	Value	0.02	0.04	0.06	0.08	0.1
	0.02	89.3	95.1	94.6	92.6	94.2
	0.04	92.5	95.3	94.9	90.4	88.2
	0.06	90.9	95.4	91.2	94.5	89
	0.08	93.1	83.2	84.6	91.2	89.3
	0.1	96.8	93.7	92.8	95.4	96.2

Table 4.1: The accuracy of anomaly detection under different value of hyperparameters λ_f , λ_s ; the dataset used is UCSD(Ped2).

memory features and query features as the hyperparameters of the loss function. We conducted a set of experiments to identify the most suitable hyperparameter values as explained in Section 4.4.3.

5. Input the obtained aggregated features into the decoder network to restore the frames.
6. Next, calculate the error between the restored video frame \hat{t} and the real t frame.
7. Uses backpropagation to update the network parameters until it minimizes the error.
8. Finally, classify the given input once the model converged to the minimum error point.

4.4.2 Dataset Description

The analyses in this work are mainly based on two different datasets: UCSD [5] and Avenue [6].

The UCSD dataset is a campus pedestrian dataset released by the University of California, San Diego in 2013, which contains two subsets called Ped1 and UCSD(Ped2). The number of training videos sets used in Ped1 and UCSD(Ped2) are 34 videos and 16 videos, respectively and this training set contains only normal frames. The test set contained both normal frames and exception frames and has 36 videos in Ped1 and 12 videos in UCSD(Ped2). Frame-level annotations are provided for all test video clips and 10 of which have pixel-level ground truth. In this research, our analyses are mainly based on UCSD(Ped2)

The Avenue is a dataset released by the Chinese University of Hong Kong in 2013, which contains 15 videos of 2 minutes each. The total number of frames is 35240 and 8478 frames from 4 videos can be used as the training set. These videos contain typical unusual events including running and throwing objects.

4.4.3 Hyperparameter selection process

We conducted several experiments to select the best values for the hyperparameter λ_f and λ_s that is used in Equation 4.13 for calculating the total loss. To verify the effectiveness of these hyperparameters with different values in our analysis, we used UCSD (UCSD(Ped2)) dataset to verify the anomaly detection accuracy when λ_s and λ_c parameters assign 0.02, 0.04, 0.06, 0.08, and 0.1 values separately for different iterations. The accuracy of the COVAD model for different experiments are shown in Table 4.1. The accuracy does not show obvious regularity, but λ_c has a great influence on the detection results. When the value of λ_c is 0.1, the experimental effect is relatively stable, and the detection accuracy is basically the highest value. Therefore, in this chapter, we set $\lambda_c=0.1$.

Table 4.1 shows the detection results under different hyperparameter values for λ_f and λ_s . The detection accuracy does not show a clear Gaussian distribution after fixing the value of one hyperparameter. The main reason behind this result may be that the relationship between the three different loss functions in equation 4.13 is nonlinear or the amount of training data is insufficient. In future, we will explore the exact reason behind this and will propose new methods to overcome this issue. Based on the results shown in Table 4.1, the two highest accuracy(average) are 96.8 and 96.2 that are obtained for different hyperparameter values. The highest accuracy is obtained when $\lambda_f=0.1$ and $\lambda_s=0.1$ or 0.02. According to the empirical values in the previous chapters [23], both these hyperparameter values are set to 0.1 and therefore, in or experiments we set them to 0.1.

4.4.4 Effectiveness of the attention mechanism in video anomaly detection

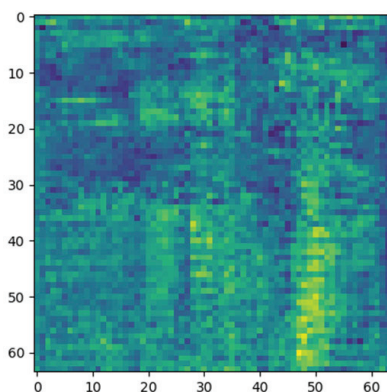
In this section, we analyze the outcome of the COVAD model to explore whether the attention mechanism improves the accuracy of video anomaly detection tasks. To verify this, we randomly selected an image (Figure 4.5(a)) from the Avenue dataset and extracted its features before and after adding coordinated attention using our COVAD model and MNAD [23] model.

The comparison process has been done using the following steps. 1. MNAD and COVAD networks are trained separately. The MNAD network is implemented without using the attention mechanism, and the COVAD is implemented using coordinated attention. 2. First, randomly select a video frame and then, input it into the above-mentioned two trained networks. 3. Generate the feature map at the end of the encoder and observe the difference between the outputs of the two networks.

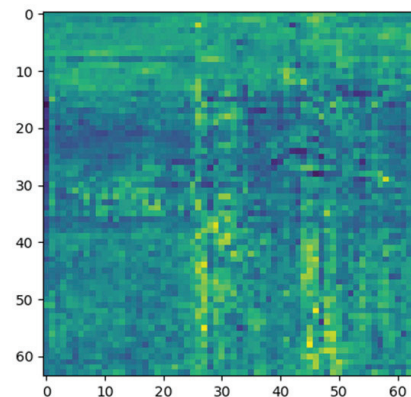
Figure 4.5 depicts how the weight redistribution of the coordinated self-attention mechanism helps the neural network to focus on meaningful targets having the effect of anti-noise, and how it helps to improve the detection efficiency.



a



b



c

Figure 4.5: The *a* is a frame in the video, *b* is the feature map generated without a coordinated attention mechanism, and *c* is the feature map generated by the coordinated attention mechanism.

Figure 4.5(b) clearly indicates that without the coordinated attention mechanism the upper part of the video frame is relatively dark. As a result, after reading this video frame, the RGB value of this area gets relatively large and hence, this will affect on the neural network training and model performances. Since this dark area, which we called as background is not important in the classification, training the neural network model using this types of frames are the best practice and it consume lots of resources as the model gets larger.

Once we apply the trained coordinated attention, we can clearly observe that the object distribution in Figure 4.5(a) is more visible on the feature map shown Figure 4.5(c) based on its the contrast and dark colors, compared with the feature map shown in Figure 4.5(b). This indicates that the model has successfully captured relatively important regions in the video based on previously trained data. Therefore, network parameters used in our chapter are more reasonable and help to obtain more effective features and more realistic video frames.

4.4.5 A Visual Test

First, since the results of unsupervised anomaly detection are still at the frame level, we propose object-oriented anomaly testing. First, the error between the predicted video frame and the real video frame is calculated to obtain the feature error map, and then the real abnormal video frame is detected through the target detector to obtain each target in the video frame. Calculate the average error within each object box. Determine a reasonable threshold through multiple tests. Object boxes whose average error is greater than the threshold are marked in red.

Figure 4.6 shows the result. In the specific implementation process, the feature error map is calculated by the feature subtraction of the restored video frame and the directly read real video frame. The object detector is implemented by retinanet single-stage object detector [101], and the network used is resnet50, and the performance is sufficient. The white bright spots in Figure 4.6.a represent areas with large errors, black represent areas with small errors, and Figure 4.6.b is the result of the object detector. The object detector is only responsible for object detection and abnormal target judgment. By calculating the object frame, the mean error within the realization.

4.4.6 Video anomaly detection using COVAD

The following section explains the different analyses we conducted on video anomaly detection using the above-mentioned datasets and the most suitable hyperparameter values.

The testing environment used in this experiments is Tesla V100 Volta P100 GPU Accelerator with a 32GB Graphics Card and few models are executed at the Google Colab.

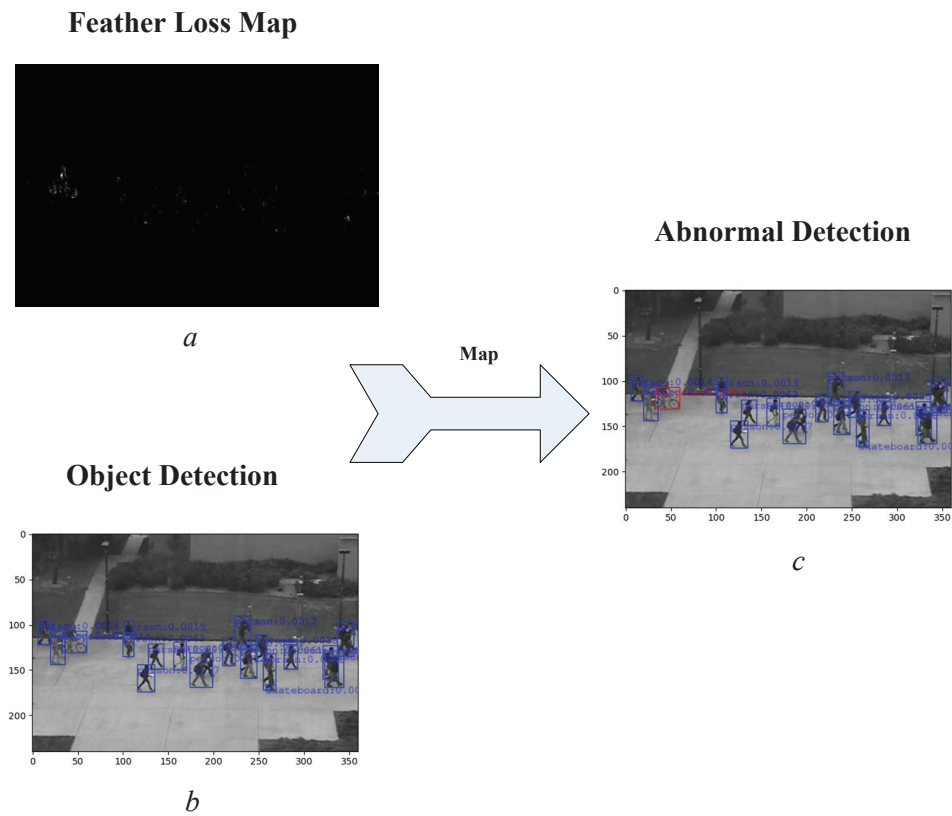


Figure 4.6: The *a* is a feather loss map, *b* is the result of object detection, and *c* is the generated result by feather loss map and objected box map .

Method	UCSD (UCSD(Ped2))	Avenue	Techniques
AMDN [38]	90.8%		DFE+SVM
Unmasking [142]	82.2%	80.6%	Unmask
StackRNN [84]	92.2%	81.7%	TSC+sRNN
MemAE [138]	91.7%	81.0%	Memory module
MNAD [23]	94.2%	80.6%	U-net
COVAD	96.5%	83.4%	CA

Table 4.2: Quantitative comparison of the frame-level AUC-PR results of our COVAD method with the state-of-the-art models. (DFE is Double Fusion Framework; Unmask is a technique previously used for authorship verification in text documents ; TSC+sRNN is Temporally-coherent Sparse Coding stacked Recurrent Neural Network; CA is the Coordinated attention).

Since we tested a large number of hyperparameters, the part of the validation experiments were run in Colab. Compared with previous networks, the network in this chapter has good time efficiency when running tests. During the testing, the COVAD model process 28 frames per second.

Table 4.2 shows the quantitative comparison results of our COVAD method and the state-of-the-art methods on frame-level AUC-PR results. Based on the obtained results, the COVAD method can effectively improve the accuracy of anomaly detection compared to other models. We can find that the COVAD method obtained the highest AUC value for both UCSD and Avenue datasets. Another most important finding in this work is on the reduction model convergence time, which is mainly due to the integration of the attention mechanisms in our COVAD approach. Since the self-attention mechanism is lightweight and mobile-level, it does not take a lot of time during training and testing. Therefore, compared with previous methods, our method is worth generalizing.

