UNIVERSITE MOHAMED KHIDER - BISKRA

FACULTÉ DES SCIENCES ET DE LA TECHNOLOGIE

DÉPARTEMENT DE GÉNIE ÉLECTRIQUE

SPÉCIALITÉ D'ÉLECTRONIQUE

Ref: .....

جامعة محمد خيضر – بسكرة
كلية العلوم والتكنولوجيا
قسم الهندسة الكهربائية
شعبة الإلكترونيك
المرجع:.....

Titre de la thèse:

Analyse du comportement des objets dans des environnements potentiellement dangereux

Par

FATMA GOUIZI

Une thèse soumise au Département de Génie Électrique en vue de l'obtention du diplôme de **Doctorat (3e Cycle)** en **Électronique (Biométrie et Télésurveillance)**

Membres du jury:

Président:	Pr. Zitouni Athmane	Prof	Université de Biskra
Superviseur:	Pr. Megherbi Ahmed Chaouki	Prof	Université de Biskra
Examinateur:	Pr. Ouafi Abdelkrim	Prof	Université de Biskra
Examinateur:	Dr. Benlamoudi Azedine	MCA	Université de Ouargla

2024/2025

MOHAMED KHIDER UNIVERSITY - BISKRA
FACULTY OF SCIENCE AND TECHNOLOGY
DEPT. ELECTRICAL ENGINEERING
DIVISION OF ELECTRONICS
Ref: .....

جامعة محمد خيضر – بسكرة
كلية العلوم والتكنولوجيا
قسم الهندسة الكهربائية
شعبة الإلكترونيك
المرجع: .....

Thesis title:

Analysis of the behavior of objects in potentially hazardous environments

By

FATMA GOUIZI

A thesis submitted to the Department of Electrical Engineering in candidacy for the Degree of **Doctorate (3rd Cycle)** in **Electronics (Biometrics and Surveillance)**.

Members of the jury:

President:	Pr. Zitouni Athmane	Prof	University of Biskra
Supervisor:	Pr. Megherbi Ahmed Chaouki	Prof	University of Biskra
Examiner:	Pr. Ouafi Abdelkrim	Prof	University of Biskra
Examiner:	Dr. Benlamoudi Azedine	MCA	University of Ouargla

2024/2025

Dedication

I dedicate this work to my family, whose unwavering support and love have been my greatest source of strength and inspiration.
First and foremost, I extend my heartfelt gratitude to my beloved parents, Houria and Salah. Their unwavering belief in me has kept my spirits and motivation. May Allah grant them health, happiness, and a long life.
To my dear brothers, Aissa and Mohammed, thank you for your constant encouragement and support. Your belief in me has been invaluable, and I sincerely appreciate all you have done to help me achieve my goals.
To my dear sisters, Meriem and Nour ElHouda, thank you for your unconditional love and constant presence.
Finally, I want to express my appreciation to my teachers and those who hold a special place in my heart.

ACKNOWLEDGEMENTS

First and foremost, I praise and thank Allah for giving me the energy and the will to finalize this thesis work. I'm incredibly grateful to my supervisor, Professor Megherbi Ahmed Chaouki, for his valuable advice and guidance, which largely contributed to the successful completion of this work.

My thanks also go to Professors Zitouni Athmane and Sbaa Salim for their commitment and guidance and to the jury members for agreeing to honor me with their presence and giving me valuable comments.

Finally, thanks to my parents and family for their encouragement and support.

Abstract

Object behavior analysis systems represent one of the most effective approaches to enhancing security, ensuring smoothness, and promoting safety in public spaces, roadways, and hazardous environments such as intersections, level crossings, and pedestrian crossings. Among the various techniques employed in behavior analysis, anomaly detection stands out as a particularly significant method. However, it presents considerable challenges due to the complexity of video environments, difficulties associated with object detection, and the ambiguous definitions of various types of anomalies.

This thesis explores behavior analysis in video, progressing from object detection to classifying these objects based on their states, ultimately determining whether they exhibit normal or disturbed behavior. Each process step is studied in detail, incorporating a comprehensive review of existing techniques, including both classical methods and those based on deep learning. The study also addresses various scientific and technical challenges and evaluation metrics. The proposed methodologies and the results obtained from experiments conducted on various datasets are presented at the end of the study.

The primary goal of the initial step is to robustly detect and localize objects while considering environmental constraints, such as variations in lighting, weather conditions, and background movement. To achieve this, we introduce a novel strategy that utilizes a novel background subtraction architecture. Our network architecture is conceptualized as a nested network, which is based on residual autoencoder blocks featuring enhanced. This structure, referred to as "Nested-net," allows residual autoencoders to improve feature generalization by extracting multi-scale features at each level, thereby effectively addressing multiple challenges.

The subsequent step focuses on detecting objects' abnormal behaviors by introducing a new approach that employs a convolutional autoencoder to extract spatial and temporal representations from the appearance and motion of object patches. This is achieved through data transformation techniques aimed at enhancing feature learning and classification accuracy.

Ultimately, the results obtained from various datasets demonstrate consistent performance, surpassing existing state-of-the-art techniques. This not only validates the proposed method's effectiveness but also makes it a highly efficient and recommended solution for practical applications, thereby enhancing the safety and security of public spaces.

Keywords: Object detection, background subtraction, behavior analysis, deep learning, video anomaly detection, video surveillance.

Résumé

Les systèmes d'analyse du comportement des objets constituent l'un des moyens les plus efficaces pour garantir la sécurité, la fluidité et la sûreté dans les espaces publics, les routes et les environnements à risque, tels que les intersections routières, les passages à niveau et les passages piétons, entre autres. La détection des anomalies est l'une des approches les plus étudiées dans l'analyse comportementale, mais elle demeure une tâche complexe en raison de la nature dynamique des environnements vidéo et des défis liés à la détection d'objets et à la définition ambiguë de l'anomalie.

Ce travail de thèse s'intéresse à l'analyse des comportements dans les vidéos, en partant de la détection des objets jusqu'à leur classification selon leur état, afin de déterminer s'ils présentent un comportement normal ou anormal. Nous avons étudié chaque étape de manière indépendante. Cette étude couvre la majorité des techniques actuelles, qu'il s'agisse de méthodes classiques ou basées sur l'apprentissage profond, en abordant également les obstacles scientifiques et techniques, ainsi que les différentes métriques d'évaluation. Enfin, nous présentons les approches proposées et les résultats obtenus sur plusieurs ensembles de données.

L'objectif de la première étape est de détecter de manière robuste les objets et de localiser leur position, en tenant compte des contraintes environnementales telles que les variations d'éclairage, les changements climatiques, les arrière-plans en mouvement, etc. Pour cela, nous avons proposé une stratégie efficace reposant sur une nouvelle architecture de soustraction de fond dans les vidéos. Notre réseau est conçu sous forme d'un réseau « imbriqué », basé sur des blocs d'auto-encodeurs résiduels intégrant davantage de connexions de type skip, d'où l'appellation "imbriqué". Ces auto-encodeurs résiduels permettent une meilleure généralisation des caractéristiques en extrayant davantage de descripteurs multi-échelles à chaque niveau, ce qui permet de relever de nombreux défis.

La seconde étape vise à détecter les comportements anormaux des objets via une nouvelle approche utilisant un auto-encodeur convolutif permettant d'extraire des représentations spatio-temporelles à partir de l'apparence et du mouvement des objets, tout en utilisant des transformations de données pour améliorer l'apprentissage des caractéristiques et la classification.

Enfin, les résultats obtenus sur plusieurs ensembles de données ont démontré des performances stables et une supériorité par rapport aux techniques existantes, rendant la méthode proposée plus efficace et fortement recommandée.

Mots-clés : détection d'objets, soustraction de fond, analyse de comportement, apprentissage profond, détection d'anomalies vidéo, surveillance vidéo.

الملخص

تعتبر أنظمة تحليل سلوك الأشياء إحدى الطرق الأكثر فعالية لتوفير الأمن، والسلاسة، والسلامة في الأماكن العامة، والطرقات، والبيئات الخطرة، مثل تقاطعات الطرق، والمعابر المستوية، ومعابر المشاة، وغيرها. يعد اكتشاف الحالات الشاذة أحد الأساليب الأكثر اهتمامًا أثناء تحليل السلوك، كما أنه من المهام الصعبة نظرًا لتعقيد بيئة الفيديو والتحديات المرتبطة باكتشاف الأشياء والتعريف الغامض لنوع الشذوذ.

تتناول هذه الأطروحة دراسة تحليل السلوكيات في الفيديو، بدءًا من اكتشاف الأشياء إلى تصنيفها حسب حالتها، ومن ثم تحديد ما إذا كانت طبيعية أو مضطربة. حيث قمنا بدراسة كل خطوة على حدة، تشمل هذه الدراسة معظم التقنيات الحالية مثل الأساليب التقليدية والأساليب القائمة على التعلم العميق، بالإضافة إلى العقبات العلمية والتقنية ومقاييس التقييم المختلفة. وانتهاءً بالمناهج المقترحة والنتائج المتحصل عليها على قواعد بيانات مختلفة.

الغرض من الخطوة الأولى هو الكشف القوي عن الكائنات وتحديد موقعها من خلال النظر في القيود المفروضة على التحديات الموجودة في البيئة، مثل تغير الإضاءة، وتغيرات الطقس، وتحرك الخلفية، وغيرها. للقيام بذلك، اقترحنا استراتيجية قوية باستخدام بنية جديدة لطرح خلفية الفيديو. تُمثل بنية شبكتنا على أنها شبكة متداخلة، تعتمد على كتل التشفير التلقائي المتبقية مع المزيد من الاتصالات التخطيطية، ولهذا سُميت "متداخلة". يمكن لأجهزة التشفير التلقائي المتبقية تحسين تعميم الميزات عن طريق استخراج المزيد من الميزات متعددة النطاق في كل مستوى، وبالتالي مواجهة العديد من التحديات.

أما الخطوة الثانية، فتتيح اكتشاف السلوكيات غير الطبيعية للأشياء من خلال اقتراح نهج جديد يستخدم مشفرًا ذاتيًا ملتفًا لاستخراج التمثيلات المكانية والزمانية من مظهر وحركة بقع الكائنات، باستخدام تحويل البيانات لتحسين تعلم الميزات والتصنيف. وفي نهاية المطاف، أظهرت النتائج التي تم تحقيقها عبر مجموعات البيانات المختلفة أداءً ثابتًا وتفوقًا على التقنيات الحديثة الموجودة، مما يجعل الطريقة المقترحة أكثر كفاءة وموصى بها بشدة.

الكلمات المفتاحية: كشف الكائنات، طرح الخلفية، تحليل السلوك، التعلم العميق، اكتشاف الشذوذ في الفيديو، مراقبة الفيديو.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Acronyms

AMSRC	Appearance-Motion Semantics Representation Consistency
ANN	Artificial Neural Network
AP	Average Precision
AUC	Area Under the Curve
BN	Batch Normalization
BSGAN	GAN based Background Subtraction
BSPVGAN	Parallel vision and Bayesian GANs based Background Subtraction
BSUV-Net	Background Subtraction of Unseen Videos Network
CA	Channel Attention
CNN	Convolutional Neural Network
CVAE	Conditional Variational Auto-Encoder
DAE	Denoising Auto-Encoder
DBNs	Dynamic Bayesian Networks
DCNNs	Deep Convolutional Neural Networks
DLA	Deep Layer Aggregation
DNNs	Deep Neural Networks
DPM	Deformable Part Model
EM	Expectation Maximization
FCESNet	Fully Convolutional Encoder-decoder Spatialtemporal Network
FCN	Fully Convolutional Networks
FgSegNet	Foreground Segmentation Network
FN	False Negative

FP False Positive

FPM Feature Pooling Module

FPR False Positive Rate

GAN Generative Adversarial Network

GMM Gaussian Mixture Model

GP Genetic Programming

GRU Gated Recurrent Unit

HMMs Hidden Markov Models

HOG Histogram of Oriented Gradients

HSV Hue Saturation Value

ILSVRC ImageNet Large Scale Visual Recognition Challenge

IOU Intersection Over Union

KDE Kernel Density Estimation

KLD Kullback Leiber Divergence

KLT Kanade-Lucas-Tomasi

LBP Local Binary Pattern

LBSP Local Binary Similarity Patterns

LRT Likelihood Ratio Test

LSTM Long Short Term Memory

mAP mean Average Precision

MFCN multiscale fully convolutional network

ML-MemAE-SC Multi Level Memory augmented Auto-Encoder with Skip Connections

MoG Mixture of Gaussians

MRAM Motion-Refined Attention Module

MRF Markov Random Field

MSE Mean Square Error

MS COCO Microsoft Common Objects in Context

MTPA	Multi-scale Temporal Pixel Aggregation
NMS	Non Maximum Suppression
OCELM	One Class Extreme Learning Machine
OCSVM	One Class SVM
OF	Optical Flow
OGAM	Object-Guided Attention Module
OMAE	Object-centric Memory-guided Auto-Encoder
OSIN	Object-centric Scene Inference Network
PAA	Probably Approximately Admissible
PAWCS	Pixel-based Adaptive Word Consensus Segmenter
PBAS	Pixel-Based Adaptive Segmenter
PCA	Principal Component Analysis
R-CNN	Region-based Convolutional Neural Network
RANSAC	RANdom Sample Consensus
RBDR	Region-Based Detection Rate
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RNNs	Recurrent Neural Networks
RM-AE	Residual Micro-AutoEncoder
RoIs	Regions of Interest
RPN	Region Proposal Network
SACON	SAmple CONsensus
SDE	Scene Dependent Evaluation
SE	Squeeze and Excitation
sEnDec	Slow Encoder-Decoder
SFPM	Spatial Feature Pooling Module
SIE	Scene Independent Evaluation

SIFT Scale-Invariant Feature Transform

SOBS Self-Organizing Background Subtraction

SSD Single Shot MultiBox Detector

SSIM Structural Similarity Index Measurement

STIT Spatial Temporal Information Transmission

SuBSENSE Self-Balanced SENsitivity SEgmenter

SVM Support Vector Machine

T2-FGMM Type-2 Fuzzy Gaussian Mixture Model

TBDR Track-Based Detection Rate

TCNN Triplet CNN

TP True Positive

TPR True Positive Rate

VAD Video Anomaly Detection

VIBE VIsual Background Extractor

VJ Viola and Jones

VOC Visual Object Classes

YOLO You Only Look Once

CHAPTER 1

INTRODUCTION

Contents

1.1 Context

Computer vision applications give remarkable importance to monitoring systems, essential to raising the safety and security of human mobility and goods transportation, thus significantly contributing to overall quality. It has become necessary to scrutinize the actions of individuals and vehicles to distinguish between typical and abnormal behaviors. However, security personnel need assistance in monitoring screens for extended periods to detect and proactively respond to potentially dangerous situations [1, 2].

In recent years, researchers and authorities have collaborated to develop intelligent traffic surveillance systems. These systems use video cameras and advanced algorithms to identify and monitor moving objects, including vehicles, pedestrians, and other objects. Identifying and monitoring moving objects is based on inferring "what is happening" within the scene, including recognizing or classifying behavior by collecting attributes such as posture, movement, gestures, and more [3].

The analysis of abnormal behavior is a significant field that involves studying and understanding the behavioral patterns of individuals or objects to detect any unusual or suspicious events and activities. The evolution of behavioral analysis began with the manual observation of human behavior and has since progressed towards developing sophisticated automated systems founded on artificial intelligence and machine learning methodologies, which have greatly improved the precision and speed of behavioral analysis. This signifies a noteworthy advancement in the ongoing process of improving the methods used in this field.

The process of detecting abnormal object behavior demonstrates proficiency in extracting valuable insights, including appearances, paths, human activities, and interactions. A consistent approach across these applications involves a two-step process: initially detecting and tracking moving objects, followed by analyzing their behaviors [4] to identify suspicious activities, anomalies, and noteworthy events.

Behavior analysis systems must possess the capability to describe and identify behaviors associated with various concepts, as outlined by [5]:

- Property refers to an attribute of an object, such as its speed, trajectory, and direction.
- State: represents a situation defining one or more objects at a specific time.

- Event: Signifies an object's situation from one instant to another.
- Situations: Comprising a mix of sub-scenarios, states, and events.

Examples of abnormal object behavior are illustrated in Figure 1.1.

Figure 1.1: Sample images from the ShanghaiTech Campus Database. From left to right: running, fighting, and non-allowed objects.

The traditional structure of Object Behavior Analysis (OBA), as illustrated in 1.2, consists of three main stages: object detection and recognition, object tracking, and analysis of their behavior. Methods that follow these steps typically extract discriminative features and use traditional machine-learning techniques to analyze the object trajectory in the video to detect abnormal behavior.

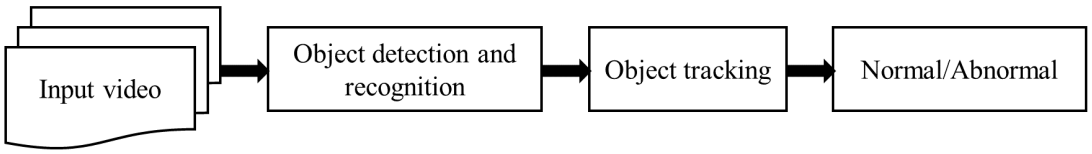


Figure 1.2: Traditional process of Object Behavior Analysis.

The first phase focuses on robustly detecting, localizing, and recognizing the objects while accommodating environment-related challenges such as varying illumination, weather conditions, and occlusions. While the second phase focuses on tracking objects extracted from the first step, modern deep learning-based methodologies have shifted away from object tracking due to occlusion issues in

crowded scenes. Subsequently, this leads to its exclusion from our dissertation research. The third and final stage is dedicated to analyzing the behavior of the detected objects to identify potentially dangerous and abnormal situations.

The advanced architecture for Object Behavior Analysis (OBA), as illustrated in 1.3 employs deep learning and advanced methods based on just two crucial stages, as : detecting objects and subsequently analyzing their behavior to infer anomalies.

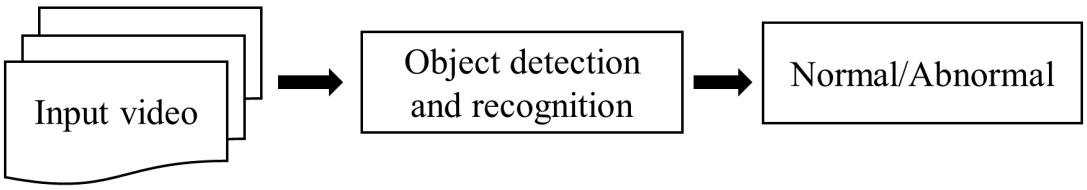


Figure 1.3: Advanced process of object behavior analysis.

1.1.1 Object detection

One important and challenging topic in computer vision is object detection. It involves locating instances of specific objects in images and videos, such as people, animals, and cars. Techniques such as frame differencing, optical flow, and background subtraction are used to detect changes or movements, which aids in effective object identification and localization in natural images. Object detection integrates machine learning (ML) and deep learning (DL) models with segmentation methods and deep convolutional neural networks (CNNs) to represent image features effectively. Background subtraction is considered the most significant technique, balancing computation time and detection quality [6–10].

1.1.2 Abnormal behavior detection

Behavior analysis in surveillance systems is a critical and complex area of computer vision research. Detecting irregular or abnormal behavioral patterns involves recognizing the distinct characteristics of objects' actions and interactions, which are highly dependent on the context and movement patterns of the objects within a scene.

A major challenge in defining abnormal behavior lies in the variability of criteria and context across different scenarios, making it difficult to establish a universal definition.

By leveraging machine learning and computer vision algorithms, it is possible to analyze the movements and interactions of objects in scenes captured by surveillance cameras. This facilitates the detection of suspicious activities, identification of behavioral patterns, and extraction of valuable insights that enhance the effectiveness of security and surveillance systems.

1.2 Problematics

The field of abnormal object behavior has seen significant advancements and has drawn considerable interest from the research community because of its diverse range of applications. For many decades, researchers have been extensively exploring this field. Numerous techniques have been introduced, including traditional and deep learning methods. Despite the creation of various approaches, each has a unique set of limitations, and numerous technical and scientific hurdles persist. The primary issues include:

- **Scene complexity and object detection constraints:** The efficiency of the object behavior analysis approach always depends on the robustness of the object detection techniques. Most of these methods are affected by the complexity of the scene or the limitations imposed by the environment, such as illumination changes, occlusions, shadows, or dynamic backgrounds like waves and moving trees. Dealing with these challenges is often quite tricky.

- **Addressing the Diversity of Abnormal Behavior:** How can we effectively analyze a wide range of abnormal behaviors, including incidents such as fighting, violence, theft, and dangerous behaviors? Unlike action recognition, anomalies lack a distinct definition that allows for clear differentiation from regular events [11]. Anomalies typically encompass a broad spectrum of activities, and their characterization may vary across different applications and datasets.

- **Limited Availability of labeled data:** Abnormal object behavior data is often scarce and may need more quantity for training and evaluation purposes. Currently, most of these methods require a labeled dataset containing

regular events, which restricts their applicability because it necessitates human intervention for continuous system retraining.

- In conventional surveillance systems, the capacity to identify and proactively address suspicious behaviors is often lacking in public places and hazardous environments such as road and rail intersections and pedestrian crossings.

1.3 Aims and objectives

The primary goal of the thesis is to develop a deep learning model designed explicitly for detecting abnormal object behavior in videos. Furthermore, the thesis outlines the following objectives:

The primary aim is to develop an intelligent video surveillance system tailored for extracting and analyzing object behavior in hazardous environments such as road intersections, level crossings (road/rail intersections), and pedestrian crossings. The expected system is divided into three main stages: object detection, object recognition, and extraction and analysis of object behavior.

The first stage involves robustly detecting and localizing objects while considering constraints such as changes in lighting, weather conditions, and occlusions.

The second stage involves object recognition. Enhancing the behavior analysis with prior knowledge is essential to improve their performance.

The third stage is dedicated to analyzing the object's behavior to detect potentially dangerous situations through spatiotemporal analysis.

1.4 Contributions

Accurate detection of moving objects is crucial for identifying abnormal object behavior in video clips, particularly considering the complexities inherent in the scene, as highlighted in the sections above.

This thesis centers on two principal components. The first component involves the exploration and advancement of background subtraction techniques, with key contributions outlined as follows:

- Undertaking a comprehensive review of object detection approaches, especially background subtraction methodologies.

- Introducing a novel supervised deep learning model developed specifically for background subtraction, showcasing a new architecture structured as a nested network with multiple skip connections between the proposed residual mini-autoencoders.
- Conducting a comparative evaluation against recent state-of-the-art methods.

The second aspect represents the exploration and enhancement of abnormal object behavior detection, with pivotal contributions delineated as follows:

- Proposing a novel unsupervised feature learning by introducing a novel method for generating data that utilizes object-centric-based irregularity creation to ensure regular feature learning without compromising inference speed. Our objective is to enhance the learning of regular features by integrating spatial and temporal transformation techniques during training, incorporating irregular information extracted from regular object boxes.
- Introducing a new frame-based abnormal behavior detection approach employing a 2D autoencoder with channel attention to learn spatial and temporal representations from the object patches.
- Proposing a new training and feature learning strategies for object-based abnormal behavior detection

1.5 Thesis Outlines

Our research in this thesis presents a coherent outline as follows:

- In Chapter 2, we conducted a state-of-the-art review in object detection with a focus on the background subtraction aspect, emphasizing key concepts, utilized approaches, and challenges encountered in this field.
- In Chapter 3, we comprehensively review the field of abnormal behavior detection in video surveillance. A review of several algorithms will be discussed.
- Chapter 4 outlines in detail the critical steps of our proposed method in the background subtraction domain, starting with feature extraction, moving

on to modeling, and finishing with a comparative evaluation against the recent state-of-the-art techniques.

- Chapter 5 is divided into two parts for clarity and in-depth exploration of our proposed methodologies where : Part 1: offers a comprehensive overview of our proposed frame-based anomaly detection and abnormal behavior identification approach. It begins by outlining the feature extraction and transformation processes, followed by the modeling phase. Part 2: delves into our proposed method for detecting abnormal behavior at the object level. This section explains the feature extraction and transformation processes, thoroughly describes our methodologies, and concludes with an experimental evaluation of the proposed approaches.
- In Chapter 6, we provide an overall conclusion regarding our work and a concept for future work.

CHAPTER 2

RELATED WORK ON OBJECT DETECTION

Contents

2.1 Introduction

Object detection is a prominent subject in scientific research, notably in computer vision applications, which involves identifying and locating instances of objects in images and videos. Our emphasis in object detection revolves around two key aspects: instance detection and foreground segmentation.

This chapter thoroughly summarizes the literature on instance detection and foreground segmentation. It summarizes the most commonly used datasets and evaluation metrics. It also provides a detailed summary of previous works, the most recent developments in this field, and a comparison of several available methodologies. The chapter also presents several limitations and general opinions in the field before concluding with a summary.

2.2 Instance Detection and segmentation

Object detection systems are designed to accurately recognize objects (such as people, cars, and bicycles) as illustrated in figure 2.1 and determine their location and coordinates by providing approximate localization and exact extent using bounding boxes [12, 13].

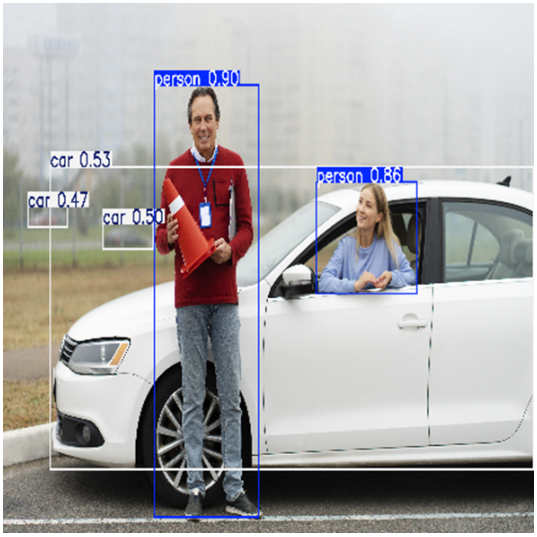


Figure 2.1: Example of instance detection.

Most object detectors that employ machine learning and deep learning techniques encounter several challenges, including difficulties in detecting smaller

objects, errors resulting from an imbalance between foreground and background classes, as well as the requirement of large datasets and significant computational power [9, 13, 14].

Efficiency and scalability challenges in object detection arise from the computational complexity of handling multiple classes, locations, and scales within images. There is also a need to manage high data rates and previously unseen objects without relying on extensive manual annotations [13].

2.2.1 Datasets

2.2.1.1 The PASCAL VOC (Visual Object Classes)

The PASCAL Visual Object Classes (VOC) dataset series[1] [15], developed from 2005 to 2012, serves as a benchmark for object detection and classification algorithms. It includes 20 common visual object categories across 11,000 images, divided into major groups such as animals, vehicles, people, and household items. These datasets feature over 27,000 labeled object instances, with nearly 7,000 having detailed segmentations as illustrated in figure 2.2. The PASCAL VOC datasets are evaluated using the interpolated average precision metric, focusing on accurate positive detections, false positives, and missed detections. Each year's dataset retains previous images, allowing for consistent year-to-year performance comparisons. Despite being foundational in establishing standardized evaluations for recognition algorithms, PASCAL VOC has been overshadowed by larger datasets like ImageNet and MS COCO in recent years.

Figure 2.2: Instance segmentation samples from Pascal VOC 2012 dataset.

[1] http://host.robots.ox.ac.uk/pascal/VOC/

2.2.1.2 ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

ImageNet[2] [16] is a crucial large-scale benchmark dataset with numerous subsets and challenges to measure the performance of object detection and classification algorithms. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Challenge is based on ImageNet and expands on PASCAL VOC's aims, offering a consistent evaluation for detection algorithms with a much wider variety of object classes and images. Specifically, ILSVRC2014 contains subsets of images around 450k for training, 20k for validation, and 40k for test images. ImageNet1000, a subset of ImageNet, features 1.2 million images across 1000 object categories, serving as a standard benchmark for image classification challenges. ImageNet ultimately encompasses tens of millions of annotated images organized semantically. This large-scale and diverse dataset is significantly more comprehensive and accurate than other image datasets.

2.2.1.3 Microsoft Common Objects in Context (MS COCO) dataset

The Microsoft Common Objects in Context (MS COCO) dataset [3] [17] is meant to recognize and segment objects in real scenarios. It includes 91 major object categories, with 82 comprising more than 5,000 labeled instances. The collection includes 2.5 million annotated instances spread over 328,000 images. Unlike ImageNet, which often has vast, well-centered objects, MS COCO presents objects in complex, cluttered scenes, more reflective of real-world scenarios. It emphasizes a wide range of object scales, including small objects, and provides richer contextual information with an average of 7.7 objects per image compared to 3.0 in ImageNet and 2.3 in PASCAL VOC. Additionally, MS COCO uses a comprehensive evaluation metric, measuring mean average precision over a range of thresholds [0.5, 0.95] and separately evaluating small, medium, and large objects to ensure accurate detector performance across various object sizes. MS COCO is a stringent benchmark for modern object detection and segmentation tasks.

2.2.2 Evaluation metrics

To emphasize the effectiveness of object detection methods, many papers such as [18, 19] have explained the performance metrics for evaluating pre-

[2]https://www.image-net.org/challenges/LSVRC/
[3]https://cocodataset.org/

trained object detection-based algorithms. The most frequently utilized metrics for measuring detection accuracy are Average Precision (AP) (eq: 2.4) and mean Average Precision (mAP) (eq: 2.8). Before defining AP, we need to establish some fundamental concepts that contribute to its definition. In order to determine what constitutes a "correct detection" it is important to employ Intersection Over Union (IOU)(eq: 2.1), which is a measurement based on the Jaccard Index, used to quantify the overlapping area between the objects being compared where detection is classified as correct if its IOU meets or exceeds the threshold t; otherwise, it is considered incorrect.

$$\text{IOU}(b, bg) = \frac{\text{area}(b \cap bg)}{\text{area}(b \cup bg)} \tag{2.1}$$

- True Positive (TP) represents accurately detecting a ground-truth bounding box.
- False Positive (FP) signifies an erroneous identification, either by detecting a nonexistent object or misplacing an object inaccurately.
- False Negative (FN) occurs when a ground-truth bounding box goes undetected.

Precision (Pr) (eq: 2.2) and Recall (Re) (eq: 2.3) are important metrics for evaluating model performance. Precision quantifies the accuracy of positive predictions, whereas recall (Re) assesses the model's ability to detect all relevant cases.

$$Pr = \frac{TP}{TP + FP} \tag{2.2}$$

$$Re = \frac{TP}{TP + FN} \tag{2.3}$$

A suitable detector should maintain high precision and recall, indicated by a high Area Under the Curve (AUC), which is smoothed using AP_k or AP_{all} interpolation to remove zigzags. The AP_k interpolation involves averaging the maximum precision values at $k = 11$ equally spaced recall levels from 0 to 1.

$$AP_k = \frac{1}{k} \sum_{Re \in \{0, 0.1, \ldots, 0.9, 1\}} Pr_{interp}(Re) \tag{2.4}$$

$$Pr_{interp}(Re) = \max_{\tilde{Re}:\tilde{Re}\geq R} Pr(\tilde{Re}) \tag{2.5}$$

$$AP_{all} = \sum_{n}(Re_{n+1} - Re_n)Pr_{interp}(Re_{n+1}) \tag{2.6}$$

$$Pr_{interp}(Re_{n+1}) = \max_{\tilde{Re}:\tilde{Re}\geq Re_{n+1}} Pr(\tilde{Re}) \tag{2.7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{2.8}$$

2.2.3 Related work

Several approaches have been presented for real-time object detection. This section covers both classic and deep learning approaches.

2.2.3.1 Traditional methods

Traditional object detection methods lean on low-level cues like color, edges, texture, and gradients. Previous approaches in this area have used hand-crafted features due to the need for advanced image representation techniques and limited computational resources. Techniques such as the Viola-Jones (VJ) detector, HOG, and SIFT have been used to identify objects. Despite some limitations, traditional methods have achieved remarkable success, especially with the PASCAL VOC dataset, by generating feature descriptions that accurately identify regions of interest in images.

