OBJECT DETECTION AND TRACKING UNDERVARYING ILLUMINATIONS

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

in

Electronics and Communication Engineering

By

CHINTHAKINDI KIRAN KUMAR

Registration Number: 41900488

Supervised By

Dr Gaurav Sethi (11106)

SEEE (Professor & HOS)

Lovely Professional University

Co-Supervised by

Dr Kirti Rawal (20248)

SEEE (Associate Professor)

Lovely Professional University

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB

2025

DECLARATION

I, hereby declared that the presented work in the thesis entitled **"Object Detection and Tracking under Varying Illuminations"** in fulfillment of degree of Doctor of Philosophy (Ph. D.) is outcome of research work carried out by me under the supervision of **Dr. Gaurav Sethi** working as Professor & HOS, in the School of Electronics & Electrical Engineering of Lovely Professional University, Punjab, India. In keeping with general practice of reporting scientific observations, due acknowledgements have been made whenever work described here has been based on findings of other investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.

(Signature of Scholar)

Name of the scholar: CHINTHAKINDI KIRAN KUMAR
Registration No.:41900488
Department/school: School of Electronics & Electrical Engineering
Lovely Professional University,
Punjab, India

CERTIFICATE

This is to certify that the work reported in the Ph. D. thesis entitled **"Object Detection and Tracking under Varying Illuminations"** submitted in fulfillment of the requirement for the reward of a degree of Doctor of Philosophy (Ph.D.) in the School of Electronics and Electrical Engineering, is a research work carried out by **Chinthakindi Kiran Kumar, 41900488**, is a bonafide record of his/her original work carried out under our supervision and that no part of the thesis has been submitted for any other degree, diploma or equivalent course.

(Signature of Supervisor)

Name of supervisor: Dr. Gaurav Sethi
Designation: HOS & Professor
Department/school: School of Electronics and Electrical Engineering
University: Lovely Professional University

(Signature of Co- Supervisor)

Name of Co-Supervisor: Dr.Kirti Rawal
Designation: Professor
Department/school: School of Electronics and Electrical Engineering
University: Lovely Professional University

ACKNOWLEDGEMENT

First and foremost, I would like to thank God Almighty for giving me the wisdom, strength, knowledge, ability, and opportunity to undertake this project. It gives me a great delight to express my gratitude to many people, without their support and inspiration; this thesis work would not have been possible.

I would like to express my sincere gratitude to my supervisor **Dr. Gaurav Sethi**, HOS & Professor, SEEE, Lovely Professional University and my co-supervisor, **Dr.Kirti Rawal**, Professor, SEEE, Lovely Professional University for the continuous support of my Ph.D. study and related research, for their patience, motivation, and immense knowledge. His/Her guidance helped me in all the time of research and writing of this thesis. It was a great pleasure for me to have a chance to work with them. I have really learned the basics of research from both of them.

I would like to thank all the **faculty and staff members** of the School of Electronics and Electrical Engineering, Lovely Professional University, Punjab, India, for their help and support. I am deeply thankful to **Chancellor, Vice Chancellor** Lovely Professional University, **Registrar**, **Dean** (Academic), **Dean** (School of Electronics & Electrical Engineering), **DRP and RDC members** for their moral support and for providing me with all the necessary facilities during my candidature.

I also express my deep sense of gratitude to all **faculty and staff members** of Electronics and Communication Engineering Department of Malla Reddy College of Engineering and Technology for their constant moral support and inspiration.

I owe to my family members for all types of support. Without their affection, there would have been nothing to write an acknowledgement. In particular, I thank my parents, **Mr Babu** and **Mrs Vijaya Laxmi** , my brother and his wife, **Mr Santosh Kumar & Niharika** and my sister and her husband **Soujanya & Abhilash**, for their wishes, affection, love, inspiration, and never-ending support that helped me in the difficult stages of my work. Very special thanks to my loving and caring wife **Sharon Shiny**, for her love, affection, continuous support and encouragement. I am thankful to my son **Snithik** for making my life filled with joy. I also thank all my family members for their moral support, well wishes, and love.

Last but not least, I express my gratitude to those who helped me directly and indirectly with the successful completion of the research work.

Thank you all for encouraging me and inspiring me to achieve my dreams.

CHINTHAKINDI KIRAN KUMAR
Registration No.:41900488
School of Electronics & Electrical Engineering
Lovely Professional University,

ABSTRACT

The advent of computer vision has revolutionized various sectors .In computer vision; object detection and tracking are fundamental tasks with a wide range of applications, from surveillance and autonomous vehicles to augmented reality and medical imaging. However, the performance of these tasks heavily relies on the quality and consistency of the input images or video frames. One of the most critical challenges faced by object detection and tracking algorithms is the adverse effects of varying illumination conditions. Illumination variations occur due to changes in lighting sources, environmental conditions, and camera settings, leading to significant alterations in the brightness, contrast, and color distribution of images. These variations can obscure objects, create false positives or negatives, and reduce the overall accuracy and robustness of detection and tracking systems. However, achieving robust performance under varying illumination conditions remains a significant challenge. Object detection and tracking under challenging lighting conditions remains a partially unresolved problem. This thesis addresses this challenge by proposing novel techniques for object detection and tracking under varying illuminations.

The key stages involved in this research work include illumination correction; object detection and tracking .we have collected different low illuminated images from various standard datasets. The primary goal of illumination correction is to reduce the impact of illumination variations while preserving the essential information within the image. Low light image enhancement is of significant importance for outdoor computer vision applications. A great deal of techniques have been designed to enhance the appearance of images, but their effectiveness is limited to specific image types such as overexposed or underexposed, and they may not produce the desired outcome for other types of images. In this we focused on boosting the clarity of images captured in low light conditions and restoring their colors. By homogenizing the illumination across images, these techniques enhance the perceptibility of objects, mitigate the influence of varying lighting conditions, and ultimately improve the performance of object detection and tracking algorithms. To

address this task, we introduce a novel illumination algorithm that effectively improves brightness, enhances contrast, and properly processes colors of low light images. The proposed method employs a concise series of steps and merges various processing principles to attain the desired outcome. The proposed algorithm is evaluated by employing the performance metrics like Lightness Order Error, Peak Signal to Noise Ratio and Mean Squared Error. The investigational results prove the substantial effectiveness of our algorithm, both visually and through numerical comparisons with existing methods. Additionally, the proposed method offers exceptional computational speed, making it a viable option for real-world applications where efficiency is a crucial factor.

After illumination correction the next stage is video object detection and tracking. Object detection and tracking are important problems in computer vision that entail automatically recognizing and localizing objects of interest in video sequences and precisely monitoring how they move over time. These tasks may be accomplished via the use of computer vision software. The process of finding objects inside individual frames of a video is referred to as video object detection. This process provides accurate annotations of bounding boxes as well as class labels. Once the objects are identified, the object tracking identifies the trajectories of the path the objects have moved. Deep learning-based techniques, like as convolutional neural networks (CNNs), have revolutionized video object detection by capitalizing on the capability of feature extraction and classification to reach high levels of accuracy and resilience in their results. We proposed a Fast-RCNN (Fast-Region-based Convolutional Neural Network) based deep learning approach for video object detection and Online Continuous-time object tracking (OCSort) from YOLOv8 to perform object tracking. The Fast-RCNN model extracts features from the input frame using pre-trained ResNet152 model and uses selective search algorithm to optimize the search process. The extracted features are then sent to RoI (Region of Interest) pooling and fully connected layers for object detection and bounding box assignment. The OCSort model tracks the detected objects even under occlusion. The proposed object detection model obtained an accuracy of 93% for Person, 88% for Car, and 43% for Truck. The proposed tracking module obtained an average accuracy of 69% in object tracking.

The research presented in this study showcases the impact of illumination correction on the accuracy, robustness, and generalizability of object detection and tracking systems. By experimentally evaluating illumination correction technique on diverse datasets, we aim to provide empirical evidence of their effectiveness in addressing illumination variations and improving the overall performance of object detection and tracking in computer vision applications.

.

LIST OF CONTENTS

LIST OF CONTENTS 1
LIST OF FIGURES 4
LIST OF TABLES 5
LIST OF ABBREVATIONS 6
CHAPTER 1 10
INTRODUCTION 10
1.1 INTRODUCTION 10
1.2 BACKGROUND AND SIGNIFICANCE 11
1.2.1 Illumination Challenges in Imaging 11
1.2.2 Need for Image Enhancement 12
1.2.3 The Application of Deep Learning in the Field of Computer Vision 13
1.2.4 Object Detection 14
1.2.5 Object Tracking 16
1.3 RESEARCH CONTEXT 17
1.4 MOTIVATION 19
1.5 RESEARCH GAPS 20
1.6 PROBLEM FORMULATION 20
1.7 OBJECTIVES 21
1.8 RESEARCH SIGNIFICANCE: 21
1.9 RESEARCH METHODOLOGY 22
1.9.1. Low Light Image Enhancement 22
1.9.2. Video Object Detection Using Deep-Learning Fast RCNN Model 23
1.9.3. Video Object Tracking Using Deep-Learning Using OC SORT Model 24
1.10 THESIS ORGANIZATION 25
CHAPTER 2 27
LITERATURE SURVEY 27
2.1 VIDEO ILLUMINATION ADJUSTMENT 27
2.2 VIDEO OBJECT DETECTION 31
2.3 VIDEO OBJECT TRACKING 39
2.4 SUMMARY 47

CHAPTER 3 48
LOW LIGHT IMAGE ENHANCEMENT: AN INNOVATIVE APPROACH FOR UNEVEN ILLUMINATION CORRECTION 48
3.1 INTRODUCTION 48
3.2 PROPOSED METHOD 49
3.2.1 Dataset 49
3.2.2 Proposed Illumination Correction Method 49
3.2.3 Input Images 50
3.2.4 Exponential Transformation of Image 52
3.2.5 Hyperbolic Tangent Profile 53
3.2.6 Logarithmic Image Processing (LIP) 54
3.2.7 Logarithmic Scaling Function 55
3.3 RESULTS AND DISCUSSION 57
3.3.1 Lightness Order Error (LOE) 58
3.3.2 Peak Signal to Noise Ratio (PSNR) 60
3.3.3 Mean Squared Error 61
3.3.4 Average Implementation Time 62
3.4 SUMMARY 63
CHAPTER 4 65
VIDEO OBJECT DETECTION USING Fast RCNN WITH ResNet MODEL 65
4.1 INTRODUCTION 65
4.2 PROPOSED MODEL 67
4.2.1 Proposed Object Detection Model 67
4.3 SIMULATION RESULTS 78
4.3.1 Datasets 78
4.3.2 Object Detection 83
4.4 SUMMARY 88
CHAPTER 5 89
VIDEO OBJECT TRACKING USING OC-SORT 89
5.1 INTRODUCTION 89
5.2 PROPOSED MODEL 91
5.2.1 Proposed Object Tracking 91
5.3 SIMULATION RESULTS 97

5.3.1 Datasets 98
5.3.2 Object Tracking 103
5.4 SUMMARY 107
CHAPTER 6 109
CONCLUSION AND FUTURE SCOPE 109
6.1 CONCLUSION 109
6.2 FUTURE SCOPE 111
LIST OF PUBLICATIONS 113
REFERENCES 114

LIST OF FIGURES

Figure 3. 1: Shows the general structure of Illumination correction 49
Figure 3. 2: shows the input images with different occlusions 51
Figure 3. 3: Shows the output of the exponential function 52
Figure 3. 4: Shows the output of hyperbolic tangent 54
Figure 3. 5: Shows the output of the LIP addition of two images 55
Figure 3. 6: Enhanced dark pixels 56
Figure 3. 7: Shows final illumination corrected images 57
Figure 3. 8: Visual Comparison of proposed method with state-of-the-art techniques 59
Figure 3. 9: Analytical graph of average comparison of LOE 60
Figure 3. 10: Analytical graph of average comparison of PSNR 61
Figure 3. 11: Analytical graph of average comparison of MSE 62
Figure 3. 12: Analytical graph of average implementation time 63
Figure 4. 1: Proposed Fast R-CNN with ResNet152 Architecture 70
Figure 4. 2: ResNet152 model 72
Figure 4. 3: Residue block 75
Figure 4. 4: Sample images in different videos 83
Figure 4. 5: Object detection results on different videos 85
Figure 4. 6: Comparative Analysis chart 87
Figure 5. 1: Object tracking framework 94
Figure 5. 2: YOLOv8 deep learning model 96
Figure 5. 3: Sample images in different videos 102
Figure 5. 4: Object Tracking results on different videos 105
Figure 5. 5: Comparative Time Analysis Chart 106
Figure 5. 6: Comparative Accuracy Analysis Chart 107

LIST OF TABLES

Table 3. 1: LOE metric score of Proposed with compared algorithms.....60
Table 3. 2: PSNR metric score of Proposed with compared algorithms 61
Table 3. 3: MSE metric score of Proposed with compared Algorithms 53
Table 3. 4: Implementation time of Proposed with compared Algorithms.....63
Table 4. 1: Comparative Analysis of proposed Object detection model 86
Table 5. 1: Comparative Analysis of proposed Object tracking model.....105

LIST OF ABBREVATIONS

CNNs	Convolutional Neural Networks
DCNN	Deep Convolutional Neural Networks
RPNs	Region proposal networks
NMS	Non-Maximum Suppression
GPUs	Graphics Processing Unit
RCNN	Region-Convolutional Neural Network
OC Sort	Observation-Centric Sort
SORT	Simple Online and Real-time Tracking
NIQE	No-Reference Image Quality Estimation
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural similarity index
IoU	Intersection over Union
MOTA	Multiple Object Tracking Accuracy
MOTP	Multiple Object Tracking Precision
DIC	Digital Image Correlation
LLIE	Low Light Image Enhancement
LVENet	Low-Visibility Enhancement Network
DVENet	Dual-View Enhancement Network
PRIEN	Progressive-Recursive Image Enhancement Network
HVR-Net	Hierarchical Video Relation Network

LSTS	Learnable Spatio-Temporal Sampling
DFA	Dense Feature Aggregation
SRFU	Sparsely Recursive Feature Updating
YOLO	You Only Look Once
MFCN	Motion-aid Feature Calibration Network
BFAN	Blur-aid Feature Aggregation Network
MRF	Markov Random Field
MAP	Maximum A Posteriori
DDM	Change Detection Mask
FFAVOD	Feature Fusion Architecture for Video Object Detection
GMPNet	Grid Message Passing Network
K-NN	K-Nearest Neighbor
AST-GRU	Attentive Spatiotemporal Transformer GRU
GRU	Gated Recurrent Unit
STA	Spatial Transformer Attention
TTA	Temporal Transformer Attention
TEN	Triple Excitation Network
VSOD	Video Salient Object Detection
GDR	Gather Diffusion Reinforcement
CRC	Cross-modality Refinement and Complement
IPF	Importance Perception Fusion
MOT	Multiple Object Tracking

SatSOT	Satellite Surveillance of objects
LaSOT	Large-scale Single Object Tracking
UAVDT	Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking
DQN	Deep Quadruplet Network
ReLU	Rectified Linear Units
Tiny-DSOD	Tiny-Deeply Supervised Object Detector
SBF-LSTM	Stacked Bidirectional-Forward Long Short-Term Memory
QDTrack	Quasi-Dense Tracking
KCF	Kernel Correlation Filter
TIR	Thermal Infrared
DASFTOT	Deep Adaptive Spatio-Temporal Feature Transformer Object Tracking
ME	Motion Estimation
CSRT	Channel and Spatial Reliability Tracking
ATOM	Accurate Tracking by Overlap Maximization
LIP	Logarithmic Image Processing
CLAHE	Contrast Limited Adaptive Histogram Equalization
LIBCP	Low-light Image Enhancement with Bright Channel Prior
MF-LIME	Multi Frame Low-light image enhancement
LECARM	Low-Light Camera Response Model
RBMP	Retinex-Based-Multiphase-Algorithm

LOE	Lightness Order Error
IMMSE	Mean-Squared Error
ResNet	Residual Network
RoI	Region of Interest
CDNET	Change Detection Dataset
SSD	Single shot detector
LSTM	Long Short-Term Memory
SiLU	Sigmoid Linear Unit
SSPF	Spatial Pyramid Pooling – Fast

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Video object detection and tracking play a crucial role in the field of computer vision [1], since they empower machines to autonomously recognize and track things in a series of video frames. These techniques play a crucial role in a wide range of applications, including surveillance systems, autonomous vehicles, video analytics, and various others. In the domain of computer vision, these entities fulfil discrete yet interrelated objectives.

The first step in video analysis involves the detection of items inside each frame of a video. The procedure is dependent on computer vision methodologies, specifically deep learning architectures such as Convolutional Neural Networks (CNNs) [2]. These models are trained to identify and classify things by acquiring knowledge of their visual characteristics through extensive exposure to labelled datasets. When the deep learning model is applied to each frame of a movie, it systematically examines the image, identifies and localizes things that are of significance, and subsequently assigns them to appropriate classes. Consequently, the outcome is the automated detection and precise determination of the objects' positions inside the frames of the movie.

Video object detection has significant versatility and applicability across a wide range of settings [3]. In the context of surveillance systems, the utilization of Artificial Intelligence (AI) technology enables the automated detection and recognition of potential hazards or objects of interest inside a designated monitored region. Autonomous vehicles assume a pivotal function in the identification and monitoring of fellow vehicles, pedestrians, and road impediments. In the context of the retail industry, computer vision technology can be effectively employed for the purpose of inventory management through its ability to autonomously identify and locate objects placed on shelves.

After the detection of objects in the initial frame, the process of video object tracking is initiated. The process of tracking entails the on-going surveillance of the objects [4] as they traverse consecutive video frames. Tracking algorithms are utilized to forecast the future position of an item in the subsequent frame by considering its previous positions, patterns of motion, and maybe other contextual information.

The primary objective of video object tracking is to sustain a constant comprehension of the object's identity and movement across a certain duration [5]. This is particularly advantageous in situations where objects are in motion, such as in the context of surveillance footage, sports analysis, or traffic monitoring. Furthermore, it is imperative to ensure the track ability of objects even in the presence of occlusions, fluctuations in lighting conditions, or alterations in object appearance.

Video object detection and tracking are fundamental procedures in the field of computer vision, facilitating the automatic recognition, localization, and ongoing surveillance of objects inside video sequences. While the task of detection involves the recognition of objects within individual frames, tracking is concerned with the continuous monitoring of objects as they traverse consecutive frames. These procedures possess a diverse array of applications, rendering them useful in numerous industries that rely on object detection and monitoring as crucial components.

1.2 BACKGROUND AND SIGNIFICANCE

This section aims to outline the relevant historical, theoretical, and empirical components of video object detection and tracking.

1.2.1 Illumination Challenges in Imaging

Uneven illumination pertains to the non-uniform dispersion of light throughout an image, leading to regions exhibiting differing levels of luminosity. This phenomenon frequently arises in photographs taken in situations when the illumination is non-uniform or when there are obstructing objects causing shadow formations. Unequal distribution of light can obscure significant features, diminish the quality of images, and hinder the effectiveness of computer vision systems. In certain scenarios, such as

security cameras or outdoor environments, the presence of uneven illumination might provide difficulties in accurately detecting or tracking objects.

Low light conditions present a distinct difficulty when it comes to capturing images [6]. In such circumstances, a lack of enough illumination might result in the production of underexposed photographs, characterized by diminished visibility and reduced discernibility of things within the frame. This phenomenon is frequently encountered in various applications such as night vision technology, astronomical observations, and indoor security systems. The improvement of image quality in low light circumstances is of utmost importance in the fields of computer vision and imaging, as it significantly enhances visibility and facilitates object recognition.