4.5 Conclusion and Future work

With the improvement of computer hardware and network bandwidth, video will definitely become the main medium for transmitting information in the future and this is one main reason to attract many researchers attract towards computer vision today. The two most popular research domains in Artificial Intelligence are computer vision and natural language processing. In the foreseeable future, computer smell and touch may become new research hot-spots. In this chapter our main focus is to detect anomalies in surveillance videos that are deployed in different locations, such as highways, schools, prisons, etc. The manual inspection of video anomaly detection in real time is not very efficient due to the discon-

tinuity of human eye monitoring over the time. The algorithm proposed in this chapter incorporates a coordinated self-attention mechanism to help the neural network to focus on meaningful objects during training by ignoring the background in the video. Based on the experimental results, our proposed algorithm can avoid the detection efficiency of unimportant background noise, that is, the algorithm in this chapter has a strong anti-noise ability. Many unsupervised video anomaly detection approaches proposed in the literature have used frame-level objective function as the training loss function, and then detect the abnormal area through the splicing Object detection algorithm. This approach seems to achieve pixel-level video anomaly detection, but this is difficult to achieve in the the actual deployment process. Compared with video anomaly detection, the network structure of video Object detection is more complex, and it is difficult to establish a joint algorithm framework to connect the two neural networks. Therefore, the best solution is to establish an anomaly detection mechanism centered on Object and Behaviors.

The direct detection of abnormal regions in the real-time video is one of our ultimate goals related to this reach. In the future, we aim to implement an unsupervised video anomaly detection network that can be jointly trained with the pixel-level object detection network. The purpose of detecting video anomalies is to solve the issues that occur in real-time, that is, to eliminate disasters that have not yet occurred. Therefore, the response mechanism in the actual deployment stage is also worthy of our consideration in future.

Consistency-constrained unsupervised video anomaly detection framework based on Co-teaching

Contents

5.1	Overview	88
5.2	Introduction	88
5.3	Duel Channel model	90
5.4	Methodology	91
5.4.1	Preliminary	92
5.4.2	Consistency-constrained Framework Based on Co-teaching	93
5.4.3	Co-teaching within Memory Module	95
5.4.4	Anomaly detection stage	100
5.4.5	Experiment 1	101
5.4.6	Experiment 2	103
5.4.7	Experiment 3	106
5.5	Conclusion and Future Work	106

5.1 Overview

This chapter introduces a novel anomaly detection framework that balance dynamic information with static information and construct a relationship between appearance features and corresponding optical flow features, where we set strong consistency constraints, which reduces the loss between dynamic information and corresponding static information, and leverages collaborative teaching network to ensure a consistent representation of both static and dynamic information for predict. The proposed framework consists of two sets of encoder-decoder pairs complemented by a memory storage module. Operating in parallel with the dual encoder network is a Co-teaching network, with the shared memory module serving as the cornerstone for collaborative training. The Consistency constrained condition guarantees the strong consistency of dynamic and static information in the learned representations. In our experimental phase, we present compelling results that showcase the superior performance of our algorithm across three publicly available datasets.

5.2 Introduction

Before the emergence of deep learning, traditional video analysis technologies primarily consisted of methods such as the frame difference method [143], color histogram [144], and HOG feature [145]. These video analysis techniques transform original video data into interpretable feature signals, aiding researchers in more effectively analyzing video data. With the advent of deep learning, video anomaly detection technology based on neural network learning can be categorized into two main groups: unsupervised learning of anomaly detection and weakly-supervised learning of anomaly detection [146] [12, 147].

The academic community has been diligently working to amalgamate the potent advantages of weakly supervised video anomaly detection with the generalization benefits of unsupervised algorithms. For instance, Wang [55] introduced a novel and robust unsupervised video anomaly detection method that incorporates a frame prediction scheme tailored for surveillance videos. Their approach employs a multipath ConvGRU-based frame prediction network, which adeptly handles semantically rich objects and regions at various scales while capturing spatiotemporal dependencies in normal videos. This algorithm enhances the representation of spatiotemporal features in unsupervised algorithms, thereby enhancing their robustness.

Similarly, Huang et al, [54],introduced the appearance-motion semantic consistency framework, which exploits the difference in appearance and motion semantic representation between normal data and abnormal data. They first designed a two-stream structure to encode the appearance and motion information representation of normal samples, and then proposed a novel consistency loss algorithm to enhance the consistency of feature semantics,

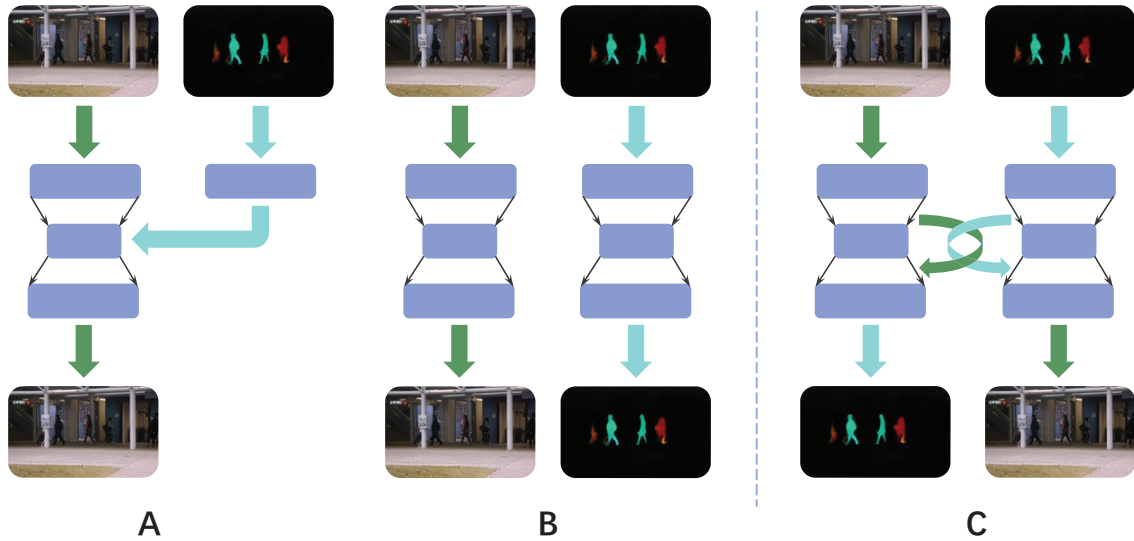


Figure 5.1: Comparison of methods: **A** uses optical flow features as a supplement to video frame appearance features to improve prediction accuracy; **B** uses parallel prediction of appearance features and optical flow features to build a joint prediction loss error; **C** is the proposed strong consistency collaborative training framework.

enabling the identification of low-consistency anomalies. This algorithm further enhances the consistent representation of dynamic and static features in unsupervised algorithms.

The most advanced semantically consistent model of appearance-motion features is the dual-channel framework proposed in 2022 [32], which proposes a spatiotemporal memory-enhanced dual-stream autoencoder framework and designs two identical and independent proxy tasks to train the dual-stream autoencoder. The structure extracts appearance and motion features separately and decodes them separately. Finally, the optical flow loss and appearance feature loss are calculated to explore the correlation between appearance and motion semantics. In this model, the only consistency constraint is the loss function, but two separate encoding-decoding processes cannot really constrain the consistency of motion features and appearance features [44, 56].

Considering the above-mentioned works, this contribution proposes a novel unsupervised learning video anomaly framework CCC-T (Consistency-constrained Framework Based on Co-teaching) as shown in Figure 5.1.C, which emphasizes the consistent representation of dynamic information and static information by utilizing carefully designed Strong consistency constraints. In this framework, dynamic information (optical flow features) and static information (appearance features) are regarded as equally important input data. The framework designed in this chapter mainly contains three parts: two sets of encoding

and decoding network structures and memory storage modules. There are two encoding and decoding structures. One is responsible for encoding the appearance features of the video frame as input, and then updating the input features in the memory module, while its decoder outputs the optical flow features corresponding to the video frame. The other encoder is responsible for encoding the optical flow of the video frame, which is used as the input feature; that input feature is updated, and finally the decoder outputs the appearance feature corresponding to the video frame. The memory storage module stores the normal pattern and updates the passing characteristics. To ensure the accuracy of optical flow features in predicting appearance features, the missing background and color information is compensated. The framework utilizes skip connections to connect the encoding layer (appearance features predict optical flow features) and the decoding layer (optical flow features predict appearance features) and reads map features from each layer as a complement. The three modules in the framework are connected through a collaborative teaching network to promote collaborative learning.

To summarize, this section makes the following three contributions

- Proposes a novel unsupervised video anomaly detection framework built using co-teaching networks;
- Achieves the first one-time collaborative training of optical flow and representational features in unsupervised video anomaly detection; and
- After testing on three datasets, the proposed model further improves the accuracy of unsupervised video anomaly detection algorithms.

5.3 Duel Channel model

Video anomaly detection algorithms within an unsupervised learning framework always focus on a single goal: improving prediction or reconstruction accuracy by extracting more precise video features. Many unsupervised methods are all dedicated to utilizing sub-tasks [57–59, 148], including identifying the order or reverse order of the sequence to extract features, thereby enhancing the extraction of dynamic features and static features. However, for video data, multi-tasking only guarantees the accuracy of extracting dynamic features and static features, it cannot constrain the consistency of dynamic features and static features.

The dual-channel unsupervised model [32, 54, 55] is a new attempt to address these issues. Differing from the framework described above, the dual-channel model attempts to directly extract dynamic features as a supplement to static features, and builds a dynamic feature- static feature constraint framework to enhance the integrity of the input features

to improve the accuracy of prediction/reconstruction. However, the existing dual-channel model, as shown in Figure 5.1.A,5.1.B, only uses dynamic features as a supplement to static features, which enhances the accuracy of input features, but does not set consistency constraints. Framework C, on the other hand, designs a completely parallel encoding-decoding structure and relies on interactive loss functions to constrain consistency. This constraint cannot affect the features extracted by the encoder, and the channels are relatively independent, that is, the processing of dynamic features and the processing of static features are independent and cannot act as a real consistency constraint on the extracted features. In addition, while mainstream methods use dynamic features as supplementary elements to enhance the representation capabilities of static features, they cannot achieve simultaneous learning of spatio-temporal features.

To solve this problem, this chapter introduces a new dual-channel video anomaly framework to enhance the detection capabilities of unsupervised learning algorithms. This framework treats dynamic information and static information as inputs of equal importance and carefully designs strong consistency constraints between dynamic information and static information to ensure consistent representation of optical flow features and appearance features, and it builds a collaborative learning and memory storage module based on co-teaching. The core of this study is collaborative learning, memory storage modules, and skip connections and other technical means, which strictly follow the consistency constraints of dynamic features and static features.

5.4 Methodology

This section provides a detailed explanation of our proposed unsupervised learning framework and the models utilized in our experiments. This includes explaining how the co-teaching architecture works in the training process of two encoder-decoder networks.

In this part, the FlowNet2 network [83] is responsible for extracting optical flow features from video frames. Subsequently, these features from video appearances and optical flow are used as the input into two encoder networks. These features are then compressed, followed by their entry into the memory storage module to update the corresponding elements of video appearances features and optical flow features.