The Viola and Jones (VJ) Detectors, as outlined in the publication by Viola and Jones [20], were the first to achieve real-time detection of human faces; their detector surpassed other algorithms in terms of speed while maintaining similar levels of detection accuracy. The VJ detector uses a straightforward sliding window technique to examine an image at all possible locations and scales to detect human faces. Despite its simplicity, the underlying calculations exceeded the computational capabilities of the time. Another approach [21] introduces a machine-learning technique for rapid visual object detection. It includes the

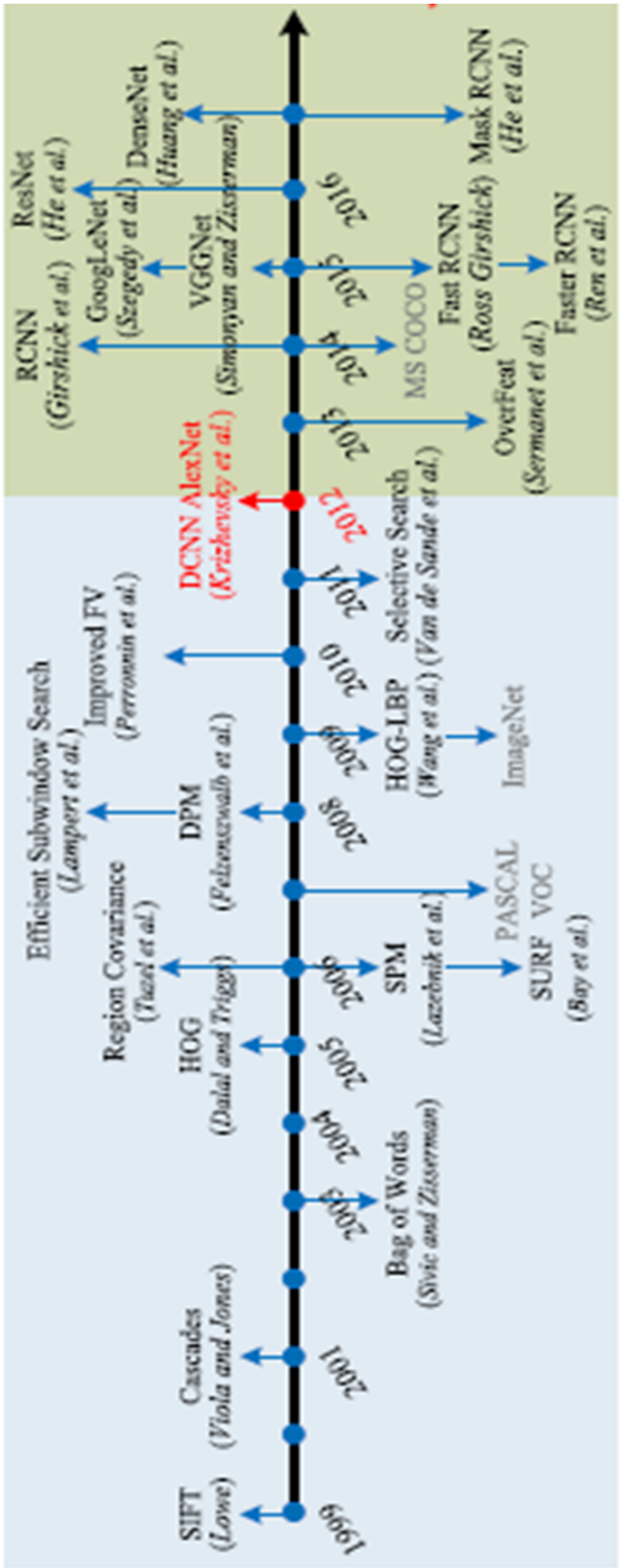


Figure 2.3: The evolution of object detection and recognition methods [13].

integral image representation for rapid feature computation, AdaBoost for efficient feature selection, and a cascade method to merge classifiers for background region elimination. This system achieves high detection rates comparable to existing systems and operates in real-time applications without image differencing or skin color detection.

The Scale-Invariant Feature Transform (SIFT) algorithm, created by David Lowe in 1999 [22] and further refined in 2004 [23], is specifically designed to identify and describe local image features in an invariant manner to scale, making it particularly suitable for object recognition. SIFT employs a multi-stage filtering process to detect key points in scale space by locating the most extreme points of a difference-of-Gaussian function. Each key point generates a feature vector that captures the scale-space coordinates of the local image region. Furthermore, by blurring the image gradient orientations, SIFT offers some tolerance to local variations such as affine or 3D projections. The resulting feature vectors, known as SIFT keys, are highly distinctive and resilient to changes in position, scale, rotation, lighting, vibration, and minor perspective shifts.

In 2005, N. Dalal and B. Triggs [24] proposed the Histogram of Oriented Gradients (HOG) feature descriptor to enhance earlier methods like the scale-invariant feature transform (SIFT) and shape contexts. HOG is specifically designed for object detection, working on a dense grid of uniformly spaced cells, with overlapping local contrast normalization applied across blocks. This approach ensures that HOG is resistant to geometric and photometric transformations, apart from object orientation, making it highly effective for detecting pedestrians. The HOG detector resizes input images to detect objects of different sizes while maintaining a constant detection window size. This technique has profoundly impacted various object detectors and computer vision applications because it can ensure reliable object detection under diverse conditions. It achieves this by dividing images into small interconnected regions called cells, creating a histogram of gradient directions for each cell, and combining these histograms into a feature descriptor. HOG's effectiveness is significantly enhanced when combined with linear SVM for human detection. This combination enables HOG to utilize fine-scale gradients, precise orientation binning, coarse spatial binning, and high-quality local contrast normalization, further highlighting its importance. The significance of HOG is underscored by its performance on a challenging dataset of human images.

A Deformable Part Model (DPM) [25] for object detection demonstrates signifi-

cant performance improvements over previous methods. P. Felzenszwalb initially proposed DPM in 2008 as an expansion of the HOG detector, using a strategy of breaking down objects into their constituent parts using "divide and conquer." The system includes new discriminative training methods, such as a margin-sensitive approach for identifying challenging negative examples and a latent SVM framework for utilizing more latent information in future improvements. DPM's effectiveness has been confirmed through its successes in the VOC detection challenges.

An improved DPM featured in [26] describes a cascade detection method that centers on star-shaped models. It presents an approach based on eliminating partial hypotheses to speed up object detection while maintaining accuracy. The method prunes partial hypotheses using a series of thresholds and introduces the idea of Probably Approximately Admissible (PAA) thresholds.

2.2.3.2 CNN-based methods

The progress in deep learning and neural networks has resulted in the creation of practical object detection models like R-CNN, YOLO, SSD, and others. These models are designed to execute tasks such as creating bounding boxes, estimating class probabilities, and determining confidence scores. These features render them indispensable in several natural environements. This part provides a brief overview of different object detection algorithms, including region-based approaches like R-CNN, Fast R-CNN, and Faster R-CNN, as well as classification-based methods like SSD, YOLOv1, YOLOv2, YOLOv3, and so on.

Region-based Convolutional Neural Network (R-CNN) : RCNN is a well-known technique for detecting objects introduced by Girshick *et al.* in 2013 [27]. Region-based Convolutional Neural Network (R-CNN) consists of three primary parts: region extractor, feature extractor, and classifier. It utilizes a selective search algorithm to divide the image, grouping nearby pixels based on color, texture, and intensity, resulting in roughly 2000 region proposals per image. These proposals are resized to a fixed size and fed into a pre-trained Convolutional Neural Network (CNN), resulting in an N-dimensional feature vector from each proposal. Subsequently, these feature vectors are classified using a Support Vector Machine (SVM), refining the localization of bounding boxes with four offset values to improve detection accuracy. The R-CNN approach is notable for utilizing multi-

layer convolutional networks to generate highly distinctive features for classifying image regions and producing bounding boxes or pixel-level segmentation masks.

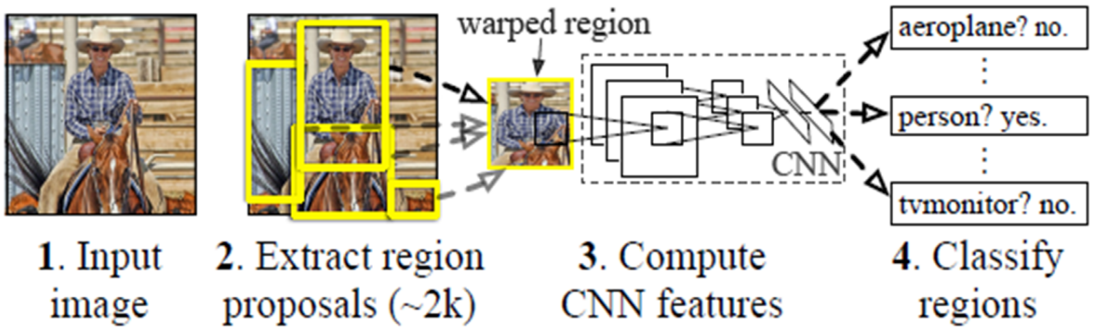


Figure 2.4: Overview of R-CNN object detection system [27].

Fast R-CNN : Proposed by Ross Girshick [28], Fast R-CNN addresses the computational efficiency and memory usage limitations of R-CNN. Unlike R-CNN, which independently computes feature vectors for each region proposal, Fast R-CNN processes the whole image through convolutional and max-pooling layers to generate a convolutional feature map. Regions of Interest (RoIs) are extracted from this feature map and transformed into fixed-length feature vectors using a RoI pooling layer. These feature vectors then enter fully connected layers, which split into two branches: one uses a softmax classifier to predict object classes, and the other outputs offset values to refine bounding boxes. This approach eliminates the need for separate training of multiple models, significantly improving speed and memory efficiency. Fast R-CNN is reported to be considerably more accurate, achieving faster training and testing times and exhibiting higher mean Average Precision (mAP).

Faster R-CNN : Faster R-CNN, introduced by Shaoqing Ren and others [29] in 2015, enhances the efficiency of real-time object detection by integrating a Region Proposal Network (RPN) into the architecture. This method starts by inputting an entire image into a deep convolutional network to produce a convolutional feature map. A mini-network, utilizing an n × n block from this feature map, generates region proposals through a sliding window mechanism. This mini-network includes two fully connected layers: a regression layer that outputs 4

× k encoded coordinates for k boxes and a classification layer that outputs 2 × k probability estimates to determine if each proposal contains an object. These proposals are parameterized relative to k anchor boxes of various sizes and aspect ratios. The RPN is a fully convolutional network that forecasts object boundaries and objectness scores at every position. This permits nearly cost-free region proposals by sharing full-image convolutional features with the detection network. This integration enables high-quality region proposals, which are then utilized by Fast R-CNN for detection, significantly increasing the efficiency and accuracy of object detection. The unified network with shared convolutional features utilizes the "attention" mechanism to guide the network on where to focus, streamlining the object detection procedure.

Single Shot MultiBox Detector (SSD) : The Single Shot MultiBox Detector (SSD) [30] is an efficient object detection technique that performs classifying and localizing objects in one pass, distinguishing itself from region proposal algorithms like Faster R-CNN. To generate feature maps from the input image, SSD uses the VGG-16 model architecture and utilizes 3 × 3 convolutional filters for predicting both bounding boxes and class scores simultaneously. By processing feature maps at various scales, SSD can effectively detect objects of different sizes, with lower-resolution maps capturing larger objects and higher-resolution maps detecting smaller ones. Default bounding boxes, with different aspect ratios to cover a broad spectrum of real-world scenarios, are compared with ground truth boxes using **IoU!** (**IoU!**). The Non Maximum Suppression (NMS) eliminates unnecessary detections, ensuring that the final output contains the most relevant bounding boxes. This multi-scale approach allows SSD to maintain high detection quality without compromising speed, making it faster than YOLO and as accurate as region-based detectors like Faster R-CNN.

You Only Look Once (YOLO) : You Only Look Once (YOLO) [31] revolutionized object detection by simplifying the process into a single step, unlike previous methods like R-CNN. YOLO processes an entire image in one step, achieving impressive speed and performance. YOLOv1 splits an input image into an s × s grid, and each grid cell serves as the basis for detecting objects. YOLOv2 improved over YOLOv1 by addressing localization and recall limitations. YOLOv3 [32] further enhances detection capabilities. YOLO's approach differs from region-based

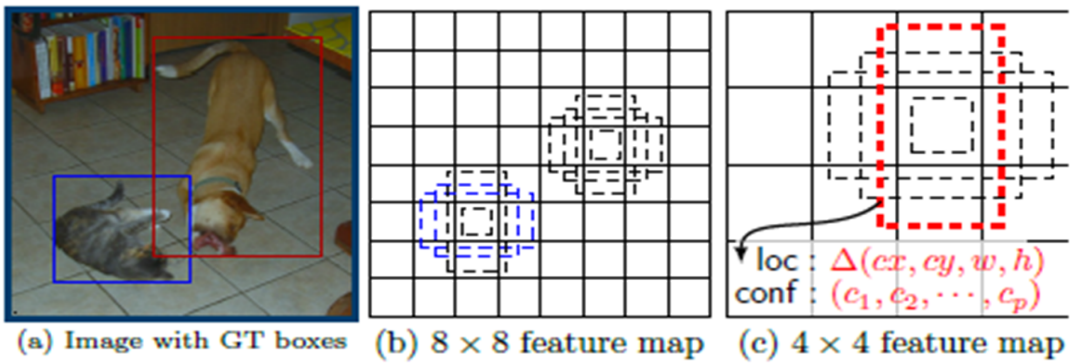


Figure 2.5: Overview of SSD object detection framework [30].

methods like Faster R-CNN, where it can struggle with localization errors, particularly for small objects and scenes with many objects. YOLOv2 and YOLO9000 [33] include improvements like batch normalization and k-means anchor boxes. YOLO9000 can detect over 9000 object categories in real time.

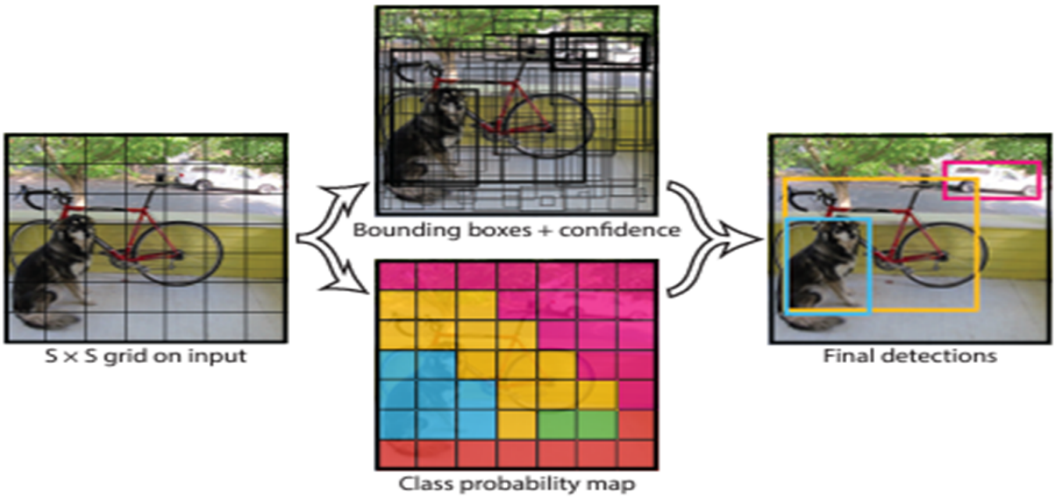


Figure 2.6: Overview of YOLO object detection model [31].

2.3 Foreground segmentation

Foreground detection is a prominent area of study in scientific research, especially in computer vision. Numerous techniques have been developed for this purpose, with the majority of object detection approaches relying on background

subtraction methods. Background subtraction is widely regarded as an efficient technique and plays a critical role in various computer vision tasks, such as object localization and behavior recognition [34, 35].

The core principle of this method involves creating a background model that aligns with the scene's characteristics based on pixel or region-level observations [36]. This is done by subtracting the static elements of the scene, referred to as the background, from the current frame (see figure 2.7).

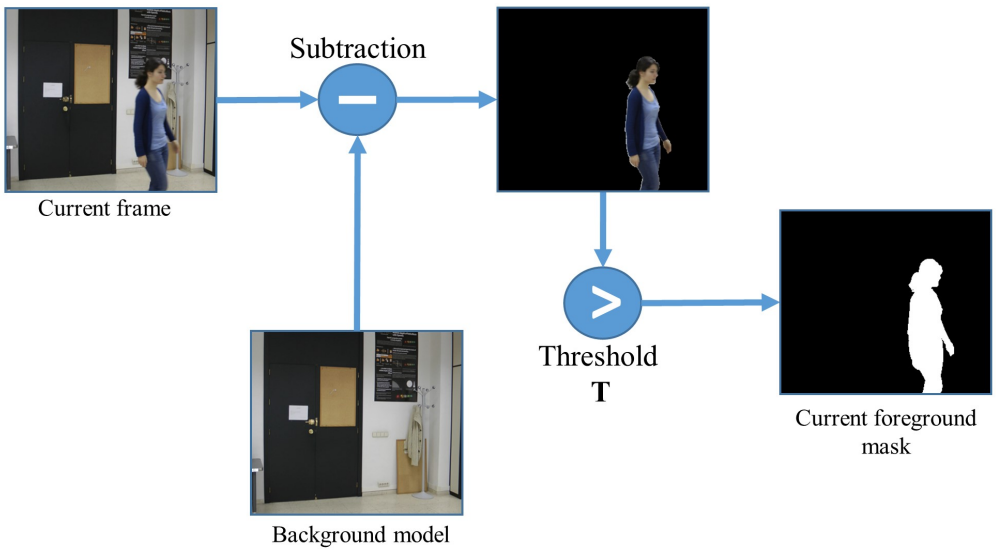


Figure 2.7: Overview of background subtraction framework.

Various background subtraction methods have been proposed, focusing on modeling the background for individual pixels and detecting movement at the pixel level. Additionally, some algorithms use regional modeling techniques. Each approach has unique features, advantages, limitations, and varying abilities to address challenges. Background subtraction algorithms face several inherent issues in video surveillance, with background motion, changing illumination, and shadows being particularly challenging for object recognition in surveillance systems [37–39].

2.3.1 Datasets

Numerous algorithms have been created for background subtraction tasks in recent years. However, the absence of a widely accepted, realistic, and large-scale

video dataset to compare these methods has been challenging. Consequently, researchers have developed several databases to fulfill this need. In the following section, we will highlight the most significant ones, such as (CDnet, SBI, LASIESTA, and UCSD).

2.3.1.1 Change Detection (CDnet)

The Change Detection CD dataset[4] is a widely recognized benchmark introduced at the IEEE Change Detection Workshops. It consists of two versions: CDnet 2012 [40] and CDnet 2014 [38]. CDnet 2012 includes nearly 90,000 frames from 31 video sequences across six categories, covering both color and thermal modalities. Each frame is meticulously annotated with ground-truth information for the foreground, background, region of interest (ROI), and shadow boundaries, facilitating accurate quantitative evaluation and comparison of change detection methods.

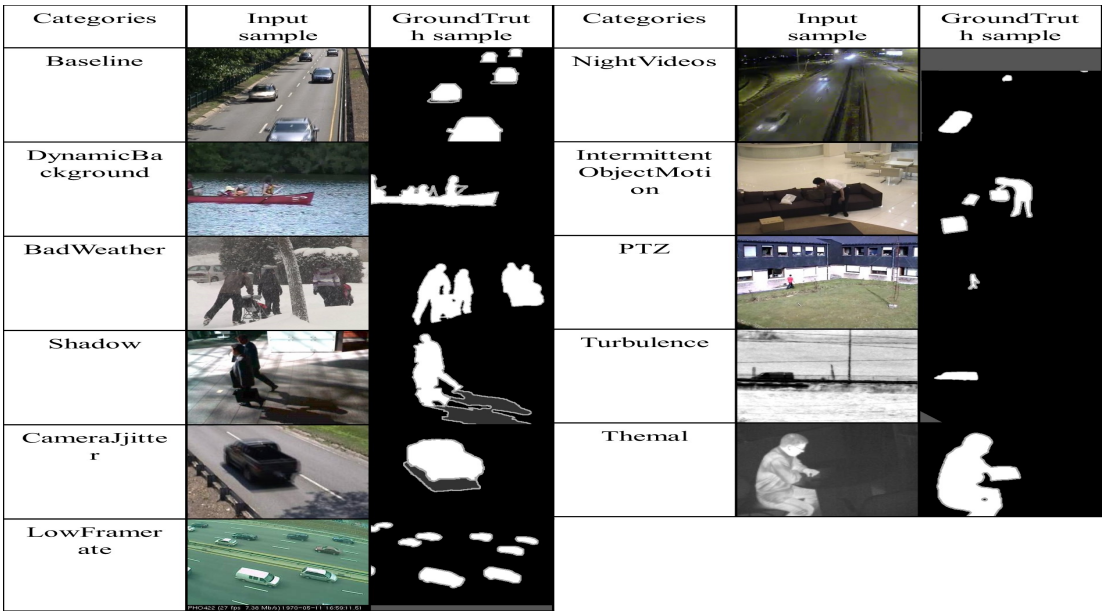


Figure 2.8: Example of video frames from CDnet 2014 dataset [38].

CDnet 2014, introduced later, addresses additional challenges in 5 categories and includes 22 videos, totaling 53 videos across 12 categories. This makes it a comprehensive resource for evaluating change detection methods in realistic scenarios. (Figure 2.8) represent examples of frame samples.

[4]http://changedetection.net/

2.3.1.2 Scene Background Initialization 2015 (SBI2015)

The Scene Background Initialization SBI 2015 dataset[5] [41], introduced by Maddalena and Petrosino in 2015, is an extensive resource, somewhat smaller than CDnet 2014. The dataset comprises 14 challenging video sequences, each with corresponding ground truths. There are nearly 5,029 frames from eight publicly available datasets developed in 2015 for assessing and contrasting different background initialization algorithms. The dataset encompasses indoor and outdoor scenes, presenting various challenges like shadows and moving backgrounds. Figure 2.9 represents example video frames from the SBI dataset featuring original frames and their corresponding ground truths.

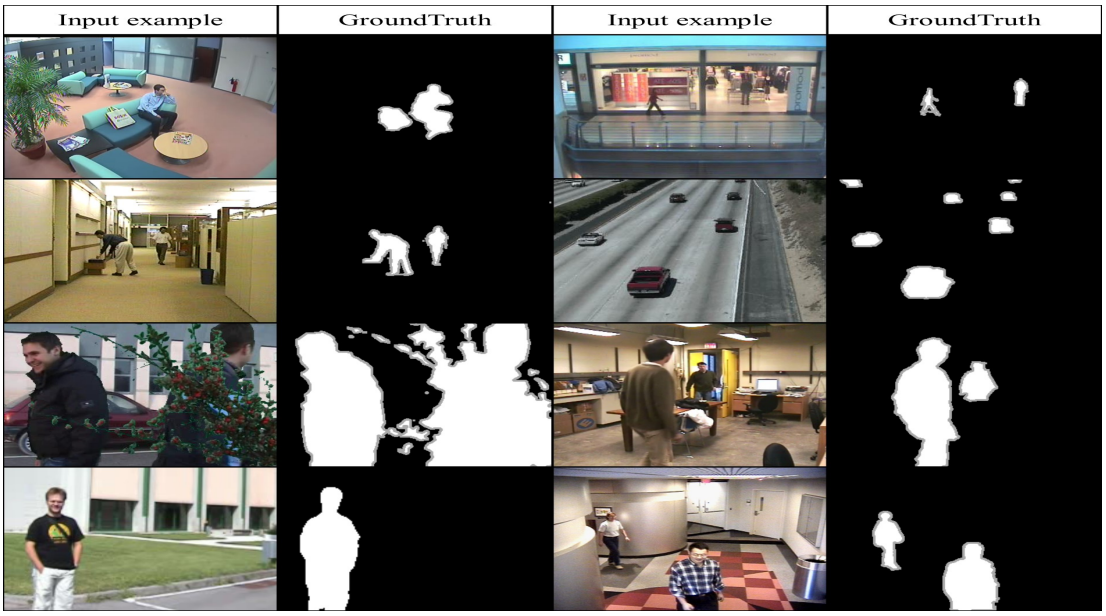


Figure 2.9: Example of video frames from SBI 2015 dataset [41].

2.3.1.3 Labeled and Annotated Sequences for Integral Evaluation of SegmenTation Algorithms (LASIESTA)

The LASIESTA dataset[6] [42], established in 2016, is an extensive collection of 48 videos recorded in various indoor and outdoor settings using both static and moving cameras with 18,425 video frames, the dataset is categorized into indoor sequences, outdoor sequences, indoor sequences with simulated motion,

[5] https://sbmi2015.na.icar.cnr.it/SBIdataset.html
[6] https://www.gti.ssr.upm.es/data/lasiesta_database/

and outdoor sequences with simulated motion. It encompasses various motion types and intensities, presenting challenges for analysis and research. Some examples from the LASIESTA dataset are shown in Figure 2.10.

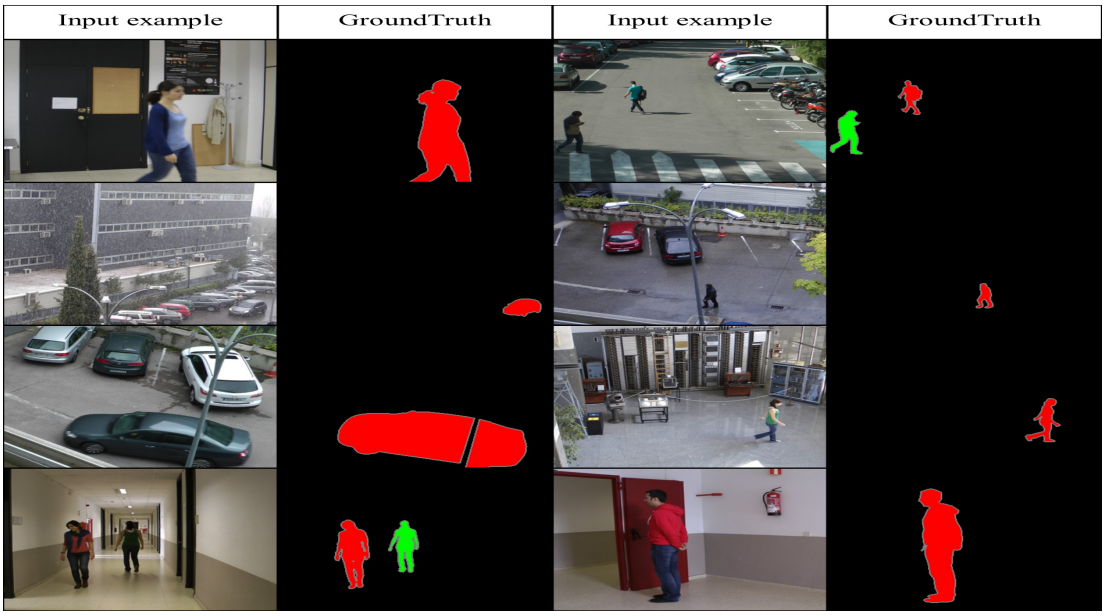


Figure 2.10: Example of video frames from LASIESTA dataset [42].

2.3.1.4 UC San Diego dataset (UCSD)

The UCSD Background Subtraction dataset[7] was developed by the Statistical Visual Computing Laboratory (SVCL). It includes 18 video sequences that were recorded in outdoor environments. The ground truth masks are given as $3D$ array variables in MATLAB. Some examples from the UCSD dataset are shown in Figure 2.11.

2.3.2 Evaluation metrics

The assessment makes use of different measurements: mean F-score 2.15, mean recall (Re) 2.9, specificity (Sp) 2.10, mean precision (Pr) 2.14, false positive rate (FPR) 2.11, false-negative rate (FNR) 2.12, percentage of incorrect classification (PWC) 2.13. These measurements are computed based on various components,

[7] http://www.svcl.ucsd.edu/projects/background_subtraction/ucsdbgsub_dataset.htm

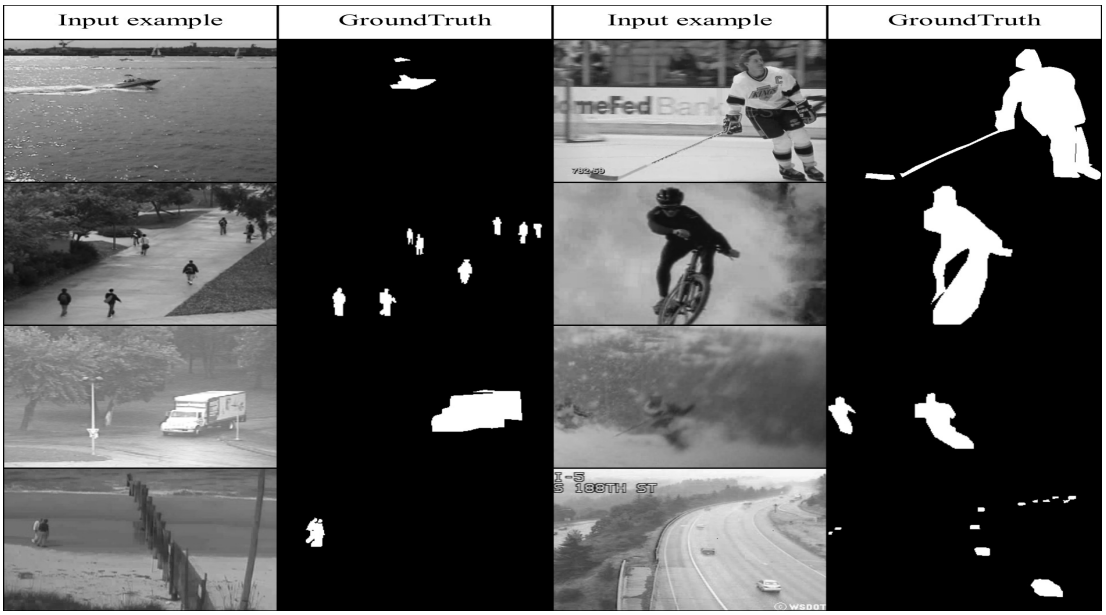


Figure 2.11: Example of video frames from UCSD dataset.

which encompass True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These components are established by comparing the actual labels to the forecasted outcomes, wherein:

$$Re = \frac{TP}{TP + FN + \alpha} \tag{2.9}$$

$$Sp = \frac{TN}{TN + FP + \alpha} \tag{2.10}$$

$$FPR = \frac{FP}{FP + TN + \alpha} \tag{2.11}$$

$$FNR = \frac{FN}{TP + FN + \alpha} \tag{2.12}$$

$$PWC = \frac{100.0 * (FP + FN)}{TP + FP + TN + FN + \alpha} \tag{2.13}$$

$$Pr = \frac{TP}{TP + FP + alpha} \tag{2.14}$$

$$Fmeasure = 2.0\frac{RePr}{Re+Pr+\alpha} \tag{2.15}$$

- TP (True Positive): The total count of pixels accurately classified as foreground.
- FP (False Positive): The overall count of pixels inaccurately classified as foreground.
- TN (True Negative): The total count of pixels accurately classified as background.
- FN (False Negative): The total count of pixels inaccurately classified as background. In addition,
- α: denotes a small threshold specified at 0.00001.

2.3.3 Related work

Many researchers have tried to develop an accurate technique for a particular scene in background subtraction. However, conventional methods are effective in some scenarios and require improvement in others. This section will outline traditional (parametric, non-parametric) and deep learning-oriented approaches for background subtraction.

2.3.3.1 Traditional methods

Over the past few decades, researchers have explored parametric methods for video background subtraction. Wren *et al.* [43] pioneered a parametric approach by modeling each pixel using a single Gaussian distribution, relying on the mean and variance to detect local pixel changes. Building on this, Stauffer and Grimson [44] introduced the Gaussian Mixture Model (GMM), which uses a set number of Gaussians to address background texture variations caused by dynamic environments. Numerous subsequent studies have further refined this model [45–48].

The authors in [45] introduced a method that enhances the adaptive background mixture model. They achieve this by utilizing various update equations at different stages, allowing quicker and more precise learning and effective

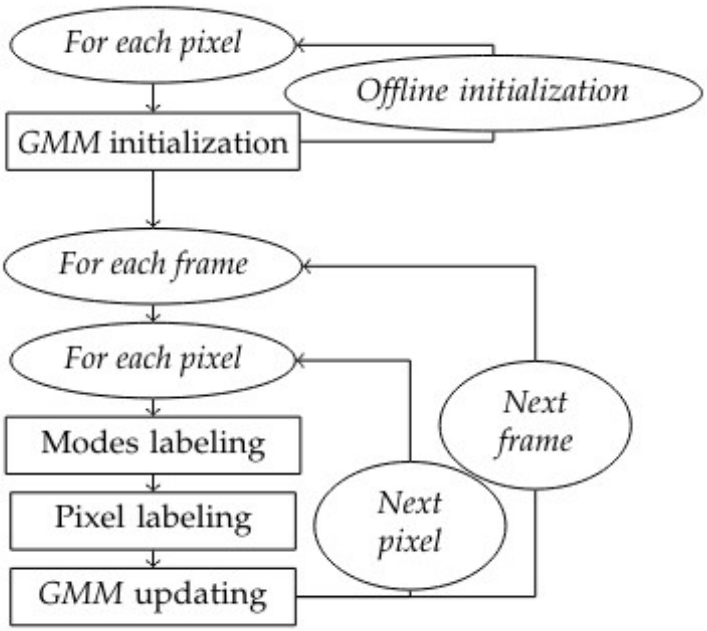


Figure 2.12: Classical GMM framework [49].

adaptation to changing environments. Additionally, they propose a shadow detection scheme based on a computational color space, which results in an improved segmentation compared to Grimson *et al.*'s method [44]. Furthermore, *Lee* [46] introduces a solution to enhance the convergence rate without sacrificing model stability. This is accomplished by substituting the global and static retention factor with an adaptive learning rate that is computed for each Gaussian at each frame. The results demonstrate significant improvements in both synthetic and real video data. The integration of this algorithm into a statistical framework for background subtraction results in improved segmentation performance compared to a standard method.