The necessity of image enhancement lies in its objective to enhance the quality and visibility of photographs. In the context of non-uniform illumination and limited luminosity, the process of picture augmentation has paramount importance in addressing these challenges. The task at hand holds considerable importance across multiple domains, encompassing medical imaging, autonomous cars, surveillance, and satellite imagery analysis. In these domains, precise object detection and tracking play a critical role in facilitating decision-making processes [7].

1.2.2 Need for Image Enhancement

Image enhancement is a key aspect of digital image processing, with the objective of enhancing the visual quality of images to facilitate improved human perception or to assist in later computer vision tasks. The imperative for the enhancement of images is motivated by a multitude of crucial variables and practical applications in the actual world.

1. **Poor Image Quality in Various Conditions**: Images obtained in real-world situations frequently exhibit substandard quality as a result of many causes, including but not limited to inconsistent illumination, noise interference, motion blur, and inherent constraints of imaging devices. The presence of flaws has the potential to hide crucial details inside photographs.
2. **Improved Interpretation**: Images that have been enhanced are more readily interpretable by both human observers and machine algorithms. This is

particularly crucial in disciplines such as medical imaging, wherein the accuracy of a diagnosis may heavily rely on the quality and precision of an X-ray or MRI scan.

3. **Low-Light Conditions**: Low-light conditions refer to situations where the level of illumination is significantly reduced, resulting in reduced visibility and potential challenges for visual perception [8]. Under conditions of reduced light, such as during nighttime or in indoor settings with little illumination, photographs may experience significant underexposure. The process of image augmentation plays a critical role in the detection and extraction of concealed information and objects under challenging circumstances. Consequently, it is an essential component in various applications, such as night vision systems and surveillance.
4. **Reducing Noise**: One potential strategy for mitigating noise is to implement measures aimed at reducing its presence or impact. Images frequently exhibit noise, resulting in a degradation in their overall quality. Noise might arise due to constraints imposed by sensors, the presence of compression artifacts, or various other contributing causes. Image enhancement techniques are utilized to reduce noise and generate photos that are clearer and include more useful information.
5. **Enhancing Object Recognition**: Image enhancement plays a crucial role in computer vision and object detection applications. Enhancing the contrast and sharpness of items inside a picture facilitates the improved accuracy of algorithms in the recognition and classification of these things.
6. **Improving the visual Appeal**: Image enhancement techniques are commonly utilized in various domains, including photography and digital art, with the aim of augmenting the visual aesthetics of photographs. This is crucial for individuals in creative professions who strive to create visually striking and aesthetically pleasant imagery.

1.2.3 The Application of Deep Learning in the Field of Computer Vision

CNNs are a class of deep learning models that have been widely used in various computer vision tasks. The field of computer vision has been significantly

transformed by the emergence of deep learning techniques, CNNs playing a pivotal role in this revolution [9]. CNNs have demonstrated significant efficacy in the domain of feature extraction and pattern recognition, rendering them well-suited for many image processing applications. CNNs have exceptional proficiency in many tasks such as object detection, picture classification, and segmentation. This is mostly attributed to their ability to autonomously acquire pertinent features from data, hence diminishing the need for manually engineered features and facilitating the execution of intricate tasks.

The significance of deep learning in the context of object detection and tracking lies in its ability to address the limitations of traditional computer vision algorithms when confronted with the intricacies of real-world settings [10]. The utilization of neural networks in deep learning facilitates the process of automated feature extraction, hence mitigating the necessity for manual feature engineering. Therefore, the aforementioned outcome leads to object identification and tracking systems that include enhanced resilience, precision, and flexibility. Deep learning models have demonstrated their efficacy in demanding scenarios, such as autonomous vehicles, surveillance systems, and robots, where there can be substantial variations in illumination conditions.

1.2.4 Object Detection

The task of object detection [11] holds significant importance in the field of computer vision, serving as a fundamental component in a wide range of applications such as autonomous vehicles, surveillance systems, facial recognition, and medical imaging. This technological advancement facilitates the ability of machines to discern and determine the identity and spatial positioning of objects contained inside photographs or frames of video [12]. The framework of object detection adheres to a generic algorithmic flow comprising multiple essential phases. Although the specific algorithms for object detection may differ, the following typical flow offers a systematic overview of the process.

The proposed algorithm for generic object detection follows the following flow:

1. **Input Data**: The initial step involves obtaining input data, commonly in the format of digital photos or video frames. These visual representations can be obtained using a range of methods, including the utilization of cameras, sensors, or pre-existing databases.
2. **Pre-processing**: Pre-processing of input data is frequently required prior to object detection in order to facilitate the subsequent detection procedure. This process may entail the modification of the image's dimensions, the manipulation of its brightness and contrast levels, and the application of noise reduction techniques. These adjustments are necessary to ensure that the images are in an appropriate format for subsequent analysis.
3. **Feature extraction**: The process of feature extraction is a crucial stage in which unique characteristics are recognized within the image [13]. The characteristics encompass many visual attributes such as edges, forms, colors, textures, and other patterns that serve to delineate and characterize objects [14]. The objective is to depict the visual representation in a manner that facilitates the process of identifying and categorizing objects.
4. **Proposal Generation for Objects**: During this stage, the system generates possible regions of interest, also known as object proposals. These regions represent potential areas where things could potentially be situated. Methods such as selective search or Region Proposal Networks (RPNs) are frequently utilized for the purpose of detecting these regions.
5. **Features Classification**: The analysis of each proposed object is conducted to assess its characteristics and ascertain the presence of a relevant object. The standard procedure entails utilizing a classifier, such as a CNN, to evaluate the presence of an object within a given region [15].
6. **Bounding box regression**: When an object is discovered within a specific region, the system proceeds to conduct bounding box regression [16]. This stage involves refining the location and dimensions of the bounding box surrounding the object in order to accurately encompass it.

7. **Non-Maximum Suppression (NMS)**: It is a technique commonly employed in computer vision and image processing tasks. NMS is employed to eliminate duplicate detections when there are instances of overlapping or redundant object proposals. This procedure guarantees that only the object detections with the highest level of confidence are preserved.
8. **Output**: The ultimate result of the object identification technique encompasses the precise coordinates of the bounding boxes surrounding the identified items, in addition to the class labels that serve to categorize the type of object contained within each bounding box.

1.2.5 Object Tracking

Object tracking is a key aspect of computer vision and holds significant importance in a wide range of applications [17], such as surveillance, autonomous cars, human-computer interaction, and visual effects [18]. The process entails the ongoing observation and tracking of a particular entity inside a series of video frames as it undergoes motion. The generic object tracking method adheres to a systematic process that facilitates the consistent tracking of an object's position and identity across a video.

1. **Initialization**: The object tracking procedure commences with the identification of the target object to be tracked in the initial frame of the video. The identification of the target is commonly achieved by delineating a bounding box around it. The bounding box serves as the first representation of the target's position and dimensions.
2. **Feature Extraction**: The process of extracting features from the target object is performed in the initial frame. The qualities may encompass attributes such as color information, texture, edges, or key points [19]. The selection of features is contingent upon the design of the tracking algorithm and the unique attributes of the object under consideration.
3. **Target Representation**: The characteristics derived from the target entity are converted into a model that serves as a representative. The present model functions as a standardized framework for the desired outcome and is employed for the purpose of comparison in later iterations.

4. **Frame-by-Frame Tracking**: After the initialization, the tracking algorithm proceeds to hunt for the target in each subsequent frame by conducting a comparison between its model and the features present in the new frame [20]. Different algorithms may employ various strategies, like correlation filters, optical flow, or deep learning networks, to execute this comparison and identify the target.
5. **Model Update**: During the motion of the object, it may be necessary for the tracking algorithm to adjust the model in order to accommodate variations in the target's visual characteristics, size, or alignment. The procedure of updating guarantees the maintenance of accurate tracking over an extended period.
6. **Position Refinement**: To improve the accuracy of tracking, it is common for the algorithm to refine the predicted position of the target item within the frame. This is commonly achieved by estimating either the centroid of the object or a more accurate location within the bounding box. The maintenance of object identity is of utmost importance in scenarios involving multi-object tracking. The tracking method guarantees the continuous identification of the tracked object throughout frames, especially in scenarios involving occlusions or objects in near vicinity.
7. **Termination**: The process of object tracking can persist until the conclusion of the video sequence or until a predetermined stopping condition is satisfied. The termination condition may be determined by various events, such as the occurrence of a tracking failure or the completion of the movie.

1.3 RESEARCH CONTEXT

The exploration of video object detection and tracking occurs within the dynamic and ever-changing field of computer vision. The field of computer vision comprises a wide range of technologies and approaches that are designed to facilitate the interpretation and comprehension of visual information by machines, enabling them to see and understand their surrounding environment. In recent years, this particular domain has witnessed notable progress, mostly propelled by the swift evolution of

deep learning techniques. This innovation has greatly enhanced the precision and resilience of object detection and tracking systems.

Video object detection and tracking are fundamental elements of computer vision, playing a crucial role in various domains. In the domains of surveillance and security, the utilization of these technologies is imperative for the automated identification and monitoring of objects or individuals that are of significance. Vehicle and pedestrian tracking systems play a crucial role in traffic management by enhancing road safety. Video object detection and tracking play a crucial role in the domain of autonomous cars since they are essential for the identification of other vehicles, pedestrians, and obstacles [21]. This capability is vital in guaranteeing the safe and effective navigation of autonomous vehicles. These technologies are also crucial in several industries, including retail, as they play a significant role in optimizing inventory management and improving consumer experiences.

Furthermore, the study framework pertaining to video object detection and tracking encompasses the wider domain of AI. The location of this phenomenon is at the convergence point of computer vision, machine learning, and deep learning. The advancement of resilient and instantaneous video object detection and tracking systems signifies a noteworthy progression in the pursuit of artificial intelligence-powered automation across diverse domains. This development has the potential to enhance safety, efficacy, and user satisfaction.

The research is undertaken within a technological environment that is marked by the continuous advancement of sophisticated hardware and software solutions. The topic of deep learning has become more accessible to a wider variety of researchers and practitioners due to the increased availability of powerful Graphical Processing Units (GPUs) and specialized accelerators, as well as the utilization of open-source deep learning frameworks. In addition, the increasing prevalence of various video data sources, including surveillance cameras, drones, and smartphones, has contributed to the need for effective techniques in video object detection and tracking.

The investigation pertaining to video object detection and tracking is situated within the dynamic environment of computer vision, deep learning, and the broader domain

of artificial intelligence. The motivation behind this endeavour is from the need to enhance automated object recognition and tracking in diverse fields such as surveillance, autonomous vehicles, retail, and other related areas. The goal is to increase safety, efficiency, and user experiences. The continuous progress in technology and the growing accessibility of video data sources emphasize the significance and pertinence of this research field.

1.4 MOTIVATION

In the contemporary landscape of rapidly evolving technology, research on video object detection and tracking is motivated by several significant factors, and it is considered a crucial pursuit. The demand for advanced computer vision systems that can interpret and understand the visual world has been increasing, and this demand is driven by various factors.

An unprecedented level of video data is being generated in our daily lives. Video streams are being generated from security cameras, dashcams, drones, and even smartphones, and they contain a wealth of information that can be harnessed for a multitude of applications. However, this vast reservoir of data remains largely untapped without effective video object detection and tracking mechanisms. Whether for surveillance, autonomous vehicles, or content analysis in the media industry, the necessity to automatically detect, identify, and follow objects within video streams is undeniable.

The requirement for enhanced video object detection and tracking techniques also arises from the urgency to bolster security and public safety. In the context of surveillance, real-time and accurate object detection and tracking are critical for identifying and responding to potential threats, intrusions, or other security breaches. The ability to monitor and track individuals, vehicles, and objects with precision is indispensable for law enforcement and various security applications.

Furthermore, the rise of autonomous systems, including self-driving cars, drones, and robotics, depends on robust video object detection and tracking. These technologies rely on computer vision to navigate and interact with the environment. Accurate

detection and tracking of objects in real-time are pivotal for ensuring the safety and efficiency of these systems, reducing accidents, and improving overall reliability.

In the realm of content creation and media, the significance of video object detection and tracking is equally pronounced. Filmmakers and content producers seek to automate labour-intensive tasks such as scene analysis, special effects, and tracking objects for storytelling purposes. Efficient video object detection and tracking can expedite the creative process, leading to innovative storytelling and captivating visual experiences.

Moreover, as significant strides have been made in recent years in the fields of machine learning and deep learning, there is a growing recognition that video object detection and tracking can be revolutionized through these technologies. The use of neural networks, particularly CNNs, has shown promise in improving the accuracy and speed of these processes. This research, therefore, capitalizes on the synergy between computer vision and advanced machine learning techniques to unlock new capabilities in video analysis.

1.5 RESEARCH GAPS

- In earlier methods detection of salient objects of interest is cumbersome, hence there is a scope for this research to detect the object of interest under different illumination condition and track it accurately.
- The multiple targets tracking problem is more complex than mono target tracking and for the efficient implementation of multiple target tracking algorithms, many problems which does not exist for mono target tracking must be resolved.
- Many of the approaches are not performed on real time video sequences

1.6 PROBLEM FORMULATION

The primary aim of this research is to develop innovative approaches for tackling these challenges in computer vision and image processing. To achieve this, we will focus on three main objectives:

1. **Low Light Image Enhancement(LLIE):** Uneven illumination often degrades the quality of images, making it challenging to analyze and interpret visual data. The problem of LLIE will be addressed by proposing a novel method to correct uneven illumination. This method aims to significantly improve the visibility of objects in low-light conditions by efficiently compensating for variations in illumination across the image.

2. **Video Object Detection:** In video analysis, detecting and identifying objects in each frame is a fundamental step. We will employ a deep-learning Fast Region-Convolutional Neural Network (F-RCNN) model to perform video object detection. This model will be designed to accurately locate and classify objects in video sequences, contributing to improved object recognition and tracking in real-time scenarios.

3. **Video Object Tracking:** Once objects are detected in a video, it is crucial to track them across frames to maintain their identities. Object tracking is an essential element of surveillance, tracking, and autonomous systems. To address this, we will utilize the OC Sort model, a deep-learning-based approach, to perform object tracking efficiently and robustly across video frames.

1.7 OBJECTIVES

Objective 1: Low Light Image Enhancement: An Innovative Approach for Uneven Illumination Correction

Objective 2: To perform video object detection using deep-learning Fast RCNN model.

Objective 3: To perform video object tracking using deep learning using OC Sort model.

1.8 RESEARCH SIGNIFICANCE:

This research project holds substantial significance for several reasons:

1. **Practical Applications:** The outcomes of this research will have broad practical applications in various fields, including surveillance, autonomous vehicles, and media production. Improving low-light image enhancement, object detection, and object tracking can enhance the performance and reliability of systems in these domains.

2. **Technological Advancements:** The proposed innovative approaches will contribute to the ongoing advancement of deep learning and computer vision techniques. By addressing the challenges associated with uneven illumination, object detection, and object tracking, we will push the boundaries of what is currently possible in the field of computer vision.

3. **Societal Impact:** The research has the potential to improve safety and security through more accurate object detection and tracking, which is essential for surveillance and autonomous systems. Additionally, it can enhance the quality of images and videos in low-light conditions, benefiting photography and media production.

4. **Academic Contribution:** The research will add to the body of knowledge in computer vision, image processing, and deep learning. The development of innovative methods for addressing these challenges can serve as a foundation for further academic research and studies in related areas.

1.9 RESEARCH METHODOLOGY

1.9.1. Low Light Image Enhancement

Data Collection:

- A diverse dataset of low-light images with varying levels of uneven illumination will be acquired.
- Ground truth images for evaluation will be ensured to be included in the dataset.

Pre-processing:

- The acquired images will be normalized and standardized.
- The dataset will be divided into training, validation, and test sets.

Innovative Approach Development:

- Existing methods for uneven illumination correction will be investigated.
- An innovative approach that combines deep learning and image processing techniques to correct uneven illumination will be proposed.
- A CNN model tailored for this task will be developed and trained.

Model Training:

- The CNN model will be trained on the training dataset.
- Appropriate loss functions and optimization algorithms for this specific task will be utilized.
- Data augmentation techniques will be implemented to increase model robustness.

Evaluation:

- The model's performance will be quantitatively evaluated using metrics like PSNR , Structural similarity index (SSIM)
- The corrected images will be qualitatively assessed for visual quality.
- Comparison of the innovative approach against baseline methods will be performed.

1.9.2. Video Object Detection Using Deep-Learning Fast RCNN Model

Data Preparation:

- Video datasets with annotated object bounding boxes for training will be collected.

- The data will be divided into training and validation subsets.

Fast RCNN Model Selection:

- A pre-trained Fast RCNN model (e.g., ResNet-based) will be chosen as the base architecture.
- The model will be fine-tuned for the object detection task.

Training:

- NMS and anchor box generation for efficient detection will be implemented.
- The Fast RCNN model will be trained on the annotated data.
- Transfer learning will be used to adapt the model to the target dataset.

Evaluation:

- The model's performance on the validation dataset will be assessed using metrics like Mean Average Precision (mAP), precision, and recall.
- Hyper parameters will be tuned for optimal results.

1.9.3. Video Object Tracking Using Deep-Learning Using OC SORT Model

Data Collection:

- Video sequences with labelled object tracks will be gathered.
- Objects with various motion patterns and occlusions will be ensured to be included in the data.

OC SORT Model Implementation:

- The OC SORT algorithm, which combines object detection and tracking, will be implemented.
- A pre-trained object detector (e.g., Fast RCNN) will be fine-tuned for improved object recognition.

- The Simple Online and Real-time Tracking (SORT) algorithm will be integrated for tracking.

Tracking in Video:

- The OC SORT model will be applied to track objects in video sequences.
- Occlusions, object identity preservation, and track maintenance will be handled.

Evaluation:

- Tracking accuracy will be quantitatively evaluated using metrics such as Intersection Of Union (IoU), Multiple Object Tracking Accuracy (MOTA), and Multiple Object Tracking Precision (MOTP).
- The model's performance under challenging scenarios like occlusions and object appearance changes will be analysed.

Integration of Object Detection and Tracking:

- The results of the object detection will be combined with the tracking to achieve robust object tracking in videos.

Cross-Objective Integration:

- The low-light image enhancement will be integrated into the video processing pipeline to improve object detection and tracking under challenging lighting conditions.

Iterative Improvement:

- The methodology will be iterated through, making necessary refinements based on evaluation results, and the models will be fine-tuned to achieve optimal performance for all three objectives.

1.10 THESIS ORGANIZATION

Chapter 1 - Introduction This chapter introduces the research, including the background and significance of the study, the challenges in imaging illumination, the

need for image enhancement, and the application of deep learning in computer vision. It also outlines the research context, motivation, problem formulation, objectives, research significance, and the research methodology, which includes LLIE, video object detection using the Fast RCNN model, and video object tracking using the OC Sort model.

Chapter 2 - Literature Survey In this chapter, a comprehensive literature survey is conducted, covering topics related to video illumination adjustment, video object detection, and video object tracking. This survey serves as a foundation for the research and provides insights into existing approaches and techniques in these areas.