The mechanism entails retrieving the features of the nearest counterpart and aggregate them into novel features. Finally, the amalgamated new features feed into the two decoder networks to predict the features of the opposing entity. For example, the optical flow features serve as the input to the encoder-decoder, resulting in the output of video frame features. Conversely, when the input is the video appearances feature, the output manifests as the optical flow feature. To address the potential information gap in video appearances

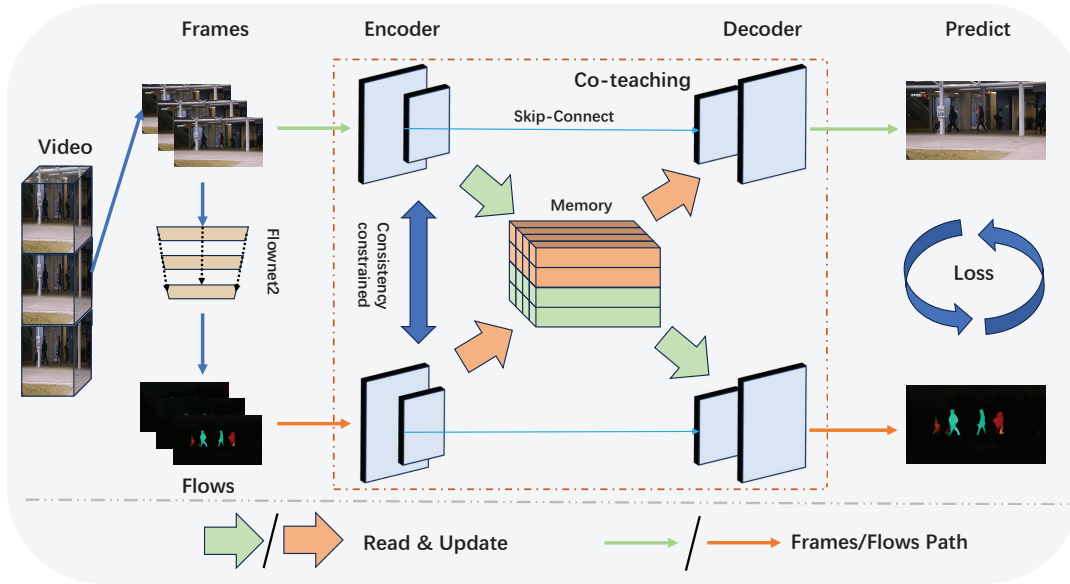


Figure 5.2: A detailed framework of CCC-T. The first step uses FlowNet2 to obtain the optical flow information of the video sequence. The second step inputs the segmented video frames and optical flow information into their respective encoding networks. The third step is to cross-read and collaborate the output of the encoding network with the memory module. Update, the fourth step, the updated input features are input to the decoder network for cross prediction.

feature prediction by optical flow features, this study integrates skip connections [135,149] that bridge the encoding map of video frames to the optical flow decoder (predictive video frames).

The loss function is comprised of the prediction loss inherent in the video appearances features and the optical flow features prediction, as well as the similarity loss in memory modules. The proposed model greatly ensures the consistent description of optical flow features and appearance features through shared memory entries.

5.4.1 Preliminary

The fundamental algorithms highlighted in this chapter contain FlowNet2, the encoder-decoder structure, the memory module, and the co-teaching framework. Notably, The encoder-decoder structure and memory module already well described in previous paper [23,138]. Consequently, the ensuing content will provide a succinct overview of FlowNet2, outlining its objectives and structural attributes, followed by an outline of the co-teaching architecture.

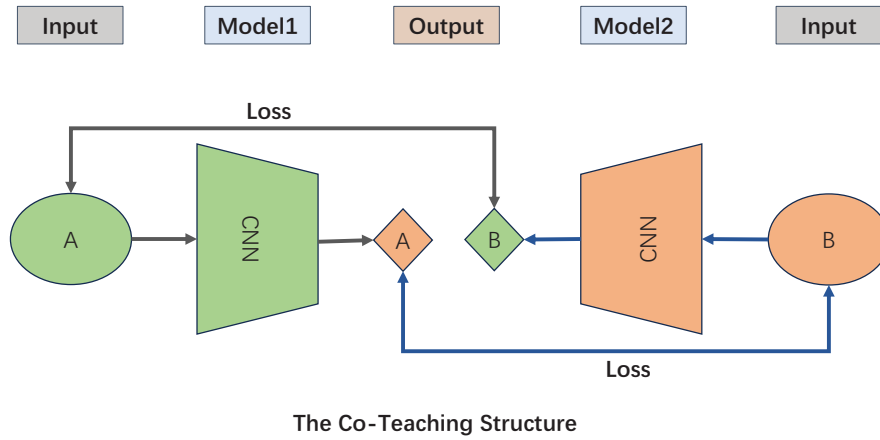


Figure 5.3: A Classify Co-teaching Structure

Co-Teaching [150]: A collaborative teaching network is a framework in which multiple neural network models collaborate to solve specific problems or achieve a common goal, as shown in Figure 5.3. For example, multiple actors merge their predictions through techniques such as voting, averaging, or weighted averaging. Classic co-teaching networks are one of the following four types: 1. Knowledge distillation [151]: A broader or more complex model (teacher model) is trained together with a smaller or simpler model (student model). The student model learns to imitate the behavior of the teacher model, reduce parameters, and/or to build multi-task models; 2. Collaborative training [152]: Multiple models are trained simultaneously and exchange training data or gradients during the optimization process; 3. An Adversarial Network [153]: Multiple models with complementary effects, such as a generator network and a discriminator network in a generative adversarial network (GAN) work together to achieve a specific result; and 4. Federated learning [154]: Many models are trained on different data subsets and then merged or averaged to generate a global model. This approach can improve privacy and data distribution issues. In this chapter, we adopt two encoder-decoder structures to share the memory module, cross-read the video frame feature pool and the optical flow feature pool, and to promote the collaborative training of the model. These two encoding structures are similar to two teacher networks, learning from each other.

5.4.2 Consistency-constrained Framework Based on Co-teaching

This section introduces the operation and interaction of each module of the framework (CCC-T: Consistency-constrained Framework Based on Co-teaching) proposed by this

chapter in detail. This CCC-T employs two interconnected encoder-decoder structures by the co-teaching network. These structures are designed to encode optical flow and video frame features separately while predicting the corresponding features of the opposite type (i.e., optical flow to video frame and vice versa). The predicted loss resulting from these predictions is then utilized to update the model. The following section outlines the detailed steps involved in the comprehensive formalization.

Formalization: There is an existing video denoted as V , which is divided into a sequence of continuous video frames: $V = v_1, v_2, v_3, \dots, v_N$, where N represents the total number of frames in the video. The optical flow features of these video frames are extracted using Flownet2, denoted as $F_{flows} = Flownet2(V)$, with individual flow features represented as $f_{flow} \in f_{flows_1}, f_{flows_2}, f_{flows_3}, \dots, f_{flows_N}$. The read library of OpenCV2 is employed to directly extract frame features from the video frames, yielding $F_{frames} = Ir(V)$, with frame features represented as $f_{frames} \in f_{frames_1}, f_{frames_2}, f_{frames_3}, \dots, f_{frames_N}$.

As stated earlier, this chapter presents a model encompasses two encoder-decoder structures, as illustrated in Figure 5.2. where ψ represent the Encoder function and ϕ the Decoder, The upper structure is the video appearances feature encoder ψ_{frames} , while the lower one is the optical flow feature encoder, referred to as ψ_{flows} . The decoder positions are the opposite: the upper one is ϕ_{flows} , and the lower one is ϕ_{frames} .

During the training phase, the extracted video frame features F_{frames} are input into the ψ_{frames} to focus and refine the quality of the appearance feature representation. Subsequently, these features are passed through a memory module. The error is calculated with the nearest video frame feature entry, leading to an update of the video frame feature storage module. Simultaneously, the module queries the optical flow entry that is closest to the input feature and then reads and updates the input feature. The updated input feature is then fed into the ϕ_{frames} to predict the optical flow feature. This process can be expressed in an equation as:

$$\begin{aligned} F_{frames}^E &= \psi_{frames}(f_{frames}) \\ &= \psi_{frames}(Ir(V)), \\ V &= v_1, v_2, v_3, \dots, v_N \end{aligned} \tag{5.1}$$

$$\hat{F}_{flows} = \phi_{frames}(\theta(F_{frames}^E, M)) \tag{5.2}$$

where, M signifies the memory storage module, and θ embodies the interaction between input data and the memory storage module, encompassing functions such as reading, updating, and the integration of novel features. Comprehensive insights into the memory storage module are expounded upon in Section 5.3.2. F_{frames}^E denotes the features ema-

nating from the encoder, while \hat{F}_{flows} encapsulates the optical flow features prognosticated by the decoder.

Conversely, the optical flow features F_{flows} , obtained from Flownet2, are input into the ψ_{flows} . This step help to refine the high-quality optical flow feature representation. These features are then processed through a memory module. Similar to the video frame features, the error is computed with the nearest optical flow feature entry, resulting in an update of the optical flow feature storage module. Furthermore, the module queries the appearance feature entry closest to the input feature, reading and updating the input feature. The updated input feature is directed into the ϕ_{flows} to predict the appearance feature.

$$\begin{aligned} F_{encoder}^{flows} &= \psi_{flows}(f_{flows}) \\ &= \psi_{flows}(F_{ownets2}(V)), \\ V &= v_1, v_2, v_3, \dots, v_N \end{aligned} \quad (5.3)$$

$$\hat{F}_{frames} = \phi_{flows}(\theta(F_{frames}^E, M)) \quad (5.4)$$

The loss function contains three components: optical flow prediction loss, appearance feature prediction loss, and memory storage module loss L_M . Optical flow prediction loss, constructed from the difference between the decoder output of appearance features \hat{f}_{flows} and the true optical flow features f_{flows} . Appearance prediction loss, constructed from the difference between the decoder output with optical flow features \hat{f}_{frames} and the real appearance features f_{frames} . Similarly, during the test phase, the anomaly score is composed of three parts.

$$loss = \begin{cases} \left\| \hat{f}_{flows} - f_{flows} \right\| \\ \left\| \hat{f}_{frames} - f_{frames} \right\| \\ L_M \end{cases} \quad (5.5)$$

The role of this module is to align input data with their corresponding entries in the memory module, thereby capturing and recording trained normal patterns. The memory module loss L_M is described in detail next.

5.4.3 Co-teaching within Memory Module

The co-teaching structure in training is designed in the memory storage module.

Following the blueprint of the conventional memory storage module [155], this unit serves two primary functions. The first involves reading, wherein the module identifies and retrieves the entry most closely aligned with the input feature, subsequently updating

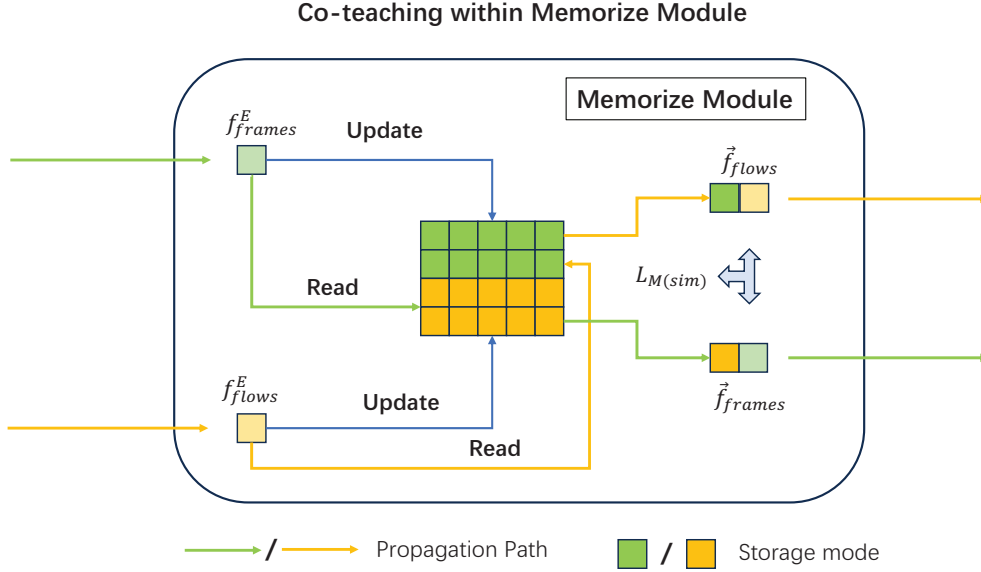


Figure 5.4: Co-teaching within a memory module: Green indicates the transfer of static features in the memory module, and orange represents the transfer of dynamic features; correspondingly, the memory module is composed of multiple static feature category entries and multiple dynamic feature category entries

the input feature. The second function entails updating, which transpires as an ongoing process throughout the training. The memory matrix continuously evolves based on the proximity between feature maps, effectively consolidating data from the training set that corresponds to the set entries.