Moreover, *Zivkovic* [47] used recursive equations to update each pixel's parameters. Additionally, Zivkovic and Van Der Heijden [48] have presented a simple non-parametric adaptive density estimation method. Furthermore, a Mixture of Gaussians (MoG) has been incorporated into several other proposed techniques, as seen in [50–54].

Nguyen *et al.* [53] proposed a new asymmetric based on a variational Bayesian learning mixture model for model detection and selection using multiple Student's-t distributions.

Akilan *et al.* [50] introduced an enhancement strategy that combines color and illumination measures. They achieved this by checking pixel matches with Gaussian distribution and adaptively setting the threshold through foreground validation. Conversely, Haines *et al.* [51, 52] proposed a novel method that integrates a Dirichlet process into a Gaussian mixture model to estimate the background distribution for each pixel, followed by applying probabilistic regularization. Additionally, Nguyen *et al.* (2014) proposed an asymmetric variational Bayesian learning mixture model for detection and selection using multiple Student's-t distributions.

Zhao *et al.* [55] introduced a novel approach for motion detection in dynamic scenes, combining a Type-2 Fuzzy Gaussian Mixture Model (T2-FGMM) with a Markov Random Field (MRF). Their method integrates spatial-temporal constraints into the T2-FGMM using a Bayesian framework.

Another Type-2 Fuzzy modeling has been presented by Darwich *et al.* [49]. Route *et al.* [54] have presented a new GMM framework that combines a Wronskian framework with the Mixture of Gaussian (MoG) process to identify local changes.

Several methods have been created for estimating background models for individual pixels in non-parametric approaches. Elgammal *et al.* [56] introduced Kernel Density Estimation (KDE) as a new background model, which estimates the probability of observing pixel intensity values for each pixel.

Maddalena *et al.*[57] introduced a Self-Organizing Background Subtraction (SOBS) using Artificial Neural Network (ANN). In this approach, every pixel is associated with a $2-D$ grid of neurons, and each neuron has a weight vector of size nn. The method utilizes the HSV color space and calculates the Euclidean distance between vectors to differentiate between foreground and background pixels. It also utilizes a dynamic threshold ϵ for this purpose. The process starts by initializing the model for the first frame and storing the Hue Saturation Value (HSV) components as weight vectors. When a new pixel p_t arrives, the model identifies the best matching based on the Euclidean distance. An established match indicates a background pixel, and a learning factor $\beta(t)$ that varies over time is used to update the weight vectors in the nn neighborhood. If there is no match, the pixel's similarity to the background model determines whether it is a shadow or foreground. This approach effectively adapts to changing backgrounds and accurately distinguishes foreground objects from shadows.

The authors have improved their method by developing a fuzzy rule-based procedure called *SOBS_CF* [58]. This enhanced SOBS algorithm integrates spatial coherence to improve background subtraction by considering adjacent pixels with slight intensity variations as coherent. This approach increases robustness against false detection. Additionally, it enhances the background model by automatically adjusting pixel contributions during the update phase using a data-dependent mechanism.

The **Codebook** algorithm, as presented by Kim *et al.* [59, 60], constructs a background model using a quantizationclustering method. Every pixel is represented by a codebook (CB) containing codewords (CW) to depict the background. The proposed Codebook is presented as follow:

The Codebook algorithm operates in two phases: a learning phase for generating initial codebooks and a subtraction phase for extracting the background. When a new pixel is encountered, its intensity is computed, and the color distortion between this pixel and a codeword is also assessed. The algorithm allows brightness changes within a predefined range, constraining the shadow and highlight levels. This range is specific to each codeword and ensures that alterations within this range are deemed valid. The logical brightness function verifies whether the pixel intensity falls within the permitted range of each codeword, thereby determining if the pixel corresponds to the background model.

Heikkila *et al.* [61] introduced a new texture-based method for background modeling in a video sequence. Every pixel is represented as a set of adaptable Local Binary Pattern (LBP) histograms calculated over a circular area surrounding the pixel. Adaptive LBP histogram models are constructed for each pixel and updated over time to represent changes in the scene. Foreground detection involves comparing the LBP histogram with the selected background histograms to distinguish between static background elements and dynamic foreground objects within the scene.

SAmple CONsensus (SACON) [62] is a new background model based on RANdom Sample Consensus (RANSAC) [63]. SACON maintains a cache of N background samples per pixel and uses a threshold T_n based on sample size and error tolerance to identify if a pixel is in the background or foreground. Unlike MOG-based models, SACON does not rely on parameters. Because of its sensitivity to lighting changes, the RGB color space can cause shadows to be misidentified as foreground. To minimize this, normalized color coordinates are used to ensure

consistent intensity ratios obtained using experimentally selected constants.

Barnich *et al.* [64, 65] developed the VIsual Background Extractor (VIBE)). This method uses pixel history values and a random strategy to estimate the background based on samples while propagating information between surrounding pixels. Figure 3.13 illustrates how each pixel $v(x)$ is compared against N background samples from previous frames and identified as background if sufficient samples fall within a predefined radius R. The method's sensitivity depends on the ratio $\frac{\#_{min}}{N}$, and fixed values of R and $\#_{min}$ have proved useful. The model starts with a single frame, uses surrounding pixel values, and updates conservatively to avoid deadlocks. Random replacement and time subsampling provide more temporal coverage, while spatial updates assure consistency. VIBE is a universal approach not affected by frame rate, color space, or scene content.

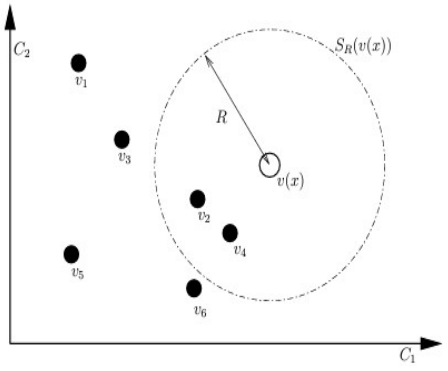


Figure 2.13: To classify $v(x)$ in a $2D$ Euclidean color space $(C1, C2)$, we count the samples of $M(x)$ within a radius R around $v(x)$ [65].

Hoffmann *et al.* [66] introduced an enhancement to the Vibe approach known as the Pixel-Based Adaptive Segmenter (PBAS). In PBAS, the background model evolves over time to account for gradual background changes using per-pixel learning parameters. The significant contribution of PBAS lies in its ability to make foreground decisions based on a threshold and to update the background through a learning parameter. Dynamic per-pixel state variables control these parameters and per-pixel thresholds, which play a role in influencing the estimation of background dynamics.

Bilodeau *et al.* [67] introduce Local Binary Similarity Patterns (LBSP), a new binary feature descriptor suitable for background removal. LBSP refers to a nn region connected to Local Self-Similarity (LSS) [68] LBSP compares a central

pixel to nearby pixels for similarity, unlike LSS, which compares individual pixels rather than patches and uses binary codes rather than vectors of integer values. The LBSP 2.14 is calculated on a nn area R, with nearby pixels being either all pixels or a subset P of pixels in R.

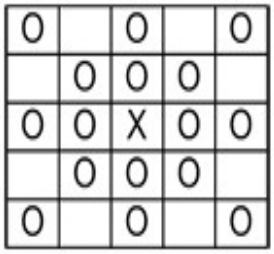


Figure 2.14: LBSP computation pattern. X: Center pixel, O: Neighbors in the computation of the binary code [67].

St-Charles *et al.* proposed an approach called Self-Balanced SENsitivity SEgmenter (SuBSENSE) [69, 70]. It incorporates pixel-level feedback loops and employs Local Binary Similarity Pattern (LBSP) features in a nonparametric feedback model. Figure 2.15 depicts the three crucial phases of the approach: sample-based modeling, pixel-level modeling, and a feedback mechanism. The initial step

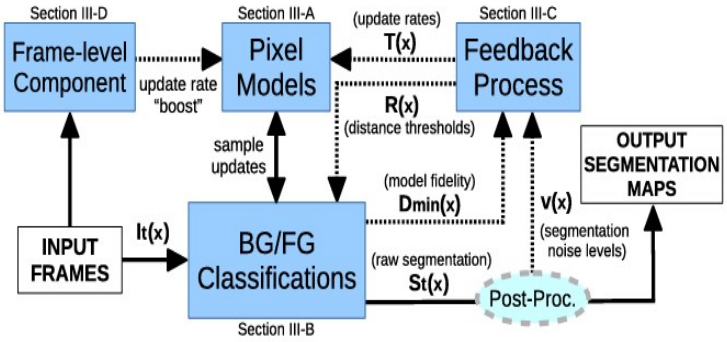


Figure 2.15: The SuBSENSE block diagram, with feedback relations shown as dotted lines [70].

of pixel-level modeling classifies individual pixels using Red Green Blue (RGB) values and LBSP features to improve the recognition of background-like objects and increase tolerance to light variations. A sample-based model randomly collects and updates pixel representations in the second step, making efficient change

detection possible. Finally, a feedback system continuously evaluates the segmentation noise and model accuracy, adjusting the algorithm's local adaptation speed and sensitivity.

The same authors have also presented a novel non-parametric pixel-level background subtraction technique called Pixel-based Adaptive Word Consensus Segmenter (PAWCS) [71] methodology. It is intended to adapt to various scenarios without requiring manual parameter modifications. This solution solves short- and long-term adaptation issues at the pixel and frame level using a persistence-based word dictionary inspired by classic codebooks and sample consensus techniques.

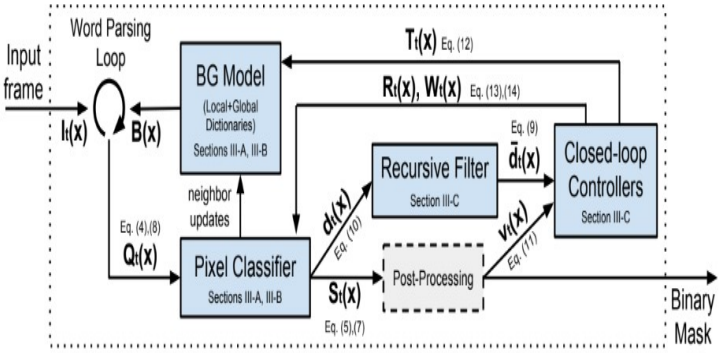


Figure 2.16: The schematic of the PAWCS approach with its main components [71].

This method evaluates each background sample (or word) online for relevance based on how frequently it appears in all local observations. The model determines how frequently background samples occur among recent observations to preserve the minimum number required for precise segmentation. By integrating these pixel-level models with a dictionary at the frame level and local feedback mechanisms, the PAWCS system efficiently handles adaptability and consistency of performance across a range of scenarios.

Bianco *et al.* [72] investigate combining advanced change detection algorithms using Genetic Programming (GP) to create a more robust algorithm. The study utilized a sophisticated combination of carefully selected top algorithms, with meticulous post-processing procedures executed using GP. This approach uses GP to automate the algorithm selection process and simultaneously achieves algorithm selection, combination, and processing.

2.3.3.2 CNN-based methods

Convolutional neural networks, or CNNs, have shown a significant advancement in computer vision, particularly in object detection, recognition, and tracking. These networks have outperformed conventional techniques in extracting numerous visual attributes and effectively adapting to the diverse difficulties the visual data poses.

Braham *et al.* [73] proposed a background subtraction algorithm using **convolutional neural networks (ConvNets)** to learn spatial features for background modeling. The algorithm uses a single grayscale background image and a scene-specific training dataset to train ConvNets. The approach eliminates the need for a complex background modeling strategy and does not aim to present a real-time adaptive technique but explores the potential of deep features learned with ConvNets for background subtraction.

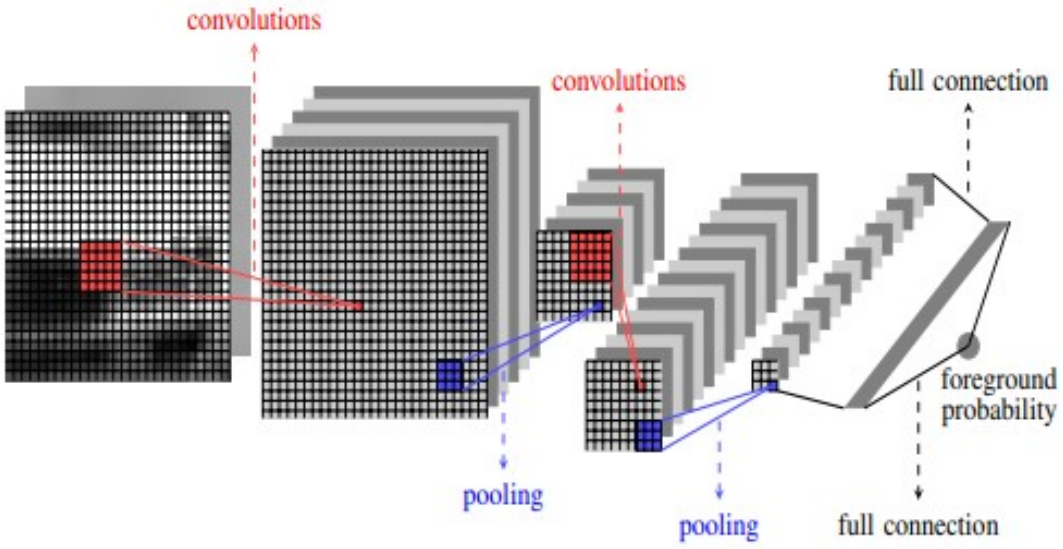


Figure 2.17: The overall ConvNet architecture [71].

Babaee *et al.* [74] introduced a background subtraction system using a deep Convolutional Neural Network (CNN) for segmentation. The method eliminates the need for feature engineering and parameter tuning. It subtracts incoming frames from a background using image patches and processes them through a CNN, with post-processing to generate the final video frame segmentation. The method integrates segmentation from the SuBSENSE algorithm [70] and the Flux

Tensor algorithm [75]) to create background images from video frames. The system also applies padding to the foreground mask to enhance the model's robustness when dealing with low-quality surveillance camera output.

Liao *et al.* [76] introduced a novel multiscale cascaded scene-specific CNNs-based background subtraction method. The scene-specific CNN structure consists of six layers, followed by multiscale cascaded CNNs and a novel training strategy to address the imbalance between positive and negative samples.

Nguyen *et al.* [77] proposed a framework for background subtraction using a sample-based background model and a data-driven feature extractor trained via a triplet network. The method involves motion feature learning, sample-based background modeling, and an adaptive feedback scheme to handle dynamic backgrounds and illumination changes.

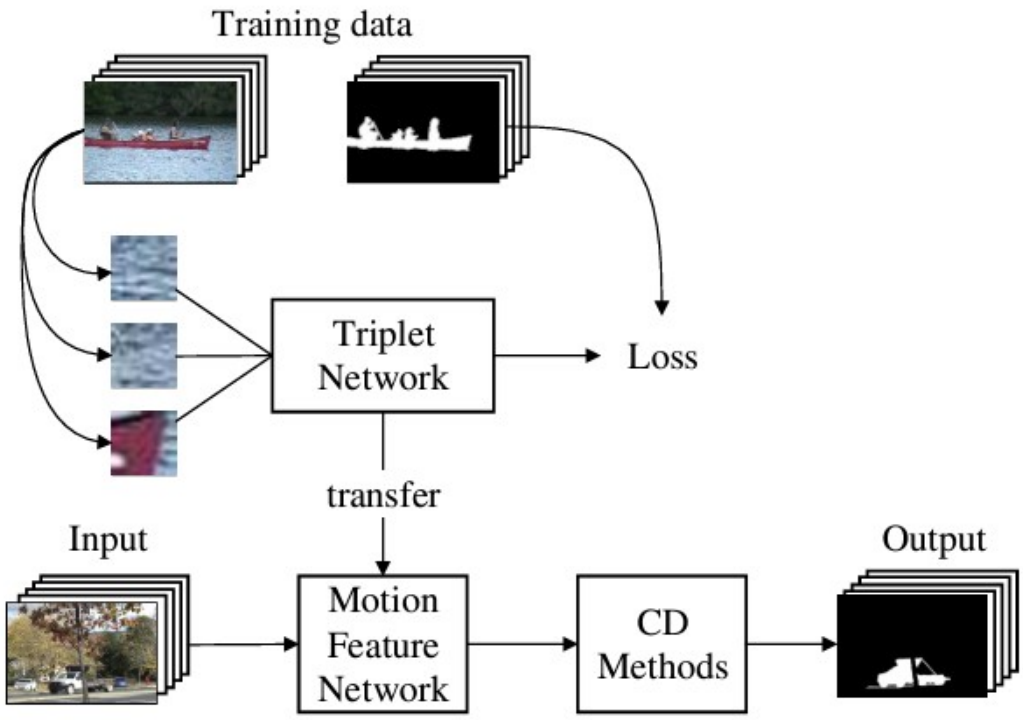


Figure 2.18: The schematic representation of triplet network [77].

The Fully Convolutional Encoder-decoder Spatialtemporal Network (FCESNet), introduced by Qiu *et al.* [78], is designed for background subtraction in video sequences. The network architecture consists of three key components: a feature encoder, a Spatial Temporal Information Transmission (STIT) module, and a

feature decoder. The encoder processes consecutive frames independently, preserving crucial spatial information. The STIT module captures spatial-temporal correlations utilizing a ConvLSTM-based architecture. Following this, the feature decoder upsamples the outputs of the STIT module to produce subtraction results for each frame. A patch-based training method addresses sparse foreground pixels and prevents overfitting. Notably, the network is capable of directly processing entire frames during inference.

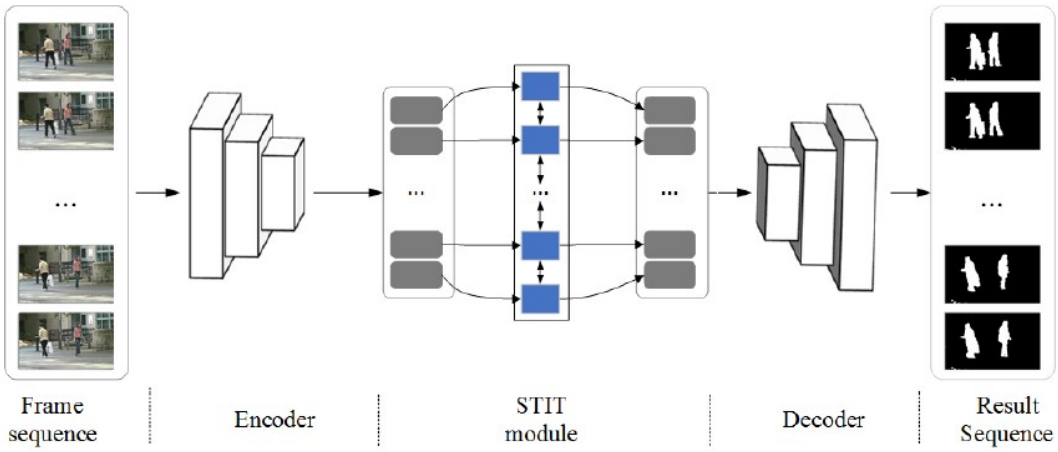


Figure 2.19: Many-to-many network architecture with STIT module for spatial-temporal information transmission and frame subtraction. [78].

Wang *et al.* [79] proposed a semi-automatic technique for segmenting moving objects in surveillance videos using a multi-resolution convolutional neural network (CNN) with a cascaded architecture. Their method aims to generate accurate segmentation maps suitable for ground truth with minimal user intervention. The model is trained using a small set of manually annotated frames and is designed with three variants to tackle different challenges. This approach efficiently utilizes the redundancy in surveillance video content to learn a foreground-background model effectively from a limited number of training samples.

Lim *et al.* [80] introduced a robust Foreground Segmentation Network (FgSegNet) that employs an encoder-decoder neural network approach for moving object segmentation. They utilized a pre-trained **VGG-16** for encoding and a transposed convolutional network for decoding. This network is designed to be trained as an end-to-end with minimal training samples, taking an RGB image in three different scales and generating a foreground segmentation probability mask.

The researchers also tackled the imbalanced class sample issue by integrating a Triplet CNN (TCNN) with a transposed convolutional neural network in an encoder-decoder structure. This approach emphasizes carefully selecting training examples, especially for dynamic backgrounds or complex scenes. It features a network architecture incorporating a TCNN for feature encoding and decoding.

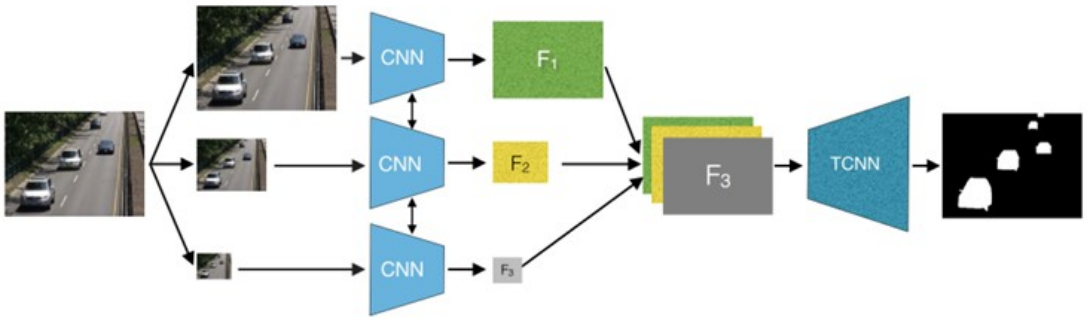


Figure 2.20: FgSegNet architecture. [80].

The same authors [81] have designed a robust encoder-decoder neural network that can be effectively trained with minimal examples as FgSegNet_V2. This approach enhances FgSegNet's Feature Pooling Module by integrating feature fusions for multiscale feature extraction. The architecture utilizes early layers of the VGG-16 network, and the modified Feature Pooling Module incorporates dilated convolutions and Instance Normalization. The decoder network utilizes Global Average Pooling to merge low-level encoder features into high-level decoder outputs, enhancing segmentation accuracy with minimal computational overhead.

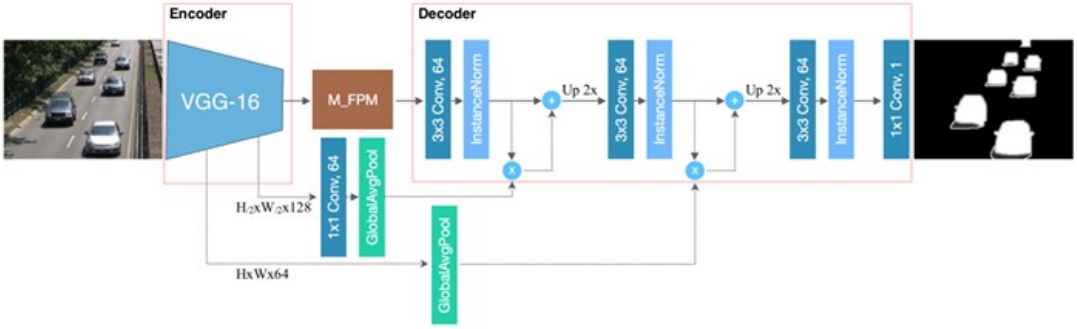


Figure 2.21: The flow of FgSegNet v2 architecture [81].

Panda *et al.* [82] introduced an enhanced version of fgsegnet_V2. They used to include the FPM with a ResNet-50 encoder-decoder network for handling complex video scenes, offering a less complex architecture compared to VGG-16. This model extracts relevant multiscale features for local change detection in complex videos and uses up-sampling in the decoder to generate a segmented probability mask for the input image. The encoder network employs a pre-trained ResNet-50 model, and the resulting feature maps are processed through a Feature Pooling Module (FPM). The decoder network then converts these pooled feature maps into a pixel-level foreground probability map, which is thresholded to create binary segmentation labels.

Zheng *et al.* [83] introduced a new multiscale fully convolutional network (MFCN) for background subtraction. They utilized transfer learning and Fully Convolutional Networks (FCN) for semantic segmentation. The authors used to improve the accuracy by restructuring and fine-tuning the VGG-16 network. The MFCN-based background subtraction framework comprises two stages: training and foreground detection. During training, input frames train the model with corresponding foreground/background label masks. The architecture incorporates a fully convolutional network with multiscale convolution and deconvolution operations, making it well-suited for background subtraction tasks.

Tezcan *et al.* [84] presented a novel supervised background subtraction algorithm Background Subtraction of Unseen Videos Network (BSUV-Net). This technique uses a fully convolutional neural network to predict foreground elements in previously unseen videos. BSUV-Net effectively tackles challenges such as changes in scene illumination and the intermittent presence of stationary objects. It accomplishes this by utilizing two reference backgrounds at different time scales and integrating semantic segmentation information to enhance accuracy. Set apart from traditional methods, BSUV-Net demonstrates true generalizability to unseen videos and incorporates a data augmentation technique to manage variations in illumination. Noteworthy components of its network architecture include residual connections, batch normalization, and spatial dropout layers. Additionally, it employs a relaxed Jaccard index as the loss function to address any class imbalance. There has also been a proposal for background modeling using a Generative Adversarial Network (GAN); Zhang *et al.* [85] proposed a GAN based Background Subtraction (BSGAN) technique. This approach employs a Bayesian GAN for background subtraction, where the background image is

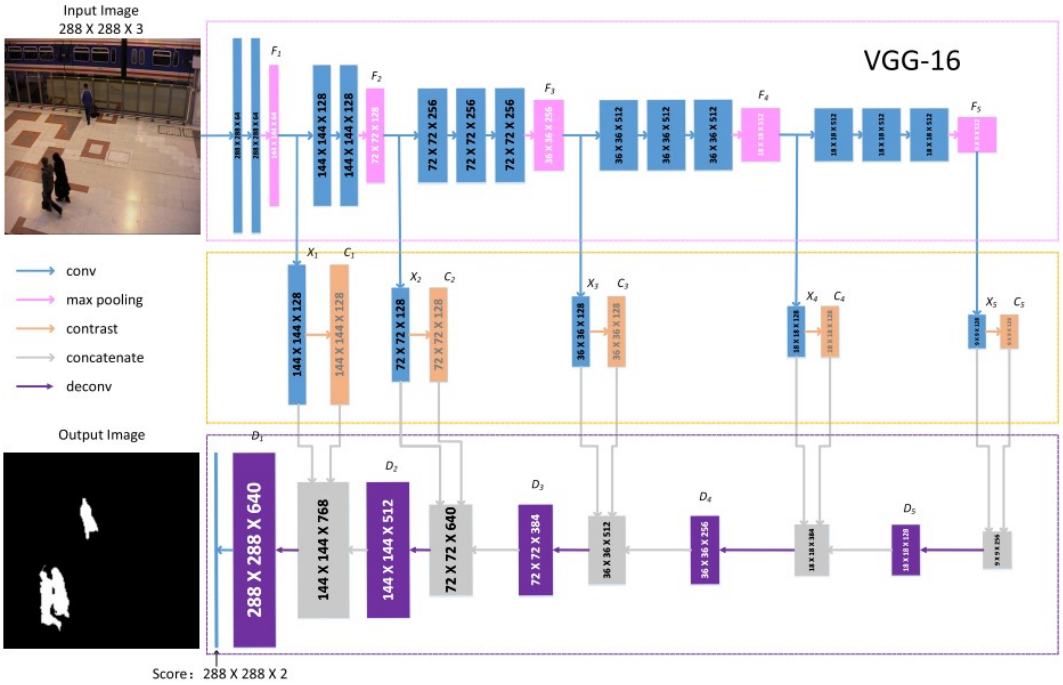


Figure 2.22: Architecture of MFCN for background subtraction [83].

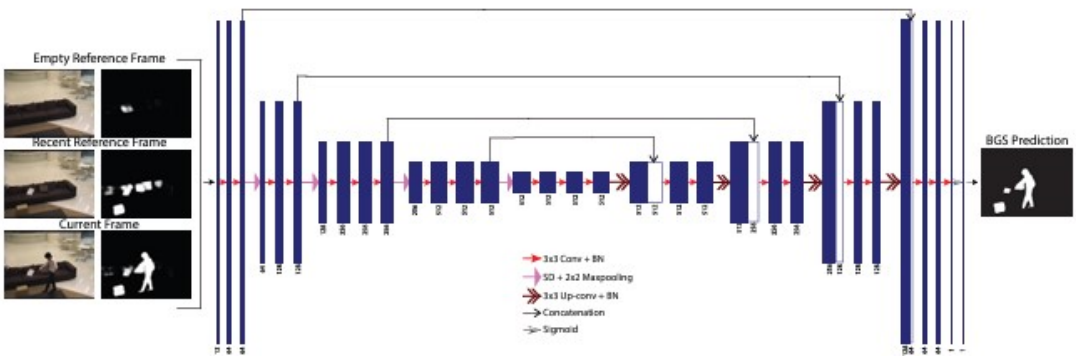


Figure 2.23: The network structure of BSUV-Net [84].

obtained through temporal median filtering. Then, the Bayesian GAN is used to detect foreground objects.

In subsequent work, the authors introduced Parallel vision and Bayesian GANs based Background Subtraction (BSPVGAN) [86], which extends BSGAN by incorporating parallel vision theory to enhance foreground detection in complex scenes. Their method involves using a median filtering algorithm to extract the background image, followed by the Bayesian GAN to classify pixels as foreground or background. The process includes background extraction, dataset generation, and training the Bayesian GAN on scene-specific datasets, with iterative training improving the model's pixel classification accuracy.

Patil *et al.* [87] introduced an end-to-end Multi-scale Temporal Pixel Aggregation (MTPA) network, incorporating adversarial learning for scene-dependent and independent object segmentation. The MTPA network is designed to capture detailed spatiotemporal features from the current and reference frames. Drawing inspiration from several studies, their framework presents a novel approach by integrating multi-scale temporal edge aggregation (MTPA) with an encoder-decoder architecture, enhanced through adversarial learning, to address the challenge of Moving Object Segmentation.

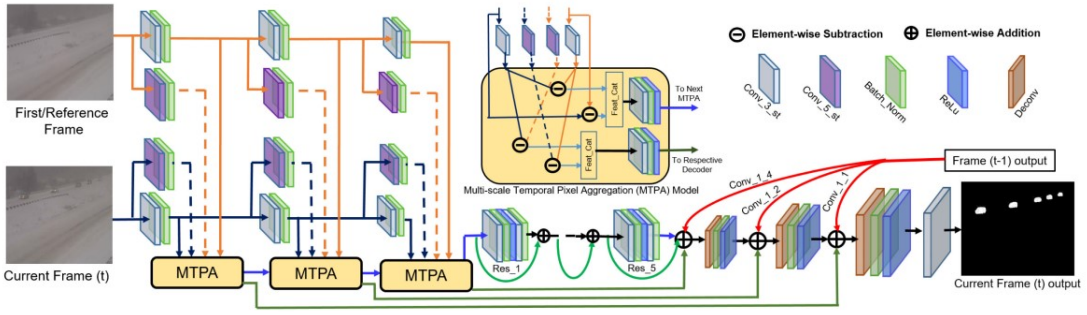


Figure 2.24: Flow-chart illustrating the MTPA structure [87].

Akilan *et al.* [88] proposed a 3D CNN-LSTM model for foreground-background (FG-BG) segmentation in video sequences. The model incorporates 3D convolutions and LSTM units to capture short-term temporal dynamics and long-short-term dependencies. It utilizes a double encoding and slow decoding strategy to enhance feature representation and localization of foreground objects, outperforming traditional Conv-LSTM networks in empirical evaluations.

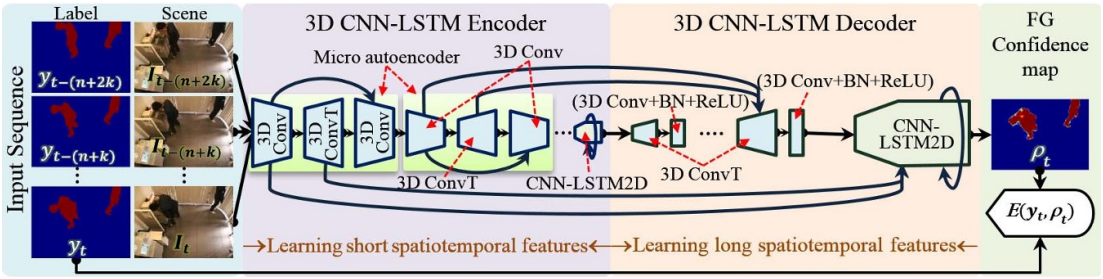


Figure 2.25: 3D CNN-LSTM structure (BN: batch normalization, 3D Conv, 3D convT: 3D convolution and deconvolution, E(·)- binary cross-entropy error [88].

The same authors introduced a novel architecture known as Slow Encoder-Decoder (sEnDec) [89] to enhance traditional image-to-image Deep Convolutional Neural Networks (DCNNs) [89]. The sEnDec architecture improves traditional image-to-image DCNNs by incorporating intermediate feature map up-sampling and residual connections, which help to recover structural details lost during spatial subsampling. Unlike the basic U-net, the sEnDec model utilizes 2D convolution for subsampling instead of max-pooling, integrates fusion in both encoding and decoding subnetworks, and employs a slow decoding strategy involving two distinct fusions of higher and lower resolution features, mediated by a layer that integrates convolution, Batch Normalization (BN), and rectification.