Chapter 3 - Low Light Image Enhancement: An Innovative Approach for Uneven Illumination Correction This chapter focuses on the proposed method for LLIE. It discusses the dataset used, the novel illumination correction method involving exponential transformation, hyperbolic tangent profile, LIP, and logarithmic scaling function. The results and discussion section evaluates the method's performance using metrics like Lightness Order Error, PSNR, Mean Squared Error, and Average Implementation Time.

Chapter 4 - Video Object Detection using Fast RCNN with ResNet model Chapter 4 introduces the research on video object detection. It presents the proposed Fast RCNN model with ResNet and details the simulation results, including the datasets used and the performance of object detection. The chapter concludes with a summary of the findings.

Chapter 5 - Video Object Tracking using OC Sort This chapter delves into video object tracking, showcasing the proposed OC Sort model for tracking. It outlines the model, the simulation results, the datasets employed, and the performance in terms of object tracking. The chapter closes with a conclusion summarizing the outcomes of the tracking experiments.

CHAPTER 2
LITERATURE SURVEY

2.1 VIDEO ILLUMINATION ADJUSTMENT

This chapter presents a comprehensive review of object detection and tracking under varying lightning conditions. The survey explores the computer vision as well as modern machine and deep learning based methods highlighting the contributions, advantages and limitations in varying illumination environment.

Da Yang et al [22] suggested a technique including global illumination adjustment aims to dynamically modify the camera's exposure duration to obtain an ideal distribution of intensity. In this study, a novel Digital Image Correlation (DIC) technique is introduced, which is designed to accurately quantify video deflection in real-time, while being resilient to variations in lighting conditions. The combined strategy enhances the resilience and accuracy of measurement via two distinct mechanisms: globally dynamic intensity adjustment and alteration of the grayscale arrays in subsets throughout the correlation process.

Bin Liao et al [23] focused to provide a technique for light control that emulates retinal processing, with the objective of mitigating variations in illumination. In this study, the authors use a weighted neighbourhood filtering approach to enhance the accuracy of optical flow estimate by introducing an edge refinement method.

Pablo Gómez et al [24] proposed a methodology for improving the quality of low-light pictures obtained from high-speed video endoscopy via the use of a CNN. In this study, the authors provide a novel approach for generating training examples with realistic darkening effects by using the Perlin noise algorithm. The use of extensive data augmentation techniques is implemented in order to address the constraint of minimal training data, hence enabling training to be conducted using a mere 55 films. The proposed methodology is evaluated against four contemporary low-light enhancement approaches, which are considered state-of-the-art. The results indicate that the proposed approach considerably outperforms each of these methods in terms

of three image quality metrics: No-Reference Image Quality Estimation (NIQE), PSNR, and SSIM.

Qing Zhang et al [25] introduced a unique method for enhancing underexposed photographs, which aims to preserve perceptual consistency. The proposed approach involves the introduction of robust criteria, known as perceptually bidirectional similarity, which provides clear guidelines for ensuring perceptual consistency. In this study, the authors utilize the Retinex theory to address the enhancement problem. They approach this problem by formulating it as an optimization task for estimating constrained illumination. Specifically, they incorporate perceptually bidirectional similarity as constraints on the illumination. By solving for the illumination that can effectively restore the desired enhancement results without any artifacts, they aim to achieve optimal enhancement outcomes. Furthermore, they provide a video improvement framework that incorporates the aforementioned illumination estimation technique to address the issue of underexposed films.

Chongyi Li et al [26] offered a comprehensive dataset including of low-light photos and movies captured by several mobile phone cameras, encompassing a wide range of lighting situations. In addition, the authors are introducing a novel online platform that offers a comprehensive range of widely used LLIE techniques. These methods may be accessed and used via a user-friendly web interface, marking a significant advancement in the services. In addition to conducting qualitative and quantitative evaluations of established methodologies using both publicly accessible datasets and the own suggested datasets, they additionally assess their efficacy in the context of low-light face identification.

Yu Guo et al [27] presented a proposal for a low-visibility enhancement network, referred to as LVENet, which utilizes the principles of Retinex theory to improve the quality of imagery in the context of marine video surveillance. The LVENet is a deep neural network that has been designed to be lightweight, using a depthwise separable convolution technique. This study introduces the concepts of synthetically-degraded picture creation and hybrid loss function as means to improve the resilience and generalization capabilities of LVENet. Both full-reference and no-reference

assessment trials indicate that LVENet has the potential to provide visual quality that are equivalent to, or even superior to, existing state-of-the-art approaches.

Jie Huang et al [28] presented a unique Dual-View Enhancement Network (DVENet) that is built around the Retinex theory. The proposed network has two distinct stages. The first phase involves estimating an illumination map in order to get a preliminary enhancement outcome, which enhances the correlation between two perspectives. Subsequently, the second phase integrates information from both perspectives to recover finer details and achieve improved picture quality, guided by the illumination map. In order to maximize the potential of the dual-view correlation, the authors have furthermore developed a view transfer module based on wavelets. This module is designed to effectively perform multi-scale detail recovery. Subsequently, an illumination-aware attention fusion module is devised in order to effectively use the complementary nature of the fused characteristics obtained from two different perspectives, as well as the features obtained from a single perspective.

Yuanyi He et al [29] examined the use of video analytics in low-light conditions and present a comprehensive system that integrates end-edge coordination, as well as joint video encoding and enhancement techniques. The system has the capability to dynamically transmit movies captured in low-light conditions from cameras, while also carrying out enhancement and inference activities at the edge. Initially, based on the empirical findings, it is evident that both the encoding and enhancement techniques used for low-light films have a substantial effect on the accuracy of inference, hence directly affecting the bandwidth utilization and computational burden. Additionally, cameras encode and send frames to the edge as a result of the inherent constraints of built-in processing resources.

Yu Guo et al [30] proposed an approach to improve the quality of low-light photographs by using regularized illumination optimization and deep noise reduction techniques. The paper introduces a hybrid regularized variational model that integrates a L0-norm gradient sparsity prior with structure-aware regularization. This model aims to enhance the coarse illumination map that was first generated using the Max-RGB method. Next, the adaptive gamma correction technique is shown as a

means to modify the improved illumination map. This study proposes a detail boosting approach based on the guided filter, with the aim of optimizing the reflection map, under the premise of Retinex theory. The process of generating improved marine photos involves the combination of modified lighting and optimized reflection maps. In order to mitigate the impact of undesired noise on the performance of imaging systems, a blind denoising framework based on deep learning is further proposed to increase the visual quality of the resulting picture.

Jinjiang Li et al [31] provided novel neural network architecture, namely the Progressive-Recursive Image Enhancement Network (PRIEN), designed specifically for the purpose of enhancing low-light photographs. The primary concept is the use of a recursive unit, which consists of a recursive layer and a residual block, in order to iteratively expand the input picture for the purpose of extracting features. In contrast to earlier methodologies, the present work employs a novel approach whereby low-light pictures are immediately fed into the dual attention model to facilitate global feature extraction. Subsequently, a fusion of recurrent layers and residual blocks is used for the purpose of local feature extraction. Ultimately, the improved picture is generated as the final product. Additionally, the global feature map of dual attention is sequentially incorporated into each step in a progressive manner. The recurrent layer in the local feature extraction module facilitates the sharing of depth characteristics across different stages.

Liu et al. [32] proposed LIEDNet**,** a lightweight model that effectively enhance and optimize low-light videos using combined histogram equalization and CNN techniques. Complementarily, Ayoub et al. [33] proposed a transmission map dehazing technique that adjusts illumination and improves clarity for both videos and images.Jiang et al[34] developed DarkSeg**,** an edge-optimized network for nighttime semantic segmentation, integrating residual blocks and illumination enhancement modules.Yue et al. [35] addressed high dynamic range (HDR) through the introduction of a novel staggered HDR video reconstruction method which employs alternating exposure and fusion and works well for night conditions.

2.2 VIDEO OBJECT DETECTION

Object detection is a basic task in the field of computer vision that includes identifying objects within an image or video frame but also accurately confining those using bounding boxes. This literature review aims to examine the development of object detection methodologies ranging from basic approaches to latest deep learning methods.

Haidi Zhu et al [36] provided an evaluation of the aforementioned works pertaining to the field of video object detection. Firstly, this paper provides a comprehensive review of the currently available datasets for video object recognition. Additionally, it discusses the generally used evaluation criteria in this field. The following categorization of video object detection techniques is shown, along by a description of each approach. Two comparative tables are presented to examine disparities in terms of both precision and computational efficiency.

Mingfei Han et al [37] propose the development of a new module that establishes an Inter-Video Proposal Relation. This module utilizes a compact multi-level triplet selection approach to get effective object representations by modeling the relationships of challenging propositions across several movies. Additionally, the authors propose the development of a Hierarchical Video Relation Network (HVR-Net) that incorporates both intra-video and inter-video proposal links in a hierarchical manner. The proposed methods leverage both intra and inter contexts in a progressive manner to enhance video object recognition.

S.Dasiopoulou et al [38] introduced a methodology for knowledge-enhanced semantic video object recognition, using a multimedia ontology framework. In the investigated domain, semantic ideas are specified inside an ontology that is enhanced with qualitative properties, such as color homogeneity. Additionally, low-level aspects, such as the distribution of color model components, are included. Furthermore, object spatial relations and multimedia processing techniques, such as color clustering, are also considered. Semantic Web technologies are used for the purpose of knowledge representation inside the Resource Description Framework RDF(S) metadata

standard. The purpose of this application is to identify video objects that align with the semantic concepts outlined in the ontology.

Long Fan et al [39] presented a technique for packet video processing. In the proposed approach, the video frames are first organized into groups. Within each group, all frames share a common optical flow feature map via the process of feature fusion. The generation of the Target Image involves enhancing object information by combining the shared feature map with the current frame. In order to allow the object detection network to concentrate more on the foreground object, the suggested strategy results in efficient background information masking.

Zhengkai Jiang et al [40] suggested to use a unique module called Learnable Spatio-Temporal Sampling (LSTS) to precisely learn semantic-level correspondences between neighbouring frame attributes. The sampled sites are first started randomly, updated repeatedly to identify improved spatial correspondences, and then gradually supervised detection is used to steer this process. Additionally, the modules for Dense Feature Aggregation (DFA) and Sparsely Recursive Feature Updating (SRFU) are provided to describe temporal relations and improve per-frame features, respectively.

Zhujun Xu et al [41] introduced a technique known as CenterNet that is based on a detector with a single stage. In order to improve the outcomes of the next picture, the authors start by propagating the prior reliable long-term detection in the form of a heatmap.

Lu Shengyu et al [42] presented a proposed technique for real-time object recognition in films, using the YOLO network architecture. The effect of the picture backdrop is mitigated by image pre-processing techniques. Subsequently, the Fast YOLO model is trained for the purpose of object recognition, enabling the extraction of pertinent object information. The YOLO network is enhanced by using a modified convolution operation, inspired by the Google Inception Net (GoogLeNet) architecture. This modification involves substituting the original convolution operation with a smaller convolution operation. This substitution effectively reduces the number of parameters and significantly accelerates the process of object identification.

Chun-Han Yao et al [43] suggested the implementation of adaptive policies via the use of reinforcement learning techniques in conjunction with basic heuristics. The proposed framework demonstrates superior performance compared to existing methods on the Imagenet VID 2015 dataset, while also achieving real-time execution on a CPU.

Zhenxun Yuan et al [44] presented a novel deep learning architecture, referred to as the motion-aid feature calibration network (MFCN), designed specifically for video object recognition. The primary concept is capitalizing on the temporal coherence of video features, taking into account their motion patterns as recorded by optical flow. This approach aims to enhance the system's ability to handle fluctuations in appearance and create more resilience. Efficient and adaptive aggregation and calibration processes are performed via an integrated optical flow network. In contrast to multi-stage approaches for video object recognition, the proposed method has an end-to-end design, resulting in significant improvements in both training and inference efficiency.

Yujie Wu et al [45] proposed an end-to-end blur-aid feature aggregation network (BFAN) for the purpose of video object recognition. The proposed BFAN primarily 2ddresses the aggregation process, which is affected by many types of blur, such as motion blur and defocus. The suggested method achieves a high level of accuracy while minimizing the additional computational burden.

Sudan Jha et al [46] offered an approach named N-YOLO, which aims to reduce the computational burden of object detection and tracking in the YOLO algorithm. Instead of resizing the image as a preprocessing step, N-YOLO divides the image into fixed-size sub-images, which are then used in the YOLO algorithm. The detection results of each sub-image are subsequently merged with the inference results obtained at different time intervals using a correlation-based tracking algorithm. By adopting this approach, the computational load for object detection and tracking can be substantially decreased.

Badri Narayan Subudhi et al [47] proposed the implementation of a novel heuristic initialization strategy for modifying information-based systems. The technique

necessitates the use of a frame that has been first partitioned. The initial frame segmentation employs a compound Markov Random Field (MRF) model to represent characteristics, and the Maximum a Posteriori (MAP) estimate is derived using a hybrid approach that combines both simulated annealing (SA) and iterative conditional mode (ICM) methods, resulting in rapid convergence. In the context of temporal segmentation, the proposal suggests using a change detection mask (CDM) based on the difference in labels between two frames, as opposed to the conventional approach of employing a gray level difference.

Wenguan Wang et al [48] presented a novel deep learning model that aims to effectively identify prominent areas within video content. The deep video saliency network presented in this study has two distinct modules, each responsible for collecting certain aspects of saliency information, namely spatial and temporal components. The dynamic saliency model incorporates saliency estimates from the static saliency model to provide spatiotemporal saliency inference without the need for computationally expensive optical flow calculation. In addition, the authors provide an innovative approach to data augmentation that emulates video training data using pre-existing annotated picture datasets. This approach enhances the ability of the network to acquire a wide range of saliency information and mitigates the risk of overfitting when working with a restricted number of training movies.

Garrick Brazil et al [49] presented a unique approach for monocular video-based 3D object recognition that utilizes kinematic motion to extract scene dynamics and enhance localization precision. In this study, the authors provide an innovative approach to decompose object orientation and introduce a self-balancing 3D confidence measure. It is shown that the inclusion of both components is essential in facilitating the optimal functioning of the kinematic model.

Hughes Perreault et al [50] introduced the Feature Fusion Architecture for Video Object Detection (FFAVOD). In this study, the authors provide a new architecture for video object recognition that enables the sharing of feature maps across adjacent frames inside a network. Furthermore, they offer a novel feature fusion module that is designed to effectively fuse feature maps in order to boost their quality. In this study,

they demonstrate that the use of the suggested architecture and fusion module yields enhanced performance outcomes for three base object detectors. These improvements were seen across two object identification benchmarks that included sequences of moving road users.

Liang Han et al [51] suggested a number of enhancements over prior research efforts. The proposed module is designed to aggregate pixel-level features in a class-aware manner. This module leverages contextual information from instances in both the current frame and other frames to describe each pixel. In contrast to the previous non-local operation, the proposed method of class-aware pixel-level feature aggregation selectively eliminates noisy information originating from the extensive background and objects belonging to various classes. Instead, it focuses on enhancing the representation of foreground pixels that pertain to the same class instances, while minimizing the inclusion of ambiguous information. Additionally, the approach incorporates a class-aware instance-level feature aggregation module, which effectively combines features from object proposals. The inclusion of the homogeneity constraint in the process of instance-level feature aggregation serves to eliminate numerous flawed proposals, thereby enhancing the accuracy of the feature aggregation. Additionally, the instance-level feature aggregation incorporates a module for feature alignment that is based on correlation, which aligns the feature maps of both the support and target proposals.

Ye Lyu et al [52] proposed the enhancement of a proficiently trained image object detector by using a class-agnostic convolutional regression tracker, aiming to improve the efficiency and effectiveness of video object recognition. The tracker acquires the ability to track objects by leveraging the characteristics extracted from the image object detector. This approach, known as a lightweight extension to the detector, incurs only a little decrease in processing performance when applied to the job of video object identification.

Mehran Yazdi et al [53] provided a comprehensive assessment on contemporary approaches for detecting moving objects in video sequences that are acquired by a camera in motion. Numerous studies and notable scholarly contributions have

examined the techniques used in object recognition and background removal for stationary cameras. However, a comprehensive overview of the various available approaches specifically tailored for scenarios involving a moving camera has yet to be presented. The approaches used in this particular domain may be categorized into four distinct groups: modeling-based background subtraction, trajectory classification, low rank and sparse matrix decomposition, and object tracking. In this discussion, the authors thoroughly examine each category and explain the primary ways that have been presented to enhance the overall notion of the techniques.

Junbo Yin et al [54] suggested the use of temporal information in numerous frames, namely point cloud films, to recognize 3D objects. The temporal information is classified into short-term and long-term patterns by empirical methods. In order to encapsulate the short-term data, the authors propose the use of a Grid Message Passing Network (GMPNet). In order to enhance the consolidation of long-term frames, they introduce a novel model called the Attentive Spatiotemporal Transformer GRU (AST-GRU). It has two components Spatial Transformer Attention (STA) to focus on small objects and Temporal Transformer Attention (TTA) to deal with motion misalignment. The model has increased complexity and evaluates solely on nuScenes, which may limit generalization claims.

Sucheng Ren et al [55] presented a novel methodology, referred to as the Triple Excitation Network (TENet), which aims to enhance the training process of video salient object detection (VSOD) by including spatial, temporal, and online excitations. The excitation mechanisms used in this study are developed in accordance with the principles of curriculum learning. Their objective is to minimize uncertainties in the early stages of training by selecting stimulating feature activations utilizing ground truth information.The use of spatial and temporal excitations has the potential to address the saliency shifting issue and reconcile the conflicting spatial and temporal characteristics seen in Video Saliency Object Detection. In addition, the semi-curriculum learning architecture facilitates the implementation of the first online refining technique for VSOD. This strategy permits the generation of stimulating and enhancing saliency responses throughout the testing phase without the need for re-training. However, it has increased model complexity due to triple excitation

components, which may lead to longer inference times and higher computational costs in resource-constrained settings.

Omar Oreifej et al [56] presented a unique technique for three-term low-rank matrix decomposition, aiming to breakdown the turbulence sequence into three distinct components: the backdrop, the turbulence, and the object. The complex issue at hand is simplified by transforming it into a minimization problem including the nuclear norm, Frobenius norm, and 21 norm. The methodology used in the study is founded upon two key observations: Initially, the turbulence generates dense and Gaussian noise, which may be effectively quantified using the Frobenius norm. Conversely, the moving objects exhibit sparsity, making them amenable to characterization using the 21 norm. Furthermore, due to the distinct nature of the object's linear motion compared to the Gaussian-like turbulence, it is possible to use a turbulence model based on Gaussian distribution to impose a supplementary restriction on the search area of the minimization process.

Mingju Chen et al [57] proposed a unique approach for selecting critical frames by integrating object identification and picture quality assessment techniques. In this study, the authors first use an object detector to identify various objects, including pedestrians and cars. Subsequently, a quality score will be applied to each training frame, with frames containing objects being granted higher quality scores. Subsequently, they use the CNN known as AlexNet architecture to extract deep features for representation purposes.