In our framework, the memory storage module is bifurcated into two distinct components. The green segment denotes the video appearance feature memory mode, while the orange segment signifies the optical flow feature memory mode. The act of reading and updating each input datum transpires in disparate sections of the memory module, so that the update operation takes place within the respective memory mode, and the reading operation unfolds in the complementary memory mode. These modes are illustrated in Figure 5.4.

5.4.3.1 Reading and Updating Mechanisms of the Memory Module

Reading Mechanism: Reading Mechanism for the Memory Module as shown in Figure 5.4: The input to the memory module is the optical flow feature. This involves calculating the cosine similarity between the query feature and the entries within the video frame appearance feature memory mode. The aim here is to identify the entry or multiple entries

with the closest proximity to the query feature, thus determining their respective distances. The softmax function is applied to establish an average probability match. Subsequently, the probability value is utilized to compute the inner product with the appearance feature entry from the memory module. This process leads to the feature update. Finally, the updated features are merged with the original query features to predict the corresponding video appearance features.

First, the cross-cosine similarity between each entry $q_{flows}^k, q_{frames}^k, q_{flows}^k \in f_{flows}^E, q_{frames}^k \in f_{frames}^E$ and memory items $p_m^{frames}, p_m^{flows}$ is calculated, where q_{flows}^k and q_{frames}^k are from the corresponding two query encoding features $f_{frames}^E, f_{flows}^E$; $p_m^{frames}, p_m^{flows}$ is set during initialization, and two two-dimensional correlation maps of size $M \times K$ are obtained. The softmax function along the vertical direction and obtain the matching probabilities $w_{frames}^{k,m}, w_{flows}^{k,m}$ as follows:

$$w_{k,m}^{frames} = \frac{\exp((p_m^{flows})^T)q_k^{frames}}{\sum_{m'}^M \exp((p_{m'}^{flows}))q_k^{frames}} \quad (5.6)$$

$$w_{k,m}^{flows} = \frac{\exp((p_m^{frames})^T)q_k^{flows}}{\sum_{m'}^M \exp((p_{m'}^{frames}))q_k^{flows}} \quad (5.7)$$

For the query items of optical flow features q_{flows}^k and appearance features q_{frames}^k , the opposite memory module is read through the calculated weight ($q_{flows}^k \rightarrow p_{m'}^{frames}, q_{frames}^k \rightarrow p_{m'}^{flows}$), which obtains the desired cross prediction information. The reading process is as follows:

$$\hat{p}_k^{frames} = \sum_{m'}^M w_{k,m'}^{frames} p_{m'}^{flows} \quad (5.8)$$

$$\hat{p}_k^{flows} = \sum_{m'}^M w_{k,m'}^{flows} p_{m'}^{frames} \quad (5.9)$$

After reading the memory module, the closest cross feature map $\hat{p}_k^{flows}, \hat{p}_k^{frames}$ is obtained, We concatenate $\hat{p}_k^{flows}, \hat{p}_k^{frames}$ with the query map $q_{flows}^k, q_{frames}^k$ along the channel dimension, and send $\vec{f}_{frames}, \vec{f}_{flows}$ into the corresponding decoder.

$$\begin{cases} \vec{f}_{frames} = \sum(\hat{p}_k^{flows}, q_{frames}^k) \\ \vec{f}_{flows} = \sum(\hat{p}_k^{frames}, q_{flows}^k) \end{cases} \quad (5.10)$$

Updating mechanism: Update mechanism of memory modules: In this case, the cosine similarity of appearance encoding features f_{frames}^E and optical flow encoding features f_{flows}^E with the corresponding memory modules is calculated. Next the probability value is calculated through the softmax function. Then read the compared memory entry by calculating the probability value, The next steps involve using the query features to increase the inner product of the obtained probability values. This sum is added to the corresponding memory entry of the original appearance feature. As a result of these operations, the memory module is effectively updated. The function of this step is to find the memory feature that is most similar to the query feature, and through its similarity loss, continuously improve the adhesion of the memory entry to the real normal pattern.

The first step is to calculate the cosine similarity between the optical flow query encoding features f_{frames}^E and the optical flow memory entries f_{flows}^E , and the appearance query encoding features and the appearance memory entries. This process is the opposite of the reading mechanism.

$$u_{k,m}^{frames} = \frac{\exp((p_m^{frames})^T q_{frames}^k)}{\sum_{k'}^K \exp((p_m^{frames})^T q_{frames}^{k'})} \quad (5.11)$$

$$u_{k,m}^{flows} = \frac{\exp((p_m^{flows})^T q_{flows}^k)}{\sum_{k'}^K \exp((p_m^{flows})^T q_{flows}^{k'})} \quad (5.12)$$

After obtaining the cosine similarity between the memory entry and the query point $u_{k,m}^{frames}$, $u_{k,m}^{flows}$, use the probability value cosine similarity to read the query entry, accumulate it with the original memory entry in the same channel, and update the memory entry. The updated storage module $\langle \tilde{p}_{flows}^m, \tilde{p}_{frames}^m \rangle$ is shown in the Figure 5.4:

$$\tilde{p}_{flows}^m = \sum (p_{flows}^m + \sum_{k=1}^K u_{k,m}^{flows} q_{flows}^k) \quad (5.13)$$

$$\tilde{p}_{frames}^m = \sum (p_{frames}^m + \sum_{k=1}^K u_{k,m}^{frames} q_{frames}^k) \quad (5.14)$$

Different from cross-reading, the memory update corresponds from optical flow to optical flow and appearance to appearance. By calculating the similarity matrix, it is accumulated to the memory module.

5.4.3.2 Strong consistency constraints

The strong consistency constraints in this chapter are mainly implemented through the reading and updating of the memory module. The reading and updating rules have been described in detail above. This chapter proposes that the memory module of the model is divided into two parts, one is optical flow feature storage, and the other is appearance feature storage; and the reading and updating of each branch are implemented for different parts of the memory module. That is, the prediction in this chapter is cross prediction, inputting optical flow features, then constructing a similarity matrix, and reading the most similar appearance feature storage entries to construct joint features to predict appearance features. When updating, only the optical flow memory entries corresponding to the optical flow features are updated. Optical flow features and appearance features are cross-read and updated. Through the loss function, appearance features and optical flow features are strengthened, and the consistent representation of dynamic information and static information is enhanced. And according to the consistent description of optical flow features and appearance features, the similarity between appearance and optical flow of videos of the same category is the highest. Based on this, this chapter sets up the consistent description probability for dynamic information and static information as $Cst_{(S,D)}$:

$$Cst_{(S,D)} = \sum_{i=1}^k \sum_{j=1}^k \langle f_{frames}^i, f_{flows}^j \rangle \quad (5.15)$$

where f_{frames}^i and f_{flows}^j are the encoded appearance features and optical flow features, and k is the number of storage entries designed by the memory module. Only when $i = j$, the consistency probability can reach the maximum value.

Therefore, the **strong consistency constraints** set as follows:

$$\begin{aligned} STC &= \sum_{i=j}^k |f_{frames}^i, f_{flows}^j| \\ &= \sum_{Max(Cst)}^k (|f_{frames}^i, M_{flows}^j| \oplus |f_{flows}^i, M_{frames}^j|) \end{aligned} \quad (5.16)$$

During the memory module reading process, this chapter sets up to read the cross feature entries that are closest to the query entries and predict the corresponding cross features $Max(Cst)$. We calculate the similarity by covariance and retrieve the intersection entries with the highest similarity. Final predicted cross correspond features (From appearances to optical flow, from optical flow to appearances).

5.4.3.3 Loss function Memorize Module

The loss function of the training process mainly consists of three parts, namely L_{flows} , L_{frames} , L_M . Among them, L_{flows} represents the error between the predicted optical flow and the real optical flow, and L_{frames} represents the error between the appearance characteristics of the predicted video frame and the real video frame. These designations are employed for partitioning the distances between distinct entries within the memory module. Here L_M loss is divided into two parts, namely Strong consistency constraint loss $L_{M(Sim)}$ and segmentation loss $L_{M(SEG)}$. $L_{M(Sim)}$ is achieved by enhancing the similarity between the optical flow features in the query entry and the most approximate flows features in the memory entry, and at the same time enhancing the appearance features in the query. The similarity between the feature and the closest optical flow feature in memory is used to ensure the consistency of optical flow features and appearance features, while the segmentation loss is used to enlarge the distance between the query point and the next closest memory entry, reducing the risk of memory entry similarity interference caused by higher

The loss function of the memory module L_M is expressed as

$$L_M = L_{M(Sim)} + L_{M(Seg)} + L_{M(STC)} = \begin{cases} < \left\| f_{frames}^E - M_{frames}(P_{nearest}, f_{frames}^E) \right\| \\ \left\| f_{flows}^E - M_{flows}(P_{nearest}, f_{flows}^E) \right\| > + \\ < - \left\| f_{frames}^E - M_{frames}(P_{sec-nearest}, f_{frames}^E) \right\| \\ - \left\| f_{flows}^E - M_{flows}(P_{sec-nearest}, f_{flows}^E) \right\| > + \\ < \left\| f_{frames}^E - M_{flows}(P_{nearest}, f_{frames}^E) \right\| \\ \left\| f_{flows}^E - M_{frames}(P_{nearest}, f_{flows}^E) \right\| > \end{cases} \quad (5.17)$$

In the equation, M represents the memory block. M_{frames} signifies the appearance pattern within the memory module, while M_{flows} represents the optical flow pattern within the same module. The variable p denotes an entry in the memory module, where $M_{frames}(P_{nearest}, f_{frames}^E)$ designates the memory entry that is nearest to the query feature, and $M_{frames}(P_{sec-nearest}, f_{frames}^E)$ denotes the second closest memory entry to the query feature.

5.4.4 Anomaly detection stage

The primary procedure of the anomaly detection stage maintains consistency with the training process.

The initial step involves preprocessing the dataset, which entails segmenting the test video into video frames and extracting optical flow features. Subsequently, the second step utilizes two distinct encoder structures to compress both the appearance and optical flow features of the video independently. In the third step, the compressed final features are directed into the memory module, where they are combined to generate novel query features. Moving on to the fourth step, these newly generated features are input into the decoder network to anticipate the corresponding optical flow and appearance representations.

The computation of the anomaly score predominantly encompasses two components: the prediction loss and the similarity loss originating from the memory module. The specific pseudo code is as follows:

Algorithm 2 Anomaly Detection Phase

- 1: Initialization:
Flownet2, Random $M \in R^{K \times 2M}$, $V = v_1, v_2, v_3, \dots, v_N$;
 - 2: $\begin{cases} F_{frames} = Ir(V) \\ F_{flows} = Flownet2(V) \end{cases}$;
 - 3: $\begin{cases} F_{frames}^E = \psi_a(F_{frames}) \\ F_{flows}^E = \psi_f(F_{flows}) \end{cases}$;
 - 4: $\vec{f}_{frames}, \vec{f}_{flows} = CoTeaching(M, F_{frames}^E, F_{flows}^E)$;
 - 5: $\begin{cases} F_{flows}^D = \phi_a(\vec{F}_{frames}) \\ F_{frames}^D = \phi_f(\vec{F}_{flows}) \end{cases}$;
- Output:** Calculate anomaly scores.
 $Score = \left\{ \alpha \|F_{flows}^D - F_{flows}\|, \beta \|F_{frames}^D - F_{frames}\| \right\}$
-

where the core part of the anomaly score is the prediction error, which includes optical flow feature prediction error and appearance feature prediction error. In the testing phase, after a large number of verification experiments, this chapter sets two prediction losses combined with hyperparameters $\alpha = 0.3$ and $\beta = 0.7$.

The experimental settings outlined in this chapter are primarily categorized into three groups. According to the experimental settings, evaluates the performance of the framework proposed in this chapter from three aspects: advancement comparison, ablation experiment, and effect display.