2.4 Conclusion

This chapter systematically explores methodologies for identifying moving and stationary objects, organizing our analysis into two primary sections. The first section is dedicated to instance segmentation, while the second section concentrates on foreground segmentation, particularly background subtraction techniques. We provide a detailed review of the predominant methods and datasets relevant to each technique, and we review the most widely used metrics for assessing the performance of these methods.

Moreover, we provide an in-depth analysis of the critical background subtraction techniques employed for detecting moving objects, tracing the evolution of these methodologies over time. This includes a thorough review of classical algorithms, especially those focused on optical flow. Additionally, we examine the background subtraction field, which is essential for behavior recognition of moving

objects in video sequences. While traditional approaches have historically delivered satisfactory results, recent advancements in deep learning have significantly improved detection performance, surpassing classical methods considerably.

Furthermore, this chapter highlights several leading deep learning models, illustrating their effectiveness in achieving superior performance. Given that the primary objective of this thesis is to analyze object behavior, this chapter offers a comprehensive overview of effective object detection methodologies. This is a foundational basis for the subsequent in-depth examination of object behavior, aiming to identify and analyze behavior within video data.

While many of these approaches have yielded satisfactory results, several limitations must be recognized. A key challenge with modern deep learning techniques is their inconsistency in producing solid results in specific scenarios but falling short in others despite the methods' complexity. This inconsistency highlights the need for improved feature generalization, a challenge we specifically addressed in Chapter 4.

CHAPTER 3

RELATED WORK ON ABNORMAL BEHAVIOR DETECTION

Contents

3.1 Introduction

Detecting suspicious activities and behavior patterns in security and surveillance is crucial, and Object Behavior Analysis (OBA) is a technique that uses computer vision and machine learning to accomplish this. Although analyzing behavior in video surveillance is an active research area, it is complex and has several challenges. The primary focus of OBA is to extract and process patterns of visual objects' behavior.

Over the years, the significance of detecting abnormal object behavior in video surveillance has increased. To this end, several intelligent video surveillance systems have been developed to monitor moving objects in the scene and analyze their behavior.

This chapter provides a comprehensive overview of the literature on detecting abnormal object behavior in video surveillance. It covers this field's fundamental challenges, feature extraction techniques, commonly used datasets, evaluation metrics, previous and recent work, general views, and limitations.

3.2 General presentation

One of the main goals of implementing an automated system for analyzing security footage is to detect unusual behavior patterns. Which requires thoroughly examining and identifying movement patterns and generating detailed descriptions of actions and interactions between objects or individuals [2, 90, 91].

3.2.1 Behavior representation

Any object, zone of interest, or static object involved in the behavior, such as individuals, crowds, or groups of people, is an actor of the behavior [5].

The low-level processing stage of behavior analysis is called behavior representation, which seeks to record pertinent details that characterize the intended object in the video. This level is quite complex and demanding as it significantly impacts how well we understand the behavior of the target object. The behavior of objects may be represented as (pixels, objects, feature representations, or a set of features, including global and local features, bounding boxes, texture for crowd monitoring, shapes for fall detection, and motion information such as trajectories

for individuals or Optical Flow (OF) for local and global motion) are extracted to represent objects in a video [90, 92, 93].

Higher-level features, such as those learned with convolutional neural networks (CNN) [94], are the most effective in many computer vision applications [95].

3.2.2 Features extraction

In order to identify abnormalities in video frames, the spatial and temporal CNN model extracts complex motion information in both space and time.

Deep learning has revolutionized video analysis by using techniques like Convolutional and Deep Neural Networks. This technology can extract insights from large, unlabeled datasets and has applications in many fields, such as computer vision, speech recognition, and healthcare.

These techniques can be grouped into two categories to address the challenges of spatial and temporal diversity in video data: Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). These methods are highly skilled in identifying complex feature interactions, learning from raw data, and working effectively with unlabeled datasets.

3.2.2.1 Spatial features extraction

Convolutional Neural Networks (CNNs) [96] are a crucial aspect of deep learning, extensively employed in image classification, natural language processing, and visual recognition. They are highly efficient and accurate models for classifying image data, which makes them the preferred solution for image-related tasks. The typical architecture of a convolutional neural network is illustrated in Figure 3.1. Convolutional Networks are multi-stage architectures made up of several layers. Each stage takes sets of arrays, known as feature maps, as input and output. They are made up of three primary layer types, namely convolution, pooling, and fully-connected layers. These layers work together uniquely to create a robust CNN architecture. Many researchers use pre-trained models to process frames in their detection framework, providing more efficient real-time results. 2D CNNs only analyze single frames/images and do not consider temporal or inter-frame motion information [95, 97–99].

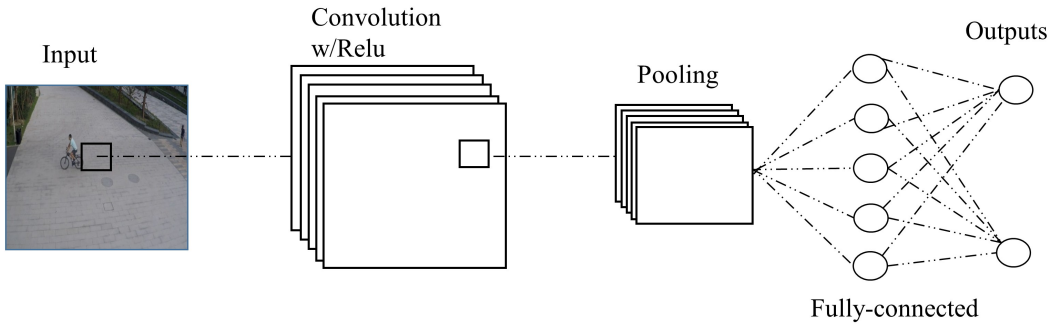


Figure 3.1: A basic CNN architecture consists of only five layers [100].

3.2.2.2 Temporal features extraction

Recurrent Neural Networks (RNNs) are a type of neural network designed to examine sequential or time-series data. Recurrent Neural Networks (RNNs) model sequential data for sequence recognition and prediction by processing real data sequences in one step at a time. RNNs feature high-dimensional hidden states with nonlinear dynamics, acting as "memory" for the network to store, remember, and process past complex signals for long periods. The state of the hidden layer at a given time is conditioned on its previous state, allowing RNNs to map an input sequence to the output sequence and predict it at the next step. The power of RNNs lies in their ability to generate new sequences by predicting them based on the training dataset [101, 102].

Several RNN-based techniques for video processing, such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been created recently and have all acquired appeal due to the spatio-temporal structure of video data. This structure refers to the spatial (visual content) and temporal (time-related changes) aspects of video data. However, among the most popular algorithms for video processing research, LSTM sticks out [98, 103].

Long Short-Term Memory (LSTM) [103] is a groundbreaking technique used to address issues related to vanishing and exploding gradients. This method employs memory cells with gates to regulate the flow of information to the hidden neurons and retain features captured from earlier time steps. Unlike traditional methods LSTM stores information in gated cells at the neurons. This makes it a widely favored and effective technique for solving a wide range of problems [101, 103, 104].

3D Convolutional Neural Networks (3D CNNs) expand upon traditional 2D CNNs to process volumetric input. While both leverage spatial hierarchies and shared weights for efficient image processing, 3D CNNs which is illustrated in Figure 3.2 uniquely incorporate depth-related and temporal information. They utilize 3D kernels to capture spatial and motion features by convolving successive frames, linking each feature map to the subsequent frames for robust analysis of volumetric and video data [97, 105, 106].

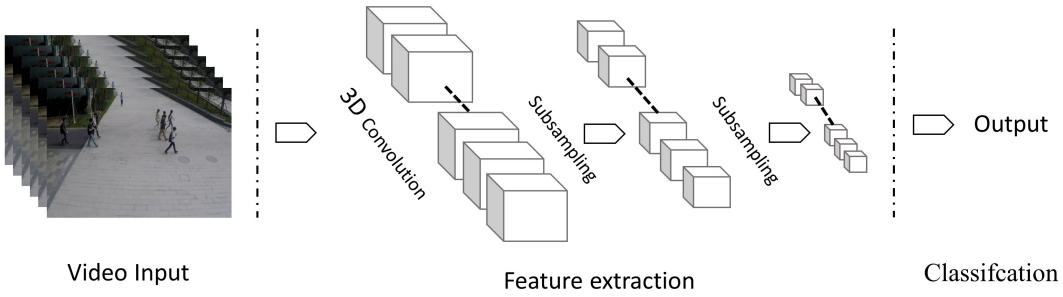


Figure 3.2: A standard structure of a 3D CNN [107].

3.2.3 Research challenges

As we progress towards building more intelligent systems for detecting anomalous object behavior in video surveillance, it is vital to address the challenges that persist in abnormal behavior identification. By using the references [108–112], these challenges can be summarized as follows:

- **Data Sparsity and Diversity**: Anomaly detection datasets often suffer from imbalances, which make it difficult to train supervised models. Furthermore, anomalies are diverse and unpredictable, making it impossible to include all potential types within a single model.
- **Noise**: Video surveillance is an essential tool for enhancing the safety of our communities. However, it often captures data from diverse settings, making it difficult to manually annotate. Additionally, footage from public areas is susceptible to various environmental factors that can compromise anomaly detection accuracy.

- **Time and Space Complexity**: The current algorithms require high computational demands, which can restrain the development of straightforward, efficient, and accurate anomaly detection systems.

  Addressing these challenges will allow us to build more effective and efficient abnormal behavior detection systems that will enhance public safety and security.

3.3 Datasets

This section provides an overview of datasets designed explicitly for detecting anomalous behaviors and events in videos. It includes a detailed summary of each dataset, along with links for easy access.

3.3.1 UCSD

The UCSD dataset[1], comprising pedestrians subsets (Ped1) and (Ped2), is a prominent resource for video anomaly detection. Created by Mahadevan *et al.* (2010) [113], it features crosswalk videos from fixed cameras capturing both normal pedestrian activities and various anomalies like unexpected objects or unconventional motions like cars, skateboarding, and biking. While the dataset offers diverse scenarios for evaluating anomaly detection approaches, Low-resolution frames and occasional crowded scenes can be challenging for the system and may affect its accuracy. Figures 3.3 and 3.4 showcase typical anomalies present in the dataset, aiding developers in refining their anomaly detection algorithms.

The Ped1 dataset comprises 70 video samples depicting individuals walking towards or away from the camera, featuring some perspective distortion. It offers 34 training and 36 testing clips, all at 234 × 159 resolution. On the other hand, the Peds2 dataset consists of 28 video samples with a higher resolution of 360 × 240, offering 16 training clips and 12 testing clips. Both subsets provide ground truth labels at both frame level and pixel level.

[1] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm/

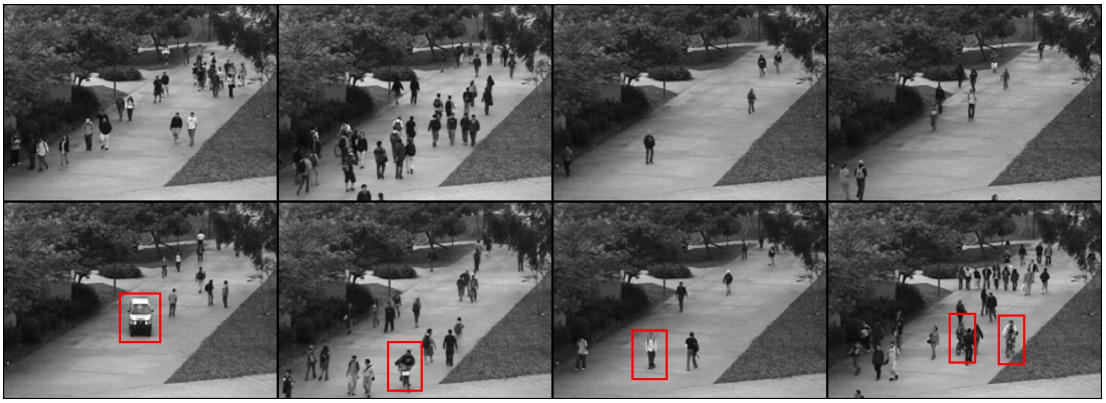


Figure 3.3: Frame samples from the UCSD Ped1 dataset (the first line shows normal instances, while the second line (from left to right) shows red boxes depicting unrecognized appearance, skateboarding, motorcyclist, biker.

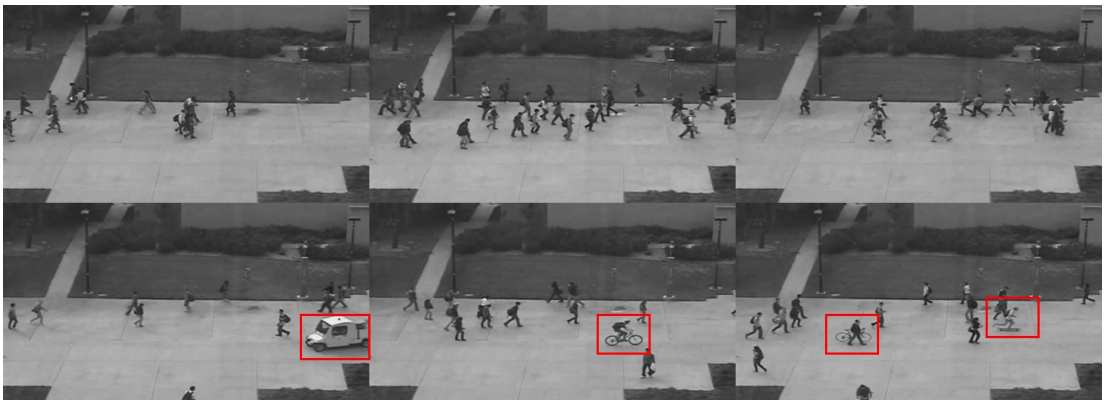


Figure 3.4: Frame samples from the UCSD Ped2 dataset.

3.3.2 CUHK Avenue

The CUHK Avenue dataset[2] is a well-known benchmark for video anomaly detection. This dataset is different from others due to camera angles and position variations. The CUHK Avenue dataset was introduced by Lu *et al.* (2013) [114] and consists of 16 training and 21 testing videos filmed on the CUHK campus avenue. The dataset offers 15,328 training frames and 15,324 testing frames. The testing set of the Avenue dataset includes 11,457 normal frames and 3,867 abnormal ones, all at a resolution of 640 × 360. This dataset showcases various anomalies, such as running, loitering, and throwing objects. Some anomalies were highlighted in the figure 3.5.

[2]https://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html/

Figure 3.5: Sample frames from the CUHK Avenue dataset are displayed as follows: the first line shows normal instances, while the second line (from left to right) shows red boxes, including objects depicting running, throwing objects, abnormal motion direction, and unrecognized appearances.

3.3.3 ShanghaiTech Campus

The ShanghaiTech Campus dataset[3] was introduced by Liu *et al.* [115] to overcome the limited scene diversity in previous benchmark datasets. This extensive dataset comprises 330 training and 107 testing videos covering 13 different scenes and various anomaly types. The videos are captured from 13 different cameras, exhibiting a wide range of lighting conditions and angles. Each video frame has a resolution of 856 × 480. The testing set includes 23,465 normal frames and 17,326 abnormal frames.

Figure 3.6 shows examples from the ShanghaiTech Campus dataset (Red boxes indicate abnormal objects).

3.3.4 UMN

The UMN dataset[4], a prominent resource for unusual crowd activity detection, offers three distinct scenes: Lawn, Indoor, and Plaza. Introduced by Hu *et al.* (2016) [116], it comprises 11 clips captured at 30 frames per second with a stationary

[3] https://svip-lab.github.io/dataset/campus_dataset.html
[4] https://mha.cs.umn.edu/

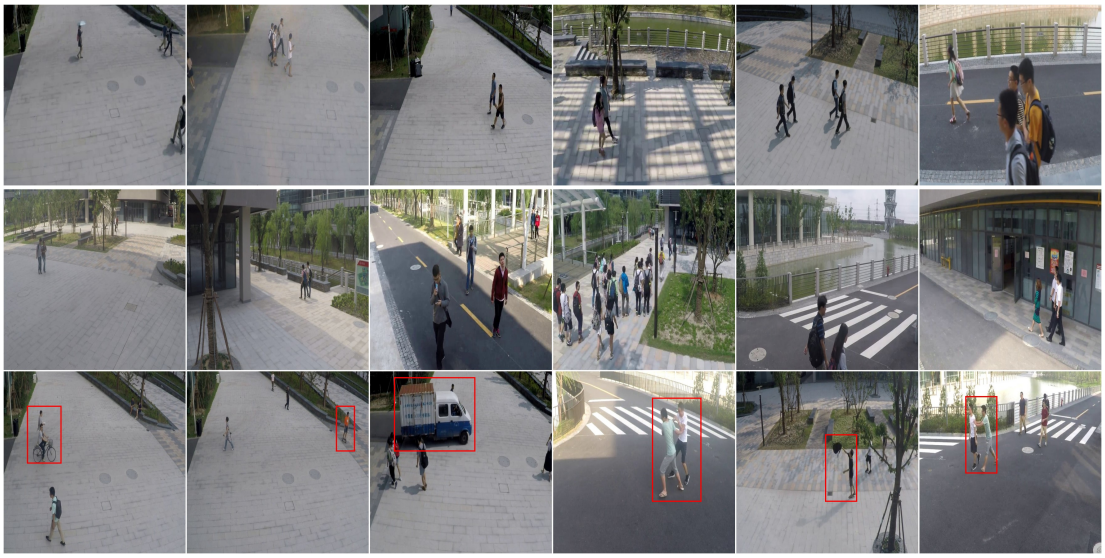


Figure 3.6: Frame examples from ShanghaiTech Campus dataset.

camera ensuring consistent illumination. The scenes contain 1,453, 4,144, and 2,144 frames, respectively. There are 7,740 frames, with 3,085 used for training and 4,656 for testing. Despite its seemingly limited training data, the UMN dataset is efficient for abnormal detection, providing accurate detection for new patterns with similar characteristics. The video resolution stands at 320 × 240 pixels, with unexpected running being recorded as an anomaly.

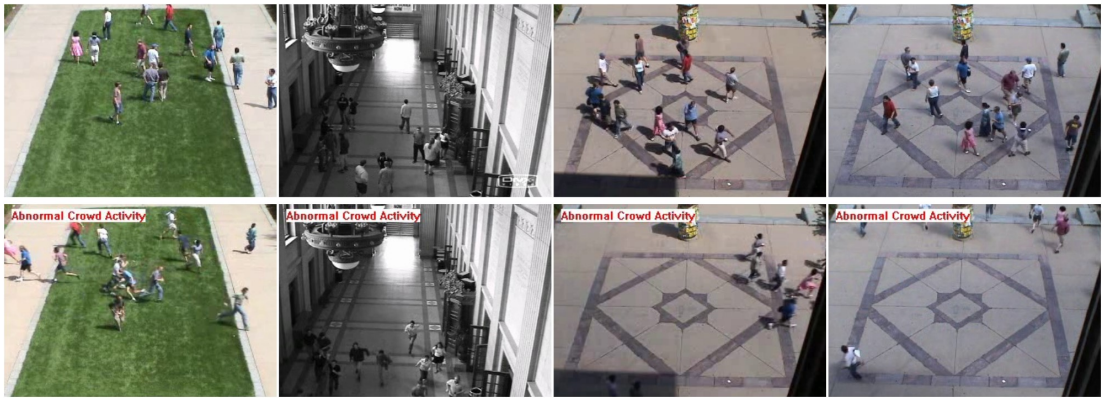


Figure 3.7: Frame examples from the UMN dataset (normal and abnormal instances are shown in the first and second rows, respectively.

3.3.5 Subway

The Subway Dataset, introduced by Adam *et al.* (2008) [117], is an invaluable resource for anyone looking to analyze human behavior in subway turnstile areas. The dataset comprises two categories of videos: "exit gate" and "entrance gate." The "entrance gate" video is two hours long, comprising 72,401 frames, and has a 512 × 384 pixels resolution. Meanwhile, the "exit gate" video spans 136,524 frames.

This dataset is a treasure trove of information, capturing a range of activities such as people evading payment, moving against the flow of the crowd, or even cleaning the walls. It is a valuable resource for researchers and analysts studying human behavior in public spaces.

Figure 3.8: Sample frames from the Subway Dataset (normal and abnormal frames are shown from left to right).

3.3.6 UCF-Crime

The *UCF-Crime* dataset[5], developed by Sultani *et al.* (2018) [118], is distinct from traditional video anomaly detection datasets as it focuses on activity detection using internet videos from various cameras. This dataset includes 13 real-world anomalies like accidents, robbery, and vandalism sourced from platforms like Youtube and LiveLeak using diverse text queries, including translations. It comprises 950 anomalous and 950 regular untrimmed surveillance videos, totaling 1,900 videos with approximately 128 hours of data at 240 × 320 pixels resolution. While the dataset offers temporal labels for testing, limiting spatial

[5]https://www.crcv.ucf.edu/research/real-world-anomaly-detection-in-surveillance-videos/

evaluation, its formulation differs significantly from the single-scene anomaly detection problem.

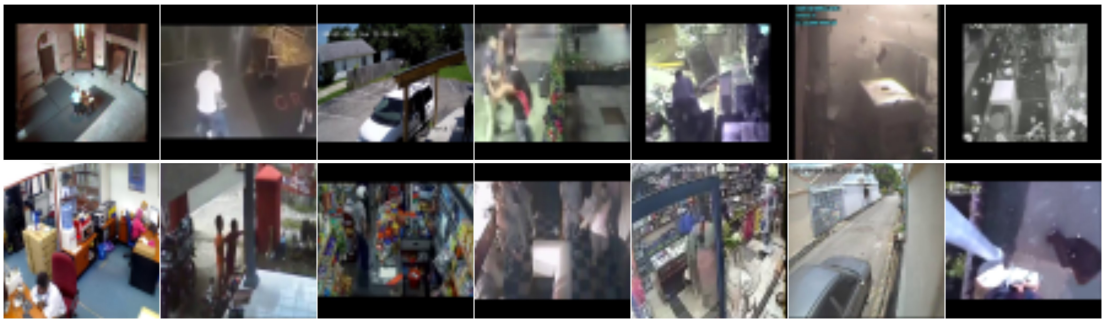


Figure 3.9: Sample normal frames (top row) and abnormal frames (bottom row) from the UCF-Crime dataset.

3.3.7 Street Scene

The Street Scene Dataset[6] was introduced by Ramachandra and Jones in 2020 [119] to overcome the limitations of older datasets. It includes more realistic and diverse anomalies. The dataset contains 46 training and 35 testing video sequences captured by a stationary camera that observes a two-lane street with bike lanes and sidewalks. The videos were recorded during the daytime and showcased various activities such as car maneuvers, pedestrian behaviors, and biking. They also include dynamic elements like shifting shadows and wind effects. The dataset contains a total of 203,257 color frames, with 56,847 frames for training and 146,410 frames for testing. The frames are at 1280 × 720 pixels, and the frame rate is 15 fps. The dataset emphasizes natural anomalies and avoids staged events. The training set follows specific criteria for normalcy, excluding activities like jaywalking and illegal parking. On the other hand, the testing sequences highlight 205 anomalous events across 17 categories.

3.4 Evaluation metrics

Most studies utilize the evaluation metrics outlined by [120], encompassing two distinct criteria: frame- and pixel-level.

[6] https://www.merl.com/research/highlights/video-anomaly-detection/

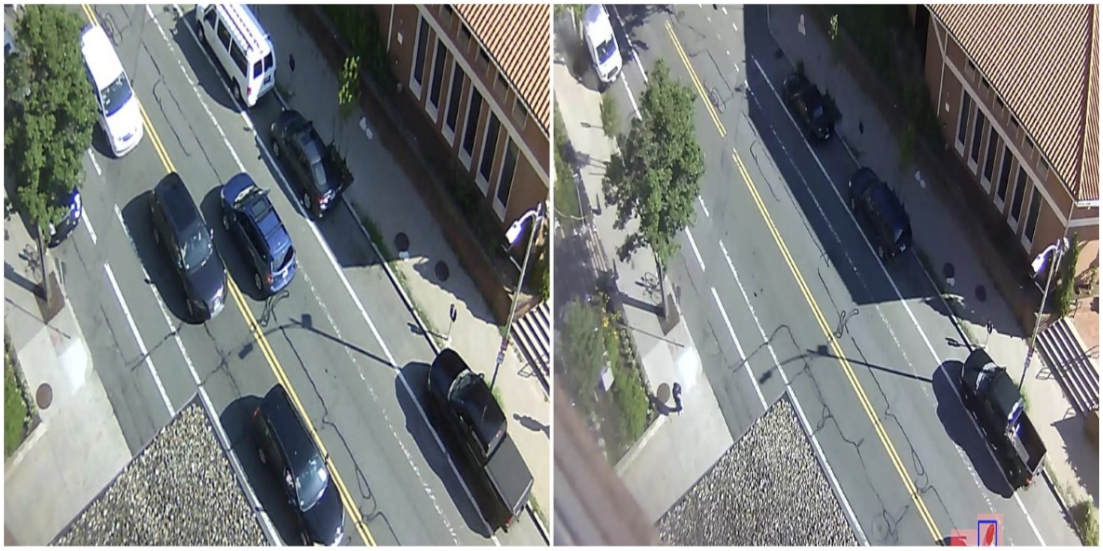


Figure 3.10: Normal and abnormal sample frames from the Street Scene Dataset.

Frame-level A frame is labeled as anomalous if at least one pixel within it is anomalous. This method does not guarantee the precise localization of anomalies. Additionally, it potentially leads to some true positives resulting from random errors rather than precise detection.

Pixel-level It considers a frame a true positive if the overlap between the detected and actual anomalous pixels exceeds 40% and a false positive if any pixels in a negative frame are incorrectly identified as anomalous. While more accurate than the frame-level approach, this method can still produce many false positives.

Two metrics for performance evaluation and comparison are essential: the equal error rate (EER) and the Area under the ROC curve (AUC). The Area under the ROC curve (ROC AUC) is computed to measure model performance. The ROC plots the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The equal error rate (EER) is also computed alongside the ROC curve, determining misclassified frames when the false positive rate equals the miss rate. A higher ROC AUC with a lower EER indicates better model performance, leveraging threshold and scale invariance strengths.

The True Positive Rate (TPR) and False Positive Rate (FPR) are defined as follows:

$$\text{TPR} = \frac{\text{\# of true-positive frames}}{\text{\# of positive frames}} \tag{3.1}$$

$$\text{FPR} = \frac{\text{\# of false-positive frames}}{\text{\# of negative frames}} \tag{3.2}$$

Historically, object-level measurements were conducted by evaluating object detection through the overlap criterion, defined as follows:

$$\text{Overlap} = \frac{\text{Area}(\text{detection} \cap \text{groundTruth})}{\text{Area}(\text{detection} \cup \text{groundTruth})} \tag{3.3}$$

A detection is classified as a true positive if the overlap exceeds a predefined threshold v. However, recent evaluations have shifted towards using the Track-Based Detection Rate (TBDR) and Region-Based Detection Rate (RBDR) criterias, which are proposed in [119].

The TBDR criterion measures the detection rate versus false positives per frame. It identifies detected tracks based on a specified threshold and calculates the ratio of detected anomalous tracks to the total number of tracks. False positives are determined based on a set threshold, and the false-positive rate (FPR) is calculated accordingly.

$$\text{TBDR} = \frac{\text{num. of anomalous tracks detected}}{\text{\# of anomalous tracks}} \tag{3.4}$$

The RBDR criterion measures the detection rate of anomalous regions compared to false positives per frame. It uses the detected regions to total anomalous regions ratio and is summarized by the Area under the false positive rates ROC curve.

$$\text{RBDR} = \frac{\text{num. of anomalous regions detected}}{\text{\# of anomalous regions}} \tag{3.5}$$

3.5 Existing works

This section outlines previous and ongoing studies on identifying irregularities, which can be categorized as traditional or modern, each having its own extraction and representation techniques, followed by modeling. However, various review papers have thoroughly examined the issue of identifying irregularities in videos [3, 121–123]. This discussion explores three key methods for detecting video irregularities: frame-level, pixel-level, and object-level descriptions. Pixel-level descriptions capture natural behavior using motion and appearance (color, gradient, texture), while object-level descriptions use basic elements such as the object's trajectory, size, shape, and speed.

3.5.1 Traditional approaches

Conventional methods generally rely on trajectory- or pixel-based approaches for detecting abnormal behavior. Trajectory analysis fundamentally entails identifying a video sequence as abnormal when the movement of objects strays from their anticipated paths. As depicted in figure 3.11 and referenced in [124, 125], trajectories underwent clustering to eliminate outliers through the anomaly detection technique based on trajectory analysis.

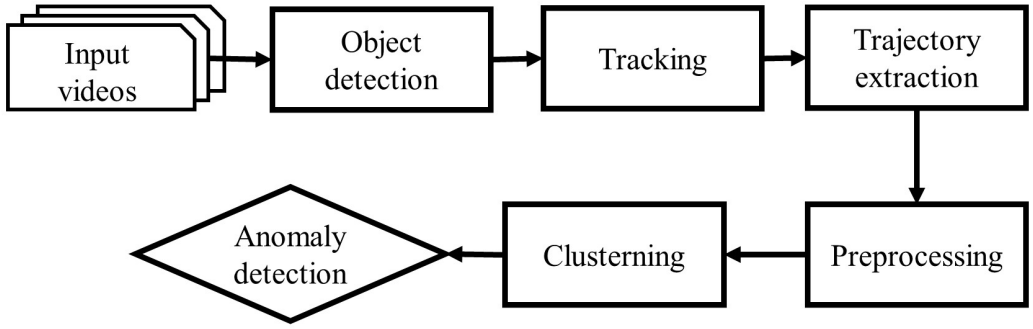


Figure 3.11: The paradigm of trajectory clustering for video anomaly detection.

3.5.1.1 Tradjectory clustering

Zhou *et al.* [126] The authors present a methodology that involves the extraction of object trajectories via a visual detection and tracking system, which integrates a Gaussian Mixture Model for background modeling, motion detection through background subtraction, and an appearance manifold-based tracking algorithm. They utilize generalized edit distance measures optimized through a supervised learning algorithm employing the Expectation-Maximization technique to effectively compare sequences with varying lengths and numbers of location samples. Anomaly detection is conducted via spectral clustering, where normal patterns emerge in large clusters, and anomalies are distinguished in smaller clusters. The system subsequently compares new trajectories against these clusters using a distance-based algorithm, an extension of the k-nearest neighbor outlier detection method.

The work of Bashir *et al.* [127] involves segmenting object trajectories into sub-trajectories, using Gaussian mixture models (GMMs) to estimate probability

density functions, employing Hidden Markov Models (HMMs) to capture temporal relations, and using Principal Component Analysis (PCA) to represent sub-trajectories for optimal energy compaction. Model-based recognition entails using GMMs to represent the probability density function of PCA coefficients for each class, where new trajectories are categorized based on maximum likelihood.

Nadeem Anjum *et al.* [128] introduced a clustering algorithm tailored explicitly for video object trajectories using multiple features to identify typical behavioral patterns and anomalies. The process involves four main steps: feature extraction, non-parametric clustering, cluster merging, and information fusion. The algorithm transforms trajectories into different feature spaces. Afterward, it uses the Mean-shift algorithm to identify clusters and modes and then consolidates clusters from all feature spaces to distinguish common and uncommon motion patterns.

The paper [129] proposes a framework for automatic behavior profiling and anomaly detection without manual labeling. It includes a compact behavior representation using Dynamic Bayesian Networks (DBNs), a natural grouping of behavior patterns through a novel spectral clustering algorithm, and a composite generative behavior model using a mixture of DBNs. For online detection, the authors introduce an accumulative anomaly measure and recognize normal behaviors using an online Likelihood Ratio Test (LRT) method.

The paper [130] proposes a framework for event detection using trajectory clustering and 4D histograms. Trajectories are clustered based on global motion flows, and 4D histograms are constructed for each cluster using the position and velocity of tracked objects. During testing, new trajectories are compared against the 4D histograms of all clusters to detect events.

The paper [131] by Claudio Piciarelli *et al.* presents a method for automatically analyzing events in video sequences, focusing on anomaly detection. The approach uses single-class Support Vector Machines (SVMs) for trajectory analysis, leveraging SVMs' capability for novelty detection. Trajectories of moving objects are represented as feature vectors and clustered using SVMs trained to identify a hyperregion in the feature space that encompasses normal trajectories, enabling the detection of anomalous trajectories that fall outside this region. The paper introduces a novel approach to parameter tuning by automatically identifying outliers in the training data based on the shrinkage rate in the hyperregion enclosing normal trajectories as the SVM's parameter varies.

Jiang *et al.* [132] proposed a method for detecting unusual video events using

unsupervised clustering of object trajectories modeled by HMMs. Their approach enhances efficiency with a dynamic hierarchical clustering algorithm and a 2-depth greedy search strategy. Unsupervised clustering is applied to distinguish abnormal events from normal ones automatically. The method integrates iterative reclassification and retraining of trajectory clusters across hierarchical levels to mitigate the overfitting issue.