Zhigang Tu et al [58] introduced a unique spatiotemporal saliency model for the purpose of object recognition in movies. In contrast to prior methodologies that prioritize the use or integration of diverse saliency signals, the suggested approach seeks to leverage object signatures that may be discerned by any kind of object segmentation techniques. The authors combine two separate saliency maps, each derived from object proposals generated by an appearance-based approach and a motion-based algorithm, in order to get an enhanced spatiotemporal saliency map. This methodology facilitates the attainment of high levels of robustness and accuracy in the identification of significant items inside movies, even when faced with diverse

and difficult settings. However, a key limitation is that its performance severely depends on the superiority of first object proposals, which may not simplify well across all video types, mainly those with delicate motion or low contrast. Moreover, the sequential processing of two saliency maps can be computationally exhaustive, limiting real-time application

Runmin Cong et al [59] examined the significance of appearance modality and motion modality. Additionally, it proposes a novel VSOD network called PSNet, which incorporates an up and down parallel symmetry architecture. The proposed approach involves the use of two parallel branches that exhibit distinct dominating modalities. These branches work collaboratively to accomplish comprehensive decoding of video saliency. This is achieved via the integration of two modules: The Gather Diffusion Reinforcement (GDR) module and the Cross-modality Refinement and Complement (CRC) module. Ultimately, the Importance Perception Fusion (IPF) module is used to integrate the characteristics derived from two parallel branches based on their varying significance in distinct situations.

Senthil Murugan et al [60] presented an examination of the many methodologies used in video summarizing, along with a comparative analysis of distinct strategies. Additionally, the literature includes many techniques for object identification, object categorization, and object tracking.

Recent reviews in image/video object recognition highlight its rapid progress caused by deep learning advances, with an increased emphasis on robustness, effectiveness, and real-time performance. Tatana et al. [61] investigate the issues of darkness, showing how image/video enhancement methods have a direct impact on detection accuracy. In parallel, Jia et al. [62] emphasized the use of Graph Neural Networks (GNNs) in object detection methods to improve spatial-temporal modelling, especially for complicated video sequences. Whereas Du et al. [63] push the limits of detection in limited-label conditions by presenting efficient methods that require few inputs. Furthermore, throughout their comprehensive study of context-aware object detection, Jamali et al. [64] make the case for models that take contextual and

environmental context into consideration in order to resolve differences in object identity and location.

2.3 VIDEO OBJECT TRACKING

Object tracking is a vital part of computer vision systems which includes continuous positioning of objects across frames of video. This literature review examines the development of object tracking methods. The study also addresses on going issues such as occlusion mitigation and real-time performance.

Gioele Ciaparrone et al [65] offered an extensive examination of literature pertaining to the use of Deep Learning models for addressing the challenge of Multiple Object Tracking (MOT) in single-camera video scenarios. This study identifies four primary stages of MOT algorithms and provides a comprehensive analysis of the use of Deep Learning in each of these phases.

Manqi Zhao et al [66] introduced SatSOT, a novel benchmark dataset for single object-tracking in satellite videos, which is characterized by extensive annotations. The Satellite Surveillance of objects (SatSOT) dataset has a total of 105 sequences, including 27,664 frames. It encompasses 11 distinct features and encompasses four distinct kinds of common moving objects seen in satellite movies, namely cars, planes, ships, and trains. Given the aforementioned dataset and the notable difficulties encountered in the domain of satellite video object tracking, including the presence of diminutive objects, interference from the backdrop, and instances of substantial occlusion, it becomes crucial to design robust tracking algorithms capable of handling such complexities

Heng Fan et al [67] introduced LaSOT, a comprehensive benchmark for Large-scale Single Object Tracking. The LaSOT dataset has a comprehensive assortment of 85 distinct object classes, including a total of 1550 videos, which together include over 3.87 million frames. The process of annotating each video frame involves meticulous and deliberate placement of a bounding box. To the best of the knowledge, LaSOT stands as the most extensive tracking benchmark with richly annotated data. The primary objective behind the release of LaSOT is to provide a specialized and

superior platform that facilitates the training and assessment of trackers. The mean duration of videos in the LaSOT dataset is around 2500 frames. Each video in the dataset encompasses a range of challenging characteristics often seen in real-world video recordings, including instances when the targets intermittently vanish and reappear. The extended durations of these videos facilitate the evaluation of trackers over extended periods of time.

Happiness Ugochi Dike et al [68] introduced a novel methodology for object tracking using a deep learning approach, which achieves exceptional performance on well recognized datasets, including the Stanford Drone dataset and the Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking (UAVDT) dataset. The suggested framework for fast RCNN, which is a region-based convolutional neural network, was improved by the integration of many activities.These activities included the calibration of key parameters, multi-scale training, hard negative mining, and feature collection. The objective of these enhancements was to boost the performance of the region-based CNN baseline. Moreover, the researchers used a deep quadruplet network (DQN) to monitor the motion of the detected objects within the densely populated setting. The DQN was designed to include a novel quadruplet loss function, enabling an examination of the feature space. The fast RCNN used a CNN with Rectified Linear Units (ReLU) to extract spatial-spectral characteristics at depth of 6.

A. Ancy Micheal et al [69] presented a unique framework that utilizes deep learning techniques to provide accurate tracking of several objects in Unmanned Aerial Vehicle (UAV) recordings. The Tiny-Deeply Supervised Object Detector (Tiny-DSOD) is used for precise object detection. This study introduces a unique tracking method called the Stacked Bidirectional-Forward Long Short-Term Memory (SBF-LSTM) tracker, which incorporates both spatial and visual data for object tracking. The spatial and visual characteristics derived by Tiny-DSOD are trained in conjunction with the tracker, which is responsible for predicting the position of objects throughout the tracking process. The use of SBF-LSTM as the tracking mechanism enables precise anticipation of the object's spatial coordinates. The process of object association is determined by factors such as the distance between bounding boxes, visual appearance, and size measurements.

Tobias Fischer et al [70] offered Quasi-Dense Similarity Learning, which intensively samples hundreds of object areas on a pair of photos for contrastive learning. In this study, the authors integrate similarity learning with several pre-existing object detectors to develop a novel tracking method known as Quasi-Dense Tracking (QDTrack). Notably, QDTrack eliminates the need for displacement regression or motion priors. It is observed that the resultant feature space has the capability to facilitate a straightforward nearest neighbor search during the inference stage for the purpose of object association.

Di Wu et al [71] presented an enhanced version of the Kernel Correlation Filter (KCF) algorithm that incorporates road information to effectively track tiny objects, particularly in scenarios when the item may be partially obscured. This paper presents a detailed account of the specific contributions it makes, which are outlined as follows: The first step involves the reconstruction of the tracking confidence module. This reconstruction incorporates both the peak response and the average peak correlation energy of the response map. The purpose of this integration is to enhance the accuracy of determining if the object is occluded. Subsequently, an adaptive Kalman filter is formulated with the purpose of dynamically modifying the parameters of the Kalman filter in response to the object's motion state. This adaptation enhances the resilience of the tracking process and mitigates the occurrence of tracking drift subsequent to object occlusion.

Qiao Liu et al [72] presented a multi-level similarity model inside a Siamese framework to enhance the robustness of thermal infrared (TIR) object tracking. The computation of various pattern similarities is carried out utilizing the multi-level similarity network that has been developed. One of the approaches concentrates on the overall semantic similarity at a global level, while the other approach calculates the local structural similarity of the TIR item. The commonalities exhibit a mutually reinforcing relationship, hence augmenting the network's ability to effectively process distractors. However, the model's complexity and reliance on a large, domain-specific dataset may limit its deployment in low-resource environments or domains with limited TIR data.

Xuan Wang et al [73] provided an empirical examination of the unforeseen result of the TFS strategy. Additionally, the authors provide general guidelines that may be used to exploit this characteristic, hence using TFS to improve the effectiveness and precision of VOT.

Shaheena Noor et al [74] provided an empirical examination of the unforeseen result of the TFS strategy. Additionally, the authors provide general guidelines that may be used to exploit this characteristic, hence using TFS to improve the effectiveness and precision of VOT.

Runlong Xia et al [75] improved the KCF algorithm. The occlusion condition has been included into the KCF method. The utilization of the KCF technique for object tracking is contingent upon the absence of occlusion. The use of the Unscented Rauch-Tung-Striebel Smoother method has been employed in cases when occlusion is present. Additionally, the anticipated location of the item has been provided as input to the KCF method.

Yifu Zhang et al [76] proposed a straightforward, efficient, and versatile association technique that involves tracking by associating almost all detection boxes rather than just focusing on those with high scores. The low score detection boxes are used in order to leverage their resemblances with tracklets, hence facilitating the identification of genuine objects and the elimination of background detections. The strategy consistently improves the IDF1 score by 1 to 10 points when applied to 9 different state-of-the-art trackers. In order to present the current leading performance of MOT the authors have developed a robust and straightforward tracker called ByteTrack. In this study, they report the attainment of 80.3 MOTA, 77.3 IDF1, and 63.1 HOTA on the MOT17 test set. These results were obtained using a single V100 GPU with a running speed of 30 frames per second.

Muxi Jiang et al [77] created a long-term tracking method based on collaborative filtering (CF). The primary approaches are outlined as follows. This study introduces a unique confidence score for assessing the reliability of tracking algorithms. Additionally, a method is presented to rectify tracking drift and maintain the long-term appearance of the target. Moreover, the multi-scale search may be used to

reposition the target if it is no longer visible. The tracker demonstrates superior performance compared to previous collaborative filtering (CF)-based trackers, particularly in terms of its high engineering applicability.

Yuantao Chen et al [78] introduced a novel visual object tracking technique that utilizes the Adaptive Combination Kernel. The process of object tracking has been divided into two distinct subtasks, namely the Translation Filter and the Scale Filter, which are used to estimate the specific characteristics of the item. The Translation Kernel Tracker has used an adaptive mix of the Linear Kernel Filter and the Gaussian Kernel Filter. The goal function has been formulated to derive the weight coefficients for the Linear Kernel filter and the Gaussian Kernel filter. This function takes into account both the empirical risk and the maximum value of the response output for each kernel. The Adaptive Combination Kernel has the benefits inherent in both local kernels and global kernels. Furthermore, the determination of the tracking position has been derived based on the response output of the adaptive combination kernel correlation filter. Furthermore, the translation filter has included a scene-adaptive learning rate based on the maximum response value. The translation filter has the capability to be enhanced by the incorporation of an adjustable learning rate. The item scale has been estimated using a one-dimensional scale filter.

Adam W. Harley et al [79] examined the methodology proposed by Sand and Teller, referred to as the "particle video" approach. Specifically, the focus will be on investigating pixel tracking as a problem of long-range motion estimation. This entails the characterization of each pixel via a trajectory that enables its identification over several future frames. The original strategy is reconstructed in this study by including components that are at the forefront of flow and object tracking research. These components include dense cost maps, iterative optimization techniques, and learning appearance updates.

Ruixu Wu et al [80] introduced a novel approach for object tracking, referred to as the Deep Adaptive Spatio-Temporal Feature Transformer Object Tracking (DASFTOT) technique. This method incorporates a backbone network, a transformer mechanism, and a bounding prediction box. Initially, a three-dimensional convolutional neural

network (3D CNN) is used to extract motion-related features. Furthermore, the authors use a dual attention spatiotemporal fused transformer (DASFT) to overlay significant temporal and geographical data. This allows us to merge local and global spatiotemporal characteristics and evaluate the association between templates and search areas. Furthermore, to enhance the resilience of the tracking process, they implement a dynamic updating mechanism for a specific portion of the template frame. Ultimately, the tracking object is situated by use of a bounding prediction box.

Xiankai Lu et al [81] Utilized adaptive area suggestions, which are successful in rescuing target objects from tracking errors brought on by severe occlusion or out-of-view movement. The authors carry out target re-detection across adaptively learned area suggestions, in contrast to conventional tracking-by-detection approaches employing random samples or sliding windows. They demonstrate that the provided adaptive area suggestions can handle the difficult scale estimation issue as well since they naturally take objectness information into consideration. Also noted are the redundant channels and chaotic feature representation, particularly for the convolutional features. As a result, they channel regularize the learning of the correlation filter.

Zhenbo Xu et al [82] proposed a novel approach for acquiring instance embedding's via segment-based learning. The proposed technique involves turning the condensed picture representation into an unordered 2D point cloud representation, resulting in a very efficient learning process. The proposed approach introduces a novel tracking-by-points framework, whereby discriminative instance embedding's are acquired during the learning process using randomly sampled points, as opposed to using whole pictures. In addition, several relevant data modalities are transformed into point-wise representations in order to enhance the characteristics at each individual point.

Jianming Zhang et al [83] provided a novel tracking framework that integrates correlation filter tracking with Siamese-based object tracking methodologies. Initially, the integration of deep features and handmade features is used in correlation filter learning. In this study, the authors provide a novel and robust criterion for assessing

the robustness of tracking outcomes. This criterion serves as a determining factor for initiating the Siamese tracking model. If the result of the robustness assessment is below the adaptive threshold, the Siamese tracking method is initiated for the purpose of object recapture.

Shiyu Xuan et al [84] focused to tackle the issue of rapid object tracking in satellite movies by the development of an innovative tracking algorithm. The proposed technique incorporates correlation filters that are integrated with motion estimates. The suggested approach offers enhancements based on the kernelized correlation filter (KCF). The present study aims to provide an innovative approach for motion estimation (ME) by integrating the Kalman filter and motion trajectory averaging techniques. Additionally, this approach seeks to address the issue of boundary effects associated with the KCF via the use of the ME algorithm. Furthermore, the study aims to provide a solution to the challenge of tracking failure that occurs when a moving object becomes partly or entirely obscured.

Bin Yan et al [85] introduced a novel approach, referred to as Unicorn, which offers a unified methodology to address four distinct tracking challenges concurrently. This technique leverages a single network architecture and shares the same set of model parameters across all four tracking issues. As a result of the lack of consensus on the precise description of the object tracking issue, many current trackers have been designed to tackle just one or a subset of activities, leading to an excessive focus on the unique features of those jobs. In comparison, the Unicorn framework offers a cohesive solution by using a consistent approach in terms of input, backbone architecture, embedding techniques, and output layers for all tracking jobs. The successful integration of the tracking network architecture and learning paradigm has been achieved for the first time.

Du Yong Kim et al [86] presented a novel online multi-object tracking system for picture observations. The program utilizes a top-down Bayesian formulation, which effectively combines state estimation, track management, and the handling of false positives, false negatives, and occlusion inside a single recursive framework. The objective is accomplished by representing the multi-object state as a labeled random

finite set and using the Bayes recursion to advance the multi-object filtering density over time. The proposed filter algorithm utilizes a combination of detection data and picture data to enhance both efficiency and accuracy. Specifically, when detection loss is detected, the algorithm seamlessly changes to using image data, therefore capitalizing on the strengths of both types of data.

Bo Du et al [87] presented a novel approach for object tracking in satellite movies by using a multiframe optical flow tracker. The integration of the Lucas-Kanade optical flow technique with the HSV color scheme and integral image was used to facilitate the tracking of objects in satellite movies. Additionally, the optical flow tracker utilized the multiframe difference approach to enhance the understanding of the results.

Xingyi Zhou et al [88] proposed an algorithm for simultaneous detection and tracking that exhibits enhanced simplicity, speed, and accuracy compared to the current state-of-the-art methods. The detection model used by the tracker, CenterTrack, is utilized to analyze a set of pictures alongside detections obtained from the preceding frame. Based on the limited data provided, the CenterTrack algorithm is used to perform object localization and anticipate the relationships of these items with the preceding frame.

Jefferson Keh et al [89] suggested the use of interactive video object tracking as a means of annotation. In response to the VOT short-term tracking challenge, the authors have devised an assessment methodology to assess the performance of three short-term trackers: KCF, Channel and Spatial Reliability Tracking(CSRT), and Accurate Tracking by Overlap Maximization (ATOM).

Deep learning developments have influenced recent advances in object tracking, with a special focus on improving accuracy and application area of expertise.Ren et al[90] explored inter-camera interactions in surveillance networks through the introduction of a multi-camera pedestrian tracking method based on differential imaging. Shajeena et al. [91] proposed a Siamese Deep Q-Learning algorithm paired with correlation filters for adaptive object tracking in challenging visual circumstances. Finally Zebarjadi et al. [92] coupled Kanade–Lucas–Tomasi (KLT) tracking with LSTM

networks to track organ movement in real-time ultrasound-guided therapies, which had an advantageous effect on medical applications.

2.4 SUMMARY

This chapter has given a comprehensive summary of previous research on the challenges of addressing object detection and tracking under varying illumination condition, with a focus on how different levels of light affect these problems. It has shown how image processing, machine learning, and deep learning have all helped to make problems with lighting less of an issue over time. This survey gives us a basic grip of the topic and shows how important it is to have robust illumination correction algorithms to improve detection and tracking accuracy. These results will instantaneously affect and shape the creation of the deep neural network-based methodologies suggested in this study. The next chapter goes into more detail on how to fix illumination problems, which is the first step toward being able to consistently detect and track objects in changing lighting circumstances.

CHAPTER 3
LOW LIGHT IMAGE ENHANCEMENT: AN INNOVATIVE APPROACH FOR UNEVEN ILLUMINATION CORRECTION

3.1 INTRODUCTION

The influence of lighting on image quality is a critical factor in the fields of visual perception and image processing. The difficulties presented by suboptimal lighting conditions and inconsistent illumination frequently lead to the production of photographs that suffer from diminished visibility, decreased level of detail, and a general absence of clarity. The domain of low light image enhancement pertains to a persistent problem that carries ramifications across various sectors and applications. The demand for solutions that can effectively address uneven illumination is evident in various contexts, such as surveillance and security, where accurate visual data is crucial for making critical decisions, as well as in photography and cinematography, where image quality plays a significant role in creative expression and storytelling.The main aim of this study is to devise novel methodologies that surpass the constraints associated with conventional picture enhancing techniques. Traditional approaches frequently prove inadequate when confronted with the complexities associated with the correction of uneven illumination in low-light environments. The conventional method of histogram equalization has the potential to excessively amplify noise, leading to visually unappealing and less informative images. Hence, this study endeavors to propose and execute innovative approaches that integrate the capabilities of deep learning and advanced image processing methodologies.

In addition to effectively tackling the technical obstacles presented by uneven illumination, this research exhibits the potential for revolutionary impact across diverse sectors. The potential implications of this discovery are extensive, encompassing various domains such as the enhancement of surveillance and security systems, the exploration of new artistic possibilities in low-light photography, and the improvement of diagnostic accuracy in medical imaging.

3.2 PROPOSED METHOD

Videos recorded in different lighting conditions often exhibit low contrast, inadequate brightness, muted colors, and elevated noise levels. To mitigate the impact of unwanted noise caused by irregular illumination and recover the visual feature of the video, we proposed an efficient illumination correction method. To achieve efficient correction of illumination effects in videos, the proposed method involves several steps. In the initial steps, the videos are divided into multiple frames, and each frame is checked to determine if it is illuminated or non-illuminated. If the frame is illuminated, the proposed technique is applied to correct the illumination. Finally, all the frames are arranged in sequence, as depicted in figure 3.1.

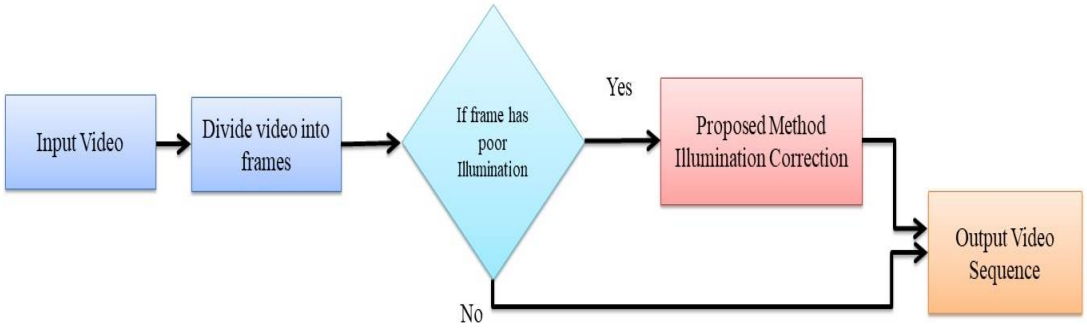


Figure 3. 1: Shows the general structure of Illumination correction

3.2.1 Dataset

To assess the proposed model using change detection dataset, this contains challenging weather, camera jitter, shadows, dynamic background, intermittent object motion, low framerate, acquisition at night. This dataset is available along with the ground truth images for object detection and tracking.