5.4.5 Experiment 1

The first group experiments pertains to a comparison of prediction accuracy with mainstream video anomaly detection algorithms. In this set of experiments, this chapter compares the detection accuracy of the model proposed in this chapter and the mainstream unsupervised model. We conducted independent comparative analyzes on three public data

Table 5.1: The result for Avenue dataset

Name	Technology	Journal	AUC
Unmasking [142]	VGG-f	ICCV2017	80.6
StackRNN [84]	Temporally-coherent	ICCV2017	81.7
MemAE [138]	Memory module	ICCV2019	81.0
MNAD [23]	Learning Memory module	CVPR2020	80.6
Covad [25]	Self-attention	CGI2022	83.4
TAC-Net [156]	Temporal-aware contrastive	IEEE TII	87.3
ITAE [60]	Two-path Generative	PR 2022	88.0
Two-P [32]	Two-path AE	ICME 2022	89.8
CCC-T	Consistency Co-teaching		89.2

Table 5.2: The result for UCSD(ped2) dataset

Name	Technology	Journal	AUC
AMDN [38]	Stacked denoising AE	CVIU	90.8
Unmasking [142]	VGG-f	ICCV2017	82.2
StackRNN [84]	Temporally-coherent	ICCV2017	92.2
MemAE [138]	Memory module	ICCV2019	91.7
STFF [157]	Fast sparse coding	PR	92.8
MNAD [23]	Learning Memory module	CVPR2020	97.0
DPU [158]	Dynamic Prototype	CVPR2021	96.9
TAC-Net [156]	Temporal-aware contrastive	IEEE TII	98.1
ITAE [60]	Two-path Generative	PR 2022	98.7
Two-P [32]	Two-path AE	ICME 2022	98.1
CCC-T	Consistency Co-teaching		99.1

sets: UCSD(ped2) [5] Avenue [6], ShanghaiTech [?]. The results are shown in the Table 5.1,5.2 and Table 5.3:

Table 5.1,5.2 and Table 5.3 shows accuracy comparisons between the framework CCC-T proposed in this chapter and mainstream algorithms across three datasets (Avenue, UCSD(ped2), ShanghaiTech). Because the ShanghaiTech dataset is too large, some of the baseline models only tested with the Avenu and UCSD(ped2), and some models use the Ped1 dataset [34]. Therefore, in this chapter we used three different tables (Table 5.1,5.2 and 5.3) to illustrates the results for each dataset with different baseline models. The result show that the prediction accuracy AUC achieved by the CCC-T algorithm has shown better performances for each dataset, thereby substantiating the effectiveness of the proposed algorithm. Specifically, while considering consistency, the way in which optical flow and appearance features are combined (either complementary or equal) becomes the primary aspect of differentiation between video data features and image data. When analyzing video data, special attention should be paid to the processing of dynamic features. From Table 5.1,5.2 and Table 5.3, it can be concluded that the CCC-T model proposed in this chapter has more advanced performance.

The second core store is the consistency constraint of optical flow features and appear-

Table 5.3: The result for ShanghaiTech dataset

Name	Technology	Journal	AUC
StackRNN [84]	Temporally-coherent	ICCV2017	68.0
MemAE [138]	Memory module	ICCV2019	69.7
BMAN [159]	Appearance-motion joint	TIP 2019	76.2
Few-Shot [160]	Few-shot scene-adaptive	ECCV2020	77.9
MNAD [23]	Learning Memory module	CVPR2020	70.5
DPU [158]	Dynamic Prototype	CVPR2021	73.8
TAC-Net [156]	Temporal-aware contrastive	IEEE TII	77.2
DissociateAE [161]	Dissociate spatio-temporal	PR 2022	73.7
ITAE [60]	Two-path Generative	PR 2022	76.3
Two-P [32]	Two-path AE	ICME 2022	73.8
CCC-T	Consistency Co-teaching		77.1

ance features. Simply making optical flow and appearance completely independent and predicting them separately does not conform to the essential characteristics of video data. Forcing the consistency of optical flow and appearance through loss functions is the key to video representation learning. The collaborative learning approach, which fuses optical flow information with appearance information, facilitates a more precise representation of video content. And because the ShanghaiTech data exceeds the limit, only the first two data sets are tested in the ablation experiment part









5.4.6 Experiment 2

The experiments are focused on ablation studies. This experiment involves the separation of various modules such as skip-connecting and Consistency Co-Teaching within the framework for distinct training tests, followed by an assessment of accuracy in the current dual-channel training approach. This chapter set up three groups of ablation experiments to study the comparison between single channel and dual channel, the performance comparison of different components of the model, and the intrinsic relationship between the dual channel loss hyperparameters. In the diagram in the Table 5.4,5.5 blue represents the propagation path of appearance features, and yellow represents the propagation path of optical flow features.

5.4.6.1 The Performance comparison of single-channel and various dual-channel models

In this experiment, we set up four groups of models to compare the performance of single-channel and dual-channel and their different variants: 1) prediction from frame appearance to frame appearance; 2) prediction from appearance features and optical flow features to appearance features; 3) prediction from appearance to appearance and optical flow to optical flow; and 4) As well as the CCC-T framework proposed in this chapter, appearance

Table 5.4: The Ablation Study : The Performance comparison of single-channel and various dual-channel models, blue arrow represents the propagation path of appearance features, and yellow arrow represents optical flow features

	Number	Input	Output	Model	AUC
UCSD(Ped2)	a	frames	frames		97.0
	b	frames,flows	frames		98.9
	c	frames,flows	frames,flows		98.7
	d	frames,flows	flows,frames		75.3
Avenue	a	frames	frames		70.5
	b	frames,flows	frames		88.0
	c	frames,flows	frames,flows		89.8
	d	frames,flows	flows,frames		73.6

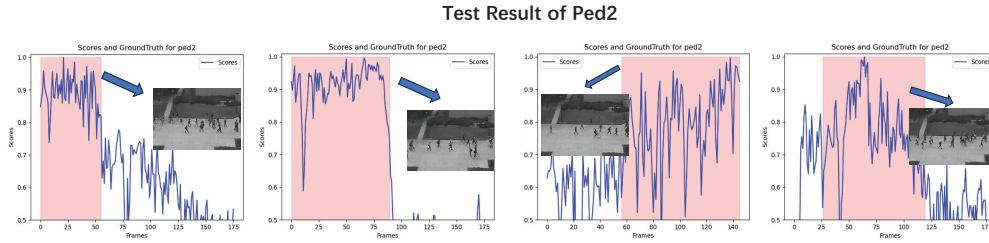


Figure 5.5: Test results on the dataset UCSD(ped2). The pink background is the area where the real anomaly occurs; The blue curve represents the change of the anomaly score with the time series. These Figure show that the model proposed in this contribution has high and stable detection capabilities

predicts optical flow, and optical flow predicts appearance. The experimental results are shown in Table 5.4.

Table 5.4 is the performance comparison between the classic single-channel model and the multi-channel dual-channel model, 1) the initial frame appearance to the prediction/reconstruction of frame appearance; 2) the optical flow as a supplementary feature, and then 3) the separate prediction of optical flow and appearance features and reconstruction; 4) the basic model proposed but without constraint. The final prediction accuracy of the model shows an upward trend. Without the assistance of the Co-Teaching module and skip connections, the performance of the model Init (Un-Constraint) proposed in this chapter is far inferior to the classic model. After analysis, it was found that this is because optical flow features lack more appearance information (such as background, color, etc.), and appearance features cannot be predicted directly from optical flow. This ablation study results are presented in Table 5.5.

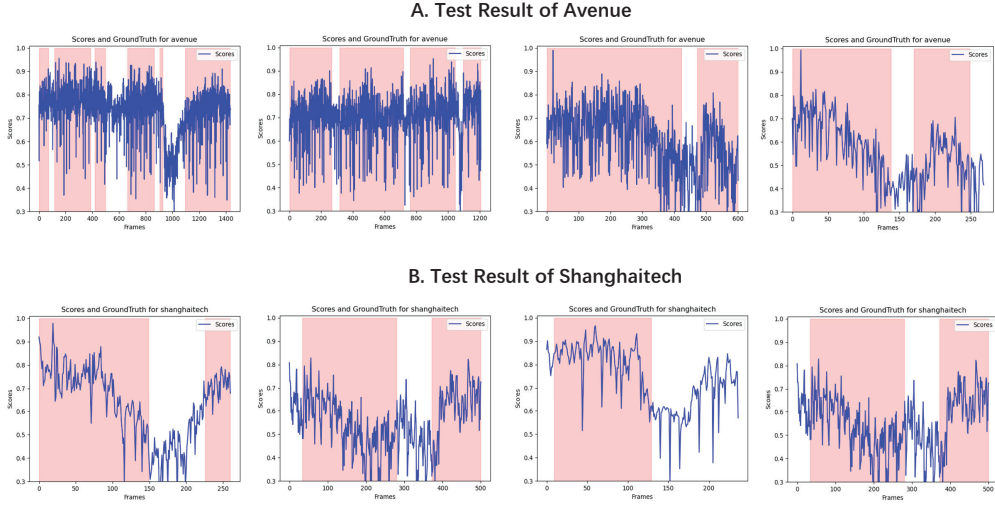


Figure 5.6: Test results on the dataset Avenue and ShanghaiTech; A represents the result of Avenue and B is the result of ShanghaiTech. The pink background is the area where the real anomaly occurs; The blue curve represents the prediction anomaly score with the time series

Table 5.5: The Ablation Study: The impact of skip connections and co-teaching on the performance of dual-channel models

Method	Model	UCSD(ped2)	Avenue
Init (Un-constraint)		75.3	73.6
Skip(Flows-Frames)		95.4	87.2
Skip(Frames-Flows)		76.2	76.7
Fully-Skip		95.6	86.9
CCC-T		99.1	89.1

5.4.6.2 The impact of skip connections and co-teaching on the performance of dual-channel models

In this set of ablation experiments, this chapter set up five sets of models: 1) the baseline model of cross prediction; 2) the skip connection model that only includes appearance to optical flow; 3) the skip connection model that only includes optical flow features to appearance features. ;4) double-skip connection model; 5) the final framework CCC-T including consistency co-teaching and double-skip connection; and compare their performances with UCSD(ped2) and Avenue datasets. The experimental results are shown in Table 5.5.

The results in Table 5.5 show the performances of different modules in the proposed CCC-T framework. From the Table 5.5, it can be concluded that the main reason for the

low cross-prediction performance is that the optical flow feature has less appearance feature information and cannot be completely restored. Therefore, in the channel where optical flow features predict appearance features, whether there are skip connections that provide appearance feature input has a greater impact on the performance of the Skip(Flows-Frames) model. Table 5.5 shows that the performance of the Skip(Flows-Frames) model containing only this core skip connection basically reaches the performance of the double-hop connection. Secondly, whether to set up a strong co-teaching network structure also has a great impact on performance. Therefore, each component of the CCC-T model performance proposed in this chapter is essential.

5.4.7 Experiment 3

Experiment 3 is a visual evaluation experiment and the fluctuation of abnormal scores between normal frames and abnormal frames.

The experimental findings from the test phase have been visually presented in Figure 5.5, Figure 5.6. From Figure 5.5, we can get that a recognizable shift in the abnormal score is observed when confronted with irregular video frames, exhibiting a significant increase. It shows that this phenomenon helps us effectively pinpoint anomalies in video data streams. Exceptions in the graphical representation include various situations, particularly the use of bicycles, skateboards, and other unconventional vehicles on sidewalks. From Figure 5.6 which shows that the current unsupervised algorithm has insufficient performance indicators in the ShanghaiTech dataset and is difficult to distinguish not obvious abnormal events. Combining the results displayed by the two effects, we can infer that the video anomaly is not for the detection of a certain frame, but for the analysis of a segment. Since the abnormality score in the picture fluctuates violently, it is difficult to locate abnormal from several other frames, but considering overall situation of video or the entire segment, abnormal events can be clearly located. This once again proves that abnormalities are continuous and indivisible.

5.5 Conclusion and Future Work

We introduces an innovative approach to unsupervised video anomaly detection framework CCC-T which is leveraging the inherent consistency between optical flow features and appearance features. The framework capitalizes on the correlation properties of these two types of features, marking the first instance of their fusion within an unsupervised algorithm.