The paper [133] presents a new method for detecting abnormal behavior in surveillance videos using sparse reconstruction analysis. It involves representing motion trajectories of objects as fixed-length parametric vectors based on cubic B-spline curves, categorizing them into behavior patterns, and distinguishing between normal and abnormal behaviors using sparse reconstruction analysis. The approach utilizes sparse reconstruction combined with L1-norm minimization to detect abnormal behavior in surveillance videos effectively.

Lee *et al.* [125] presented a technique for detecting abnormal behavior by extracting trajectories of moving objects in video using a Gaussian Mixture Model (GMM) and the Kanade-Lucas-Tomasi (KLT) algorithm. The system then analyzes behavior based on predefined scenarios to detect abnormal activities, focusing on the similarity of trajectories to identify abnormal behavior, specifically targeting running individuals. The system detects moving objects using GMM to separate the foreground from the background and tracks feature points using the KLT algorithm. Trajectory similarity is assessed based on distance, coordinates, and direction, while trajectory clustering is applied to identify abnormal behavior.

The paper [134] presents a novel approach for abnormal behavior detection in videos using Trajectory Sparse Reconstruction Analysis (SRA). The proposed method leverages trajectory data to detect abnormal behaviors and constructs a dictionary set using control point features from cubic B-spline curves. This dictionary set plays a crucial role in the method, representing the various paths or trajectories observed in normal behaviors. The set is further divided into Route sets. Using SRA, the approach calculates sparse reconstruction coefficients and residuals for a test trajectory. The minimal residual obtained from this process is used to classify the test behavior as normal or abnormal.

Biswas *et al.* [135] presented a methodology for identifying abnormal objects in a video stream using short local trajectories (SLTs). This approach involves extracting SLTs from super-pixels associated with foreground objects, utilizing spatial and temporal data. The trajectory extraction method showcases resilience

across videos with varying crowd densities and occlusions. Hidden Markov models (HMMs) capture typical trajectory patterns in the training phase. During the detection phase, SLTs observe each super-pixel, and their likelihood of being anomalies is assessed using the learned HMMs. Furthermore, a spatial consistency measure is introduced for each SLT based on neighboring trajectories to ensure the localization of the abnormal objects.

The paper [124] presents a novel approach for anomaly detection by utilizing a modified Hausdorff distance to analyze sub-trajectories, referred to as ADB-STR. The methodology is divided into two primary phases. In the first phase, the model undergoes training to determine appropriate thresholds where the trajectories are grouped into clusters, and the median route for each cluster is computed. In the second phase, the ADB-STR algorithm is applied for anomaly detection. This algorithm iterates over sub-trajectories, calculating the minimum Hausdorff distance between each trajectory and its corresponding sub-trajectories. A trajectory is flagged as anomalous if the calculated distance exceeds a predefined threshold.

The paper [136] introduces a unified approach for analyzing behavior in video scenes. This approach combines object trajectory analysis and pixel-based features to identify abnormal behaviors related to speed, direction, and finer motion patterns. The method can detect various abnormal behaviors with fewer false alarms. Noteworthy contributions include the proposal of snapped trajectories, an unsupervised method for discovering significant motion regions, and the introduction of a unified pipeline for comprehensive abnormal behavior detection.

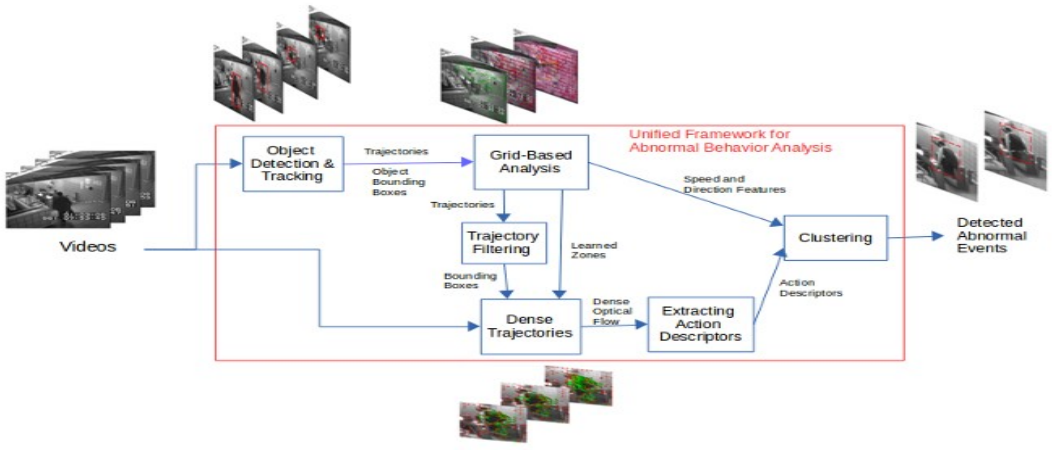


Figure 3.12: Diagram of [136] combining trajectory and pixel features.

The authors in [137] have introduced a new feature known as point trajectory-based histogram of optical flow (PT-HOF) to detect and locate irregular behaviors in surveillance videos. The framework consists of two main phases: anomaly trajectory estimation and consistency motion object construction. PT-HOF captures dynamic motion along point trajectories in surveillance videos, while consistency motion objects (CMOs) cluster similar trajectories within local regions to enhance anomaly detection accuracy.

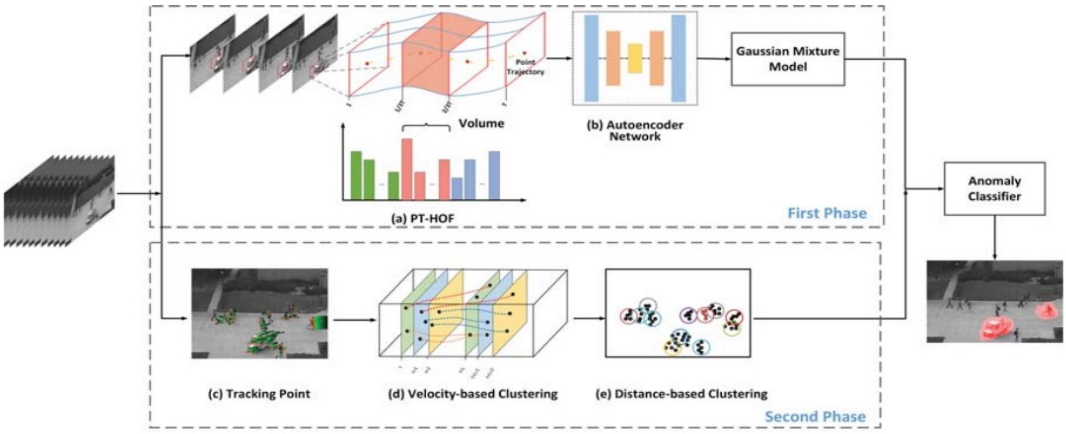


Figure 3.13: The outline of the two-phase anomaly detection framework [137].

3.5.2 Deep learning based approaches

In recent years, deep learning techniques have become essential for enhancing outcomes in both supervised and unsupervised learning. These methodologies are generally categorized into two main frameworks: frame-based and object-based approaches, both of which utilize spatial and temporal feature representations. Additionally, these frameworks can be further divided into prediction-based and reconstruction-based methods.

3.5.2.1 Frame-based appraoches

Two methods based on autoencoders were presented by Hasan *et al.* [138] to find temporal regularities in video sequences. The first approach trains a fully connected autoencoder using manually created spatiotemporal local features. The second approach uses a fully convolutional autoencoder to simultaneously learn local features and classifiers within an end-to-end framework. Their algorithm

works well on anomaly detection tasks and effectively captures regularities from various datasets. They initialized the weights using the Xavier technique and utilized stochastic gradient descent with the AdaGrad method for autoencoder optimization to ensure stability.

The methodology presented in [139] leverages more straightforward discriminative learning techniques, bypassing the need for traditional density estimation. This approach introduces a permutation-based framework for anomaly detection that operates independently of temporal sequencing, supported by a guiding theory for selecting key parameters. The system is designed as a feature-agnostic framework, allowing it to be adapted to any relevant feature set within the field.

Feng *et al.* [140] introduced an unsupervised technique for automatic video event representation and modeling. They utilized a PCANet [141] to extract appearance and motion features from 3D gradients and then employed a deep Gaussian mixture model (GMM) to characterize normal event patterns. The method involves training the PCANet, extracting high-level features, and modeling normal event patterns with a deep GMM. Subsequently, during testing, features are computed using the trained PCANet, and an event is classified as abnormal if its probability falls below a predefined threshold.

The paper [142] introduces a novel method for detecting video anomalies by leveraging a convolutional neural network (CNN) for encoding appearance in each frame and a convolutional long short-term memory (ConvLSTM) network for capturing motion. This method incorporates an auto-encoder to learn regular events. The proposed architecture involves encoding video frames with CNNs, processing them through ConvLSTM to capture temporal dependencies, and reconstructing frames with DeconvNet. The model operates on T consecutive frames, with the ConvLSTM capturing historical information to aid in frame reconstruction. The reconstruction error for each frame is computed, and high errors serve as indicators of anomalies.

Anomaly detection using a convolutional winner-take-all autoencoder, which is proposed by Tran *et al.* [143], is a new approach for video anomaly detection that combines a convolutional autoencoder with a one-class SVM. It uses motion-feature encoding from the autoencoder as input to the SVM and introduces a spatial winner-take-all step to achieve high sparsity. The Convolutional Winner-Take-All Autoencoder (Conv-WTA) learns hierarchical unsupervised sparse representations and uses a One Class SVM (OCSVM) for outlier detection. This

approach captures normal flow pattern variations and enhances anomaly detection accuracy.

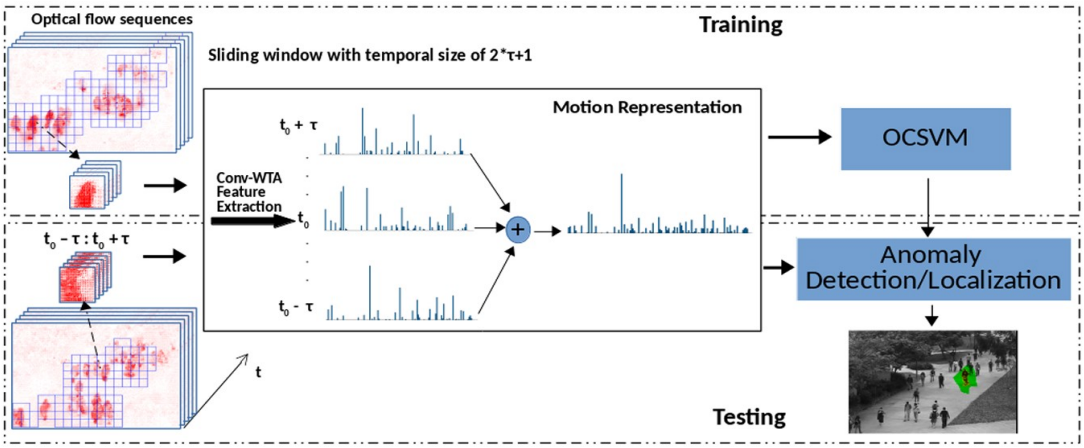


Figure 3.14: An outline of convolutional winner-take-all autoencoder [143].

In their paper, Tudor *et al.* [144] introduced a novel framework that eliminates the need for training sequences in video anomaly detection. The framework employs an unmasking technique, where a binary classifier is iteratively trained to distinguish between two consecutive video sequences. At each iteration, the most discriminative features are progressively removed, with the accuracy of the intermediate classifiers serving as indicators of abnormal events. The framework independently utilizes both motion and appearance features, calculating average scores for each frame. Motion features are derived from 3D gradient features of spatiotemporal cubes, while appearance features are extracted from a pre-trained VGG-f CNN model. The unmasking process measures the degree of difference between consecutive events, with a gradual decline in classifier accuracy signaling the presence of anomalies.

The paper [145] presents a comprehensive framework for detecting video anomalies using a Restricted Boltzmann Machine (RBM). The RBM directly processes image pixels, learning new data representations without the need for labels and using reconstruction errors to identify abnormal events. This method can be used for both offline and streaming scenarios, and the framework is trained in an entirely unsupervised manner. The system manages high-dimensional video data by dividing images into patches and clustering similar patches to train RBMs.

The article by Liu *et al.* [115] presents a novel approach to video anomaly detection. The method compares predicted future frames with their ground truth to identify anomalies. It incorporates motion and appearance constraints to improve the accuracy of predicting future frames for normal events. Additionally, the method ensures consistency in the optical flow between predicted and ground truth frames, contributing to its anomaly detection effectiveness. The framework includes a Generative Adversarial Network (GAN) module and employs adversarial training using the Least Squares GAN to enhance the realism of predicted frames..

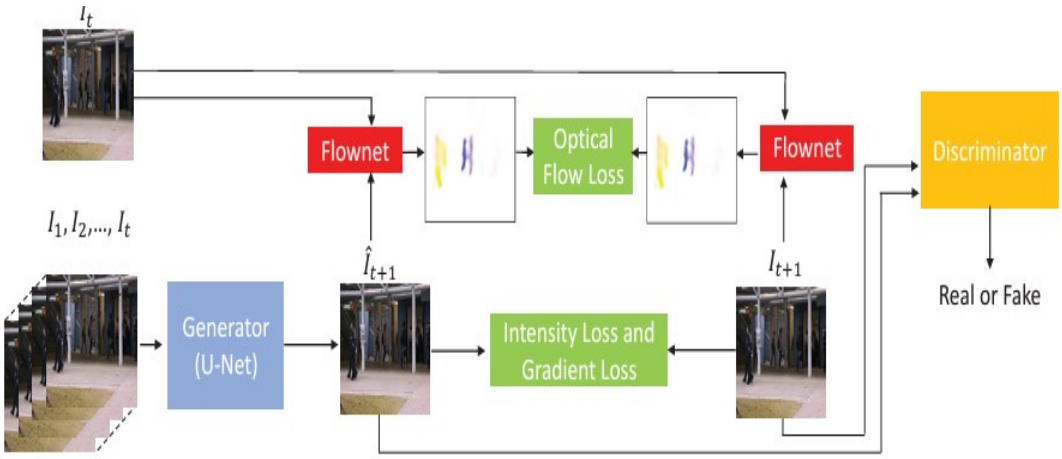


Figure 3.15: Future frame prediction network pipeline [115].

The research paper [114] presents a method for identifying abnormal events in surveillance videos, which has achieved impressive detection rates on standard datasets. The approach operates at an impressive speed of 140-150 frames per second using MATLAB. It effectively harnesses the redundancy in video structure and simplifies the problem to a series of straightforward least square optimization steps, ensuring swift detection without compromising the quality of results. The method utilizes a sparse combination learning technique, taking advantage of the high redundancy in surveillance videos by predefining potential combinations and selecting the best one through the least square error evaluation. The training process involves segmenting video frames into patches, computing 3D gradient features, and learning a sparse basis combination set to minimize reconstruction error.

The method [146] utilizes sparse denoising autoencoders to identify and pinpoint abnormalities by representing each video as cubic patches. It utilizes local

and global descriptors to capture spatial and temporal changes and combines them using Gaussian classifiers to detect anomalies and their positions within frames. A dropout layer is added to the autoencoder to prevent overfitting, transforming it into a denoising autoencoder. Additionally, Kullback Leiber Divergence (KLD) Divergence regularization is applied to generalize the model. Local descriptors evaluate spatiotemporal relations between patches using the Structural Similarity Index Measurement (SSIM), while global descriptors are acquired through autoencoders that compress input data to reveal underlying structures. Gaussian classifiers then differentiate between normal and anomalous patches by computing the Mahalanobis distance, utilizing a threshold derived from training data.

The paper of wang *et al.* [147] presents a method for automatic video anomaly detection and localization using two novel motion-based video descriptors: SL-HOF and Uniform Local Gradient Pattern-Optical Flow (ULGP-OF). The method utilizes the One Class Extreme Learning Machine (OCELM) algorithm and a Robust PCA-based scheme for foreground localization.

The deep spatiotemporal translation network (DSTN) [148] is an innovative method for unsupervised anomaly detection and localization. It utilizes a generative adversarial network (GAN) and edge wrapping (EW) to detect anomalous events by considering appearance and motion characteristics. During the training phase, only frames of regular events are used to create their corresponding dense optical flow as temporal features.

Chang *et al.* [149] presented a convolutional autoencoder framework for detecting abnormal video events. Their approach involves the separate modeling of spatial and temporal information to enhance the identification of abnormal events. Specifically, a spatial autoencoder is employed to characterize normal appearance, while a temporal autoencoder processes consecutive frames to capture motion. In addition, they have incorporated a variance-based attention module in the motion autoencoder to highlight areas with significant movement and utilized a deep K-means clustering strategy to derive concise representations. This method leverages unsupervised learning on normal videos to learn regular patterns and identify anomalies by comparing the final prediction with the actual frame and subsequently determining anomaly scores based on prediction accuracy and cluster distance.

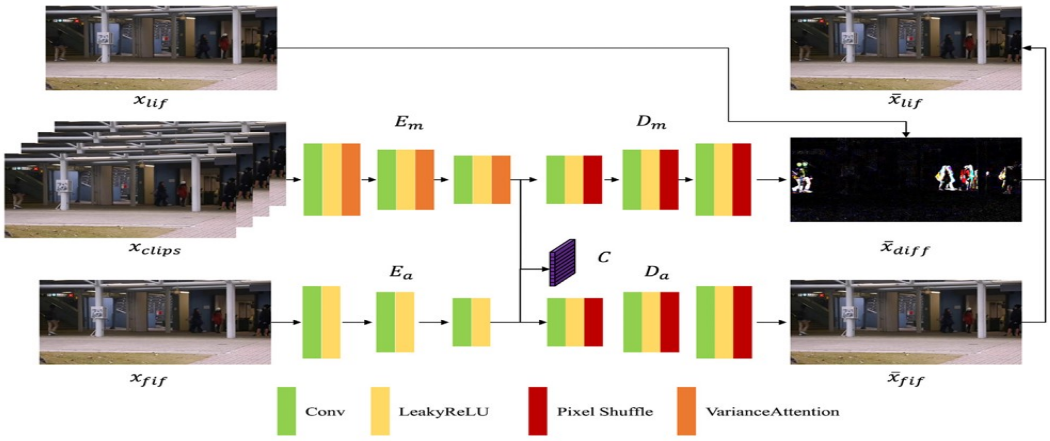


Figure 3.16: Overview of the spatio-temporal dissociation [149] showcasing spatial-temporal modules and deep K-means clustering.

The paper proposed by wang *et al.* [150] introduces a new method called STR-VAD (Spatio-Temporal Relationships for Video Anomaly Detection). It uses a fully convolutional encoder-decoder network with symmetric skip connections and an attention mechanism to understand object spatiotemporal relationships. The method also incorporates a dynamic pattern generator to distinguish anomalies by reinforcing normal pattern reconstruction while making abnormal patterns stand out. Anomalies are identified by analyzing spatio-temporal relationships and movement patterns among objects. The proposed framework consists of three major components: the encoder, the decoder, and the predictor. It includes a two-part loss function: future frame prediction loss and dynamic pattern reconstruction loss.

3.5.2.2 Object-centric based approaches

Object-centric based abnormal behavior detection is the process of detecting strange or abnormal objects by focusing on individual objects rather than carefully examining each frame or video.

The publication by Reiss *et al.* [151] introduces a technique for Video Anomaly Detection (VAD) based on attribute-based representations. The method consists of pre-processing, feature extraction, and density estimation phases. It utilizes existing motion estimators for optical flow prediction and object detection to identify and categorize objects. An anomaly score is calculated for each frame using density estimation and then refined using a temporal Gaussian filter. The

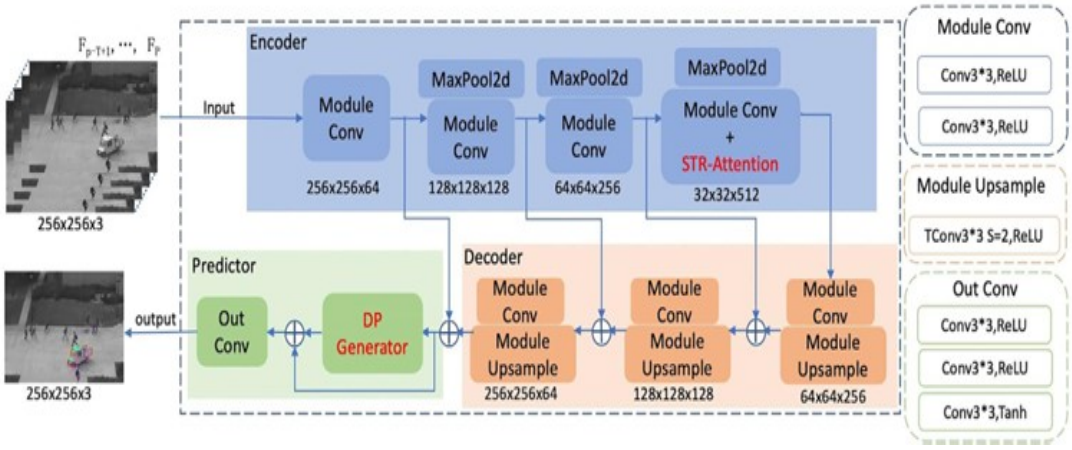


Figure 3.17: Framework of the proposed STR-VAD method [150].

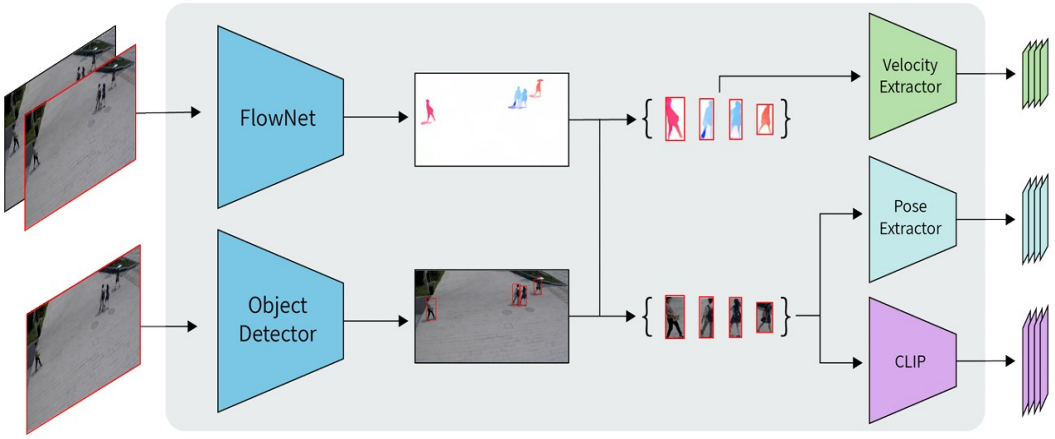
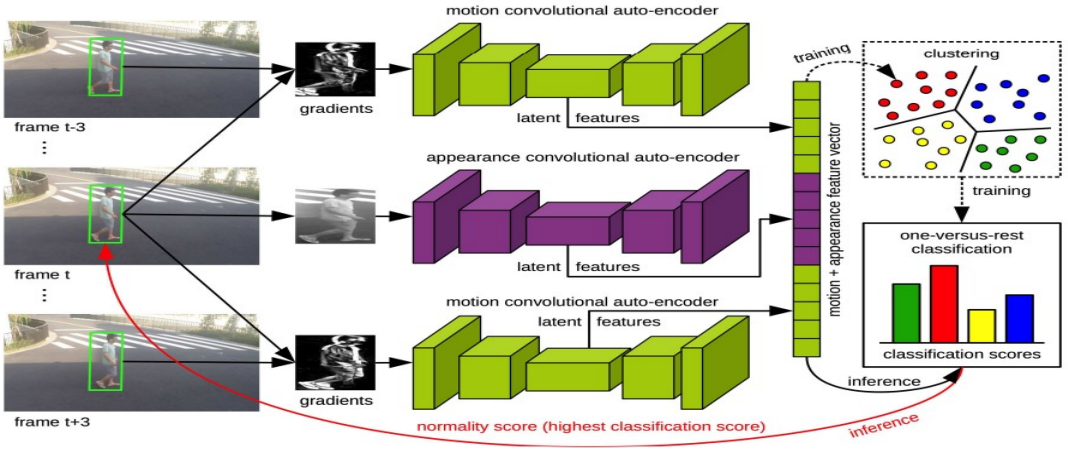


Figure 3.18: Detailed flowchart of [151] approach, illustrating the process from object extraction to feature representation.

authors in [152] present a novel framework for detecting anomalies in video using object-centric auto-encoders and dummy anomalies. They redefine abnormal event detection as a binary classification problem and propose unsupervised feature learning with object-centric convolutional auto-encoders. Moreover, they introduce a supervised classification approach employing one-versus-rest abnormal event classifiers. The proposed framework consists of four sequential stages: object detection, feature learning, model training, and inference. The model training approach formulates abnormal event detection as a multi-class classification task by clustering normal training samples using k-means and a one-versus-rest scheme. During inference, each test sample is classified by k binary SVM models

and temporal smoothing is used using a Gaussian filter to enhance the robustness of frame-level anomaly predictions.



;

Figure 3.19: Convolutional autoencoders are trained on object detection, integrating motion and appearance representations to classify anomalies [152].

The paper [153] introduces a multi-timescale model for precise prediction of pose trajectories and robust anomaly detection. The model comprises two modules, one for predicting future trajectories and another for reconstructing past trajectories. It uses hierarchical prediction frameworks and 1D convolutional filtering to capture temporal dynamics effectively. The model leverages a sliding window approach and a weighted Mean Square Error (MSE) loss function for anomaly detection.

Video Event Completion (VEC) has been proposed [154]. VEC aims to achieve precise and comprehensive localization of video activities using appearance and motion cues. VEC employs Deep Neural Networks (DNNs) to complete visual cloze tests, and it incorporates ensemble strategies to enhance VAD performance. This approach combines appearance and motion cues to localize video activities and extract video events, thereby overcoming the "closed world" problem. Critical components of VEC include using a pre-trained object detector to identify and filter regions of interest (ROIs) based on appearance cues, utilizing temporal gradients to detect motion, and erasing specific patches from spatio-temporal cubes (STCs) to create incomplete events (IEs) for the DNNs to complete.

A deep probabilistic GMM-DAE model by Ouyang *et al.* [155] introduces a model that combines a Denoising Auto-Encoder (DAE) for density estimation with

soft-clustering on deep features. The GMM-DAE model incorporates approximate rank pooling to capture motion information and utilizes a DAE to learn spatiotemporal representations. The model consists of four main parts: image patch generation, encoding and decoding using two deep DAEs, density estimation, and anomaly inference. YOLOv3 is employed for patch extraction, and the DAE utilizes convolutional layers for encoding and transposed convolutional layers for decoding. Density estimation involves clustering latent representations using a Gaussian Mixture Model (GMM) trained with Expectation Maximization (EM), which produces soft clusters.

A new **Hybrid Video Anomaly Detection Framework (HF2-VAD)** [156] has been developed for video anomaly detection. It combines flow reconstruction and frame prediction. The framework includes two main components: the Multi Level Memory augmented Auto-Encoder with Skip Connections (ML-MemAE-SC) for optical flow reconstruction and the CVAE for predicting future video frames. The ML-MemAE-SC network is designed to learn normal patterns in optical flow using memory modules and skip-connections, resulting in better reconstruction of normal flows and higher errors for abnormal ones. The Conditional Variational Auto-Encoder (CVAE) model predicts the next video frame by considering previous frames and the reconstructed flows, taking advantage of the strong correlation between video frames and optical flows. This dual-error approach enhances the accuracy of anomaly detection in videos.

The paper of Georgescu *et al.* [157] introduces a new method for detecting anomalous video events using self-supervised and multi-task learning techniques. It utilizes a pre-trained object detector and a 3D CNN that learns multiple proxy tasks, including self-supervised and knowledge distillation tasks. The approach integrates these tasks into a single architecture, making it the first for video anomaly detection. The shared 3D CNN architecture and four independent prediction heads allow for effective learning of the combined tasks. During inference, anomaly scores are calculated for each task and averaged for precise localization and addressing potential false negatives.

The paper **Predicting Next Local Appearance for Video Anomaly Detection (NLAPnet)** [158] introduces a new method for detecting video anomalies. NLAPnet uses an adversarial framework to predict the appearance of normally behaving objects in a scene and detect deviations from expected behavior. It features a computationally efficient single generator with a U-Net architecture,

skip connections, and an adversarial loss component. The method incorporates the structural similarity index measure (SSIM) and pre-trained Deep Layer Aggregation (DLA) [159] backbone for object detection for real-time video anomaly detection. Additionally, it utilizes grayscale pixel-level intensity images of objects in past, current, and next frames and a pre-trained multi-class object detector called CenterNet.

The paper "Local Anomaly Detection in Videos using Object-Centric Adversarial Learning" [160] proposes a novel unsupervised method for detecting local anomalies at the frame level in videos. It introduces a two-stage object-centric adversarial framework, utilizing a GAN called Gradient-Appearance Relation Discovery Network (GARDiN), which focuses on object regions for detection. In the first stage, the method models the relationship between the current appearance and the past gradient images of objects in normal scenes. The second stage computes region-level anomaly detection scores by analyzing partial reconstruction errors between real and generated images. Regions of interest are extracted from each frame using a pre-trained object detector, and spatial gradients are calculated from the previous frame. Additionally, two cross-domain generators and discriminators are trained to predict past gradients from appearances and vice versa, enabling the detection of anomalies by differentiating real from generated images.

The paper of Wang *et al.* [161] introduces **Spatio-Temporal Auto-Trans-Encoder (STATE)**, an autoencoder designed for improved consecutive frame reconstruction. It integrates a learnable convolutional attention module for efficient temporal learning and proposes a novel reconstruction-based input perturbation technique to enhance the differentiation of anomalous frames. An anomaly score combines raw and motion reconstruction errors with perturbed inputs. The approach leverages pre-trained object detection models to extract object-centric patches for sequential patch reconstruction. STATE employs a unique architecture combining transformer-like self-attention with convolutional auto-encoders for robust temporal reasoning and reconstruction.

The paper [162] proposes a video anomaly detection (VAD) approach that solves **spatiotemporal jigsaw puzzles** as a multi-label fine-grained classification task. This approach decouples spatio-temporal jigsaw puzzles, utilizes total permutations, and offers advantages over existing methods. It simplifies self-supervised learning by focusing on a single pretext task—decoupled spatial

and temporal jigsaw puzzles. The method avoids pre-trained models, relying solely on challenging pretext tasks to learn rich representations. This work presents a challenging pretext task: solving spatiotemporal jigsaw puzzles. The approach breaks down 3D spatiotemporal puzzles into spatial and temporal components corresponding to learning appearance and motion patterns. Object-centric cubes are constructed during training, and spatial or temporal shuffling is applied to create jigsaw puzzles.

The paper [163] presents a new method for video anomaly detection that identifies anomalous objects using spatiotemporal relationships. It uses a convolutional encoder-decoder network with skip connections, an attention mechanism, and a dynamic pattern generator. The model Object-centric Memory-guided Auto-Encoder (OMAE), first detects objects, computes optical flow, extracts appearance and motion features, encodes them, and uses them as a query to retrieve similar prototypes from a memory module. An anomaly score is computed based on feature reconstructions and query similarity. The model is trained using various loss functions, including reconstruction loss and memory loss, to learn normal patterns.

The paper [164] introduces a new **multi-level attention network** for video anomaly detection, incorporating frame-level and object-level semantics through the Object-Guided Attention Module (OGAM) and the Motion-Refined Attention Module (MRAM). This approach aims to enhance future frame prediction by highlighting salient objects within frame features and aligning local object features with global frame features. The network effectively integrates object-focused attention mechanisms into frame prediction tasks, enhancing prediction accuracy and enabling precise anomaly localization.

The paper of Doshi *et al.* [165] introduces a framework for **interpretable video anomaly detection**. It monitors object interactions to detect anomalous events and provide contextual interpretations utilizing scene graphs. The framework comprises global and local object monitoring branches, feeding into a sequential anomaly detection module. It leverages YOLO-v4 for object detection, VGG-16 for appearance feature extraction, and AlphaPose for pose estimation during implementation. The training process involves relationship detection and pose data augmentation to optimize models for robust anomaly detection across different surveillance scenarios.

An Object-centric Scene Inference Network (OSIN) [166] has been proposed

for video anomaly detection tasks. This model comprises three streams - temporal, spatial, and object - which capture motion information, static scene features, and object interactions within the overall scene. The integrated streams are fed into a decoder for future frame prediction, where significant prediction errors during testing indicate anomalies. The OSIN model effectively integrates spatial-temporal normality and object-scene interactions within a unified framework, utilizing a three-stream network to address different aspects of normality. The training loss encompasses prediction loss, scene loss, and object loss. At the same time, the anomaly score is calculated based on prediction errors and the distance between the current scene and the learned scene memory.