3.2.2 Proposed Illumination Correction Method

The main objective of the proposed illumination correction method is to recuperate the natural quality images from several images and video sequences acquired under varying illumination conditions.

The proposed algorithm delivers the required results by applying different processing ideas. The proposed method, involves the following steps:

(1) Performing exponential transformation on the input image,

(2) Computing the hyperbolic tangent function on the exponential transformed image,

(3) Calculating the LIP model for the two images obtained from the previous steps,

(4) Determining the logarithmic scaling function of the resultant image, and

(5) Multiplying the exponential transformed image pixel-wise with the output of the logarithmic scaling function. Each key operation is thoroughly described in the following sections with accompanying images.

Pseudocode of proposed method

Input: Low illuminated image (X)
Compute the exponential transformation of the image X to obtain y1.
Compute the hyperbolic tangent function of the y1as shown in equation 3 to obtain y2.
Use the adapted LIP model on y1 and y2 to obtain y3.
Compute the logarithmic scaling function of y3 to obtain y4.
Calculate the final output by performing the pixel-wise multiplication of the exponential transformed image (y2) and the logarithmically scaled image (y4).
Output: illumination corrected image y

3.2.3 Input Images

We illustrate the step-by-step process of the proposed method using frames from various video sequences, as depicted in figure 3.2.

(a)

(b)

(c)

(d)

(e)

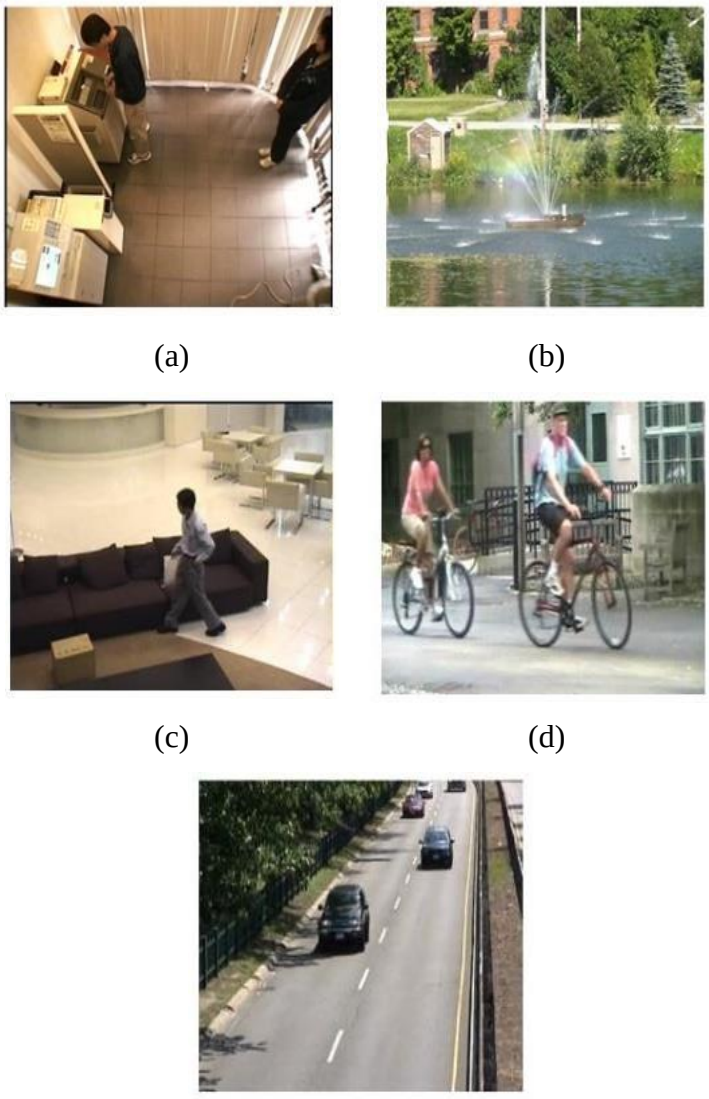
Figure 3. 2: shows the input images with different occlusions

3.2.4 Exponential Transformation of Image

The input image is the input exponential function which is represented as shown in the equation (3.1). $I_{input}(i\,j)$ represents specific pixel value. Here we use modified version of equation (3.1) which can be represented as in equation (3.2).

$$I_{Output} = \exp(I_{input}) \tag{3.1}$$

$$y_1 = c((1+\alpha)^X - 1 \tag{3.2}$$



(a) (b)

(c) (d)

(e)

Figure 3. 3: Shows the output of the exponential function

Equation (3.2), where y_1 represents the output image and '*X*' represents the input image The exponential function is represented using (1+α) where α takes the values in the range of [0.3 -0.5]. Where 'c' is a scaling factor which represents the output will lie in appropriate range. Here we have considered the scaling factor to be 4 and the 'α' as 0.3.

3.2.5 Hyperbolic Tangent Profile

The hyperbolic tangent function is applied to increase the brightness of the dim pixels, which can recover the pictorial quality of images taken in low light condition. To attain this, the output of the exponential transformed image y1 is given as input to the hyperbolic tangent, which is adopted as a nonlinear mapping as shown in equation (3.3). As a result of the applied function the low intensity pixels are amplified, leading to an increase in resulting output values, while high intensity pixels will become saturated. The below figure 3.4 shows the boosted output of equation 3.3.

$$y2 = \tanh\left(\frac{1-\exp(-2*y1)}{1+\exp(-2*y1)}\right) \quad (3.3)$$

(a)

(b)

(c)

(d)

(e)

Figure 3. 4: Shows the output of hyperbolic tangent

3.2.6 Logarithmic Image Processing (LIP)

To combine images y1 and y2 here appropriate LIP model is used. Various LIP models have been proposed to merge the features form two different images, but a specific model highlighted has shown significant potential in producing excellent outcomes across a wide range of low light images. The used LIP model can be calculated as follows in equation 3.4. The output of the LIP model for y1 and y2 are presented visually in the below figure 3.5.

$$y3 = \frac{(y1+y2)}{(1+y1*y2)} \tag{3.4}$$

(a)

(b)

(c)

(d)

(e)

Figure 3. 5: Shows the output of the LIP addition of two images

3.2.7 Logarithmic Scaling Function

For further processing the proposed image enhancement model uses logarithmic scaling function for the above output y3. 'y3' represents the output of the LIP model. This type of transformation enlarges of dim pixels, while simultaneously compressing the values of brighter pixels to prevent excessive increment. Figure 3.6 displays the visuals produced by the logarithmic scaling function of different images. *y*4 is the resultant image from the logarithmic scaling function [93].

$$y4 = \frac{\max(y3)}{\log(\max(y3+1))} * \log(y3 + 1) \quad (3.5)$$

(a)

(b)

(c)

(d)

(e)

Figure 3. 6: Enhanced dark pixels

To reduce the overall enhancement finally we do the pixel wise multiplication of the exponential transformed image y1 with the logarithmic scaled image y4, which reduces the overall brightness and retains the true color and contrast of the image. The illumination corrected image is represented using the equation (6). The illumination corrected images are shown below figure 3.7.

$$y = y1 * y4 \qquad (3.6)$$

(a)

(b)

(c)

(d)

(e)

Figure 3. 7: Shows final illumination corrected images

3.3 RESULTS AND DISCUSSION

The effectiveness of the proposed technique was assessed through extensive testing using Matlab R2021a on a PC equipped with an Intel(R) Core(TM) i5-6200U CPU @ 2.30 GHz and 8 GB of installed RAM. The results were illustrated and discussed by conducting several comparisons with other methods. Here we considered the change detection dataset (CDNET) [94] for the purpose of evaluation of matrices and to compare with the start-art-methods. For the purpose of comparison, five advanced techniques were selected, namely Contrast Limited Adaptive Histogram Equalization (CLAHE) [95], Low-light Image Enhancement with Bright Channel Prior (LIBCP) [96], Multi Frame Low-light image enhancement algorithm (MF-LIME) [97], Low-Light Camera Response Model (LECARM) [98], and Retinex-Based-Multiphase-Algorithm (RBMP) [99]. Quantitative measures are essential for objectively comparing the performance of various image enhancement algorithms. The results obtained from the assessments are evaluated using three advanced metrics: Lightness Order Error (LOE), PSNR, and Mean-squared error(IMMSE). The figure 3.8 displays the pictorial comparison between the proposed technique and prevailing state-of-the-art techniques. The input images are labeled from a1 to a5, while b1 to b5 represent the outputs processed by CLAHE. The outputs processed by LIBCP are labeled from c1 to c5, and the outputs generated by MF-LIME are labeled from d1 to d5. The results produced by LECARM are shown from e1 to e5, and the outputs of RBMP are labeled from f1 to f5. Finally, the proposed method outputs are represented from g1 to g5.

3.3.1 Lightness Order Error (LOE)

LOE is a metric used to evaluate the accuracy of color transformations in image processing. It measures the dissimilarity between the lightness values of the original image and the transformed image. The LOE value provides information about how well the colors in the transformed image maintain their relative lightness values compared to the original image. The LOE is defined as:

$$LOE = \frac{1}{m}\sum_{x=1}^{m} RD(x) \tag{3.7}$$

Where m represent the number of pixels count in the image. RD(x) represents the change of relative light order among the original image and the improved image at the x^{th} pixel.

The quality of an enhancement effect is indicated by a low value of LOE.A value of zero for LOE means that the algorithm has not made any improvement to the original image. The results from testing the proposed algorithm on baseline images showed that the LOE value was zero, demonstrating that the proposed method is efficient in correcting images with illumination issues. Figure 3.9 shows the average analytical graph of LOE where the proposed method has the least LOE value. Table 3.1 show that the proposed algorithm has the lowermost LOE among all the state-of-the-art methods, indicating that it is superior to the existing approaches.

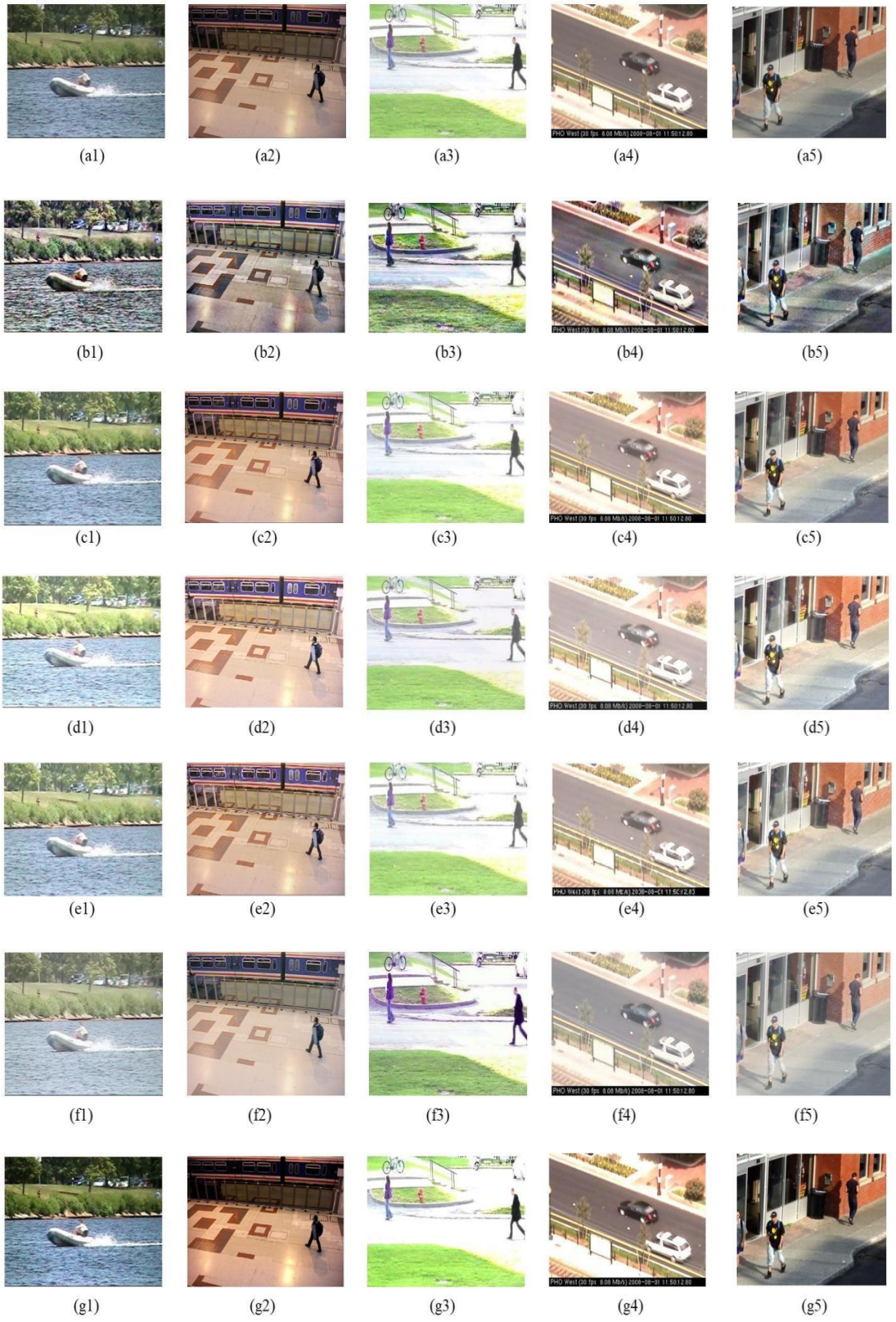


Figure 3. 8: Visual Comparison of proposed method with state-of-the-art techniques

. Table 3. 1: LOE metric score of Proposed with compared algorithms

METHOD	Img-1	Img-2	Img-3	Img-4	Img-5	Img-6	Img-7	Img-8	Img-9	Img-10	AVG
CLAHE[82]	452.9	577.2	570.3	468.8	519.7	0.2	463.5	573.2	334.8	134.26	410
BCP [83]	461.3	522.5	657.3	697.4	503.1	3	506.1	653.5	673.7	38.71	472
MF-LIME [84]	956.7	557.9	549.5	361.6	999.7	102.2	791.1	705.9	532.5	239.35	626
LECARM [85]	462.1	562	920.2	728.3	525	4	556.1	706.4	742.6	38.71	525
RBMP [86]	461.3	504.4	457.4	673.6	503.8	0.5	490.1	621.8	675.8	30.82	442
PROPOSED	**46.54**	**30.72**	**20.91**	**88.09**	**58.27**	**0.1**	**38.68**	**61.27**	**70.07**	**33.78**	**44.8**

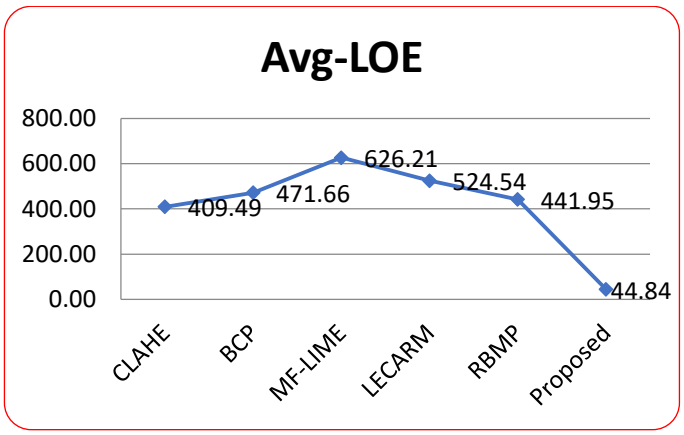


Figure 3.9: Analytical graph of average comparison of LOE

3.3.2 Peak Signal to Noise Ratio (PSNR)

The PSNR is a widely used metric for determining the difference in quality between two images. The PSNR block calculates this measurement, which is often used to assess the quality of a reconstructed or compressed image in comparison to the original. A higher PSNR score signifies a development in the superiority of the reconstructed image.

Table 3.2 presents a comparison of the proposed technique with state-of-the-art systems, and it can be seen that the reconstructed images produced by the proposed technique exhibit a high PSNR. This highlights the exceptional performance of the proposed algorithm in enhancing the superiority of the reconstructed images. Furthermore, the average PSNR values, as shown in Figure 3.10, when related to the

state-of-the-art methods, demonstrate that the proposed algorithm surpasses these methods with outstanding results.

Table 3.2: PSNR metric score of Proposed with compared algorithms

METHOD	Img-1	Img-2	Img-3	Img-4	Img-5	Img-6	Img-7	Img-8	Img-9	Img-10	AVG
CLAHE [82]	15.96	15.93	19.17	18.15	16.39	16.33	17.17	**19.87**	18.19	**18.997**	17.6
BCP [83]	17.27	17.25	16.59	16.91	16.61	26.34	16.46	16.66	15.98	18.94	17.9
MF-LIME [84]	10.57	11.34	11.49	12.15	11.1	**26.46**	10.72	12.57	11.81	14.163	13.2
LECARM [85]	13.48	13.84	13.45	14.68	12.47	25.1	13.25	14.16	13.25	17.111	15.1
RBMP [86]	13.87	14.71	15.16	14.55	13.7	25.97	14.34	16.83	13.77	18.427	16.1
PROPOSED	**23.15**	**23.25**	**21.77**	**21.53**	**24.33**	15.47	**22.38**	18.07	**23.06**	18.53	**21.2**

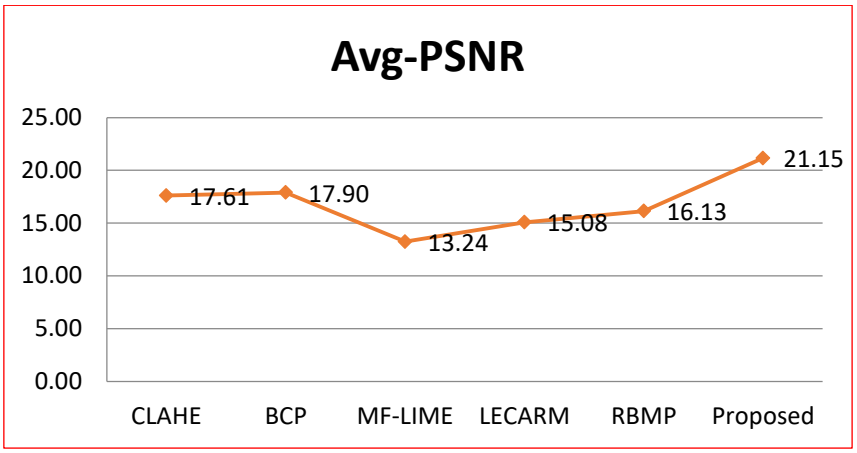


Figure 3.10: Analytical graph of average comparison of PSNR

3.3.3 Mean Squared Error

Mean Squared Error (MSE) is a popular error metric in image processing and computer vision that is often utilized as a loss function in various image processing techniques to optimize the quality of reconstructed or compressed images. It calculates the average difference between the original image and the reconstructed or compressed image by summing the squared differences between each corresponding pixel in the two images and dividing the result by the total number of pixels in the image. The final output is a scalar value that represents the overall quality of the reconstructed or compressed image, with a lower MSE indicating a higher quality, meaning that the differences between the original and reconstructed images are smaller.