In this framework, we set strong consistency constraints to achieve consistent alignment of appearance features and motion features, and introduce a novel prediction mechanism.

This mechanism is bidirectional predicting both optical flow from appearance and appearance from optical flow. This ingenious strategy effectively mitigates the robustness challenges that typically afflict unsupervised learning, thereby generates enhancements in algorithmic performance. Furthermore, the framework employs a co-teaching network, which fosters coordination between the two channels. This approach skillfully averts distortions that can arise from the neural network potent representation capacity. The empirical findings the superior and more resilient performance of the algorithm proposed in this chapter, as compared to conventional methods for predicting video frames.

In future, our team is committed to delving deeper into the placement of optical flow features within unsupervised anomaly detection algorithms. We aim to explore the potential synergies between a broader range of unsupervised and weakly supervised algorithms, with the goal of pushing the boundaries of anomaly detection even further. We will also further explore the connection between abnormal events and context, hoping to combine text sentiment analysis and video understanding to explore more accurate definitions of abnormalities in videos.

Chapter **6**

Discussion

Contents

6.1	Implications of Findings	110
6.2	Challenges and Limitations	111
6.3	Future Research Directions	113
6.4	Conclusion	115

6.1 Implications of Findings

Video anomaly detection is a key issue in computer vision in the security field. In recent years, due to the rise of deep learning, especially the rapid development of deep neural networks, unsupervised video anomaly detection algorithms have made a series of new progress. However, there are still many problems due to the current unsupervised video anomaly detection techniques. For example, the number of normal behavior pattern types in the normal data set is uncertain, the data quantity distribution of the normal pattern is unbalanced; the detection and recognition of specific target behavior and appearance are not high; the abnormal detection of traffic accidents is still in the problem of postmortem detection. Therefore, anomaly detection for video data is still the mainstream research direction in the field of computer vision. Improving the real-time performance and accuracy of anomaly detection and designing an anomaly detection algorithm with good performance are still important issues. The existing mainstream algorithms are divided into unsupervised algorithms and weakly supervised algorithms, which have their own advantages in robustness and versatility. How to improve the performance of video anomaly detection algorithms is the key research content of this project. Weakly supervised algorithms are more robust than unsupervised algorithms, but less versatile, while unsupervised algorithms are the opposite; therefore, this thesis further optimizes and improves the existing video anomaly detection model.

Contribution I, due to the randomness of the start of multi-instance learning, that is, to randomly optimize the fragments with large initial outliers, and to give priority to short-term features in temporal feature extraction, this chapter proposes weak supervision based on temporal convolutional network optimization. detection scheme. The framework considers the information of video sequences holistically and incorporates them into the components of input features, providing a more reliable reference for the initiation of multi-instance learning optimization. The TCN network structure adopted makes up for the deficiency that C3D and i3d can only capture short video action features and effectively improve the performance of the weakly supervised algorithm.

Contribution II, the bottleneck of the unsupervised algorithm is that it spends a lot of weight and attention on the meaningless background information, which greatly interferes with the extraction of normal patterns, and ignores the importance of dynamic targets. The algorithm proposed in this chapter uses a coordinated self-attention mechanism to help the neural network focus on meaningful objects by ignoring the background in the video during training. Compared with the traditional unsupervised algorithm, the algorithm proposed in this chapter is more reasonable by using a small number of important features in the video frame as the main basis for the abnormal score. According to the experimental results, the

algorithm we propose can avoid the detection efficiency of unimportant background noise, that is, the algorithm in this chapter has strong anti-noise ability.

Contribution III proposes an innovative approach for unsupervised video anomaly detection by exploiting the intrinsic consistency between optical flow features and appearance features. This method further relieves unsupervised algorithms from focusing on large areas of meaningless background information. The introduction of optical flow features and appearance feature consistency signifies that the proposed framework pays great attention to dynamic objects. The most important thing is that the framework utilizes the related attributes of these two types of features to build a co-teaching learning framework, which helps the model to establish a dual-channel prediction mechanism, in which optical flow features and appearance features are considered equally important information, the model can either predict the optical flow based on the appearance, or predict the appearance based on the optical flow. This ingenious strategy effectively alleviates the robustness challenges that usually plague unsupervised learning, thereby significantly improving algorithm performance. In both contributions, we used the three datasets (two is same as contribution II) for analysis and contribution III improved the detection performance by 5% compared to contribution II.

In conclusion, the three innovative models proposed in this thesis bring performance improvements in the field of data-driven video anomaly detection. Chapter 3 optimizes the error at the initial startup of the neural network in contrastive learning and strengthens the temporal features. chapter 4 optimizes the unsupervised model and uses a novel self-attention mechanism to optimize the weight distribution rules in the model. Chapter 5 proposes a new two-channel unsupervised video anomaly detection framework, which treats dynamic features and static features equally, and co-trains the two channels through the co-teaching network to ensure dynamic features and static features The consistency further improves the performance. With the deepening of research, our team has gradually discovered the limitations of data-driven models, which will profoundly affect our future research.

6.2 Challenges and Limitations

Video anomaly detection has always been a challenging research field. The core limitation and challenges that existed before was the problem of abnormal data collection in application scenarios. In fact, this problem has not been solved, but the problem has been circumvented under the **preliminary definition of abnormality** (unsupervised learning framework and weakly supervised learning framework). The two detection frameworks decrease reliance on anomalous data within the training dataset by formulating rules for ab-

normalities. Therefore, when the weakly supervised framework based on multiple instances and the unsupervised framework based on representation learning appear, the main challenge is shifted from abnormal data collection to accurate extraction of video data features.

Existing unsupervised models and weakly supervised models are guided by the two frameworks to reduce the dependence on datasets. However, the emphasis of the two is different. Weakly supervised contrastive learning emphasizes the error between normal events and abnormal events, so its core is to segment video sequences and compare scores, while unsupervised representation learning emphasizes the difference between abnormal patterns and normal patterns. The core of the deviation is the comparison between the features learned from normal data and its own real features. This leads to new limitations of the two algorithms. For the weakly supervised algorithm, since the video packets are packaged into instances by slices, the training features are discontinuous, which also leads to inaccurate feature extraction, and due to the optimization goal of contrastive learning is generally consistent. If the abnormal instance is not targeted by the neural network in the first round, the subsequent optimization will be invalid optimization. In response to this problem, my thesis proposes a weakly supervised framework based on the TCN network. This framework calculates the score of the entire video sequence as a whole, and generates a pseudo-label for each instance based on the score, reducing the occurrence of misjudgment and improving the precision of model feature extraction. Based on my further research we explored that the weakly supervised algorithm still needs some labelled data (which is limited in video anomaly detection datasets); therefore, we shifted the research focus to the unsupervised field.

The main problem I face when using unsupervised algorithms is how to accurately represent the features of video data. The previous model has realized the prediction/reconstruction of appearance features from only using appearance features, and mixing appearance features and dynamic features to predict/reconstruct appearance features. Due to the lack of consistency between appearance features and dynamic features, these methods cannot accurately extract the features of video sequences. In order to solve this problem, in my thesis I propose a new unsupervised framework based on the consistency of dynamic features and static features in contribution III, which further improves the accuracy of unsupervised model video feature extraction.

With the advancement of computing power and video processing technology, real-time online processing of video data has emerged as a popular application in several domains. Researchers will no longer be limited by computing power and precise feature extraction. Nevertheless, there is a need to further strengthen data dependencies, redefine anomaly criteria, and refine evaluation methods. Because the traditional unsupervised framework and weakly supervised framework still have a strong dependence on data, they belong to the

field of data-driven deep learning. Therefore, one of the future challenges is about **the deep definition of abnormality** which can be identified as transforming the framework from data dependence to knowledge dependence. Hence, researchers must reemphasize solid AI and knowledge-driven approaches, treating knowledge as an essential part of solving statistical problems.

In the field of video anomaly detection, another important consideration for anomaly detection is knowledge discovery. The video anomaly detection problem relies on the knowledge of how anomaly patterns should be defined in the current environment. This is also the core challenge in video anomaly detection. In different scenarios, the definition of exception may be different. For example, when a car is driving on a road, it is considered a normal event, but if the same car is driving on a sidewalk, it is considered an abnormal event. Even though we use a lot of video processing techniques in this scene to extract features, build an anomaly detection model to calculate anomaly scores, and finally get correct results, this process consumes a lot of computing resources. More importantly, most of the computing resources are not effectively used in the core definition of the event.

With the deepening of video anomaly detection research, more and more researchers have found that in order to achieve the preparation definition of anomaly patterns, the most important prerequisite is to complete the extraction of contextual information, including background information, video targets, target motion, etc. Therefore, it becomes increasingly difficult to completely rely on computer vision technology to complete the definition and detection of anomalies. The fusion of interdisciplinary cross-fusion technologies such as computer vision and natural language processing is the key to solving more practical video anomaly detection in the future.

6.3 Future Research Directions

As we focus on technological advancements, it becomes increasingly clear that the deep learning landscape is about to change. This transformation hinges on a shift from the prevailing data-driven paradigm to a knowledge-driven approach. In addition, the fusion of different disciplines is bound to be an effective and reliable solution to future multifaceted challenges. An illustration of this trajectory can be found in the field of video anomaly detection, where the synergy of computer vision, natural language processing, and other fields promises to unlock unprecedented capabilities.

Currently, video anomaly detection stands as a point of research, capturing the imagination of experts across numerous domains. Video classification is a classic problem within the field of computer vision, and many representative results have been obtained. But recent breakthroughs in video classification technology, anchored in video understanding,

have graced the pages of prestigious journals and conferences, including the revered CVPR and TIP. It is within the domain of video understanding that a novel technical route appeared - a pathway that 'What happened within this video?' replaces 'What's wrong in this video?'

Video anomaly detection based on video scene understanding is completely different from the definition of traditional video anomaly detection. Traditional video anomaly detection focuses on the extraction, combination and definition of abnormal patterns of video optical flow and appearance features, while anomaly detection based on video scene understanding pays more attention to the context information. This exceptional paradigm shift not only redefines the direction in which problems are solved but also change the technology our address these question. Compared with the vision-based scheme, the scene understanding-based scheme is more complex, but greatly reduces the dependence on data labels and avoids noise interference. This boost ultimately translates computer vision problems into natural language processing problems and regenerates a novel technological trajectory.

The promise of video anomaly detection is doomed to increase complexity over time. Cross-domain joint modeling will become the mainstream of problem-solving, and the intersection of NLP and CV technology will become the key to solving video anomaly detection. The supplementary support of NLP to CV comes from two directions. The first is that the definition of abnormality gets rid of the absolute dependence on appearance features, and the second is that the definition of abnormality and the interpretability of the model are enhanced. The combination of cross-fields has accelerated the application progress of video anomaly detection technology. This fusion brings a paradigm shift, not only enabling the migration from vision to semantics but also understanding the contextual basis of video scenes, pushing anomaly detection to a more flexible and free level. This process forms the cornerstone of accurate and general anomaly detection techniques by understanding the subtle interplay between what the eye perceives and what speech conveys.

However, the process of achieving this convergence has not been without challenges. It includes three aspects. First, the integration of NLP and CV needs to create a collaborative architecture that can realize the transformation of visual data into text semantic features. The second is to design a coding framework that can effectively encapsulate the data characteristics of these two fields and realize reverse restoration. Third and most importantly, defining anomalies using contextual semantic information also requires the formulation of specific anomaly semantic rules and semantic libraries. This effort requires novel semantic architecture paradigms. In addition, the data environment for integrated models in complex environments needs to be carefully set up. Annotated datasets encapsulate exceptional complexity across visual and contextual dimensions and are the lifeblood

of knowledge-driven models. Creating semantic repositories of interdisciplinary benchmark datasets suitable for this complex environment becomes critical.