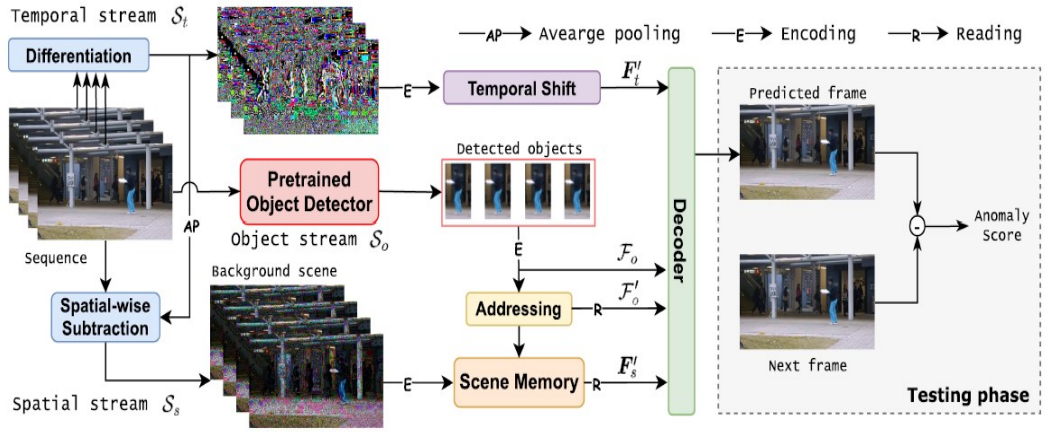


Figure 3.20: Architecture of the proposed OSIN model [166], combining temporal, spatial, and object streams to capture motion, global scene appearance, and object-scene interactions for video anomaly detection.

The paper [167] proposes a framework known as Appearance-Motion Semantics Representation Consistency (AMSRC) that aims to tackle the challenges of video anomaly detection. Its primary focus is improving the consistency of feature semantics between appearance and motion information in normal samples to facilitate anomaly detection. The AMSRC framework comprises a two-stream encoder, a decoder, and a gated fusion module. The gated fusion module processes selectively to generate outputs that differ significantly from pre-fusion representations, thereby enhancing anomaly detection. During testing, the anomaly score combines inconsistency in appearance-motion feature semantics with the prediction error of future frames using a weighted sum strategy.

3.6 Limitations and Considerations for Improvement

Although most of these methods have provided satisfactory results, some limitations must be considered. The limitations of contemporary deep learning-based techniques have necessitated the consideration of certain constraints. In particular, issues such as occlusion within a scene have been observed in existing research, which complicates the task of tracking objects and leads to suboptimal results when analyzing their trajectories. While frame-based algorithms may initially yield promising outcomes, it is crucial to ensure that detected anomalies are correctly associated with the intended objects. This can only be achieved by accurately localizing the anomalous objects and subsequently evaluating them after determining their precise location. Additionally, many algorithms rely on the availability of labeled data, thereby restricting their applicability to datasets where such labeled data is abundant. Furthermore, the complexity of these algorithms and the methods used to extract temporal and spatial features may negatively impact computational efficiency and performance speed. These challenges have motivated us to propose an unsupervised, low-complexity approach that can be effectively applied in contexts where labeled data is scarce.

3.7 Conclusion

This chapter explores video anomaly detection methodologies, covering classical, object-tracking, and modern approaches. It highlights vital datasets and commonly used performance metrics, while also addressing the limitations of current methods.

We provide a comprehensive and meticulous analysis of traditional techniques, such as tracking and path analysis, followed by a review of the deep learning-based approaches. Special emphasis is placed on frame-based abnormal behavior detection techniques, including optical flow and background subtraction. Additionally, we examine algorithms targeting anomalies in objects extracted from video.

The chapter concludes with a review of leading deep learning models, evaluating their strengths and limitations in analyzing object behavior in video data.

CHAPTER 4

FOREGROUND SEGMENTATION: A DEEP NESTED NETWORK FOR BACKGROUND SUBTRACTION (NESTED-NET)

Contents

4.1 Introduction

This chapter presents an advanced background subtraction method that utilizes residual micro-autoencoder blocks. This method is based on a U-net architecture, with additional skip connections to enhance performance. The method effectively captures essential multi-scale features from complex scenarios by incorporating residual connections between micro-autoencoders. Our evaluations demonstrate that this technique outperforms other state-of-the-art methods on benchmark datasets, even under challenging conditions. Notably, the model operates without the need for temporal data or post-processing, and it can achieve high performance with only a few training examples.

4.2 The proposed method: Deep nested network for background subtraction (Nested-Net)

4.2.1 Introduction

Convolutional neural networks (CNNs) have become the preferred method for addressing computer vision challenges, such as object detection and background subtraction. We have developed a robust approach for video background subtraction using a unique architecture. Our model incorporates a nested network with multiple skip connections that link residual mini-autoencoders to enhance feature generalization by extracting multi-scale features at each level. In addition, we utilize a pre-trained VGG-16 model and introduce a new supervised deep-learning model within the nested network.

4.2.2 Methodology

Our U-net-like network depicted in figure 4.1 consists of interconnected residual fine-grained autoencoders (Residual Micro-AutoEncoder (RM-AE)) blocks with additional residual connections. To begin, we will outline the complete architecture of the proposed RM-AE and then discuss the Spatial Feature Pooling Module (SFPM) module, which serves as an intermediate feature map of our network.

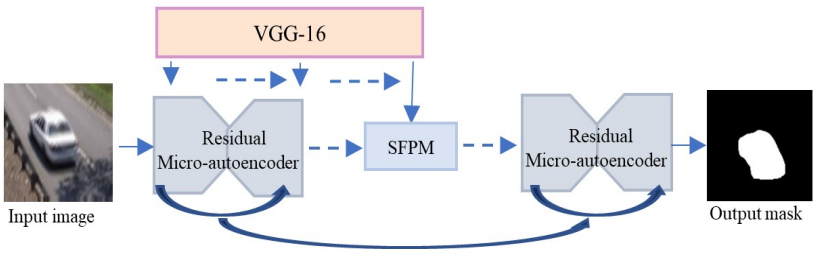


Figure 4.1: Architecture Overview of the Nested-Net Model.

4.2.2.1 Residual micro-autoencoder (RM-AE)

Our proposed enhanced residual micro-autoencoder (RM-AE) is designed to improve feature generalization by capturing a broader range of multi-scale features. The RM-AE architecture is shown in Figure 4.2. It consists of mini-encoder and mini-decoder levels, each having a single convolutional layer. These levels are combined to form an RM-AE. Specifically, the mini-encoder (E) utilizes a convolutional layer to convert the input feature into an intermediate map, progressively reduced by half through a downsampling operation ($MaxPooling2D$). The mini-decoder (D) processes the intermediate feature map using an upsampling layer with bilinear interpolation. This process ensures that the mini-ecoder's last layer produces a feature map with the same spatial dimension as the mini-encoder. The output layers of the mini-encoder and mini-decoder are combined into a single feature map via a residual link $E(x)+D(x)$. Finally, the resulting feature map is passed through a Rectified Linear Unit (ReLU) activation layer and batch normalization.

Our proposed RM-AE is constructed based on the research by Akilan *et al.* [88, 89]. Akilan *et al.* [88] introduced a time-dependent background subtraction technique in 3D, utilizing a series of micro-autoencoders based on Conv-LSTM, 3D convolution, and transposed 3D convolution layers for spatiotemporal cues. Their subsequent work, sEnDec [89], improved this approach by using $2D$ CNN modeling with $2D$ convolutions and $2D$ transposed convolutional layers instead of $3D$ convolutions. However, these techniques led to unsatisfactory outcomes and increased model complexity. In contrast, the structure of our RM-AE, detailed in Figure 4.2, featuring 3×3 kernel sizes with a consistent number of 64 filters across all layers. Additionally, we utilized pooling and upsampling layers with residual summation connections instead of convolution and transposed convolution with two strides and concatenation connections to reduce model complexity.

The structure of our RM-AE is shown in Figure 4.2. As illustrated, all layers utilize a 3 × 3 kernel size, with the number of filters consistently set to 64.

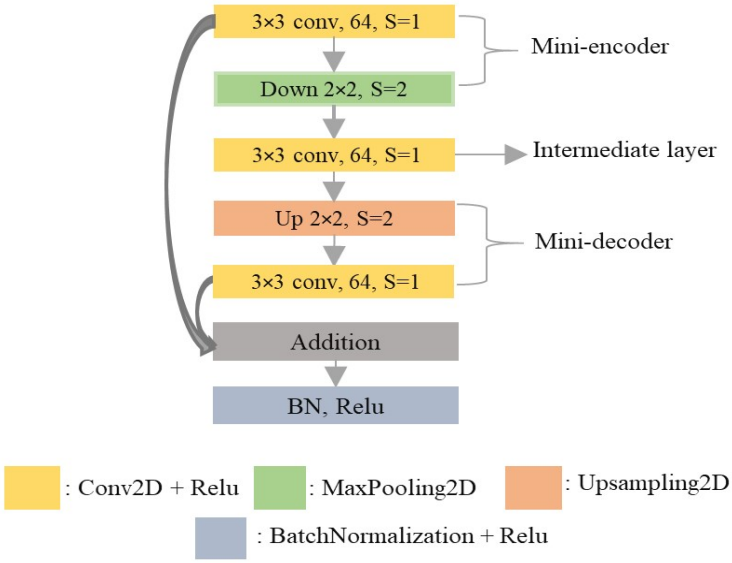


Figure 4.2: The flow of micro-autoencoder (RM-AE)

4.2.2.2 Spatial feature pooling module (SFPM)

Our objective in introducing SFPM is to enrich our Nested-Net by integrating various spatial patterns and combining convolutions with different kernel sizes. This concept resembles Szegedy *et al.* [168], but we employ a more extensive convolution using a 7 × 7 kernel size. To achieve the sparse structure of the resulting feature map, we utilized a straightforward spatial feature pooling module illustrated in Figure 4.3. We employed multiple convolutions with different kernel sizes of 1 × 1, 3 × 3, 5 × 5, 7 × 7, in addition to a max-pooling layer with a size of 2 × 2. These spatial scales were applied to the feature map F derived from the output of the pre-trained VGG-16. In Figure 4.3, we utilized a 1 × 1 convolution for dimensionality reduction before implementing the more computationally intensive 3 × 3, 5 × 5, and 7 × 7 convolutions. Subsequently, a 2 × 2 pooling layer followed by a 1 × 1 convolution was employed to obtain a smaller output dimension. The resulting layers were merged through a concatenation cue along the depth channel. All the convolution layers comprised 64 depth filters.

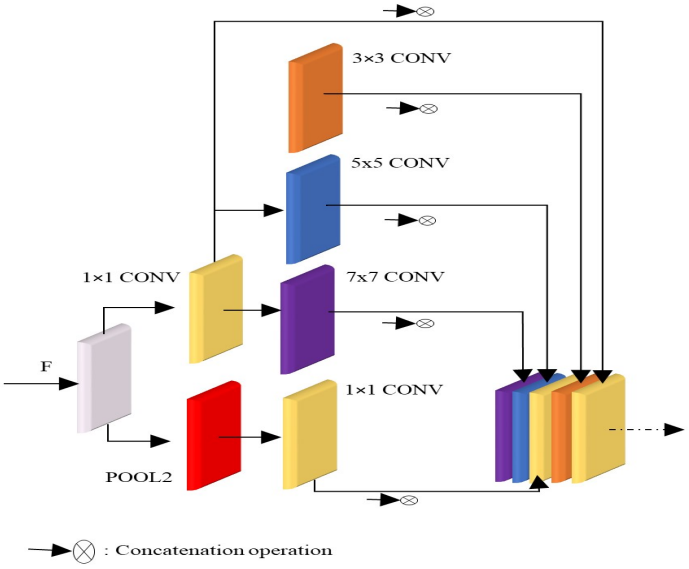


Figure 4.3: The flow of spatial feature pooling module (SFPM).

4.2.2.3 Architecture of Nested-Net

Indeed, the key components of our network consist of the RM-AE blocks and SFPM. The proposed Nested-Net comprises four main parts: encoder, intermediate map represented as an SFPM module, decoder, and classifier, requiring six RM-AEs, three for the encoder and three for the decoder. The encoder section is built upon the pre-trained VGG-16 [169], previously trained on ImageNet [170]. We advocate for using VGG-16 for several reasons:

1. VGG-16 shares a U-Net-like structure, which aligns conveniently with our network architecture in Nested-Net.

2. Its relatively smaller model size than other pre-trained networks translates to superior computational efficiency.

3. The abundance of filters and iterative structure of VGG-16 allows for robust feature extraction from input images, resulting in excellent segmentation performance. This becomes particularly advantageous when using its outputs as inputs for each RM-AE, where we employ a small number of filters in each RM-AE.

The network structure utilized the first four blocks of VGG-16 but excluded the fifth block. Specifically, only the first layer of the fourth block was used to

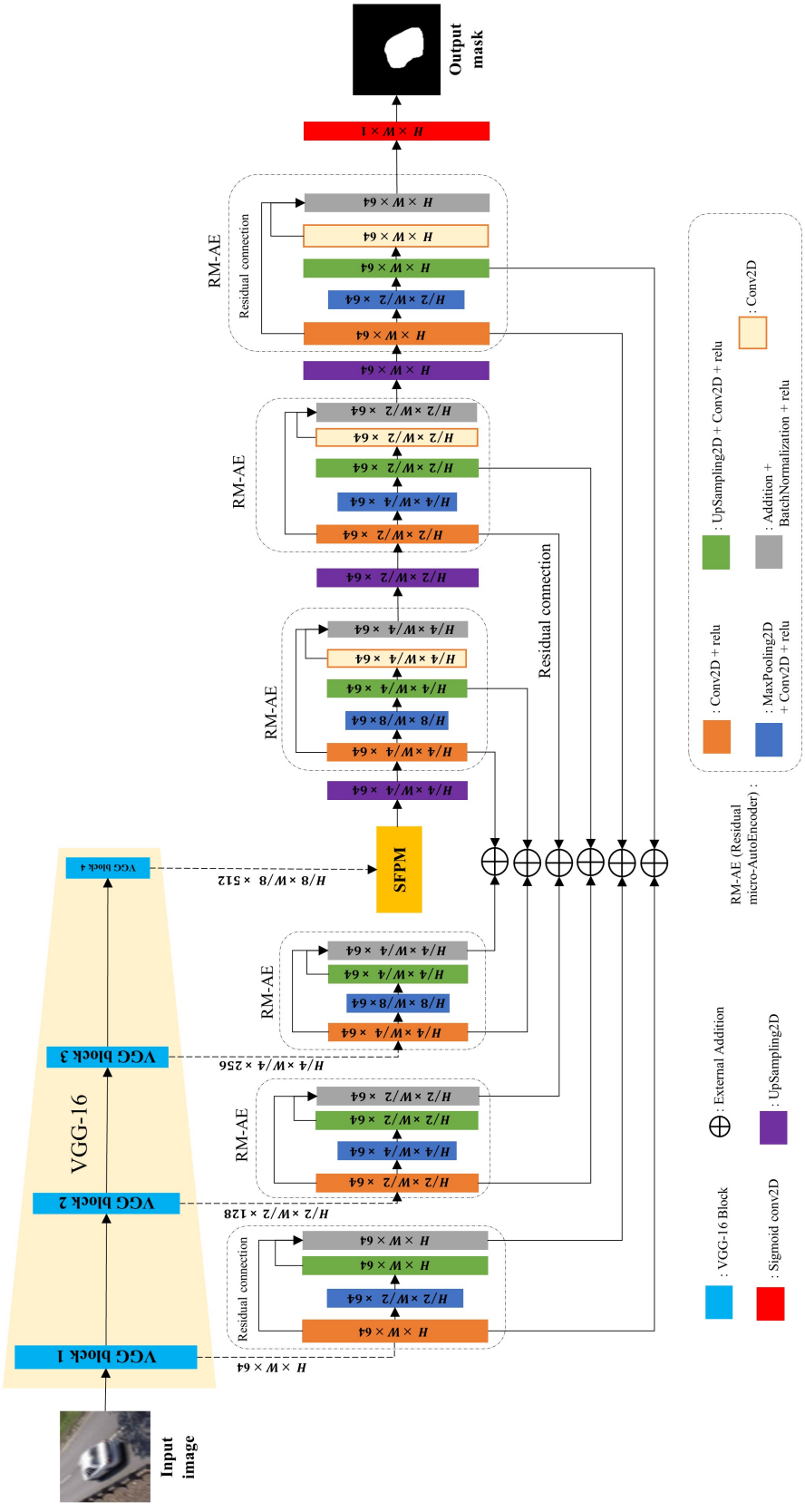


Figure 4.4: Comprehensive structure of the Nested Network.

simplify the model and reduce the number of parameters, as this block serves as an input to the SFPM. A 3×3 kernel size was employed for feature extraction in each RM-AE block, reducing space usage and enhanced scalability. The decoder network comprised three sequentially linked RM-AE blocks. The feature maps of the decoder were guided by the low-level features from the encoder's mini-encoder and mini-decoder blocks.

The structure of the RM-AE blocks in the decoder mirrored that of the encoder, with the distinction that each RM-AE in the decoder contained more layers. Like the U-net, each RM-AE in the decoder received connections from the encoder, including summation operations applied to the output of additional convolutional layers used in the mini-encoder and mini-decoder of the network. Specifically, all mini-encoder and mini-decoder outputs were added to the encoder's mini-decoder and mini-encoder output features. This process was repeated for each RM-AE. The stacks of three RM-AE blocks concluded with a 1×1 convolutional layer with a single feature slice and a sigmoid function as the classifier to generate the final probability map. Figure 4.4 presented an overview of the entire Nested-Net architecture.

4.3 Training details

Our model utilizes two inputs: an RGB input frame and its corresponding ground truth. We retain the weights for the first two blocks from the pre-trained VGG-16 model with the same coefficient to leverage transfer learning and fine-tune the rest. The proposed Nested-Net model is trained using the RMSprop gradient optimizer with a maintained rho of 0.9 and an initial learning rate 0.0001. If the validation loss does not improve after ten epochs, the learning rate is decreased by a factor of 10. We set the maximum number of epochs to 200 and the batch size to 1. Concurrent training and validation are conducted using 20% of the training dataset for validation and 80% for training. An early stopping function is implemented, causing the model to cease learning if the validation loss does not improve after 20 epochs. To prevent bias from repeating information in consecutive frames, the training samples are randomly shuffled before each step and before each epoch. After each epoch, our network is trained on each scene. The validation loss is calculated using a binary cross-entropy loss function (Ł), which is utilized to compare each pixel's accurate and predicted label. The function is

defined by equation Eq. 4.1 as follows:

$$\mathscr{L} = -\frac{1}{N_p}\sum_{p=1}^{N_p}\left[P_t \log\left(\overline{P}_t\right) + (1-P_t)\log\left(1-\overline{P}_t\right)\right] \tag{4.1}$$

where N_p is the total number of trained pixels, P_t is the current foreground pixel probabilities $P_t \in [0,1]$, and $\overline{P}_t$ is its corresponding ground truth label.

4.4 Comparative analysis and evaluation schemes

This section presents a detailed comparison and summary of our proposed network to the latest techniques. This encompasses vital features, evaluation frameworks, training data, and significant contributions outlined in Table 5.2. Our proposed network has practical implications in the field of background subtraction, which we will discuss in detail. In background subtraction, most methods are trained using scene-dependent (Scene Dependent Evaluation (SDE)) and scene-independent evaluations (Scene Independent Evaluation (SIE)). SDE involves assessing the method's performance in a specific video sequence, considering each video's unique properties. On the other hand, SIE evaluates the method's effectiveness across multiple sequences without considering the specific characteristics of each video. Scene-independent evaluation is valuable for determining the method's applicability across diverse scenes. At the same time, scene-dependent assessment is crucial for realistically measuring the method's performance and highlighting its strengths and limitations in a particular scene.

4.5 Experimental settings, results, and discussions

Nested-Net was tested using the CDnet 2014 and SBI 2015 datasets discovered in Chapter 2. It has also been compared to recent supervised methods based on empirical data. Quantitative and qualitative analyses are included for each dataset's results using the performance metrics Assessment, which involves using various metrics identified in Chapter 2, such as average F-measure, recall, specificity, precision, false-positive rate, false-negative rate, and percentage of the

Method	Input	BE	Network type	Pretrained weights	Main contributions	Train/Val data	Testing data	Evaluation framework
FgSegNet_S [80]	Single RGB scale	N	CNN	VGG-16	- Utilizing the VGG-16 as an encoder. - Introducing an FPM module made up of convolutions with varying dilation rates.	Selective 200 frames for CDnet 2014 Selective 20% of frames for SBI 2015	Remaining frames	SDE
FgSegNet_V2 [81]	Single RGB scale	N	CNN with skip connections	VGG-16	- Improved over FgSegNet_S by enhancing the FPM module and introducing global average pooling layers to improve the features generalization in the decoder.	Selective 200 frames for CDnet 2014 Selective 20% of frames for SBI 2015	Remaining frames	SDE
ResNet50_FPM [82]	Single RGB scale	N	CNN with skip connections	Resnet50	- Improved over FgSegNet_V2 - Resnet50 is used as an encoder instead of VGG-16, followed by an improved FPM module.	Selective 200 frames for CDnet 2014	Remaining frames	SDE
sEnDec [89]	2 gray frames	Temporal median	CNN with skip connections	N	- An autoencoder-like micro modules is proposed to perform a slow-decoding approach.	70% of frames	Remaining frames	SDE
BSPVGAN [86]	2 frames	Temporal median	Generative adversarial networks	N	- A background image is extracted via median filtering, and then a bayesian GAN is utilized to identify the foreground objects in video sequences.	Random 200 frames for CDnet 2014.	100%	SDE
Cascade_CNN [79]	Single frame patch based	N	CNN	N	- A cascaded architecture is used to train a scene-specific multi-resolution CNN.	Selective 200 frames for CDnet 2014. Selective 20% of frames for SBI 2015	Remaining frames	SDE
BSUV-Net [84]	3 frames	Temporal median	CNN with skip connections	DeepLabv3	- It feeds a CNN with the current and two reference background images, along with their semantic segmentation and novel data augmentation models.	18 different sets	100%	SIE
DeepBS [74]	2 frame patch based	SuBSENSE	CNN	N	- SubSENSE is used to produce a background image, which is then partitioned into patches and concatenated with the current frame to form the input layer.	Random 5% of frames	Remaining frames	SDE
The proposed Nested-Net	Single RGB scale	N	CNN with skip connections	VGG-16	- Proposing a new deep-learning model for background subtraction using a novel architecture. - VGG-16 is used as an adhered input to the encoder - Introducing an efficient feature extraction module (RM-AE). - Introducing an SFPM module made up of various convolution filters with different kernel sizes.	Selective 200 frames for CDnet 2014 Selective 20% of frames for SBI 2015	Remaining frames	SDE

Table 4.1: An In-Depth Comparison of Our Proposed Nested Network with SOTA Approaches, Evaluating Both Network Architecture and Experimental Configurations (BE: Background Estimator).

wrong classification. These metrics are derived from true positive, true negative, false positive, and false negative.

4.5.1 Experiment on CDnet 2014

In this section, we followed the same methodology outlined in Lim and Keles [80, 81] and Wang *et al.* [79] for the CDnet 2014 database. A separate model was generated for each scene, using a set of 200 frames for both training and validation. These frames were manually selected and arranged randomly, focusing on frames containing the labeled ground truth. Following the testing phase, the model's output was presented as probability masks with values ranging from 0 to 1, which were then transformed into binary images using a consistent threshold of 0.8.

4.5.1.1 The spatial feature pooling module (SFPM) experiments

In this study, we assessed the effectiveness of the SFPM module on CDnet 2014. We followed the same protocol as in previous experiments for this experiment, allocating 200 frames for training and validation and using the remainder for testing. Our analysis, detailed in Table 5.3, focused on the performance of the test frames that did not include SFPM in the training set. Our results revealed a modest 0.15% improvement in utilizing the network with SFPM compared to the network without SFPM across most categories. However, we observed a notable 0.44% performance advantage in the PTZ category for the network with SFPM. These findings suggest that employing convolutions with diverse kernel sizes enhances the efficacy of foreground segmentation masks.

	F-measure											Overall
	badWeat	baseline	camJit	dynBg	intermit	lowFrame	nightVid	PTZ	shadow	thermal	turbul	
no_SFPM	0.9805	0.9972	0.9919	0.9832	**0.9939**	0.8951	0.9739	0.9769	0.9955	0.9912	0.9751	0.9777
with_SFPM	**0.9819**	**0.9973**	**0.9933**	**0.9856**	0.9938	**0.8974**	**0.9759**	**0.9813**	**0.9957**	**0.9929**	**0.9756**	**0.9792**

Table 4.2: Quantitative Evaluation of Nested-Net With/Without SFPM Integration (Top Results Highlighted in Bold).

4.5.1.2 Sanity check

We conducted a comprehensive comparison between our Nested-Net (without SFPM) and the slow Encoder-Decoder (sEnDec) [89] method, which served as the

Method	F-measure								Overall
	Hw	Tr	Sk	Ops	Blz	Blv	Tst	Shd	
sEnDec [89]	0.9673	0.8701	0.9610	0.9491	0.9608	0.9300	0.8899	0.9549	0.9354
Nested-Net (proposed)	**0.9979**	**0.9887**	**0.9949**	**0.9955**	**0.9889**	**0.9954**	**0.9597**	**0.9984**	**0.9899**

Table 4.3: Analysis of Results: Comparison of Nested-Net Without SFPM to sEnDec. The selected scenes for comparison are: Highway (Hw), Traffic (Tr), Skating (Sk), Overpass (Ops), Blizzard (Blz), Boulevard (Blv), TramStation (Tst), and PeopleInShade (Shd).

inspiration for our proposed residual micro-autoencoder (RM-AE). The evaluation used the CDnet 2014 dataset, explicitly focusing on the video sequences utilized in their paper. It is important to note that we conducted an independent comparison, as the original assessment only covered a portion of the dataset. A detailed comparison of our Nested-Net and their findings is provided in Table 4.3. Our analysis revealed that our network surpasses sEnDec by 5.45 percentage points, solidifying the outstanding performance of our approach.

4.5.1.3 Quantitative results

We assessed the performance of our proposed network by using standard evaluation measures such as average F-measure, average recall, average precision, specificity, false-negative rate (FNR), percentage of wrong classification (PWC), and false-positive rate (FPR) listed in Table 4.4. These measures were computed using only the test frames. The results revealed an average F-measure of 97.92% and a percentage of wrong classification of 3.57%, demonstrating strong performance across all categories in the CDNet 2014 benchmark dataset. Additionally, our proposed network, Nested-Net, was compared with eight existing supervised state-of-the-art approaches. The comparative analysis in Table 4.5 showcased that Nested-Net achieved a higher average F-measure value (98.91%) for the CDnet 2014, outperforming other methods by margins ranging from 0.26% to 22.98%. Nested-Net also exhibited competitive performance in some categories while yielding better results in most categories, including low frame rate, night videos, and turbulence categories. Notably, all methods could have performed better at low frame rates, with our proposed network achieving a score of 96.17%. This limitation is attributed to the challenge of identifying small moving objects in the port_0_17fps scene, which proves difficult even for human observers. In summary, these experiments highlight the effectiveness of our proposed technique

in detecting foreground in scenes with various challenging conditions.

Category	Recall	Specificity	FPR	FNR	PWC	Precision	F-Measure
baseline	0.9962	0.99996	0.00004	0.0038	0.0123	0.9984	0.9973
cameraJit	0.9907	0.9999	0.00015	0.0093	0.0473	0.9958	0.9933
badWeather	0.9741	0.9999	0.0001	0.0259	0.0330	0.9900	0.9819
dynamicBa	0.9914	0.99997	0.00003	0.0086	0.0074	0.9802	0.9856
intermitt	0.9911	0.9998	0.0002	0.0089	0.0731	0.9968	0.9938
lowFramer	0.9174	0.9999	0.0001	0.0826	0.0274	0.8819	0.8974
nightVide	0.9690	0.9997	0.0003	0.0310	0.0774	0.9829	0.9759
PTZ	0.9833	0.99996	0.00004	0.0167	0.0141	0.9796	0.9813
shadow	0.9944	0.9999	0.0001	0.0056	0.0285	0.9967	0.9957
thermal	0.9889	0.9999	0.0001	0.0111	0.0490	0.9970	0.9929
turbulence	0.9744	0.9999	0.00008	0.0256	0.0231	0.9770	0.9756
Overall	0.9792	0.9999	0.0001	0.0208	0.0357	0.9797	**0.9792**

Table 4.4: Performance Evaluation of Nested-Net on CDnet 2014 Dataset Test Frames.

4.5.1.4 Qualitative results

The following information should be noted: In this section, we compared the segmentation masks generated by our proposed network with those produced by existing methods. We evaluated seven random scenes from the CDnet 2014 dataset: office, traffic, skating, fluidHighway, intermittentPan, busStation, and diningRoom, shown from left to right, respectively. We visually assessed the effectiveness of our Nested-Net in comparison to earlier supervised methods such as FgSegNet_V2 [81], FgSeg-Net_S [80], BSPVGAN [86], and cascade_CNN [79] on the mentioned scenes. The segmentation masks provided by our method and the techniques above are displayed in Figure 4.5. Our Nested-Net showcased outstanding performance in foreground detection, particularly in the NightVideo category, where our model consistently detected objects that other approaches struggled to identify accurately. However, in the port_0_17fps scene, categorized as having a low frame rate, our approach encountered challenges in identifying small objects, as illustrated in Figure 4.6, resulting in slightly unsatisfactory results in this particular scene. Compared to other methods, our approach demonstrated the ability to detect foreground objects with fewer artifacts, as seen in the office scene. As a result, the results produced by our Nested-Net provide reliable foreground detection.

Method	badWeat	baseline	camJit	dynBg	intermit	lowFrame	nightVid	PTZ	shadow	thermal	turbul	overall
BSUV-Net+Semantic [84]	0.8671	0.9639	0.7788	0.8176	0.7601	0.6945	0.6635	0.6595	0.9664	0.8455	0.7961	0.8012
BSUV-Net [84]	0.8649	0.9693	0.7743	0.7967	0.7498	0.6967	0.6885	0.6320	0.9233	0.8581	0.7513	0.7914
DeepBS [74]	0.8647	0.9580	0.8990	0.8761	0.6097	0.5900	0.6359	0.3306	0.9304	0.7583	0.8993	0.7593
Cascade_CNN [79] #	0.9451	0.9786	0.9758	0.9658	0.8505	0.8804	0.8926	0.9344	0.9593	0.8958	0.9215	0.9272
BSPVGAN [86] #	0.9619	0.9836	0.9893	0.9849	0.9366	0.8960	0.9018	0.9522	0.9849	0.9764	0.9539	0.9565
FgSegNet_S [80]#	0.9868	0.9977	0.9957	0.9958	0.9940	0.9326	0.9628	0.9888	0.9927	0.9937	0.9708	0.9829
FgSegNet_V2 [81] #	0.9884	0.9978	0.9971	0.9951	0.9961	0.9433	0.9662	0.9904	0.9955	0.9938	0.9762	0.9854
Resnet50_FPM [82] #	0.9888	0.9977	0.9878	0.9947	0.9952	0.9575	0.9796	0.9831	0.9962	0.9922	0.9787	0.9865
Nested-Net (proposed)	0.9883	0.9979	0.9961	0.9938	0.9947	0.9617	0.9831	0.9914	0.9965	0.9947	0.9813	0.9891

Table 4.5: F-measure Performance Analysis on CDnet 2014 (Red and Blue Highlight the Top Two Results). # All compared SOTA methods were trained and evaluated under the same SDE conditions as the Nested-Net.

Figure 4.5: Visual Analysis Outcomes on CDNet 2014. The rows indicate, from top to bottom: the input frame, ground truth, our proposed Nested-Net, FgSegNet_V2 [81], FgSegNet_S [80], BSPVGAN [86], and Cascade_CNN [79]. Columns show seven examples from CDnet 2014: Baseline, CameraJitter, BadWeather, NightVideos, PTZ, Shadow, and Thermal, arranged from left to right.

Figure 4.6: Visual results for Our Method in the Low Frame Rate Category (port_0_17fps Scene), Where Performance Issues Occur in Some Sequences. Left to Right: Input Frame, Ground Truth, and Our Proposed Nested-Net.

4.5.2 Experiment on SBI 2015 and UCSD dataset

The Scene Background Initialization (SBI) and UC San Diego dataset (UCSD) databases were used and handled according to a procedure similar to that in references [79–81], wherein 20% of the frames were used for training and validation, while 80% were reserved for testing. A threshold of 0.2 and 0.6 was implemented to convert the predicted masks into binary masks for the SBI and UCSD datasets, respectively.

4.5.2.1 Quantitative results

We conducted an additional evaluation of our model using the SBI 2015 and UCSD benchmarks. We calculated the average F-measure metric, excluding the training frames from the assessment. The resulting scores are outlined in Table 4.6, displaying the F-measure for each scene and the overall average score derived only from the test frames.