The results of comparing the MSE scores of the proposed algorithm with five advanced techniques are presented in Table 3.3. These outcomes reveal that the images reconstructed using the proposed method exhibit a low MSE and reveal extraordinary performance in refining the quality of the reconstructed images. Additionally, when comparing the average MSE values, as shown in Figure 3.11, to state-of-the-art methods, it is clear that the proposed algorithm outperforms these techniques with remarkable results.

Table 3. 3: MSE metric score of Proposed with compared Algorithms

METHOD	Img-1	Img-2	Img-3	Img-4	Img-5	Img-6	Img-7	Img-8	Img-9	Img-10	AVG
CLAHE[82]	0.025	0.026	0.012	0.015	0.023	0.023	0.019	**0.01**	0.015	**0.013**	0.02
BCP[83]	0.019	0.019	0.022	0.02	0.022	**0.002**	0.023	0.022	0.025	**0.013**	0.02
MF-IME [84]	57.87	47.85	46.45	39.16	50.72	14.94	55.97	35.09	42.6	24.275	41.5
LECARM [85]	0.045	0.041	0.045	0.034	0.057	0.003	0.047	0.038	0.047	0.019	0.04
RBMP [86]	0.041	0.034	0.03	0.035	0.043	0.003	0.037	0.021	0.042	0.014	**0.03**
PROPOSED	**0.005**	**0.005**	**0.007**	**0.007**	**0.004**	0.028	**0.006**	0.016	**0.005**	0.014	**0.01**

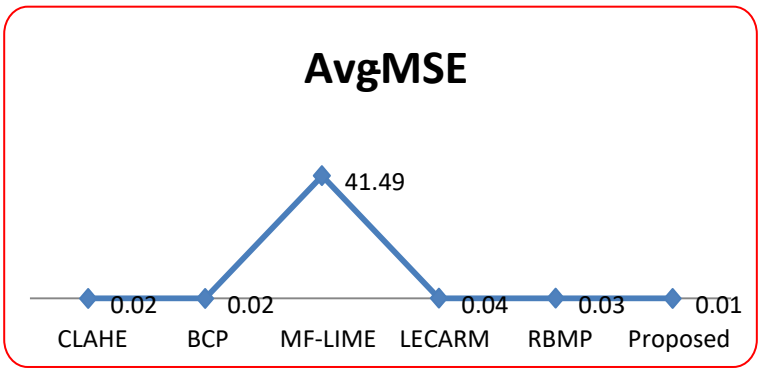


Figure 3. 11: Analytical graph of average comparison of MSE

3.3.4 Average Implementation Time

The evaluation of the time of implementation of the proposed method with other techniques is a crucial factor to consider in evaluating its efficiency and practicality. A shorter time of implementation makes the proposed algorithm a more viable option for real-time applications. As shown in Table 3.4, the comparison results specify that the proposed algorithm has a faster implementation time compared to state-of-the-art

techniques. Additionally, the LECAM and RBMP methods have similar implementation times to the proposed algorithm. However, the BCP method has the longest implementation time, while the CLAHE and MF-LIME methods have moderate implementation times. These findings, as shown in Figure 3.12, demonstrate that the proposed algorithm offers superior computational efficiency compared to state-of-the-art methods. This makes it a suitable candidate for real-world applications where efficiency is critical.

Table 3.4: Implementation time of Proposed with compared Algorithms

METHOD	Img-1	Img-2	Img-3	Img-4	Img-5	Img-6	Img-7	Img-8	Img-9	Img-10	AVG
CLAHE [82]	2.583	1.199	2.612	2.532	2.498	2.53	1.015	2.405	1.047	0.9498	1.937
BCP [83]	3.944	11.34	2.993	10.64	3.715	2.961	3.641	3.504	3.713	3.8713	5.033
MF-LIME [84]	1.059	1.798	1.256	1.453	1.402	0.852	0.983	0.911	0.838	0.8977	1.145
LECARM [85]	0.496	0.79	**0.34**	0.552	**0.247**	**0.284**	**0.277**	0.378	0.291	**0.2933**	0.395
RBMP[86]	0.073	**0.089**	0.417	0.317	0.683	0.487	0.604	0.375	0.592	0.0775	0.372
PROPOSED	**0.064**	0.579	0.366	**0.261**	0.352	0.379	0.365	**0.376**	**0.076**	0.3446	**0.316**

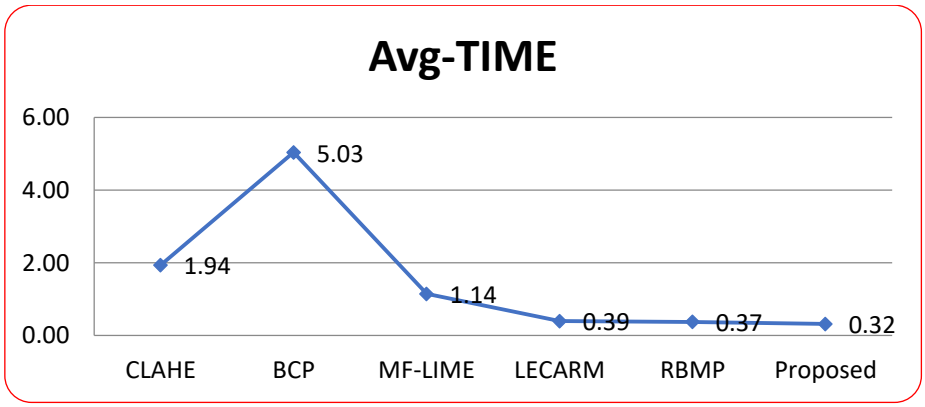


Figure 3.12: Analytical graph of average implementation time

3.4 SUMMARY

In this chapter a new illumination correction algorithm for enhancing low light or non-uniform illuminated images has been presented, which is characterized by its simplicity and ease of implementation. The algorithm uses a combination of exponential and hyperbolic tangent functions, an adapted LIP method, and a logarithmic scaling function to achieve improved brightness, contrast, and color preservation. Upon testing the algorithm against five other specialized algorithms, the

proposed method was found to perform better in terms of LOE, MSE, PSNR, and processing time. Moreover, the images produced using the proposed algorithm were found to have a better perceptual quality, featuring clearer edges and brighter details in low-energy regions without sacrificing information in high-energy areas. The results demonstrate the exceptional performance and efficiency of the proposed system, making it a strong contender for real-world applications where image enhancement is a critical factor. Whereas chapter 4 focuses on object detection which serves as a crucial step for object tracking. Accurate detection of objects within the image or video frame forms the foundation for reliable tracking. Following the illumination correction technique discussed in this chapter, the next chapter investigates into the methodology used to identify and localize objects.

CHAPTER 4
VIDEO OBJECT DETECTION USING Fast RCNN WITH ResNet MODEL

4.1 INTRODUCTION

The demand for rapid and reliable video analysis has experienced significant growth in a world that is progressively digital and visual. Video object identification plays a crucial role in the field of computer vision, encompassing the task of detecting and tracking items of interest inside video streams. This technology facilitates the comprehension and interpretation of visual data by machines, rendering it highly valuable in many domains such as surveillance, driverless cars, video content analysis, and augmented reality.

Video object detection expands upon the functionalities of conventional image object detection by enabling systems to analyze video sequences in real-time or almost real-time. This technology enables the identification, localization, and tracking of objects inside individual frames of a movie. The utilization of deep learning, specifically CNNs, is employed to extract spatial and temporal data from video footage. This enables the recognition of objects, monitoring of their motions, and even prediction of their future placements. The aforementioned technology possesses the capacity to revolutionize various sectors and augment multiple facets of everyday existence.

This introduction aims to explore the fundamental concepts and issues associated with video object identification, along with its diverse range of applications and the latest advancements in this subject. By delving into the fundamental principles of video object detection, we may enhance our comprehension of how this technological advancement is revolutionizing our interactions with and utilization of the extensive quantities of visual data present in our environment.

The recognition of objects in videos holds significant significance across multiple domains, particularly in the fields of computer vision and video analysis. This technological advancement facilitates the instantaneous recognition and monitoring of objects within video feeds, presenting a wide range of potential applications.

Surveillance and security represent a prominent domain wherein the utilization of video object detection technology becomes important in the monitoring of public places, discerning potentially dubious behaviors, and augmenting the overall level of safety. In the context of the automotive sector, the ability of autonomous vehicles to accurately identify and monitor things present on the road is of utmost importance. This capability plays a vital role in safeguarding the well-being of both occupants of the vehicle and individuals traversing the roadways. Moreover, video object detection has various uses in the field of healthcare, as it may be utilized to aid in the surveillance of patients, the tracking of surgical tools, and the analysis of medical imaging. Video object detection is a crucial component in improving the effectiveness and security of various industries through the provision of real-time analysis of the dynamic environment depicted in video recordings.

Moreover, video object detection has become increasingly prominent in the domains of entertainment and advertising. This technology facilitates the creation of interactive and immersive experiences, such as augmented reality (AR) and virtual reality (VR) apps, by the flawless tracking and integration of digital items inside the physical environment. In the field of marketing and advertising, the utilization of object and scene recognition technology facilitates the implementation of targeted advertising strategies. This technology enables advertisers to identify and interpret items and scenes in real-time, thereby empowering them to send personalized content to individual consumers. Video object detection plays a crucial role in the field of sports analytics by aiding coaches and teams in the monitoring of player movements, enhancing strategic decision-making, and offering valuable data for performance analysis. In brief, video object detection is a multifaceted technology that exhibits extensive utility across a range of disciplines, encompassing security, healthcare, entertainment, marketing, and sports analytics. Its widespread implementation fosters innovation and facilitates enhanced outcomes in diverse fields.

4.2 PROPOSED MODEL

4.2.1 Proposed Object Detection Model

R-CNN utilizes a method known as Selective Search to generate approximately 2000 region suggestions for every image. These proposals are subsequently fed into the underlying network architecture. As a result, this implies that 2000 forward passes would be executed on a single image. Now, let's imagine training the network with a dataset consisting of one thousand images.

The Fast R-CNN model, which incorporates the ResNet-152 architecture, is a prominent deep learning approach employed in the field of computer vision for the purpose of object identification. Object detection refers to the process of recognizing and precisely localizing various objects present in an image. Fast R-CNN represents a notable advancement in comparison to its precursor, R-CNN, primarily attributed to its enhanced speed and efficiency. This improvement is mostly attributed to the replacement of selective search for region proposal, which was a contributing factor to the slower and less efficient performance of R-CNN. The Fast R-CNN algorithm incorporates a number of significant advancements aimed at enhancing the speed and precision of object detection. On the contrary, ResNet-152 is a highly potent deep neural network renowned for its outstanding efficacy in many computer vision applications.

In this analysis, we will deconstruct the Fast R-CNN architecture, incorporating the ResNet-152 model, and elucidate its distinct modules, providing a comprehensive explanation for each.

- Input Image: -

The model receives a color image as input, typically in the RGB format, with no specific size requirements. The Fast R-CNN algorithm exhibits the ability to accommodate varying image dimensions, hence eliminating the need for a constant input size.

- ResNet-152 backbone: -

The ResNet-152 backbone is utilized. The ResNet-152 is utilized as the primary feature extraction framework in the Fast R-CNN. The deep convolutional neural network (DCNN) design under discussion is renowned for its use of residual connections, which effectively address the challenge of training extremely deep networks by minimizing the issue of vanishing gradients. ResNet-152 is composed of a series of residual blocks, which are constructed using convolutional layers, batch normalization, and non-linear activation functions such as ReLU.

- Region Proposal Network (RPN): -

The RPN is a key component in object detection systems. The inclusion of the RPN is a fundamental component inside the Fast R-CNN framework. The operation is performed on the feature maps that have been extracted by the ResNet-152 backbone. The RPN produces region proposals, which are bounding boxes that have a high probability of encompassing items that are of significance. The ideas are sorted based on their objectness scores, and the highest-ranking candidates are advanced to the subsequent step.

- RoI Pooling Layer: -

The RoI Pooling Layer is a component that utilizes the concept of Region of Interest (RoI). Following the acquisition of region proposals from the RPN, the RoI pooling layer is employed to effectively align these areas, which possess irregular shapes, to a feature map of defined dimensions. This method guarantees that the features obtained from each RoI are standardized in terms of size, enabling them to be seamlessly inputted into the succeeding layers of the model.

- Fully Connected Layers (FC): -

The RoI-pooled features obtained from the preceding layer are propagated through one or more fully connected layers. The aforementioned layers are employed to do additional processing on the features, hence facilitating their readiness for the ultimate tasks of classification and bounding box regression.

- Classification and Bounding Box Regression Heads: -

The classification and bounding box regression heads in Fast R-CNN are implemented using two distinct branches.

  - The classification head is tasked with the classification of the items contained inside the region of interests (RoIs). The system generates class scores for every candidate object.
  - The Bounding Box Regression Head is responsible for estimating the precise coordinates of the bounding boxes associated with each RoI The algorithm refines the placements of the suggested bounding boxes in order to achieve a more precise match around the objects.

- Non-Maximum Suppression (NMS): -

NMS is a technique used in computer vision to eliminate redundant or overlapping bounding boxes or regions of interest (ROIs) in an image. Once the class scores and bounding box coordinates for all the region suggestions have been acquired, the NMS technique is employed to eliminate redundant and low-confidence detections. This process guarantees that only the object detections with the greatest scores and relevance are preserved.

The operational mechanism of Fast R-CNN with ResNet-152 is as follows:

- The input image undergoes processing using the ResNet-152 backbone, which facilitates the extraction of a feature map. This feature map contains encoded information pertaining to the objects present in the image as well as their spatial arrangement.
- The RPN utilizes the aforementioned feature map to produce region proposals, which represent potential bounding boxes for objects. Every proposal is accompanied by a numerical score that indicates the probability of it including an object.
- The region suggestions undergo processing in the RoI pooling layer, wherein each RoI is converted into a feature map of a consistent size.

- The feature maps obtained by RoI pooling are subsequently inputted into the fully connected layers, where the model carries out object categorization and bounding box regression.
- The classification branch is responsible for assigning class labels to the Regions of Interest (RoIs), whereas the regression branch is responsible for refining the bounding box coordinates.
- NMS is a technique employed to eliminate redundant and low-confidence detections, hence insuring the retention of just the most precise object detections.

In brief, the Fast R-CNN model, augmented with the ResNet-152 architecture, is a sophisticated deep learning framework designed to recognize objects in images. This model effectively leverages the ResNet-152 backbone to process images, incorporates a RPN to generate potential bounding boxes, and further refines these candidates through fine-tuning, ultimately yielding precise and reliable object recognition outcomes. The system has various components, including as feature extraction, region proposal, RoI pooling, classification, and bounding box regression, which collaborate to detect and precisely locate objects within images.

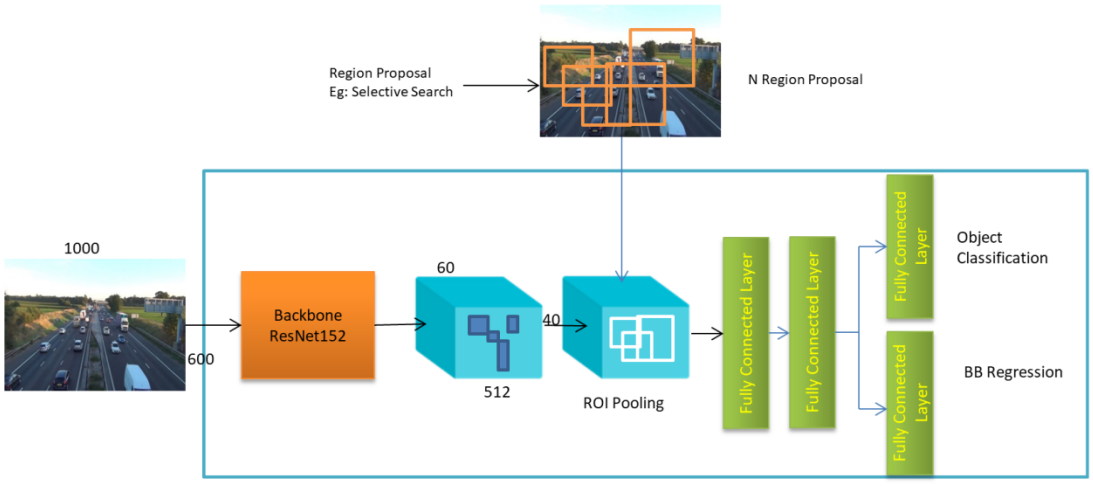


Figure 4.1: Proposed Fast R-CNN with ResNet152 Architecture

In Figure 4.1, the depicted model takes an input image and passes it through the ResNet152 backbone network to extract a set of high-level features from the image. Based on these extracted features, the RPN generates a set of object proposals, also

known as bounding boxes. Each proposal represents a potential object within the image. The RoI pooling layer is responsible for extracting features from each proposal, which are then used for object detection. The features obtained from the RoI pooling layer are fed into a sequence of fully connected layers that classify the objects and predict the bounding box coordinates for each object. Once the objects are identified, NMS is applied to eliminate duplicate detections and select the detections with the highest confidence.

Compared to other models for object detection, Fast R-CNN with ResNet152 offers several unique advantages. ResNet152 is a Deep Residual Network that excels at extracting high-level features from images with very deep architectures. This enables the model to accurately recognize complex objects. The RPN efficiently generates proposals by utilizing the feature maps provided by ResNet152, and the RoI pooling layer reduces the computational requirements for object detection. Furthermore, the entire network can be trained end-to-end, allowing for simultaneous optimization of the entire system.

ResNet152 Model

ResNet152's uses a residual learning architecture as a solution to the issue of vanishing gradients that may occur in deep neural networks. This framework implements skip connections, sometimes referred to as "shortcut" connections, which allow one or more levels to be skipped over. These skip connections make it easier to learn residual mappings since they let the flow of information from older levels straight to subsequent layers. The discrepancy between the intended output and the input is represented via residual mappings, which makes it simpler for the network to learn the required mapping. When it comes to the extraction of features from images, ResNet152 is especially good in capturing high-level semantic information. Each layer of the network retrieves elements that are increasingly more abstract and complicated as the image that is being processed moves through the network's levels. The first few layers are responsible for capturing lower-level characteristics like edges and textures, while the deeper layers are responsible for capturing higher-level notions like object forms and semantic properties.

Because of its deep architecture, ResNet152 is able to acquire rich and discriminative features, which makes it an excellent candidate for applications such as object identification. The network is capable of successfully capturing the granular features and variances that are present in a wide variety of objects, which results in feature representations that are more accurate and resilient.