In conclusion, the future of deep learning is poised for a major transformation, one that forsakes the rigid confines of data-driven approaches in favour of knowledge-driven methodologies. This shift is intricately linked with the integration of multiple disciplines, a phenomenon that holds unparalleled potential in tackling complex problems. Video anomaly detection stands as a beacon of this interdisciplinary synergy, showcasing how the integration of NLP and CV can redefine traditional CV problems, ushering them into the realm of textual understanding. This shift, however, requires the forging of new paths, the development of pioneering architectures, and the cultivation of curated datasets. As these pieces fall into place, the union of knowledge-driven deep learning and interdisciplinary integration will undoubtedly chart the course for groundbreaking advancements in the landscape of video anomaly detection and beyond.

6.4 Conclusion

Video anomaly detection is an application-oriented research direction, and it is a research framework for the fusion and intersection of various visual tasks. In the process of studying this problem, researchers should not specialize in one or several specific technologies but should pay attention to the integration of multiple technologies. Just like the future research direction mentioned in the previous section, explainable (reliable) artificial intelligence will definitely become the trend of future research. Therefore, with a new energy injection, video anomaly detection is still a vibrant research field, as evidenced by the continuous publication of relevant research in top journals and conferences every year. Currently, video classification mainly relies on traditional data mining techniques supplemented by existing deep learning advances; however, it is difficult to go beyond the limitations of being fully data-driven. Looking ahead, a promising avenue is the development of knowledge content-driven neural network models, which marks an emerging avenue for future technological exploration.

References

- [1] Y. Myagmar-Ochir and W. Kim, "A survey of video surveillance systems in smart city," *Electronics*, vol. 12, no. 17, p. 3567, 2023.
- [2] W. Ullah, A. Ullah, T. Hussain, K. Muhammad, A. A. Heidari, J. Del Ser, S. W. Baik, and V. H. C. De Albuquerque, "Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data," *Future Generation Computer Systems*, vol. 129, pp. 286–297, 2022.
- [3] E. Şengönül, R. Samet, Q. Abu Al-Haija, A. Alqahtani, B. Alturki, and A. A. Alsulami, "An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey," *Applied Sciences*, vol. 13, no. 8, p. 4956, 2023.
- [4] S. Zhu, C. Chen, and W. Sultani, "Video anomaly detection for smart surveillance," in *Computer Vision: A Reference Guide*. Springer, 2020, pp. 1–8.
- [5] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.
- [6] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.
- [7] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.
- [8] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.
- [9] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018.
- [10] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4487–4495, 2020.
- [11] G. Ding, Q. Wu, L. Zhang, Y. Lin, T. A. Tsiftsis, and Y.-D. Yao, "An amateur drone surveillance system based on the cognitive internet of things," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 29–35, 2018.
- [12] S. Chandrakala, K. Deepak, and G. Revathy, "Anomaly detection in surveillance videos: a thematic taxonomy of deep models, review and performance analysis," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3319–3368, 2023.
- [13] J. J. P. Suarez and P. C. Naval Jr, "A survey on deep learning techniques for video anomaly detection," *arXiv preprint arXiv:2009.14146*, 2020.
- [14] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

-
- [15] D. R. Patrikar and M. R. Parate, "Anomaly detection using edge computing in video surveillance system," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 2, pp. 85–110, 2022.
- [16] K. Liu and H. Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1490–1499.
- [17] F. Landi, C. G. Snoek, and R. Cucchiara, "Anomaly locality in video surveillance," *arXiv preprint arXiv:1901.10364*, 2019.
- [18] K. Doshi and Y. Yilmaz, "Multi-task learning for video surveillance with limited data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3889–3899.
- [19] H. Mu, R. Sun, M. Wang, and Z. Chen, "Spatio-temporal graph-based cnns for anomaly detection in weakly-labeled videos," *Information Processing & Management*, vol. 59, no. 4, p. 102983, 2022.
- [20] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [22] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2293–2312, 2020.
- [23] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 372–14 381.
- [24] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4975–4986.
- [25] W. Shao, P. Rajapaksha, Y. Wei, D. Li, N. Crespi, and Z. Luo, "Covad: Content-oriented video anomaly detection using a self-attention based deep learning model," *Virtual Reality & Intelligent Hardware*, vol. 5, no. 1, pp. 24–41, 2023.
- [26] W. Shao, R. Xiao, P. Rajapaksha, M. Wang, N. Crespi, Z. Luo, and R. Minerva, "Video anomaly detection with ntcn-ml: A novel tcn for multi-instance learning," *Pattern Recognition*, p. 109765, 2023.
- [27] Y. Zhao, W. Wu, Y. He, Y. Li, X. Tan, and S. Chen, "Good practices and a strong baseline for traffic anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3993–4001.
- [28] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV 12*. Springer, 2012, pp. 702–715.
- [29] A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "Ubnorm: New benchmark for supervised open-set video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 143–20 153.
- [30] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, vol. 46, pp. 235–262, 2013.
- [31] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 588–13 597.
- [32] Y. Liu, J. Liu, M. Zhao, D. Yang, X. Zhu, and L. Song, "Learning appearance-motion normality for video anomaly detection," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

-
- [33] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, and Y. Wang, "Abnormal event detection using deep contrastive learning for intelligent video surveillance system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5171–5179, 2021.
- [34] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.
- [35] T. Jiang, Y. Li, W. Xie, and Q. Du, "Discriminative reconstruction constrained generative adversarial network for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4666–4679, 2020.
- [36] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.
- [37] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1933–1941.
- [38] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.
- [39] N. Li and F. Chang, "Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, 2019.
- [40] E. Duman and O. A. Erdem, "Anomaly detection in videos using optical flow and convolutional autoencoder," *IEEE Access*, vol. 7, pp. 183 914–183 923, 2019.
- [41] N. Madan, A. Farkhondeh, K. Nasrollahi, S. Escalera, and T. B. Moeslund, "Temporal cues from socially unacceptable trajectories for anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2150–2158.
- [42] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2138–2148, 2019.
- [43] C. Sun, Y. Jia, H. Song, and Y. Wu, "Adversarial 3d convolutional auto-encoder for abnormal event detection in videos," *IEEE Transactions on Multimedia*, vol. 23, pp. 3292–3305, 2020.
- [44] T.-N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1273–1283.
- [45] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "Anopc: Video anomaly detection via deep predictive coding network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1805–1813.
- [46] Y. Lu, K. M. Kumar, S. shahabeddin Nabavi, and Y. Wang, "Future frame prediction using convolutional vrnn for anomaly detection," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.
- [47] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, "Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5546–5554.
- [48] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An efficient anomaly recognition framework using an attention residual lstm in surveillance videos," *Sensors*, vol. 21, no. 8, p. 2811, 2021.
- [49] C. Park, M. Cho, M. Lee, and S. Lee, "Fastano: Fast anomaly detection via spatio-temporal patch transformation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2249–2259.
- [50] J. Lee, W.-J. Nam, and S.-W. Lee, "Multi-contextual predictions with vision transformer for video anomaly detection," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1012–1018.

-
- [51] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, "Transanomaly: Video anomaly detection using video vision transformer," *IEEE Access*, vol. 9, pp. 123 977–123 986, 2021.
- [52] M. Adnan, S. Kalra, and H. R. Tizhoosh, "Representation learning of histopathology images using graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 988–989.
- [53] Y. Kim, J. Yun, H. Shon, and J. Kim, "Joint negative and positive learning for noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9442–9451.
- [54] X. Huang, C. Zhao, and Z. Wu, "A video anomaly detection framework based on appearance-motion semantics representation consistency," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [55] X. Wang, Z. Che, B. Jiang, N. Xiao, K. Yang, J. Tang, J. Ye, J. Wang, and Q. Qi, "Robust unsupervised video anomaly detection by multipath frame prediction," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 6, pp. 2301–2312, 2021.
- [56] S. Sun and X. Gong, "Hierarchical semantic contrast for scene-aware video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 846–22 856.
- [57] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 742–12 752.
- [58] H. Deng, Z. Zhang, S. Zou, and X. Li, "Bi-directional frame interpolation for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2634–2643.
- [59] C. Chen, Y. Xie, S. Lin, A. Yao, G. Jiang, W. Zhang, Y. Qu, R. Qiao, B. Ren, and L. Ma, "Comprehensive regularization in a bi-directional predictive network for video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 230–238.
- [60] M. Cho, T. Kim, W. J. Kim, S. Cho, and S. Lee, "Unsupervised video anomaly detection via normalizing flows with implicit latent features," *Pattern Recognition*, vol. 129, p. 108703, 2022.
- [61] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4975–4986.
- [62] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 009–14 018.
- [63] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE transactions on image processing*, vol. 30, pp. 4505–4515, 2021.
- [64] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1237–1246.
- [65] B. Wan, Y. Fang, X. Xia, and J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.
- [66] B. Ramachandra, M. Jones, and R. Vatsavai, "Learning a distance function with a siamese network to localize anomalies in videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2598–2607.
- [67] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.

-
- [68] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, “Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection,” in *European Conference on Computer Vision*. Springer, 2020, pp. 358–376.
- [69] J. Wu, W. Zhang, G. Li, W. Wu, X. Tan, Y. Li, E. Ding, and L. Lin, “Weakly-supervised spatio-temporal anomaly detection in surveillance video,” *arXiv preprint arXiv:2108.03825*, 2021.
- [70] P. Wu and J. Liu, “Learning causal temporal relation and feature discrimination for anomaly detection,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3513–3527, 2021.
- [71] S. Li, F. Liu, and L. Jiao, “Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection,” *Proceedings of the AAAI, Virtual*, vol. 24, 2022.
- [72] W. Liu, W. Luo, Z. Li, P. Zhao, S. Gao *et al.*, “Margin learning embedded prediction for video anomaly detection with a few anomalies.” in *IJCAI*, 2019, pp. 3023–3030.
- [73] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 322–339.
- [74] H. Zhou, J. Yu, and W. Yang, “Dual memory units with uncertainty regulation for weakly supervised video anomaly detection,” *arXiv preprint arXiv:2302.05160*, 2023.
- [75] M. Cho, M. Kim, S. Hwang, C. Park, K. Lee, and S. Lee, “Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 137–12 146.
- [76] C. Wu, S. Shao, C. Tunc, P. Satam, and S. Hariri, “An explainable and efficient deep learning framework for video anomaly detection,” *Cluster computing*, pp. 1–23, 2021.
- [77] T. Reiss and Y. Hoshen, “Attribute-based representations for accurate and interpretable video anomaly detection,” *arXiv preprint arXiv:2212.00789*, 2022.
- [78] K. Doshi and Y. Yilmaz, “Towards interpretable video anomaly detection,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2655–2664.
- [79] D. Fortun, P. Bouthemy, and C. Kervrann, “Optical flow modeling and computation: A survey,” *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, 2015.
- [80] J. Hur and S. Roth, “Iterative residual refinement for joint optical flow and occlusion estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5754–5763.
- [81] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, “Learning to estimate hidden motions with global motion aggregation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9772–9781.
- [82] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, “Flownet: Learning optical flow with convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.
- [83] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.
- [84] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.
- [85] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, “Activity recognition using temporal optical flow convolutional features and multilayer lstm,” *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.
- [86] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, “Exploring temporal coherence for more general video face forgery detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 044–15 054.