Scene	F-measure
Board	0.9979
Candela_m1.10	0.9932
CAVIAR1	0.9989
CAVIAR2	0.9806
CaVignal	0.9895
Foliage	0.9775
HallAndMonitor	0.9902
HighwayI	0.9931
HighwayII	0.9958
HumanBody2	0.9910
IBMtest2	0.9829
PeopleAndFoliage	0.9896
Snellen	0.9727
Toscana	0.9272
Nested-Net (proposed)	0.9843
Cascade_CNN [79] #	0.8932
BSPVGAN [86]	0.9190
FgSegNet_S [80] #	0.9831
FgSegNet_V2 [81] #	0.9838

Table 4.6: Quantitative Evaluation of the Proposed Nested-Net on Test Frames from the SBI 2015 Dataset, Compared to Four Recent Methods (Red Represents the Best Result, and Blue Represents the Second Best). # These approaches were trained and assessed using the identical SDE setup as the proposed Nested-Net.

According to Table 4.6, the proposed Nested-Net outperforms previous studies, namely FgSegNet_V2 [81], FgSegNet_S [80], BSPVGAN [86], and cascade_CNN [79], by 0.05%, 0.12%, 5.73%, and 9.11% respectively. Having fewer frames leads to inferior results, as observed in Toscana, which only comprises six frames, with two used for training and validation. While for the UCSD dataset, as shown in Table 4.7, the proposed method outperforms FgSegNet_V2 [81], FgSegNet_S [80] and FgSegNet_M [80] by a small margin.

Scene	F-measure
birds	0.8649
boats	0.9210
bottle	0.9561
chopper	0.9146
cyclists	0.9216
flock	0.9389
freeway	0.7783
hockey	0.9167
jump	0.9358
landing	0.9244
ocean	0.8932
peds	0.8876
rain	0.9177
skiing	0.9119
surfers	0.8891
surf	0.7317
traffic	0.9081
zodiac	0.8990
Nested-Net (proposed)	0.8950
FgSegNet_v2 [81] #	0.8945
FgSegNet_S [80] #	0.8822
FgSegNet_M [80] #	0.8948

Table 4.7: Qualitative results on UCSD dataset with 20% split. (Red Represents the Best Result, and Blue Represents the Second Best). # These approaches were trained and assessed using the identical SDE setup as the proposed Nested-Net.

4.5.2.2 Qualitative results

We present our network's visual results by showing illustrated examples from the generated masks, as shown in Figure 4.7. Our Nested-Net effectively delivers foreground detection results by accurately removing shadows, gaps, and artifacts.

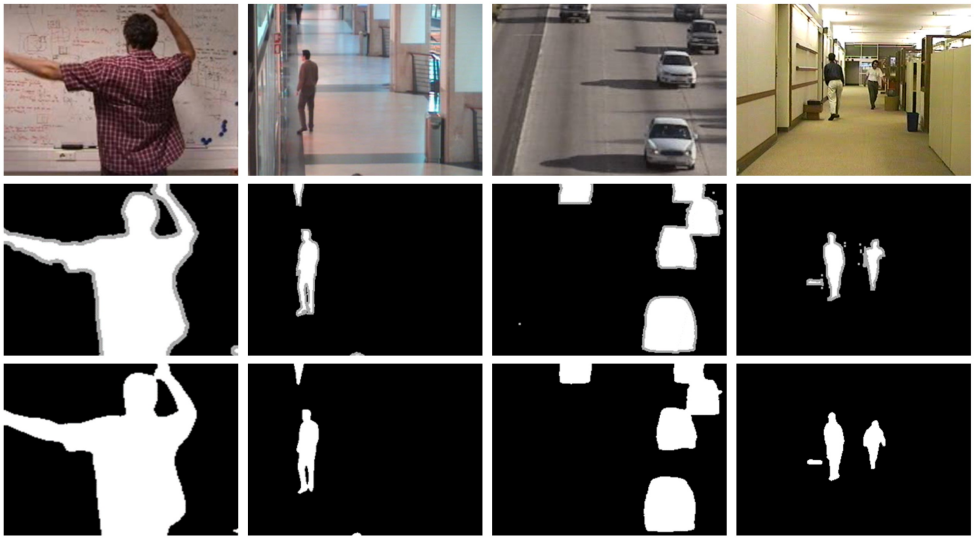


Figure 4.7: Visual Analysis of the Proposed Nested-Net on SBI 2015. Rows depict the input frame, ground truth, and results from our Nested-Net, while columns illustrate examples from the SBI 2015 dataset: Board, Caviar, Highway, and HullAndMonitor, respectively.

4.5.2.3 Implementation and execution time

We have developed and trained the Nested-Net using the Keras software library with the TensorFlow backend on a system equipped with a single NVIDIA GeForce RTX 2060 GPU paired with an Intel (R) Core (TM) i7-10870H processor running at CPU @2.20 GHz and 16.0 GB of RAM. Our Nested-Net achieves an average execution time of around 34 frames per second for images with a spatial resolution of 320×240. Additionally, our Nested-Net boasts fewer parameters than the state-of-the-art techniques, as illustrated in Table 4.8.

Method	Total parameters	Trainable parameters	Non-trainable parameters
Nested-Net (proposed)	**4,393,473**	**4,132,545**	260,928
FgSegNet_V2 [81]	9,225,161	7,489,673	1,735,488
FgSegNet_S [80]	9,358,593	7,622,465	1,736,128
sEnDec [89]	5,432,097	5,376,977	55,120

Table 4.8: Comparative Analysis of Parameters for the Proposed Nested-Net Versus Existing Approaches.

4.6 Conclusion

In this chapter, we present a novel deep-learning model for background subtraction. The model features a unique architecture with a nested network that includes multiple skip connections between the proposed residual small autoencoders (RM-AE). These residual micro-autoencoders are designed to extract additional features at each scale, enhancing the features' generalization. Moreover, pre-trained VGG-16 blocks are used as inputs to the proposed residual small autoencoders in the encoder section. Our network effectively detects moving objects in challenging scenes with minimal training instances. Experimental evaluation shows that our model outperforms existing state-of-the-art methods without post-processing. We plan to incorporate temporal data for future research to enhance the robustness of the model further.

CHAPTER 5

PROPOSED UNSUPERVISED FRAME-BASED AND OBJECT-BASED ABNORMAL BEHAVIOR DETECTION IN VIDEO SURVEILLANCE

Contents

5.1 Introduction

Abnormal behavior identification and video anomaly detection (VAD) focus on classifying each frame in a video as either normal or abnormal, using a training set that contains only normal frames. Abnormal frames are identified by a generated score exceeding a predefined threshold.

The 'FastAno' [112] approach introduces a novel technique to video anomaly detection focused on enhancing both speed and accuracy. It tackles critical challenges, where the authors propose two techniques for feature learning: Spatial Rotation Transformation (SRT) and Temporal Mixing Transformation (TMT). These transformations generate artificial anomalies during training, enabling the model to learn regular spatial and temporal patterns more effectively.

Built on an Autoencoder (AE) architecture, the model predicts normal frames from transformed input sequences and detects anomalies by comparing the predicted and current frames.

While 'FastAno' demonstrates competitive results and improved speed compared to previous approaches, certain limitations persist. 3D convolutions can impact computational efficiency, and the random selection of patches for spatial and temporal correction may reduce training precision. To mitigate these issues, we propose two key contributions to further refine the method.

5.2 Methodologies

Our proposed approaches follow three key stages: preprocessing, feature learning, model training, and inference.

5.2.1 Frame-based unsupervised abnormal behavior detection in video surveillance via improved patch transformation

5.2.1.1 Data preprocessing

The initial captured frames from the input videos are transformed to a fitting resolution, with the pixel values adjusted to conform within the range of $[-1, 1]$, guaranteeing consistent representation across all frames. Moreover, the frames are converted to grayscale and adjusted to the size of 240×360 pixels to simplify

computational processes by decreasing dimensionality. To enhance the contrast and improve the visual features of the input images, we used the Contrast Limited Adaptive Histogram Equalization (CLAHE) [171] as an essential step in our method.

5.2.1.2 Object-centric based patch irregularity generation

In order to effectively identify unusual visual patterns and motion behaviors in typical scenarios, it is essential to have a deep understanding of both spatial and temporal contexts. We start by collecting a series of T consecutive frames from time steps $t - T/2$ to $t + T/2$ and stacking them along the channel axis to form a 3D box $(B_I) \in \mathbb{R}^{H \cdot W \cdot T}$. We then extract object patches using a pre-trained YOLOv8 [31] model as indicated in figure 5.1, focusing on foreground objects instead of backgrounds to prioritize essential scene elements. Next, we randomly select patches equal to $\lambda \cdot n_p$ and apply temporal and spatial transformations to them, where n_p represents the total number of extracted objects. The selected patches P'_t undergo random spatial rotations of 90°, 180°, or 270°, and we introduce skipping frames between (1 to 4) to modify the motion regularity of objects in the sequence, aiding in reconstructing irregular patches in the temporal context. During the training process, the model is exposed to frame sequences that have undergone the described transformations, enabling the network to concentrate on detecting irregular regions and recognizing regular features. As a result, the network can identify and reconcile abnormal occurrences with normal ones, refining its predictions to align with the ground truth frames closely. The patch anomaly generation phase provides efficient feature learning compared to other spatiotemporal feature extraction techniques. Significantly, this phase does not impact detection speed during inference, ensuring that our model maintains low complexity and computational costs.

5.2.1.3 Network architecture

In our model, we utilized a $2D$ U-Net with squeeze and excitation. This differentiates it from other approaches employing complex network architectures like convolutional-LSTM and 3D Convolutions. Our network has a straightforward yet effective structure.

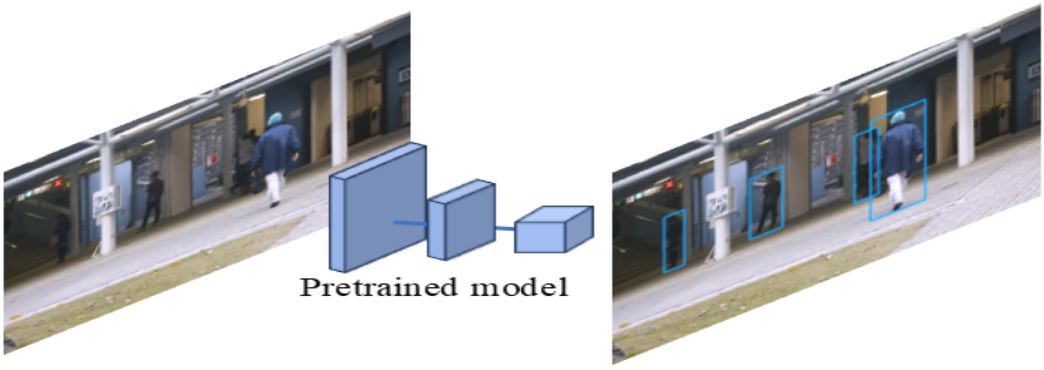


Figure 5.1: Object detection framework utilizing a pre-trained model.

Channel attention The attention mechanism, which is also known as the Squeeze and Excitation (Squeeze and Excitation (SE)) block [172] as shown in figure 5.2, allows neural networks to focus on essential areas of their feature representations [173]. Our research includes arranging frames along the channel axis. Each channel of the feature vector represents different aspects of each frame, making this mechanism very efficient.

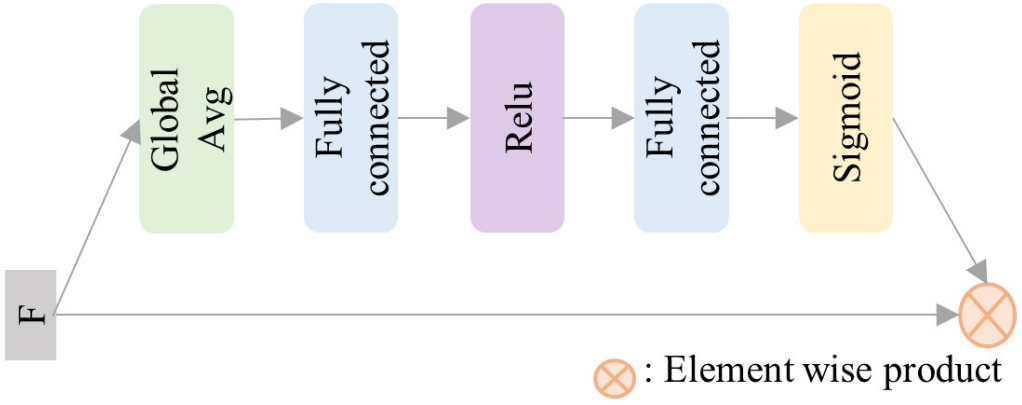


Figure 5.2: Channel Attention Block (CA).

We apply global average pooling (G.Avg) to the feature map F to capture channel dependencies, resulting in a feature vector v with C values. Next, a fully connected layer reduces the dimensionality by a factor of r and applies rectified linear unit activation δ. Then, a second fully connected layer with sigmoid activation $s(F)$ restores the original channel dimension. The output F is the element-wise product of the original F and the sigmoid output $s(F)$, representing the channel attention applied to the encoder's convolutional output. The channel

attention (Channel Attention (CA)) structure is illustrated in Figure 5.2.

Overall structure The proposed network consists of an encoder and a decoder, connected by a skip connection and improved by a squeeze-and-excitation block. These components work together to extract compact representations from the input data and reconstruct the original data by minimizing the reconstruction error of these representations. The encoder consists of three $2D$ convolutional layers, each with a 2×2 stride and a 3×3 kernel size. Batch normalization and LeakyReLU (LRelu) activation which are applied after each convolution. In contrast, the decoder follows a structure similar to the encoder but employs $2D$ deconvolutional layers to expand the feature map size. Skip connections, preceded by channel attentions, are incorporated into the feature map $F \in R^{H \cdot W \cdot C}$ following every convolutional layer. Figure 5.3 depicts the overall framework design.

5.2.1.4 Anomaly detection and objective function

To identify irregularities in a video, we calculate the anomaly score $S(t)$ (eq. 5.3). This score is obtained by comparing the reconstructed frames I' with the ground truth frames I using the Mean Squared Error (MSE). To determine the MSE (eq. 5.2), we initially measured the squared differences between corresponding pixel values of the ground truth and reconstructed frames across T frames in the sequence. Subsequently, we average these squared differences across all frames. Lower MSE values correspond to higher frame quality. Therefore, ensuring the similarity of all pixels can be achieved through an intensity constraint, which assesses each pixel value between the ground-truth frame and the reconstructed frame, as illustrated in L_{recon} (eq. 5.1). The following functions are utilized:

$$L_{recon}(I_t, I'_t) = ||I_t - I'_t||_2^2 \tag{5.1}$$

$$MSE(I, I') = \frac{1}{T} \sum_{i=1}^{T} ||I_t - I'_t||_2^2 \tag{5.2}$$

We employ the MSE (eq. 5.2) to compute the anomaly score $S(t)$ (eq. 5.3) for each frame in the test video sequence. This score is derived by standardizing

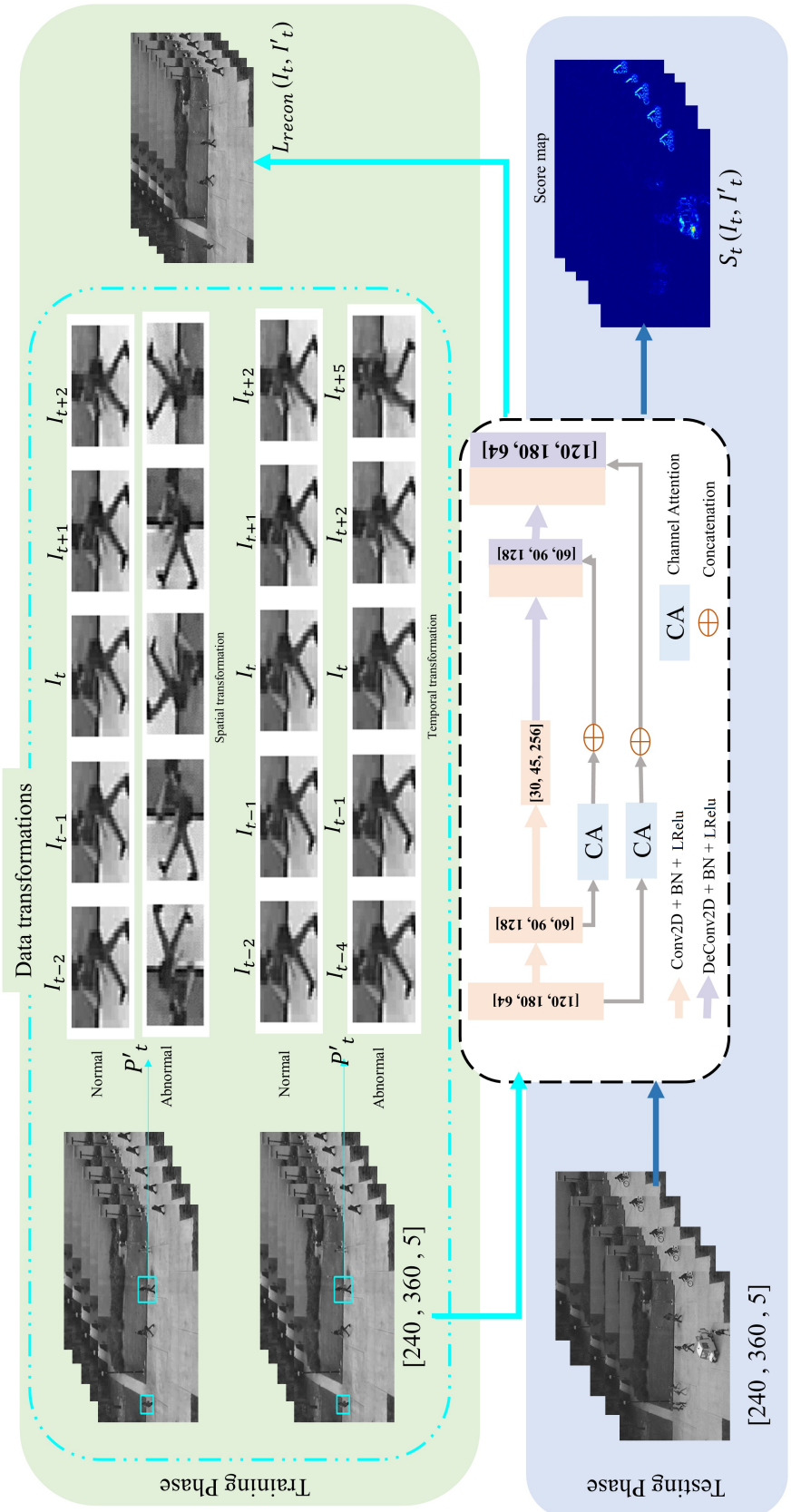


Figure 5.3: A detailed structure of the proposed frame-based method, including both training and testing stages.

the MSE values of all frames in the sequence. Nevertheless, the score indicates whether each frame is typical or atypical.

$$S(t) = \frac{MSE(I_t, I'_t) - min \quad MSE(I_t, I'_t)}{max \quad MSE(I_t, I'_t) - min \quad MSE(I_t, I'_t)} \tag{5.3}$$

5.2.2 Object-centric abnormal behavior detection in video surveillance via Feature Learning and Pseudo-anomaly generation

We will follow the same steps as previously mentioned, focusing only on highlighting the differences.

5.2.2.1 Data preprocessing

We first extract objects of interest by applying the same steps as outlined in the previous section, but instead of processing the entire frame, we focus on individual objects. For each frame i, we use the selected bounding boxes to clip the objects. The clipped objects from temporally adjacent frames $i-t,\ldots,i-1,i,i+1,\ldots,i+t$ are stacked to form object-centric boxes using the same bounding boxes. All patches are resized to a fixed resolution of 48Ö 48 pixels to reduce dimensionality and simplify calculations.

5.2.2.2 RoIs based irregularity generation for feature learning

During training, for each batch of data, a subset of the constructed bounding boxes B_I, equal to $\lambda \cdot bz$ where bz is the batch size, is randomly selected for temporal and spatial transformations. These selected Regions of Interest (RoIs) undergo random spatial rotations of 90°, 180°, or 270°. Additionally, we introduce frame skipping (ranging from 1 to 4 frames) to alter the motion regularity of objects within the sequence, simulating irregular motion patterns in the temporal context.

By exposing the model to these transformed RoIs, it learns to detect irregular regions and distinguish between normal and anomalous patterns. This process enhances the model's ability to refine its predictions, improving alignment with the ground truth.

5.2.2.3 Network architecture

We employ the same $2D$ Autoencoder with channel attention described above, distinguishing our approach from more complex models. Our network, while remaining simple, delivers a highly effective performance. As illustrated in Figure 5.4, it comprises an encoder and decoder connected by skip connections and channel attention.

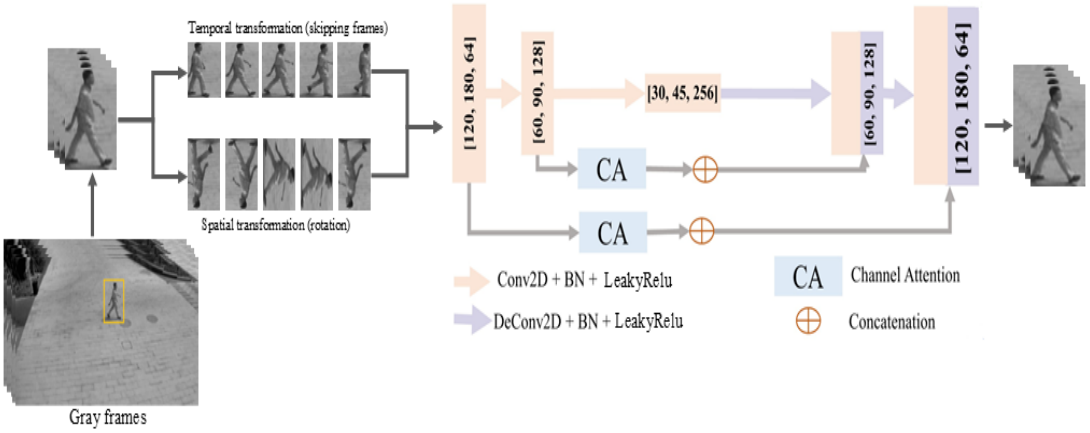


Figure 5.4: A detailed structure of the proposed object-centric based method

5.2.2.4 Anomaly detection and objective function

In order to detect the anomalies in a video, we calculate the anomaly score $S(t)$ (eq. 5.3) based on the highest object score at time t. The score is derived from both the Mean Squared Error (MSE) (eq. 5.2) between reconstructed RoI I' and ground truth RoI I

5.3 Experiments

5.3.1 Experimental Setup

Datasets and evaluation metrics: We have validated our approach using four well-known benchmark datasets: UCSD Pedestrians 2 [113], CUHK Avenue [114], UMN [116], and ShanghaiTech Campus [115]. These datasets were chosen to cover a wide range of real-world scenarios, featuring various resolutions, camera angles, and types of anomalies. Ground truth binary flags were used for every frame to

evaluate our model, ensuring accurate identification of anomalous occurrences. Frame-level AUC was used to quantitatively evaluate the model's performance as the primary assessment statistic. We also provided qualitative and quantitative assessments of the results for each dataset.

Implementation details Our experiments utilized Pytorch [174] on a single $Nvidia\ \ GeForce\ \ RTX$ 2060. The Adam optimizer [175] was employed with an initial learning rate of 1e−4.

For the frame-based method, we use $T = 5$. The batch size was set to 4, and the maximum number of epochs was 30 for Ped2 [113], 20 for Avenue [114], and 10 for the ShanghaiTech [115] datasets. The hyperparameter λ was set to 0.3.

For the object-based method, we use $T = 7$. The batch size was set to 64 for Ped2 [113] and Avenue [114], 128 for the ShanghaiTech, while the maximum number of epochs was 30 for UCSD Ped2 [113], and Avenue [114], and 20 for the ShanghaiTech [115] datasets. The hyperparameter λ was set to 0.5.

In the first stage of our architecture, we employ YOLOv8 [31] for object detection. We retain detections with a confidence level exceeding 0.2, 0.8, and 0.5 for UCSD Ped2, CUHK Avenue, and ShanghaiTech, respectively. This criterion remains consistent during both the training and inference phases.

5.3.2 Experimental results

5.3.2.1 Quantitative analysis

To evaluate our proposed methods' effectiveness, we use the Area Under the Curve (AUC) based on frame-level ground-truth annotations and report both micro- and macro-averaged AUC metrics as illustrated in table 5.1. We compared the results with the recent approaches on the UCSD Pedestrians 2 (Ped2), CUHK Avenue, and ShanghaiTech datasets. As shown in Table 5.2.

Our evaluation results for the frame-based method show that our proposed achieved the highest average AUC value (98.72%) and outperformed the state-of-the-art [188] by 0.27%. On CUHK Avenue, our method surpassed all earlier works, achieving a Frame AUC of 89.55%. In contrast, on the ShanghaiTech dataset, our method demonstrated competitiveness with prior techniques and even exceeded TI-VAD [165], which used optical flow for motion estimation.

Dataset	Frame-based approach		Object-centric based approach	
	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC
UCSD Ped2 [113]	98.72	98.53	99.35	99.87
CUHK Avenue [114]	89.55	89.20	88.10	90.50
Shanghai Tech [115]	73.84	83.50	76.47	83.48
UMN [116]	99.70	99.80	99.6	99.83

Table 5.1: Our Micro and Macro AUC scores on four well-known benchmark datasets.

Our analysis of the object-based method reveals that the proposed strategy outperformed the state-of-the-art method Liu et al. [156] by 0.05%, achieving the highest average AUC score of 99.35%. Our approach performed competitively on the CUHK Avenue and ShanghaiTech datasets; however, our results on the ShanghaiTech dataset surpass HF2-VAD [156], which used both RGB and optical flow-based modality.

In summary, our proposed techniques deliver significant efficacy and efficiency as an anomaly detection solution in terms of accuracy and ease of use. Furthermore, our methods demonstrated superior performance speed, as shown in the table 5.3.

5.3.2.2 Qualitative analysis

The outcomes of unusual events occurring at time t are illustrated in Figures 5.5 and 5.6. The model consistently produces low anomaly scores during regular periods, indicating its adaptability. On the other hand, when anomalies occur, these scores rapidly increase and remain consistently high for the duration of the anomaly. As depicted in these figures, the notable anomalies for the Ped2 and ShanghaiTech datasets are unfamiliar objects, such as a car, a bicycle, and a motorcycle. In contrast, the Avenue dataset identifies motion anomalies, such as a person running, jumping, and throwing objects. Additionally, Figures 5.7 and 5.8 show the method's prediction results and errors for anomalous frames, further highlighting the model's versatility. This behavior demonstrates the method's ability to pinpoint and identify abnormal events efficiently and accurately. This is a critical aspect of the proposed models, which aims to precisely locate and classify anomalies in video surveillance.

Method	UCSD Ped2	CUHK Avenue	ShanghaiTech
Frame or Pixel-based			
Hasan et al. (2016) [176]	90.0	70.2	60.9
Lu et al. (2019) [177]	96.1	85.8	-
Nguyen et al. (2019) [178]	84.3	82.8	-
Luo et al. (2019) [179]	96.9	86.6	73.8
Tang et al. (2020) [180]	96.2	83.7	71.5
Chang et al. (2020) [181]	96.5	86.0	73.3
Park et al. (2020) [182]	97.0	88.5	70.5
Li et al. (2020) [183]	95.4	86.0	71.4
Zheng et al. (2020) [184]	95.4	86.8	73.6
Cai et al. (2021) [185]	96.6	86.6	73.7
Huang et al. (2022) [186]	97.6	88.8	74.3
Zhao et al. (2022) [187]	97.1	89.3	73.0
Park et al. (2022) [112]	96.3	85.3	72.2
Wang et al. (2023) [150]	98.4	86.1	73.2
Astrid et al. (2023) [188]	98.44	87.1	73.7
Yang et al. (2024) [189]	97.9	88.5	74.1
Proposed (frame-based)	98.72	89.55	73.84
Object-based			
Hinami et al. (2017) [190]	92.2	89.8	-
Ionescu et al. (2019) [152]	94.30	87.40	78.70
Morais et al. (2019) [191]	-	-	73.4
Sun et al. (2020) [192]	-	89.60	74.70
Doshi et al. (2020) [193, 194]	97.80	86.40	71.62
Yu et al. (2020) [154]	97.3	89.6	74.8
Liu et al. (2021) [156]	99.3	91.1	76.2
Wang et al. (2022) [161]	-	89.8	73.7
Doshi et al. (2023) [165]	-	85.78	71.18
Huang et al. (2023) [195]	97.7	89.7	75.8
Proposed (Object-based)	99.35	88.1	76.47

Table 5.2: Frame level AUC comparison between the proposed methods and the most recent VAD techniques (red and blue values represent the first and second best results, respectively).

Method	FPS
Frame-based	
Park et al. (2022) [112]	101
Wang et al. (2023) [150]	60
Proposed (frame-based)	183
Object-centric	
Ionescu et al. (2019) [152]	13.5
Liu et al. (2021) [156]	10
Proposed (Object-centric)	42 (CUHK Avenue)

Table 5.3: Frame Per Second (FPS) comparison among the top methods.

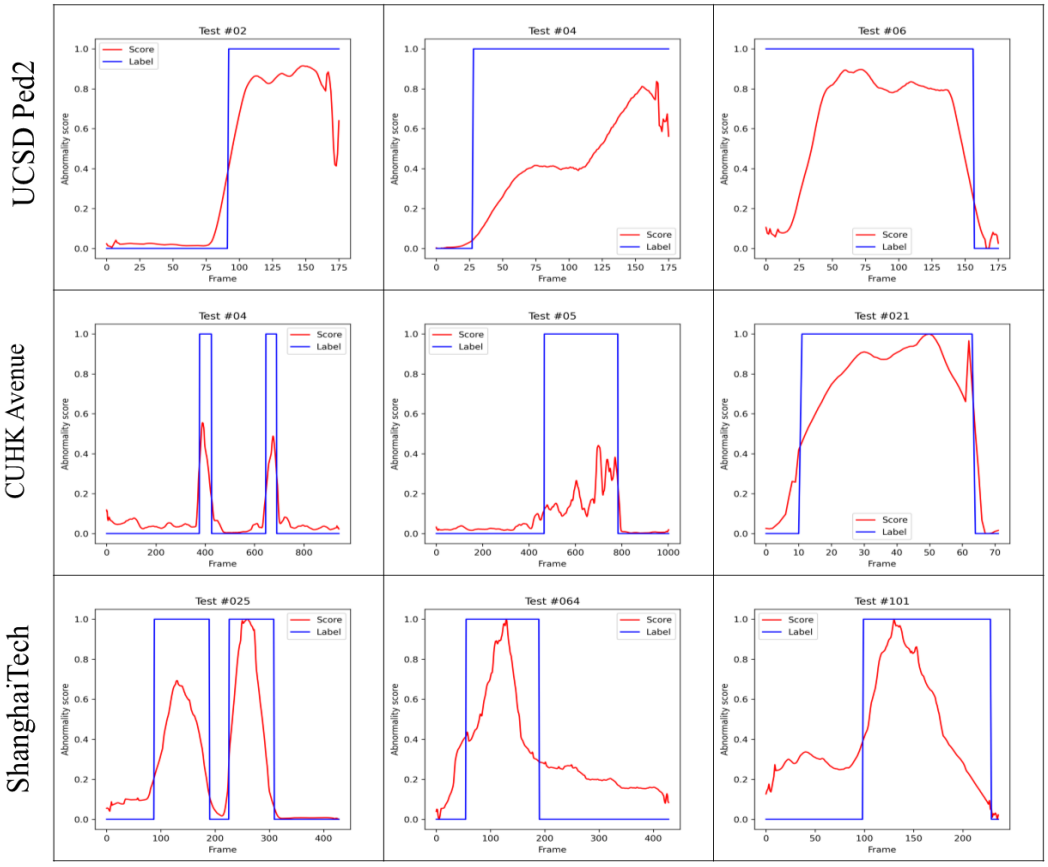


Figure 5.5: The score curves that were acquired via assessment on frame-based proposed approach. The anomaly score S_t is shown by the red line in the plot, while the blue line represents the labels. Higher values indicate a higher occurrence of anomalies.