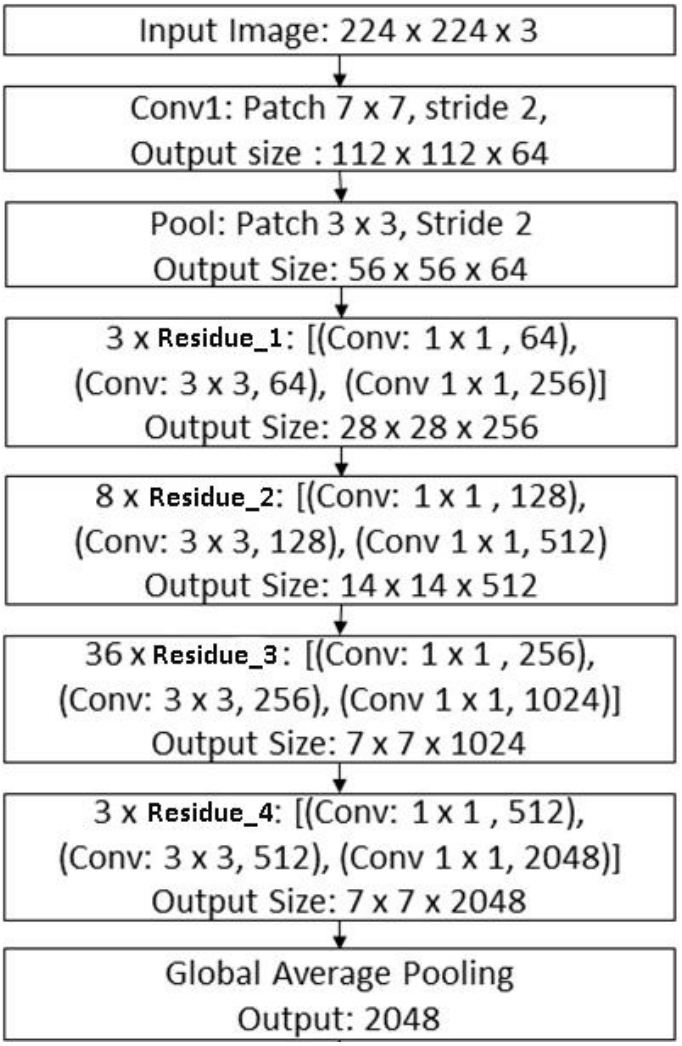


Figure 4. 2: ResNet152 model

- **Convolutional layer**

The convolutional layer serves as the fundamental component of CNNs and is specifically engineered to extract feature representations from input data. The utilization of this technique has demonstrated notable efficacy in the processing of data structured in a grid-like format, such as photographs.

The process of convolution entails the application of a tiny filter, sometimes referred to as a kernel, to the input data. This requires executing element-wise multiplications and subsequently summing the results. The aforementioned procedure produces feature maps that effectively capture various patterns and features present in the input.

The convolutional layer is composed of a set of filters, with each filter containing adjustable parameters. The parameters undergo updates during the training process in order to facilitate the network's adaptation to identify distinct characteristics.

Convolutional layers commonly incorporate activation functions, such as ReLU, to bring non-linear characteristics into the network, so enabling it to acquire knowledge of intricate patterns.

Within the realm of picture classification, the initial convolutional layer is likely to acquire rudimentary features such as edges and corners, whilst subsequent layers are expected to encapsulate more intricate features such as textures or constituent elements of objects.

- **Rectified Linear Unit (ReLU)**

The purpose of using the ReLU activation function is to perform an element-wise operation on the output of convolutional or fully connected layers. The incorporation of non-linearity into the model serves to enhance its ability to capture intricate patterns.

The operation known as ReLU involves the substitution of negative numbers with zero, while preserving positive values without alteration. Mathematically, the ReLU function can be formally defined as follows: $f(x) = \max(0, x)$

One of the advantages of the ReLU activation function is its computational efficiency. This characteristic allows for faster computations, which is particularly beneficial when training deep neural networks. Additionally, ReLU helps alleviate the vanishing gradient problem, which can hinder the training process. By mitigating this issue, ReLU facilitates the training of deep neural networks, making it easier to optimize their performance.

Various variants of the ReLU have been developed, including Leaky ReLU, which introduces a minor gradient for negative values, and Parametric ReLU (PReLU), where the slope of the negative component is a parameter that may be learned.

The ReLU is extensively employed in CNNs due to its ability to facilitate the learning of complex patterns and features within data through its non-linear characteristics.

- **Pooling layer**

The purpose of the pooling layer is to decrease the spatial dimensions of the feature maps generated by convolutional layers. This approach aids in reducing the computing burden and mitigating the issue of overfitting.

In the context of operations, two often employed pooling techniques are max-pooling and average-pooling. In the context of CNNs max-pooling is a pooling operation that involves selecting the largest value within a specific local region. On the other hand, average-pooling is a pooling operation that calculates the average value within the same local region. In the process, a pooled window is moved across the input, and within each window, a chosen operation is executed.

Pooling layers are commonly characterized by their lack of parameters, as they just impact the spatial dimensions of the data.

Pooling is advantageous since it enhances the invariance of representations to minor translations and fluctuations in the input data, hence increasing the robustness of the network.

Downsampling, specifically through the process of pooling, serves to diminish the dimensions of feature maps. This reduction in size holds significance in certain applications, such as picture classification, where the inclusion of high-resolution details may not always be necessary.

- **Residue block**

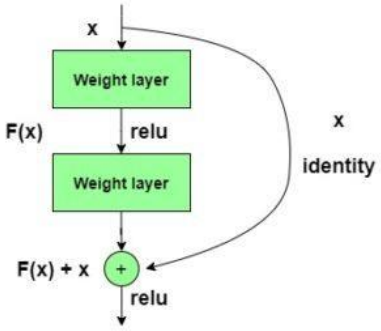


Figure 4. 3: Residue block

A residue block, sometimes referred to as a residual block or a residual unit, serves as a key component within DCNNs , specifically in architectural designs such as Residual Network (ResNet). The utilization of residue blocks was implemented as a solution to mitigate the issue of diminishing gradients that arises during the training process of exceedingly deep neural networks. Fundamentally, they facilitate the training of considerably deeper networks by permitting the transmission of information to bypass one or more layers, hence facilitating the propagation of gradients and simplifying the training process for deep models.

The fundamental principle underlying a residue block pertains to the notion of residual learning. The primary function of conventional neural network layers is to effectively approximate a desired mapping from the input to the output. Within a residue block, the primary objective is not to directly approximate the underlying mapping, but rather to learn how to approximate the residual. The residual refers to the disparity between the desired output and the present input. In a mathematical context, a residue block can be mathematically expressed as:

$$Output = Input + F(Input)$$

The input to the block is denoted as "$Input$", and the function "$F(Input)$" represents the learned residual function. Through the utilization of this "shortcut" link, the network is capable of acquiring the ability to adapt the input in order to approach the desired output more closely. The aforementioned technique exhibits significant

efficacy in deep networks due to its ability to mitigate the issue of vanishing gradients during the backpropagation process.

A standard residue block is composed of the subsequent elements:

- Convolutional layers are responsible for conducting the fundamental process of feature extraction and transformation on the input data. Convolution operations, batch normalization, and activation functions can be incorporated.
- The shortcut link is a direct pathway that enables the input to bypass one or more layers and be directly transmitted to the output of the block. This allows for the information to be expedited and not go through the usual sequential processing. The link between two layers in a neural network is frequently established as an identity mapping if the size of the layers is the same. However, if the dimensions need to be modified, additional transformations such as a 1x1 convolution may be employed.
- The addition operation is responsible for merging the outputs of the convolutional layers and the shortcut connection in order to generate the ultimate output of the residual block.

Residual blocks exhibit a range of complexities, wherein certain instances incorporate supplementary elements such as dropout, skip connections with intricate transformations, or even stacked residual blocks. The specific configuration of a residue block may exhibit variability contingent upon the particular neural network architecture and the challenge at hand.

The utilization of residue blocks has a substantial influence on the domain of deep learning, as they enable the construction of exceptionally deep neural networks capable of capturing complicated elements within the dataset. The capacity to effectively train these deep networks has resulted in enhanced performance across various domains, including computer vision, natural language processing, and other related tasks.

- **Global Average Pooling**

Global Average Pooling is a widely employed strategy in CNNs that aims to decrease the spatial dimensions of feature maps prior to the ultimate classification or regression layer. Computer vision tasks often exhibit a high degree of prevalence in this context. In contrast to conventional fully linked layers, global average pooling presents a computationally efficient and regularization-friendly alternative that retains the fundamental information contained within the feature maps, without introducing a substantial number of parameters.

The Global Average Pooling procedure calculates the mean value of each feature channel throughout the whole spatial dimensions of the feature map. The process operates as follows: for every channel inside the feature map, the operation computes the mean value of all the elements included within that specific channel. Consequently, a singular value is derived for each channel, so transforming the spatial information into a comprehensive summary across channels. The result is a one-dimensional vector consisting of averages calculated for each channel individually.

One of the primary advantages of Global Average Pooling is its ability to substantially decrease the quantity of parameters within the network. The process of reducing spatial information involves the conversion of said information into a vector of fixed size. This vector is thereafter inputted into the final layer of classification or regression. A reduction in the number of parameters within the model decreases its susceptibility to overfitting, a particularly favorable characteristic when confronted with a scarcity of training data.

An additional benefit of Global Average Pooling is its enhancement of the interpretability of the network's output. The presence of specific features in the input image can be inferred from the important information included in each channel, as each channel corresponds to a distinct feature or concept. This capability facilitates comprehension of the network's acquired knowledge, rendering it valuable for purposes such as visual representation and feature extraction.

4.3 SIMULATION RESULTS

This section provides an overview of the simulation results obtained from the Moving Object detection model on diverse videos. The research methodology involved applying an object detection model to the videos, followed by the implementation of a tracking model. The proposed detection and tracking models were then evaluated and compared to existing state-of-the-art models using different videos. Detailed information regarding the videos used in the study and a comprehensive analysis of the performance of the proposed object detection and tracking models will be presented in subsequent sections.

4.3.1 Datasets

In the first stage of this research work, Various video sequence images are considered for object detection and continuous video footage is applied for object tracking. The Video-1 is considered from the CDNET [94], which stands for change detection dataset, is a benchmark that is used for object identification, instance segmentation, and several other computer vision applications. However, the CDNET dataset that was used initially did not have a category that was dedicated to the identification of pedestrians. The data collection largely concentrates on 80 different types of common object categories, including things like people, animals, automobiles, and goods found in the home, among other things. From this huge data set, pedestrians' footage is considered as Video-1 and vehicles on highway is considered as Video-2. The Video-3 and 4 are considered from real video footages. The Video-3 is the M6 motorway which is the longest and most significant route in Britain, spanning from the Midlands all the way up to the west coast. It is also the nation's longest motorway. Each and every day, tens of thousands of cars make use of it. This video-3 shows the traffic on the M6 between the Knutsford Services junction and junction 19. Since the camera is pointing south, the roadway heading south is on the left, while the carriageway heading north is on the right. In terms of automobiles, this video has close to two hundred huge articulated vehicles. In addition to that, there are eight Eddie Stobart trucks. In addition to the heavy goods vehicles, there are around 30 coaches, several thousand automobiles and vans, a small number of motorcycles, an ice cream truck,

and an ambulance. Similarly, Video-4 is the traffic jam located in Dhaka, Bangladesh. This video footage is live traffic jam captured during morning time. The image samples which are considered for proposed object detection tracking models are depicted in Figure 4.4.

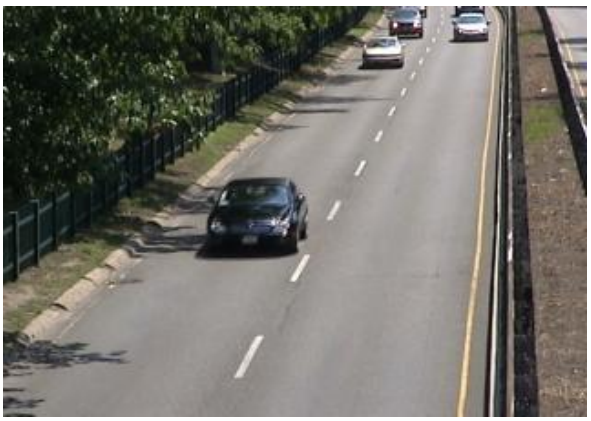
(i)

(ii)

(iii)

(a) Video-1

(i)



(ii)

(iii)

(b) Video-2

(i)

(ii)

(iii)

(c) Video-3

(i)

(ii)

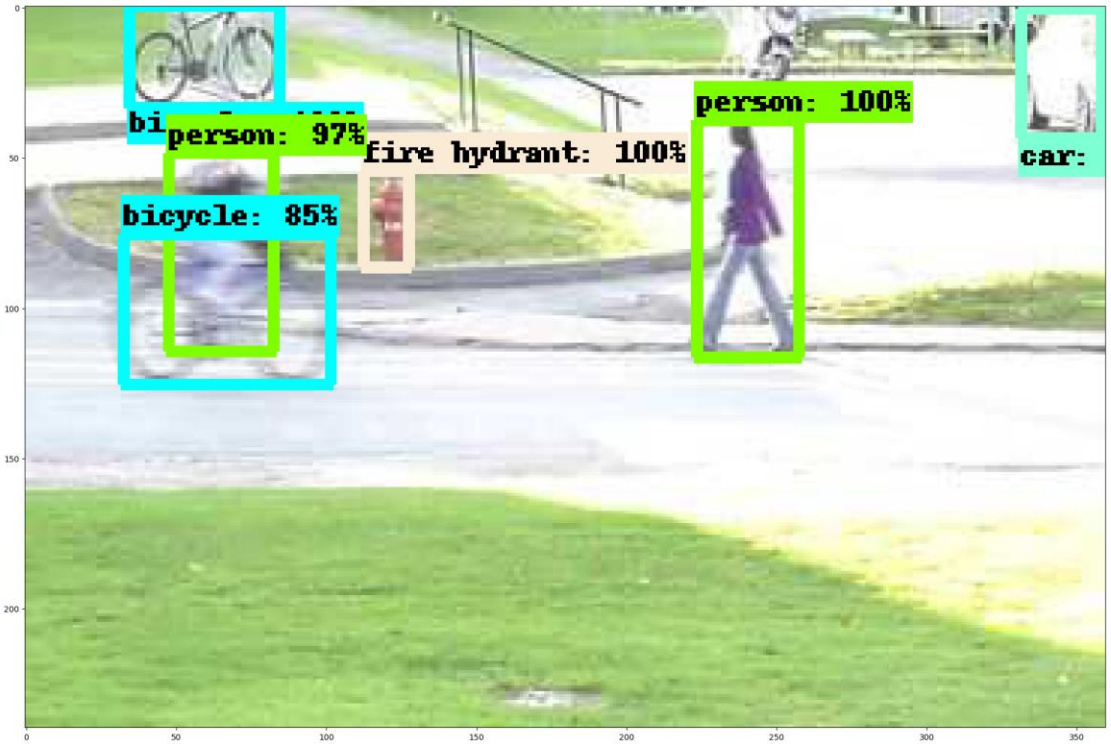
(iii)

(d) Video-4

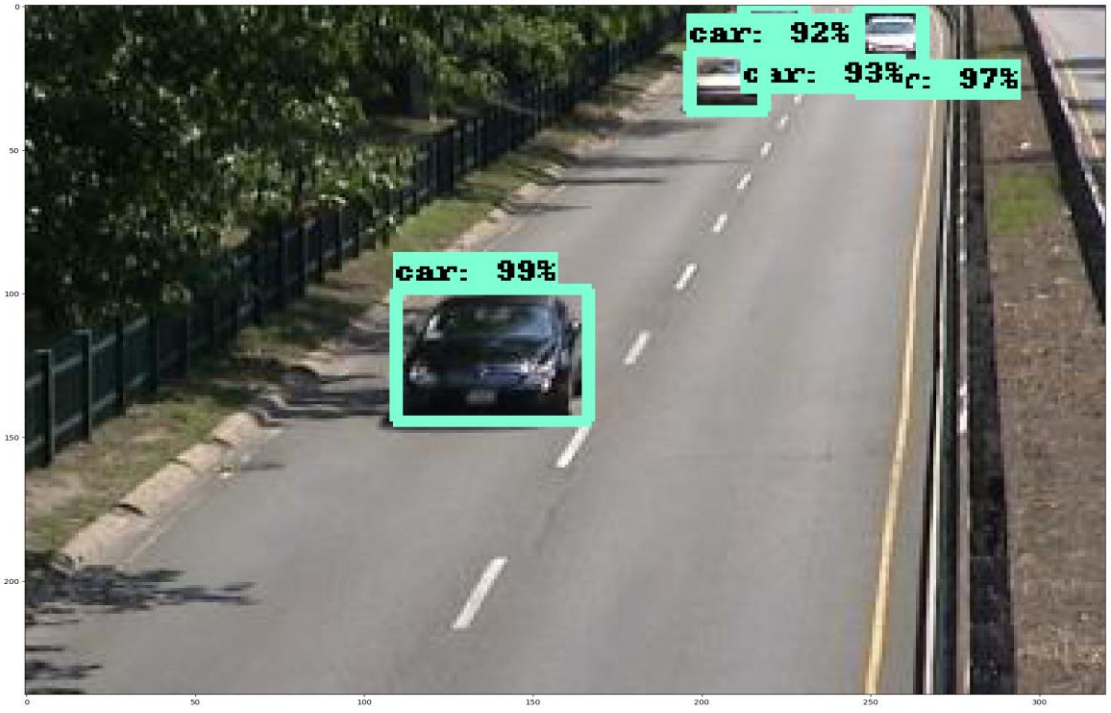
Figure 4.4: Sample images in different videos

4.3.2 Object Detection

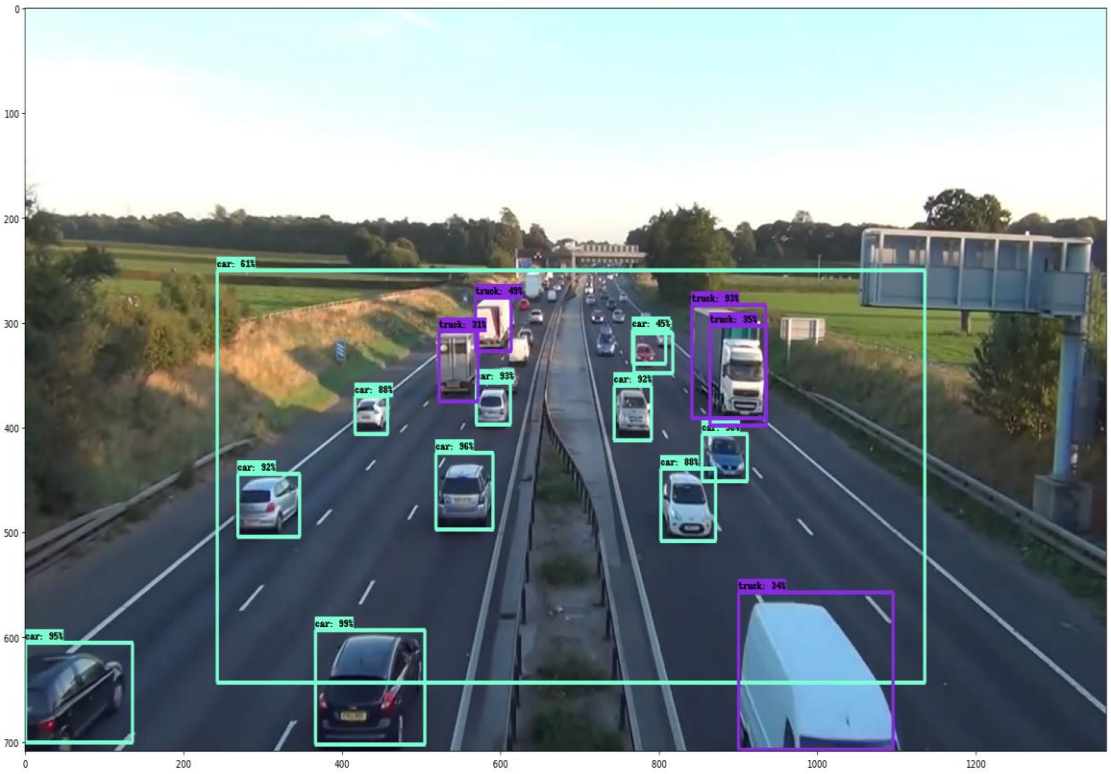
The Proposed model showed exceptional performance in object detection. It got a score of 100% accuracy for person-1 and 2, which indicates a good level of accuracy in both the localization and classification of objects. The model was able to recognise a wide variety of items, including automobiles, people, bicycles, and traffic signs, with high accuracy. The visualization results of the proposed object tracking results are depicted in Figure 4.5.