-
- [87] D. Liu, J. Liang, T. Geng, A. Loui, and T. Zhou, "Tripartite feature enhanced pyramid network for dense prediction," *IEEE Transactions on Image Processing*, 2023.
- [88] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.
- [89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.
- [90] C. Liu, J. Liu, S. Xu, J. Wang, C. Liu, T. Chen, and T. Jiang, "A spatiotemporal dilated convolutional generative network for point-of-interest recommendation," *ISPRS International Journal of Geo-Information*, vol. 9, no. 2, p. 113, 2020.
- [91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [92] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [93] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8857–8866.
- [94] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.
- [95] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, vol. 10, 1997.
- [96] J. Cooper, "Comparative learning theory and its application in the training of horses," *Equine Veterinary Journal*, vol. 30, no. S27, pp. 39–43, 1998.
- [97] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [98] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.
- [99] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 076–10 085.
- [100] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.
- [101] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [102] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.
- [103] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [104] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, pp. 13–es, 2006.
- [105] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

-
- [106] Y. Zhang, J. Li, and C. Zhang, "Accumulation of adversarial examples for underwater visual object tracking," in *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2022, pp. 986–990.
- [107] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "Dcfnet: Discriminant correlation filters network for visual tracking," *arXiv preprint arXiv:1704.04057*, 2017.
- [108] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
- [109] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the roc curve," *Machine learning*, vol. 77, no. 1, pp. 103–123, 2009.
- [110] D. K. McClish, "Analyzing a portion of the roc curve," *Medical decision making*, vol. 9, no. 3, pp. 190–195, 1989.
- [111] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.
- [112] A. C. J. Janssens and F. K. Martens, "Reflection on modern methods: Revisiting the area under the roc curve," *International journal of epidemiology*, vol. 49, no. 4, pp. 1397–1403, 2020.
- [113] X.-L. Zhang and M. Xu, "Auc optimization for deep learning-based voice activity detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–12, 2022.
- [114] F. V. Massoli, F. Falchi, A. Kantarci, Ş. Akti, H. K. Ekenel, and G. Amato, "Mocca: Multilayer one-class classification for anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [115] S. Huang, Z. Liu, W. Jin, and Y. Mu, "Bag dissimilarity regularized multi-instance learning," *Pattern Recognition*, vol. 126, p. 108583, 2022.
- [116] Y. Zou, R. V. Donner, N. Marwan, J. F. Donges, and J. Kurths, "Complex network approaches to nonlinear time series analysis," *Physics Reports*, vol. 787, pp. 1–97, 2019.
- [117] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [118] C. Cheng, C. Zhang, Y. Wei, and Y.-G. Jiang, "Sparse temporal causal convolution for efficient action modeling," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 592–600.
- [119] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.
- [120] N. R. Chilukuri and C. Eliasmith, "Parallelizing legendre memory unit training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1898–1907.
- [121] S. Nah, S. Son, and K. M. Lee, "Recurrent neural networks with intra-frame iterations for video deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8102–8111.
- [122] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [123] W. Hao, R. Zhang, S. Li, J. Li, F. Li, S. Zhao, and W. Zhang, "Anomaly event detection in security surveillance using two-stream based model," *Security and Communication Networks*, vol. 2020, 2020.
- [124] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4030–4034.
- [125] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Processing Letters*, vol. 27, pp. 1705–1709, 2020.
- [126] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *European conference on computer vision*. Springer, 2020, pp. 322–339.

- [127] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "Dance with self-attention: A new look of conditional random fields on anomaly detection in videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 173–183.
- [128] H. Sapkota and Q. Yu, "Bayesian nonparametric submodular video partition for robust anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3212–3221.
- [129] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in *2017 IEEE International conference on multimedia and expo (ICME)*. IEEE, 2017, pp. 439–444.
- [130] —, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.
- [131] W. Wang, F. Chang, and C. Liu, "Mutuality-oriented reconstruction and prediction hybrid network for video anomaly detection," *Signal, Image and Video Processing*, vol. 16, no. 7, pp. 1747–1754, 2022.
- [132] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [133] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-unet: A novel architecture for semantic segmentation in unstructured environment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 358–359.
- [134] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.
- [135] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.
- [136] S. H. Hashemi, S. Abdu Jyothi, and R. Campbell, "Tictac: Accelerating distributed deep learning with communication scheduling," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 418–430, 2019.
- [137] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [138] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.
- [139] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
- [140] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.
- [141] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.
- [142] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2895–2903.
- [143] D. A. Migliore, M. Matteucci, and M. Naccari, "A reevaluation of frame difference in fast and robust motion detection," in *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, 2006, pp. 215–218.
- [144] J.-H. Han, S. Yang, and B.-U. Lee, "A novel 3-d color histogram equalization method with uniform 1-d gray scale histogram," *IEEE Transactions on Image Processing*, vol. 20, no. 2, pp. 506–512, 2010.
- [145] T. Zhang, X. Zhang, X. Ke, C. Liu, X. Xu, X. Zhan, C. Wang, I. Ahmad, Y. Zhou, D. Pan *et al.*, "Hog-shipclsnet: A novel deep learning network with hog feature fusion for sar ship classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–22, 2021.

-
- [146] W. Shao, R. Xiao, P. Rajapaksha, M. Wang, N. Crespi, Z. Luo, and R. Minerva, "Video anomaly detection with ntcn-ml: A novel tcn for multi-instance learning," *Pattern Recognition*, p. 109765, 2023.
- [147] Z. K. Abbas and A. A. Al-Ani, "A comprehensive review for video anomaly detection on videos," in *2022 International Conference on Computer Science and Software Engineering (CSASE)*. IEEE, 2022, pp. 1–1.
- [148] J. Wu, S. Pan, X. Zhu, C. Zhang, and X. Wu, "Multi-instance learning with discriminative bag mapping," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1065–1080, 2018.
- [149] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4799–4807.
- [150] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.
- [151] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.
- [152] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 691–706.
- [153] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.
- [154] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106854, 2020.
- [155] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.
- [156] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, and Y. Wang, "Abnormal event detection using deep contrastive learning for intelligent video surveillance system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5171–5179, 2021.
- [157] P. Wu, J. Liu, M. Li, Y. Sun, and F. Shen, "Fast sparse coding networks for anomaly detection in videos," *Pattern Recognition*, vol. 107, p. 107515, 2020.
- [158] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 425–15 434.
- [159] S. Lee, H. G. Kim, and Y. M. Ro, "Bman: Bidirectional multi-scale aggregation networks for abnormal event detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 2395–2408, 2019.
- [160] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, "Few-shot scene-adaptive anomaly detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 125–141.
- [161] Y. Chang, Z. Tu, W. Xie, B. Luo, S. Zhang, H. Sui, and J. Yuan, "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognition*, vol. 122, p. 108213, 2022.

List of figures

2.1	Classification of Algorithms in Video Anomaly Detection [22]	29
3.1	A representation of the spatiotemporal dimension of anomalous events.	45
3.2	Feature similarity analysis of positive and negative instances: A means abnormal data, N means normal data. $A1$ represents normal instances in abnormal data, $A2$ represents abnormal instances, and $N1$ represents instances in normal data, S_F represents similarity of features between two instances and P_{Ab} is the probability of an anomalous instance.	45
3.3	The structure of TCN and Application, (a) The proposed Temporal Convolutional Network structure, under $d = 2$, $k = 3$, the input is $X = x_1, x_2, x_3, \dots, x_T$, $k = 3$ is the number of kernels that three upper-layer neurons map a neuron of the current layer, $d = 2$ is stride representing the distance between two kernel units; (b) The novel TCN application in video processing. The red line represents the feature of the current node to be extracted, the blue line represents the feature of the previous node, and the red curve represents a 1×1 convolution unit that retains the most original features of the current node, and the output of the TCN is z , which is the probability value that the input node is an abnormal instance.	49
3.4	The NTCN-ML framework: The model training process is divided into two phases. The first phase is composed of a vanilla discriminator and novel TCN. The training purpose of this phase is to extract temporal features; the second phase is composed of a vanilla discriminator and TCN. The training model is composed of the MIL module, and the novel TCN module, this stage is to improve the classification ability of the MIL network.	51
3.5	The distribution of the ShanghaiTech dataset, (a) denotes the abnormal distribution of abnormal data entries (63 videos) in training, x-axis represents the video number and y-axis represents the total number of frames. Yellow colour is the location for abnormal frames; (b) denotes the abnormal distribution of abnormal data entries in testing for 44 videos.	56
3.6	Visual effects of the anomaly detection phase, (a) the detection results of anomalous events with a short duration, (b) the detection results of anomalous events with a longer duration d , the yellow line indicates the correctly detected samples, and the red line indicates the detection results is wrong.	61
3.7	Visual effects of the anomaly detection phase. Red is the area where real anomalies occur, and the curve is the anomaly score.	62
4.1	Algorithm Framework: 1. Extract video features through an encoder, 2. Then input collaborative attention mechanism to redistribute weights, 3. Read memory module and update, 4. Restore the aggregated query features and memory module features to video frames, 5. Calculate the loss, back-propagate, and update parameters	68
4.2	The basic U-Net: The U-Net network is composed of convolution, pooling, upsampling, and skip connections, where convolution and pooling are used to extract input features, upsampling is to restore the pooled and scaled features, and skip connections are feature splicing, trying to use a wider range of information to help restore video frames	69
4.3	This is the algorithm flow of the memory module, including the flow chart of reading and updating memory	71
4.4	Coordinate Attention C is the number of channels; H, W represent the length and width of the current feature, respectively	76
4.5	The a is a frame in the video, b is the feature map generated without a coordinated attention mechanism, and c is the feature map generated by the coordinated attention mechanism.	81

4.6	The a is a feather loss map, b is the result of object detection, and c is the generated result by feather loss map and objected box map	83
5.1	Comparison of methods: A uses optical flow features as a supplement to video frame appearance features to improve prediction accuracy; B uses parallel prediction of appearance features and optical flow features to build a joint prediction loss error; C is the proposed strong consistency collaborative training framework.	89
5.2	A detailed framework of CCC-T. The first step uses Flownet2 to obtain the optical flow information of the video sequence. The second step inputs the segmented video frames and optical flow information into their respective encoding networks. The third step is to cross-read and collaborate the output of the encoding network with the memory module. Update, the fourth step, the updated input features are input to the decoder network for cross prediction.	92
5.3	A Classify Co-teaching Structure	93
5.4	Co-teaching within a memory module: Green indicates the transfer of static features in the memory module, and orange represents the transfer of dynamic features; correspondingly, the memory module is composed of multiple static feature category entries and multiple dynamic feature category entries	96
5.5	Test results on the dataset UCSD(ped2). The pink background is the area where the real anomaly occurs; The blue curve represents the change of the anomaly score with the time series. These Figure show that the model proposed in this contribution has high and stable detection capabilities	104
5.6	Test results on the dataset Avenue and ShanghaiTech; A represents the result of Avenue and B is the result of ShanghaiTech. The pink background is the area where the real anomaly occurs; The blue curve represents the prediction anomaly score with the time series	105

List of tables

2.1	Unsupervised Video anomaly detection classification	33
2.2	Unsupervised learning-based video anomaly detection models	34
2.3	Weakly supervised algorithm classification	35
3.1	Dataset Overview: Nor and Abnor are normal and abnormal videos; Atype is the number of abnormal types; N/A denotes the number of normal videos / abnormal videos	56
3.2	The TCN classification performance analysis under the UCF-Crime	57
3.3	Accuracy test of current mainstream algorithms on the UCF-Crime dataset	58
3.4	Accuracy test of current mainstream algorithms on the ShanghaiTech dataset	58
3.5	Ablation study: Divided two datasets into four groups: I3D+MIL, C3D+MIL, I3D+TCN+MIL, C3D+TCN+MIL, to evaluate the TCN module.	59
3.6	The Study of Loss Function: Set different loss function combination modes to explore the impact of different loss functions	59
4.1	The accuracy of anomaly detection under different value of hyperparameters λ_f, λ_s ; the dataset used is UCSD(Ped2).	79
4.2	Quantitative comparison of the frame-level AUC-PR results of our COVAD method with the state-of-the-art models. (DFE is Double Fusion Framework; Unmask is a technique previously used for authorship verification in text documents ; TSC+sRNN is Temporally-coherent Sparse Coding stacked Recurrent Neural Network; CA is the Coordinated attention).	84
5.1	The result for Avenue dataset	102
5.2	The result for UCSD(ped2) dataset	102
5.3	The result for ShanghaiTech dataset	103
5.4	The Ablation Study : The Performance comparison of single-channel and various dual-channel models, blue arrow represents the propagation path of appearance features, and yellow arrow represents optical flow features	104
5.5	The Ablation Study: The impact of skip connections and co-teaching on the performance of dual-channel models	105