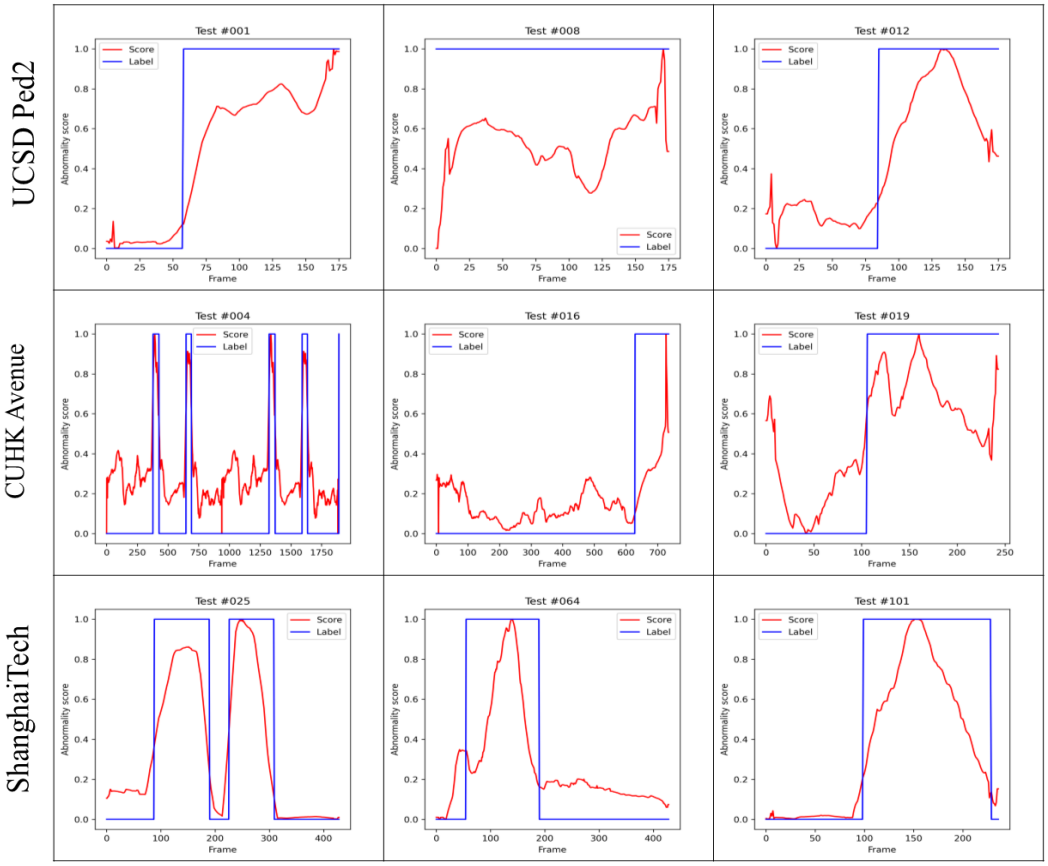


Figure 5.6: The score curves that were acquired via assessment on object-based proposed approach. The anomaly score S_t is shown by the red line in the plot, while the blue line represents the labels. Higher values indicate a higher occurrence of anomalies.

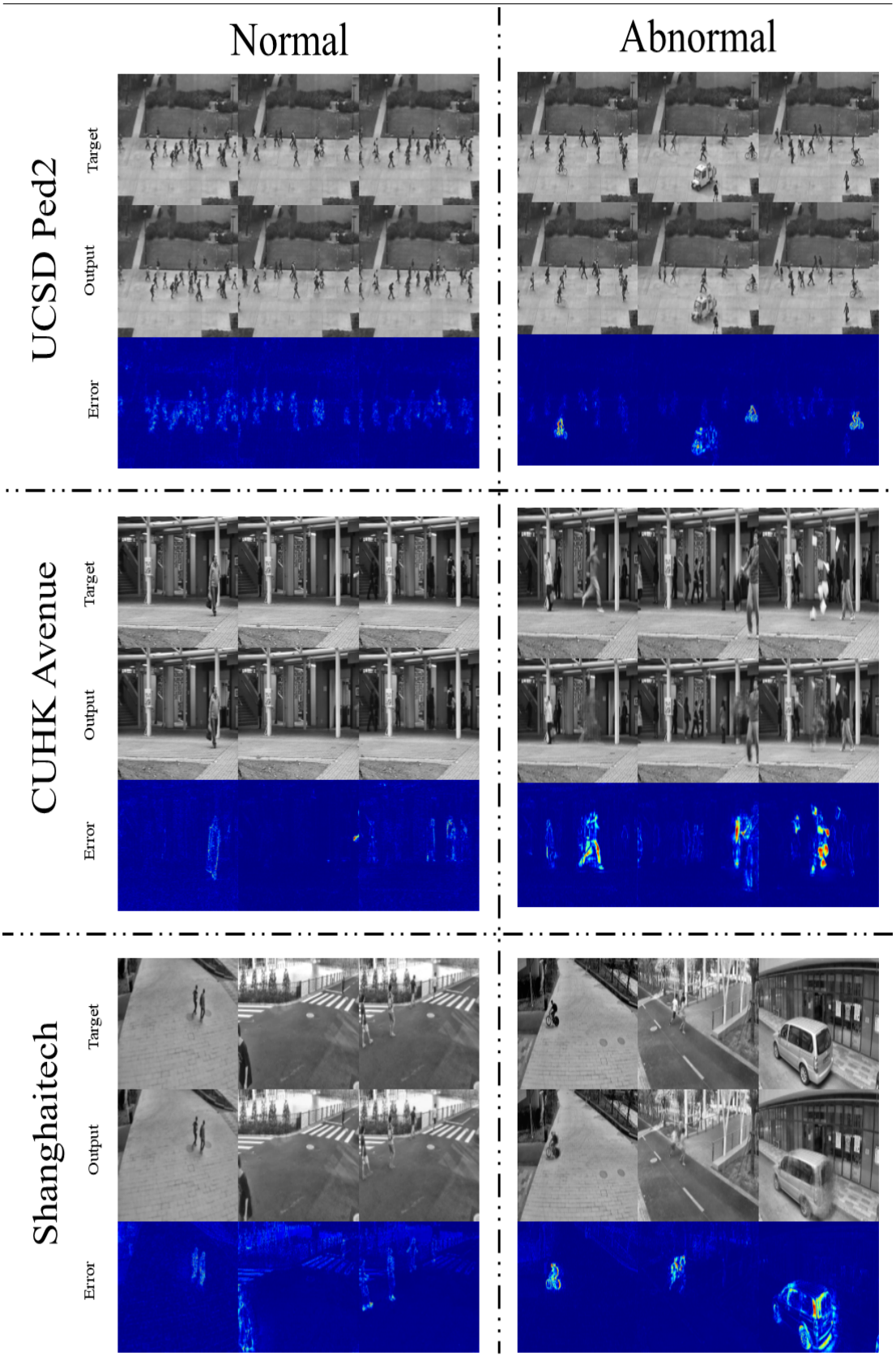


Figure 5.7: Normal and abnormal Reconstructed frame examples and its error maps on UCSD Ped2, CUHK Avenue, and ShanghaiTech benchmark datasets.

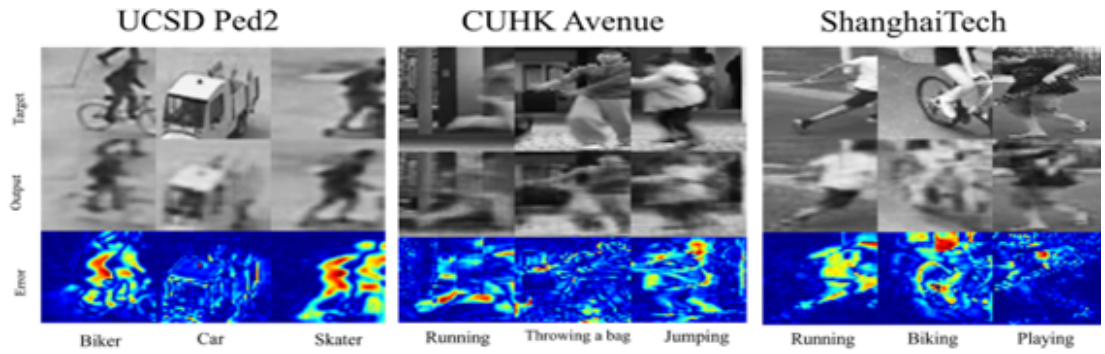


Figure 5.8: Abnormal Reconstructed objects examples and its error maps on UCSD Ped2, CUHK Avenue, and ShanghaiTech benchmark datasets.

5.4 Conclusion

This chapter presents two novel strategies for identifying unusual behavior in videos. One is merging the generation of object-centric patch anomalies on standard frames with a reconstruction network to improve the detection of irregular regions and boost the learning of normal features. The second is based on an object-centric level while maintaining pseudo anomaly generation at each batch of data. We employ a $2D$ U-Net model with attention blocks at the channel level to capture fundamental patterns crucial for abnormality detection. To enhance the model's generalization of normal features, we train it to focus on typical appearance and motion patterns using spatial and temporal transformation techniques. We assess our methods on three standard datasets, and the outcomes demonstrate the proficient performance of our network in identifying anomalies at the frame and object levels.

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

Contents

6.1 Conclusions

This research highlights the critical need to monitor public areas and provides a practical methodology for doing so. Observing and extracting valuable insights from these environments allows us to recognize and classify objects based on their behaviors, which is essential for applications such as security and surveillance. This comprehensive approach, which involves detecting, identifying, and tracking objects in real time, can significantly enhance the effectiveness of surveillance systems in real-world scenarios.

This thesis tackles two critical components of behavior analysis: object detection and behavior classification. These processes are fundamentally related, as accurate behavior analysis hinges on successfully identifying and tracking objects. The first step involves detecting and identifying objects, including determining their locations within a scene. The second step focuses on analyzing the behavior of these detected objects to identify any anomalies or potential threats. These processes work together to create a more complete knowledge of the scenario and allow for prompt reactions to aberrant or dangerous circumstances.

Our methodologies for addressing these two critical steps is outlined in distinct sections of the thesis. First, for detecting moving objects in video, we propose a novel deep-learning model for background subtraction. With its novel architecture incorporating multiple skip connections between micro-autoencoders and pre-trained VGG-16 blocks, this model represents a significant advancement in the field. This network's architecture enables detecting moving objects in complex scenes with minimal training data, making it highly efficient and adaptable.

Second, we introduce an advanced anomaly detection network to detect suspicious behaviors. This single-stream network leverages a 2D attention U-Net and pseudo anomaly generation to enhance spatiotemporal feature learning in surveillance videos. The framework detailed in the thesis represents a significant advancement in automated surveillance systems, enabling more accurate detection of objects and their abnormal behaviors.

Each of these unique perspectives is discussed in detail below:

In Chapters 2 and 4, we provide a comprehensive review of object detection, with a particular focus on background subtraction techniques. This review, which results from extensive research and analysis, outlines key concepts, methodologies, and challenges within the field, offering a critical assessment of existing

approaches. Chapter 4 introduces a novel deep learning model for background subtraction, featuring a novel architecture that incorporates multiple skip connections between new micro-autoencoder units and pre-trained VGG-16 blocks. This model demonstrates high efficiency in detecting moving objects within complex and dynamic scenes, even with minimal training data, showcasing its potential in real-world applications.

Chapters 3 and 5 present an in-depth analysis of abnormal object behavior detection in video surveillance, thoroughly examining various algorithms and techniques used in this domain. Chapter 5 defines the proposed single-stream network for abnormal behavior detection, which leverages object-centric spatiotemporal transformations. This approach introduces a novel reconstruction-based method for generating irregularities, enhancing the network's capacity to identify abnormal behaviors. Additionally, the network employs a two-dimensional U-Net with channel attention to optimize video analysis efficiency alongside an unsupervised spatiotemporal learning technique. This enables robust detection of abnormal object behavior in surveillance video, significantly advancing this area's state of the art.

6.2 Limitations

The proposed methods have several limitations. In Chapter 4, while our nested network delivers reliable results in background subtraction, it struggles to detect tiny objects, particularly in scenarios with low frame rates. Additionally, the model's performance is highly dependent on the number of training instances. A limited number of training cases can lead to suboptimal detection of moving objects, reducing the overall effectiveness of the approach.

In Chapter 5, while the results demonstrate strong performance, there are some limitations in handling anomalies. The method operates under an unsupervised framework, meaning no explicit anomalies were defined during training. Instead, the system assumes anomalies are any actions, movements, or objects not encountered during training. While effective in some contexts, this unsupervised approach cannot be universally applied to all datasets, particularly those requiring a predefined specification of anomaly types. However, potential solutions, such as explicit definitions of anomalous behaviors, can be explored to ensure accurate detection. Furthermore, in this chapter, we did not employ a dedicated moving

object detector, such as those relying on optical flow or background subtraction prior to extracting abnormal behaviors. As a result, our approach may occasionally misclassify or overlook objects that belong to the background. Additionally, some false positives and missed detections still occur, primarily due to the occlusion of moving objects. These challenges highlight areas for improvement, particularly in object differentiation and handling complex scenes with overlapping objects.

6.3 Future Work

Each of the proposed frameworks might be modified and expanded to address the restrictions mentioned above. This section presents some directions for further development. To enhance the method's performance presented in Chapter 4, we plan to employ independent scene evaluation to better assess its quality. This will allow it to be applied to other datasets.

In Chapter 5, incorporating weak supervision may uncover additional identifiable cases, allowing the proposed method to be applied systematically to larger datasets, thereby broadening its applicability.

To reduce false detections, we aim to integrate the method with optical flow techniques and introduce approaches such as pose estimation to detect extreme human movements. These enhancements significantly improve performance and increase the model's overall effectiveness.

BIBLIOGRAPHY

[1] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, pp. 983–1009, 2013.

[2] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 3, pp. 334–352, 2004.

[3] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012.

[4] Y. Yang, J. Liu, and M. Shah, "Video scene understanding using multi-scale analysis," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1669–1676, IEEE, 2009.

[5] F. Brémond, M. Thonnat, and M. Zúniga, "Video-understanding framework for automatic behavior recognition," *Behavior Research Methods*, vol. 38, no. 3, pp. 416–426, 2006.

[6] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Computer Science Review*, vol. 35, p. 100204, 2020.

[7] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digital Signal Processing*, vol. 132, p. 103812, 2023.

[8] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[9] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using yolo: Challenges, architectural successors, datasets and applications," *multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, 2023.

[10] K. Ragland and P. Tharcis, "A survey on object detection, classification and tracking methods," *Int. J. Eng. Res. Technol*, vol. 3, no. 11, pp. 622–628, 2014.

[11] K. Liu, M. Zhu, H. Fu, H. Ma, and T.-S. Chua, "Enhancing anomaly detection in surveillance videos with transfer learning from action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4664–4668, 2020.

[12] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pp. 297–312, Springer, 2014.

[13] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, pp. 261–318, 2020.

[14] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 390–391, 2020.

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[18] S. Sanchez, H. Romero, and A. Morales, "A review: Comparison of performance metrics of pretrained models for object detection using the tensorflow framework," in *IOP Conference Series: Materials Science and Engineering*, vol. 844, p. 012024, IOP Publishing, 2020.

[19] R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 international conference on systems, signals and image processing (IWSSIP)*, pp. 237–242, IEEE, 2020.

[20] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137–154, 2004.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, Ieee, 2001.

[22] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, Ieee, 2005.

[25] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Ieee, 2008.

[26] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *2010 IEEE Computer society conference on computer vision and pattern recognition*, pp. 2241–2248, Ieee, 2010.

[27] 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.

[28] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[30] 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*, pp. 21–37, Springer, 2016.

[31] 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*, pp. 779–788, 2016.

[32] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[33] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[34] S. H. Shaikh, K. Saeed, N. Chaki, S. H. Shaikh, K. Saeed, and N. Chaki, *Moving object detection using background subtraction*. Springer, 2014.

[35] B. Karasulu, S. Korukoglu, B. Karasulu, and S. Korukoglu, *Moving object detection and tracking in videos*. Springer, 2013.

[36] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer science review*, vol. 11, pp. 31–66, 2014.

[37] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 43–60, 2016.

[38] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 387–394, 2014.

[39] S. Brutzer, B. Höferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *CVPR 2011*, pp. 1937–1944, IEEE, 2011.

[40] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 1–8, IEEE, 2012.

[41] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *New Trends in Image Analysis and Processing–ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proceedings 18*, pp. 469–476, Springer, 2015.

[42] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta," *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, 2016.

[43] C. Wren, "Real-time tracking of the human body. in photonics east," in *SPIE*, vol. 2615, 1995.

[44] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[45] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," *Video-based surveillance systems: Computer vision and distributed processing*, pp. 135–144, 2002.

[46] D.-S. Lee, "Effective gaussian mixture learning for video background subtraction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 827–832, 2005.

[47] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2, pp. 28–31, IEEE, 2004.

[48] Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, no. 7, pp. 773–780, 2006.

[49] A. Darwich, P.-A. Hébert, A. Bigand, and Y. Mohanna, "Background subtraction based on a new fuzzy mixture of gaussians for moving object detection," *Journal of Imaging*, vol. 4, no. 7, p. 92, 2018.

[50] T. Akilan, Q. J. Wu, and Y. Yang, "Fusion-based foreground enhancement for background subtraction using multivariate multi-model gaussian distribution," *Information Sciences*, vol. 430, pp. 414–431, 2018.

[51] T. S. Haines and T. Xiang, "Background subtraction with dirichlet processes," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pp. 99–113, Springer, 2012.

[52] T. S. Haines and T. Xiang, "Background subtraction with dirichletprocess mixture models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 4, pp. 670–683, 2013.

[53] T. M. Nguyen, Q. J. Wu, and H. Zhang, "Asymmetric mixture model with simultaneous feature selection and model detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 400–408, 2014.

[54] D. K. Rout, B. N. Subudhi, T. Veerakumar, and S. Chaudhury, "Spatio-contextual gaussian mixture model for local change detection in underwater video," *Expert Systems with Applications*, vol. 97, pp. 117–136, 2018.

[55] Z. Zhao, T. Bouwmans, X. Zhang, and Y. Fang, "A fuzzy background modeling approach for motion detection in dynamic backgrounds," in *Multimedia and Signal Processing: Second International Conference, CMSP 2012, Shanghai, China, December 7-9, 2012. Proceedings*, pp. 177–185, Springer, 2012.

[56] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Computer Vision—ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6*, pp. 751–767, Springer, 2000.

[57] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on image processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[58] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing and Applications*, vol. 19, pp. 179–186, 2010.

[59] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Background modeling and subtraction by codebook construction," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, vol. 5, pp. 3061–3064, IEEE, 2004.

[60] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[61] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[62] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern recognition*, vol. 40, no. 3, pp. 1091–1105, 2007.

[63] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[64] O. Barnich and M. Van Droogenbroeck, "Vibe: a powerful random technique to estimate the background in video sequences," in *2009 IEEE international conference on acoustics, speech and signal processing*, pp. 945–948, IEEE, 2009.

[65] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2010.

[66] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 38–43, IEEE, 2012.

[67] G.-A. Bilodeau, J.-P. Jodoin, and N. Saunier, "Change detection in feature space using local binary similarity patterns," in *2013 International conference on computer and robot vision*, pp. 106–112, IEEE, 2013.

[68] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2007.

[69] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Flexible background subtraction with self-balanced local sensitivity," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 408–413, 2014.

[70] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2014.

[71] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Universal background subtraction using word consensus models," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4768–4781, 2016.

[72] S. Bianco, G. Ciocca, and R. Schettini, "Combination of video change detection algorithms by genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 6, pp. 914–928, 2017.

[73] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *2016 international conference on systems, signals and image processing (IWSSIP)*, pp. 1–4, IEEE, 2016.

[74] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for background subtraction," *arXiv preprint arXiv:1702.01731*, 2017.

[75] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split gaussian models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 414–418, 2014.

[76] J. Liao, G. Guo, Y. Yan, and H. Wang, "Multiscale cascaded scene-specific convolutional neural networks for background subtraction," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*, pp. 524–533, Springer, 2018.

[77] T. P. Nguyen, C. C. Pham, S. V.-U. Ha, and J. W. Jeon, "Change detection by training a triplet network for motion feature extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 433–446, 2018.

[78] M. Qiu and X. Li, "A fully convolutional encoder–decoder spatial–temporal network for real-time background subtraction," *IEEE Access*, vol. 7, pp. 85949–85958, 2019.

[79] Y. Wang, Z. Luo, and P.-M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognition Letters*, vol. 96, pp. 66–75, 2017.

[80] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognition Letters*, vol. 112, pp. 256–262, 2018.

[81] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1369–1380, 2020.

[82] M. K. Panda, A. Sharma, V. Bajpai, B. N. Subudhi, V. Thangaraj, and V. Jakhetiya, "Encoder and decoder network with resnet-50 and global average feature pooling for local change detection," *Computer Vision and Image Understanding*, vol. 222, p. 103501, 2022.

[83] D. Zeng and M. Zhu, "Background subtraction using multiscale fully convolutional network," *IEEE Access*, vol. 6, pp. 16010–16021, 2018.

[84] O. Tezcan, P. Ishwar, and J. Konrad, "Bsuv-net: A fully-convolutional neural network for background subtraction of unseen videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2774–2783, 2020.

[85] W. Zheng, K. Wang, and F. Wang, "Background subtraction algorithm based on bayesian generative adversarial networks," *Acta Automatica Sinica*, vol. 44, no. 5, pp. 878–890, 2018.

[86] W. Zheng, K. Wang, and F.-Y. Wang, "A novel background subtraction algorithm based on parallel vision and bayesian gans," *Neurocomputing*, vol. 394, pp. 178–200, 2020.

[87] P. W. Patil, A. Dudhane, S. Murala, and A. B. Gonde, "Deep adversarial network for scene independent moving object segmentation," *IEEE Signal Processing Letters*, vol. 28, pp. 489–493, 2021.

[88] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3d cnn-lstm-based image-to-image foreground segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 959–971, 2019.

[89] T. Akilan and Q. J. Wu, "sendec: an improved image to image cnn for foreground localization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4435–4443, 2019.

[90] A. Taha, H. H. Zayed, M. Khalifa, and E.-S. M. El-Horbaty, "Exploring behavior analysis in video surveillance applications," *International Journal of Computer Applications*, vol. 93, no. 14, pp. 22–32, 2014.

[91] S. Roka, M. Diwakar, P. Singh, and P. Singh, "Anomaly behavior detection analysis in video surveillance: a critical review," *Journal of Electronic Imaging*, vol. 32, no. 4, pp. 042106–042106, 2023.

[92] 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.

[93] D. Durães, F. S. Marcondes, F. Gonçalves, J. Fonseca, J. Machado, and P. Novais, "Detection violent behaviors: a survey," in *Ambient Intelligence–Software and Applications: 11th International Symposium on Ambient Intelligence*, pp. 106–116, Springer, 2021.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[95] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe, "Deep appearance features for abnormal behavior detection in video," in *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19*, pp. 779–789, Springer, 2017.

[96] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[97] X. Yan, H. Gong, Y. Jiang, S.-T. Xia, F. Zheng, X. You, and L. Shao, "Video scene parsing: An overview of deep learning methods and datasets," *Computer Vision and Image Understanding*, vol. 201, p. 103077, 2020.

[98] P. Dhruv and S. Naskar, "Image classification using convolutional neural network (cnn) and recurrent neural network (rnn): A review," *Machine learning and information processing: proceedings of ICMLIP 2019*, pp. 367–381, 2020.

[99] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*, pp. 253–256, IEEE, 2010.

[100] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[101] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," *arXiv preprint arXiv:1801.01078*, 2017.

[102] Q. Abbas, M. E. Ibrahim, and M. A. Jaffar, "Video scene analysis: an overview and challenges on deep learning algorithms," *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 20415–20453, 2018.

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

[104] R. Zhao, H. Ali, and P. Van der Smagt, "Two-stream rnn/cnn for action recognition in 3d videos," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4260–4267, IEEE, 2017.

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

[106] S. Tiwari, G. Jain, D. K. Shetty, M. Sudhi, J. M. Balakrishnan, and S. R. Bhatta, "A comprehensive review on the application of 3d convolutional neural networks in medical imaging," *Engineering Proceedings*, vol. 59, no. 1, p. 3, 2023.

[107] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, "3d deep learning on medical images: a review," *Sensors*, vol. 20, no. 18, p. 5097, 2020.

[108] S. Anoopa and A. Salim, "Survey on anomaly detection in surveillance videos," *Materials Today: Proceedings*, vol. 58, pp. 162–167, 2022.

[109] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep video anomaly detection: Opportunities and challenges," in *2021 international conference on data mining workshops (ICDMW)*, pp. 959–966, IEEE, 2021.

[110] K. M. Biradar, A. Gupta, M. Mandal, and S. K. Vipparthi, "Challenges in time-stamp aware anomaly detection in traffic videos," *arXiv preprint arXiv:1906.04574*, 2019.

[111] S. Bhakat and G. Ramakrishnan, "Anomaly detection in surveillance videos," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pp. 252–255, 2019.

[112] 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*, pp. 2249–2259, 2022.

[113] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes. computer vision and pattern recognition (cvpr)," in *2010 IEEE Conference on*, vol. 1981, 1975.

[114] 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*, pp. 2720–2727, 2013.

[115] 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*, pp. 6536–6545, 2018.

[116] N. Papanikolopoulos and V. Morellas, "Unusual crowd activity dataset of university of minesota," 2015.

[117] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[118] 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*, pp. 6479–6488, 2018.

[119] M. Jones and B. Ramachandra, "Street scene: A new dataset and evaluation protocol for video anomaly detection," tech. rep., Tech. Rep. TR2018-188, MERL-Mitsubishi Electric Research Laboratories . . . , 2019.

[120] 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.

[121] M. Baradaran and R. Bergevin, "A critical study on the recent deep learning based semi-supervised video anomaly detection methods," *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 27761–27807, 2024.

[122] Y. A. Samaila, P. Sebastian, N. S. S. Singh, A. N. Shuaibu, S. S. A. Ali, T. I. Amosa, G. E. M. Abro, and I. Shuaibu, "Video anomaly detection: A systematic review of issues and prospects," *Neurocomputing*, p. 127726, 2024.

[123] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image and Vision Computing*, vol. 106, p. 104078, 2021.

[124] D. V. Ngo, N. T. Do, and L. A. T. Nguyen, "Anomaly detection in video surveillance: A novel approach based on sub-trajectory," in *2016 International Conference on Electronics, Information, and Communications (ICEIC)*, pp. 1–4, IEEE, 2016.

[125] J.-J. Lee, G.-J. Kim, and M.-H. Kim, "Trajectory extraction for abnormal behavior detection in public area," in *2012 9th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT)*, pp. 1–5, IEEE, 2012.

[126] Y. Zhou, S. Yan, and T. S. Huang, "Detecting anomaly in videos from trajectory similarity analysis," in *2007 IEEE international conference on multimedia and expo*, pp. 1087–1090, IEEE, 2007.

[127] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Object trajectory-based activity classification and recognition using hidden markov models," *IEEE transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[128] N. Anjum and A. Cavallaro, "Multifeature object trajectory clustering for video analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1555–1564, 2008.

[129] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 5, pp. 893–908, 2008.

[130] C. R. Jung, L. Hennemann, and S. R. Musse, "Event detection using trajectory clustering and 4-d histograms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1565–1575, 2008.

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

[132] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 907–913, 2009.

[133] C. Li, Z. Han, Q. Ye, and J. Jiao, "Abnormal behavior detection via sparse reconstruction analysis of trajectory," in *2011 Sixth International Conference on Image and Graphics*, pp. 807–810, IEEE, 2011.

[134] C. Li, Z. Han, Q. Ye, and J. Jiao, "Visual abnormal behavior detection based on trajectory sparse reconstruction analysis," *Neurocomputing*, vol. 119, pp. 94–100, 2013.

[135] S. Biswas and R. V. Babu, "Short local trajectory based moving anomaly detection," in *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, pp. 1–8, 2014.

[136] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 683–695, 2016.

[137] K. Zhao, B. Liu, W. Li, N. Yu, and Z. Liu, "Anomaly detection and localization: a novel two-phase framework based on trajectory-level characteristics," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2018.

[138] 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*, pp. 733–742, 2016.

[139] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 334–349, Springer, 2016.

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

[141] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[142] 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)*, pp. 439–444, IEEE, 2017.

[143] H. T. Tran and D. Hogg, "Anomaly detection using a convolutional winner-take-all autoencoder," in *Proceedings of the British machine vision conference 2017*, British machine vision association, 2017.

[144] 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*, pp. 2895–2903, 2017.

[145] H. Vu, T. D. Nguyen, A. Travers, S. Venkatesh, and D. Phung, "Energy-based localized anomaly detection in video surveillance," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 641–653, Springer, 2017.

[146] M. G. Narasimhan, "Dynamic video anomaly detection and localization using sparse denoising autoencoders," *Multimedia Tools and Applications*, vol. 77, pp. 13173–13195, 2018.

[147] S. Wang, E. Zhu, J. Yin, and F. Porikli, "Video anomaly detection and localization by local motion based joint video representation and ocelm," *Neurocomputing*, vol. 277, pp. 161–175, 2018.

[148] T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised anomaly detection and localization based on deep spatiotemporal translation network," *IEEE Access*, vol. 8, pp. 50312–50329, 2020.

[149] 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.

[150] Y. Wang, T. Liu, J. Zhou, and J. Guan, "Video anomaly detection based on spatio-temporal relationships among objects," *Neurocomputing*, vol. 532, pp. 141–151, 2023.

[151] T. Reiss and Y. Hoshen, "Attribute-based representations for accurate and interpretable video anomaly detection. arxiv 2022," *arXiv preprint arXiv:2212.00789*.

[152] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7842–7851, 2019.

[153] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2626–2634, 2020.

[154] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 583–591, 2020.

[155] Y. Ouyang and V. Sanchez, "Video anomaly detection by estimating likelihood of representations," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8984–8991, IEEE, 2021.

[156] 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*, pp. 13588–13597, 2021.

[157] 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*, pp. 12742–12752, 2021.

[158] P. R. Roy, G.-A. Bilodeau, and L. Seoud, "Predicting next local appearance for video anomaly detection," in *2021 17th International Conference on Machine Vision and Applications (MVA)*, pp. 1–5, IEEE, 2021.

[159] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2403–2412, 2018.

[160] P. R. Roy, G.-A. Bilodeau, and L. Seoud, "Local anomaly detection in videos using object-centric adversarial learning," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pp. 219–234, Springer, 2021.

[161] Y. Wang, C. Qin, Y. Bai, Y. Xu, X. Ma, and Y. Fu, "Making reconstruction-based method great again for video anomaly detection," in *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1215–1220, IEEE, 2022.

[162] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, and D. Huang, "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles," in *European Conference on Computer Vision*, pp. 494–511, Springer, 2022.

[163] K. Bergaoui, Y. Naji, A. Setkov, A. Loesch, M. Gouiffès, and R. Audigier, "Object-centric and memory-guided normality reconstruction for video anomaly detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2691–2695, IEEE, 2022.

[164] W. Zhou, Y. Li, and C. Zhao, "Object-guided and motion-refined attention network for video anomaly detection," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2022.

[165] K. Doshi and Y. Yilmaz, "Towards interpretable video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2655–2664, 2023.

[166] Y. Liu, Z. Guo, J. Liu, C. Li, and L. Song, "Osin: Object-centric scene inference network for unsupervised video anomaly detection," *IEEE Signal Processing Letters*, vol. 30, pp. 359–363, 2023.

[167] 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)*, pp. 1–5, IEEE, 2023.

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

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

[170] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[171] E. D. Pisano, S. Zong, B. M. Hemminger, M. DeLuca, R. E. Johnston, K. Muller, M. P. Braeuning, and S. M. Pizer, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital imaging*, vol. 11, pp. 193–200, 1998.

[172] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[173] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10663–10671, 2020.

[174] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

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

[176] 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*, pp. 733–742, 2016.

[177] 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), pp. 1–8, IEEE, 2019.

[178] T. N. Nguyen and J. Meunier, "Hybrid deep network for anomaly detection," *arXiv preprint arXiv:1908.06347*, 2019.

[179] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao, "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1070–1084, 2019.

[180] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognition Letters*, vol. 129, pp. 123–130, 2020.

[181] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pp. 329–345, Springer, 2020.

[182] 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*, pp. 14372–14381, 2020.

[183] Q. Li, R. Yang, F. Xiao, B. Bhanu, and F. Zhang, "Attention-based anomaly detection in multi-view surveillance videos," *Knowledge-Based Systems*, vol. 252, p. 109348, 2022.

[184] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3694–3706, 2020.

[185] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 938–946, 2021.

[186] C. Huang, J. Wen, Y. Xu, Q. Jiang, J. Yang, Y. Wang, and D. Zhang, "Self-supervised attentive generative adversarial networks for video anomaly detection," *IEEE transactions on neural networks and learning systems*, 2022.

[187] M. Zhao, X. Zeng, Y. Liu, J. Liu, D. Li, X. Hu, and C. Pang, "Lgn-net: Local-global normality network for video anomaly detection," *arXiv preprint arXiv:2211.07454*, 2022.

[188] M. Astrid, M. Z. Zaheer, and S.-I. Lee, "Pseudobound: Limiting the anomaly reconstruction capability of one-class classifiers using pseudo anomalies," *Neurocomputing*, vol. 534, pp. 147–160, 2023.

[189] Z. Yang and R. J. Radke, "Context-aware video anomaly detection in long-term datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4002–4011, 2024.

[190] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proceedings of the IEEE international conference on computer vision*, pp. 3619–3627, 2017.

[191] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11996–12004, 2019.

[192] C. Sun, Y. Jia, Y. Hu, and Y. Wu, "Scene-aware context reasoning for unsupervised abnormal event detection in videos," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 184–192, 2020.

[193] K. Doshi and Y. Yilmaz, "Any-shot sequential anomaly detection in surveillance videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 934–935, 2020.

[194] K. Doshi and Y. Yilmaz, "Continual learning for anomaly detection in surveillance videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 254–255, 2020.

[195] H. Huang, B. Zhao, F. Gao, P. Chen, J. Wang, and A. Hussain, "A novel unsupervised video anomaly detection framework based on optical flow reconstruction and erased frame prediction," *Sensors*, vol. 23, no. 10, p. 4828, 2023.