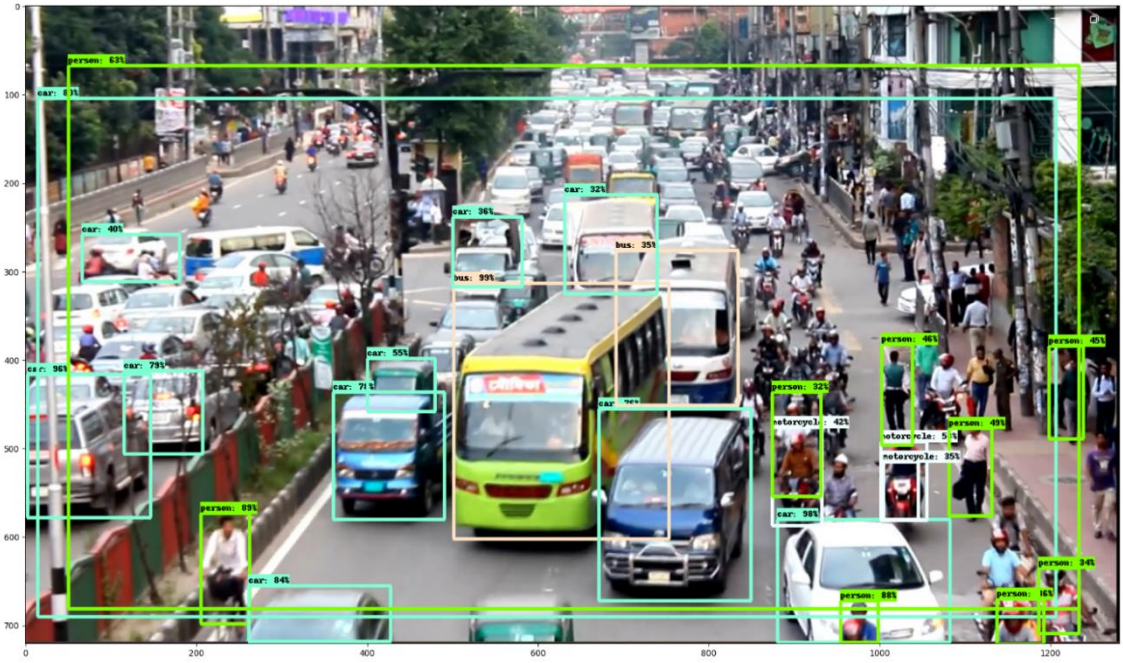
(a) Object detection on Video-1



(b) Object detection on Video-2



(c) Object detection on Video-3



(d) Object detection on Video-4

Figure 4. 5: Object detection results on different videos

The proposed model did an excellent job of detecting objects at a variety of sizes and orientations, as well as in demanding conditions including occlusions and cluttered surroundings. The proposed model is compared with other well-known models and corresponding results are reported in Table 4.1.

Table 4.1: Comparative Analysis of proposed Object detection model

Model Name	Time	Accuracy			
		Person-1	Person-2	Car	Truck
Efficient D1 640x640	4.82	65	39	86	56
Efficient D7 1536x1536	28.29	73	33	89	82
SSD MobileNetV2 320x320	13.76	78	41	63	36
SSD MobileNetV1 FPN 640x640	5.95	41	55	63	36
SSD MobileNetV2 FPN lite 640x640	5.96	55	42	75	33
SSD ResNet50 V1 FPN 640x640	5.96	34	78	78	48
SSD ResNet50 V1 FPN 1024x1024	7.62	Not detected	30	88	57
SSD ResNet101 V1 FPN 640x640	4.71	46	42	83	35
SSD ResNet101 V1 FPN 1024x1024	4.79	31	83	88	66
SSD ResNet152 V1 FPN 640x640	8.3	39	46	80	46
SSD ResNet152 V1 FPN 1024x1024	5.52	Not detected	35	86	77
Fast RCNN ResNet50 V1 640x640	20.382	97	99	100	46
Fast RCNN ResNet50 V1 1024x1024	16.42	35	32	98	78
Fast RCNN ResNet101 V1 640x640	4.97	92	100	99	79
Fast RCNN ResNet101 V1 1024x1024	4.88	96	99	99	73
Fast RCNN ResNet152 V1 1024x1024	5.8	Not detected	91	99	80
Fast RCNN Inception ResNet V2 604x604	12.86	98	97	99	84
Fast RCNN Inception ResNet V2 1024x1024	10.67	88	90	98	95
Proposed Fast RCNN ResNet152 V1 640x640	**6.13**	**97**	**100**	**99**	93

The table 4.1 provides a comparison of several object detection deep learning models on different videos reported in Figure 4.5, along with their respective performance metrics about the amount of time required for inference as well as their level of accuracy for certain classes of objects. The performance of each model is judged according to how well it can recognise certain items, such as Person-1, Person-2, Car, and Truck. The Fast RCNN ResNet152 V1 640x640 model is the only one in the table that displays a reasonably quick inference time of 6.13 seconds. It reaches a high level of accuracy, with rates of 97% for Person-1, 100% for Person-2, 99% for Car, and 93% for Truck. This demonstrates that the model is effective at identifying these objects, and as a result, it is suited for use in applications that call for object detection to take place in real time. The Single Shot Detector (SSD) ResNet101 V1 FPN 640x640 model, on the other hand, obtains a quicker inference time of 4.71 seconds but displays substantially lower accuracy rates. It obtains an accuracy of 83% for cars,

35% for trucks, 46% for persons 1 and 42% for person 2. In comparison to other models, it could be quicker, but it might also be less dependable when it comes to precisely recognising things. The comparative analysis chart is shown in Figure 4.6.

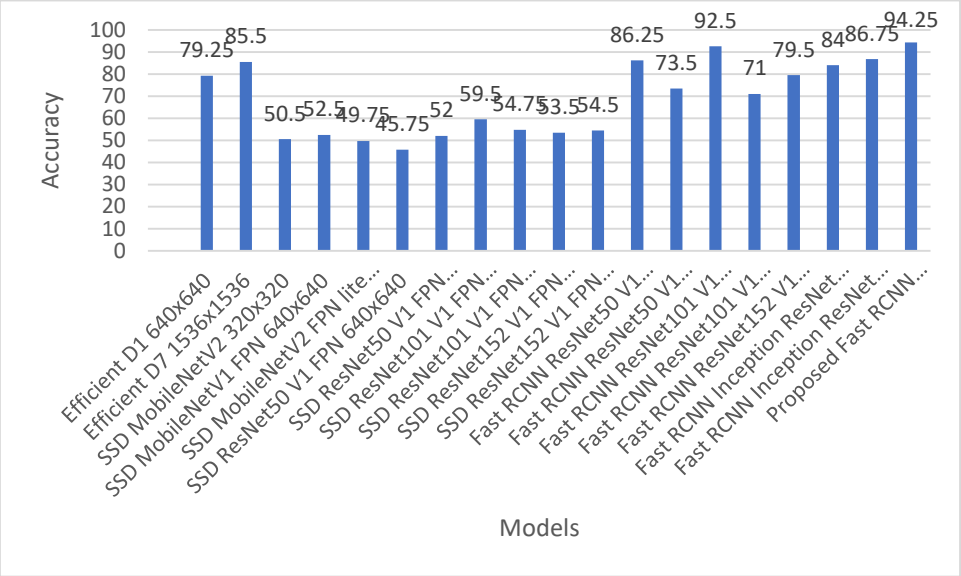


Figure 4.6: Comparative Analysis chart

The Fast RCNN ResNet50 V1 640x640 model detects the person 1 with accuracy of 97% , high accuracy rates of 99% for Person-2, and accuracy for Car at 100%. Nevertheless, its accuracy falls to 46% when applied to Truck. It has a longer inference time of 20.382 seconds, which indicates that it may not be suited for real-time applications. However, it may be useful in undertakings in where high accuracy for certain object classes is an absolute must. The proposed Fast RCNN ResNet152 V1 640x640 model has 97 % accuracy for Person-1, high accuracy rates of 100% for Person-2, and accuracy for Car at 99%. However, when applied to Truck, its accuracy drops all the way down to 93%. The proposed model obtains a quicker inference time of 6.13 seconds, it is possible that it is suitable for use in applications that need performance in real time.

These test cases provide insight on the tradeoffs that exist between object detection models' inference times and their levels of accuracy. Models with higher processing speeds, such as the Efficient D1 640x640 and the SSD MobileNetV2 320x320, provide results more quickly, but at the low accuracy. On the other hand, models such

as Fast RCNN ResNet50 V1 640x640 provide greater accuracy while requiring more time to complete the inference process. In the end, the selection of an object detection model is determined by the particular needs of the application. When speed in real time is of the utmost importance and that need a greater level of precision, models such as Fast RCNN ResNet152 V1 640x640 could be the best option, despite the fact that their inference time is minimal and highest accuracy.

4.4 SUMMARY

Within the ever-evolving domain of computer vision, the endeavor to achieve precise and effective video object detection through the utilization of the Fast RCNN model has been a very consequential undertaking. This chapter has examined the fundamental ideas, techniques, and actual implementations of video object detection using the Fast RCNN approach. It has provided insights into the extensive possibilities and promising prospects of this technology across several areas. The primary recognition is the central significance of object detection in the field of video analytics. In the contemporary era inundated with vast amounts of video data, the capacity to autonomously discern and determine the presence and position of objects inside dynamic environments is of utmost importance. The Fast RCNN model, due to its ability to incorporate CNNs and sophisticated RPNs, represents a significant advancement in this field. The technology not only provides exceptional precision but also demonstrates prompt or nearly prompt execution, a vital necessity in numerous modern applications. While chapter 5 focuses on Object Tacking, it is important to note that accurate object detection within an image or video fame serves as the foundation for effective and reliable tracking. Building on the object detection technique in this chapter, the next chapter explores into the methodology used to track objects across frames addressing the challenges posed by motion estimation, occlusion and trajectory maintenance.

CHAPTER 5
VIDEO OBJECT TRACKING USING OC-SORT

5.1 INTRODUCTION

The task of video object tracking is a crucial undertaking in the fields of computer vision and machine learning, which entails the process of tracking and localizing a designated object throughout a series of video frames. It serves as a crucial component in a diverse array of applications, encompassing surveillance, driverless cars, augmented reality, and video analysis.

The basic goal in video object tracking is to ensure the consistent identification of a selected item as it traverses different frames. This task involves addressing several problems, including alterations in lighting conditions, background elements, occlusions, scale variations, and changes in viewpoint. The task can be classified into numerous fundamental components.

The method commences by initially identifying the object to be tracked in the initial frame of the video. Object identification techniques such as YOLO or Faster R-CNN are frequently employed for this specific task. After the detection of the object, it is imperative to ensure its effective representation for the purpose of tracking. Various forms of representation are commonly utilized in tracking applications, such as bounding boxes, keypoints, or masks.

The fundamental aspect of video object tracking involves the prediction of an object's location in upcoming frames in order to effectively monitor its movement. This entails the estimation of the object's motion, which encompasses a spectrum of models ranging from basic linear or constant velocity models to more intricate methods such as optical flow or Kalman filters. The precise measurement of motion is of utmost importance in order to accurately anticipate the trajectory of an object.

Within a given video sequence, it is possible for several objects to exhibit occurrences of appearance and disappearance, as well as instances of overlapping with the target item. In order to preserve the integrity of the object, it is imperative to accurately

establish its association with the object representation in every frame. This task can be accomplished by the utilization of techniques such as closest neighbor assignment or data association methods, such as the Hungarian algorithm.

Real-world situations present difficulties arising from occlusions, variations in lighting conditions, presence of background clutter, and deformations of objects. It is imperative for tracking algorithms to possess a high level of robustness in order to effectively manage the intricacies associated with tracking tasks. This robustness is essential to ensure that the identification of the tracked object remains intact, even when confronted with adverse environmental conditions.

The classification of video object tracking techniques can be divided into two main categories: online and offline approaches. The method of online tracking involves the sequential analysis of each frame as it is received, hence enabling the achievement of real-time tracking capabilities. On the other hand, offline tracking takes into account the complete video sequence in order to enhance the accuracy of tracking. Both approaches possess their respective advantages, contingent upon the specific demands of the application and the available computational resources.

In recent years, the field of video object tracking has witnessed substantial advancements due to the enormous progress made in deep learning techniques. The utilization of deep neural networks, such as siamese networks or LSTM networks has been employed to enhance tracking performance. These strategies utilize the capabilities of deep learning to efficiently address intricate tracking problems.

The application of video object tracking is of utmost importance in diverse fields, such as surveillance for the purpose of monitoring and tracking subjects of interest, robotics for autonomous navigation, sports analysis for tracking players and balls, and other related domains. The field of computer vision remains in a state of constant evolution due to the advancement of novel algorithms and the incorporation of artificial intelligence methodologies. This ongoing progress contributes to the dynamic and captivating nature of this domain.

Video object tracking is a crucial component in a wide range of applications, such as surveillance, driverless vehicles, augmented reality, and sports analysis. The significance of this technology resides in its capacity to promptly identify and monitor objects or individuals in video streams, thereby facilitating heightened security through the tracking of suspicious activities. Additionally, it enhances the navigational capabilities of autonomous vehicles, enabling them to effectively maneuver and evade obstacles. Moreover, it enriches user experiences in augmented reality applications by establishing a connection between virtual objects and real-world objects. Lastly, it offers valuable insights into athlete performance and strategy in the realm of sports analysis. In a broad spectrum of sectors, video object tracking technology exhibits versatility by augmenting safety, ease, and comprehension.

5.2 PROPOSED MODEL

5.2.1 Proposed Object Tracking

Online Continuous-time object tracking (OC Sort), is a computer vision approach that is used to track objects in a video stream throughout the course of time. Continuous-time tracking, as opposed to the more typical frame-by-frame tracking systems, which work on discrete time frames, takes into consideration the movement of objects in a continuous way. This strategy involves the object tracker making an estimate of the state of the object (including its location, size, and orientation) at each given instant in time. This results in tracking that is both more precise and smoother. In order for the tracker to function properly, its estimate must be routinely revised in light of any newly acquired data from the video frames.

1. **Initialization phase**: In the Object Detection initialization process begins by capturing a screenshot of the initial frame from the video stream. This frame serves as the starting point for detecting objects in the subsequent frames. The captured frame is then transmitted across the network to undergo object detection. The frame is analyzed to identify and locate objects present within it. This analysis results in the retrieval of important information associated with each identified item, including the bounding boxes that indicate the object's position and shape, the class probabilities that represent the likelihood

of the object belonging to a specific class, and the confidence scores that indicate the level of certainty in the detection. Each identified object is assigned a unique identification. This identification helps in accurately tracking and distinguishing individual objects throughout the video stream, enabling further processing and analysis of their movements and behaviors. By assigning unique IDs, the system can maintain consistency and reliability in identifying and monitoring objects over time.

2. **Motion Estimation**: The procedure involves developing an estimate of the motion of the objects that have been observed between the previous frame and the present frame. The purpose of this stage is to monitor the motion of objects over a period of time and estimate where they will be located in the current frame. The system determines the apparent motion of objects by assessing the movement of pixels from one frame to the next using optical flow. This process takes place between successive frames. On the other hand, the algorithm uses probabilistic models to estimate the motion of an item based on prior observations and projected dynamics. After an estimate of the motion has been made, the bounding box coordinates of the objects are then modified to reflect the new information.

3. **Object tracking**: Once the object detections have been obtained, the system will retrieve important information for each detected object. This information will include the bounding boxes, which will indicate their positions and sizes, the class probabilities, which will represent the likelihood of belonging to specific classes, and the confidence scores, which will indicate the reliability of the detections. It is necessary to maintain continuity from one frame to the next, therefore the object detections from the current frame are compared with the tracked objects from the previous frame. This process of matching may be carried out utilizing methods such as the Intersection over Union (IoU) methodology or the appearance-based matching method. The system may identify the correspondences between the items by comparing the overlapping

regions of the objects or the visual attributes of the items. After the matching has been finished, the system will update the bounding box coordinates of the objects that are being tracked based on the detections that have been matched. This modification guarantees that the bounding boxes appropriately surround the objects in their present places, taking into account any changes to their locations or sizes that may have occurred.

4. **Occlusion handling**: In order for the system to deal with occlusion, it may concentrate its object identification efforts on certain places inside the frame. The system is able to increase the likelihood of successfully recognizing obstructed objects by reducing the search area to make the search more specific. In addition, signals based on an object's appearance or motion may be used to help identify potentially obscured objects and make it easier to see them. When occluded objects are found, the system will start the tracking procedure for these objects over again from the beginning. This re-initialization entails resetting the tracking settings and updating the object's location depending on the newly discovered or re-detected occluded object. Additionally, this re-initialization may require resetting the tracking parameters. After then, one-of-a-kind IDs are given to these occluded objects in order to ensure that their distinct identities are preserved throughout the monitoring procedure. The system is able to efficiently manage the issues posed by occlusion in object tracking by using occlusion handling methods such as re-detection or tracking-by-detection, as well as particular search approaches and motion cues.

5. **Trajectory maintenance**: The system maintains trajectories by forming connections between object detections or tracking outcomes over successive frames in order to guarantee continuity and track objects over time. This allows the system to track objects across time. The system connects the objects that have been identified or tracked across frames by linking object detections or tracking results. This results in the creation of a trajectory that depicts the object's movement over the course of time. This connection may be

made by a variety of approaches, such as matching based on space overlap, appearance similarity, or motion consistency, among other possibilities. The model will update the trajectories by adding the most recent bounding box coordinates and IDs of the objects when the objects move and their bounding box coordinates change. This guarantees that the trajectories provide an accurate representation of the locations and identities of the objects as they change over the course of time.

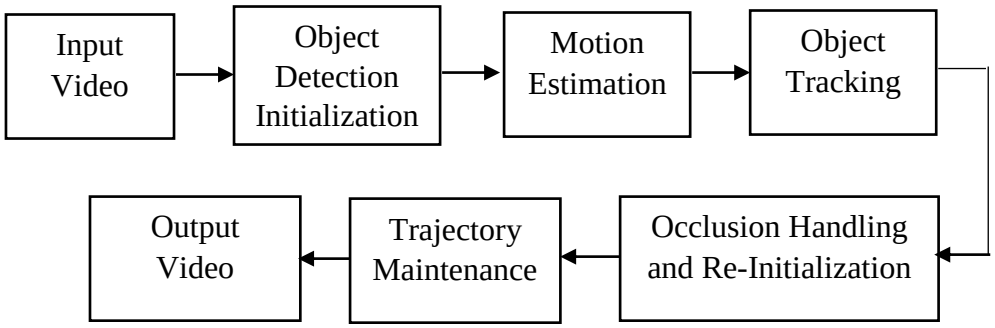
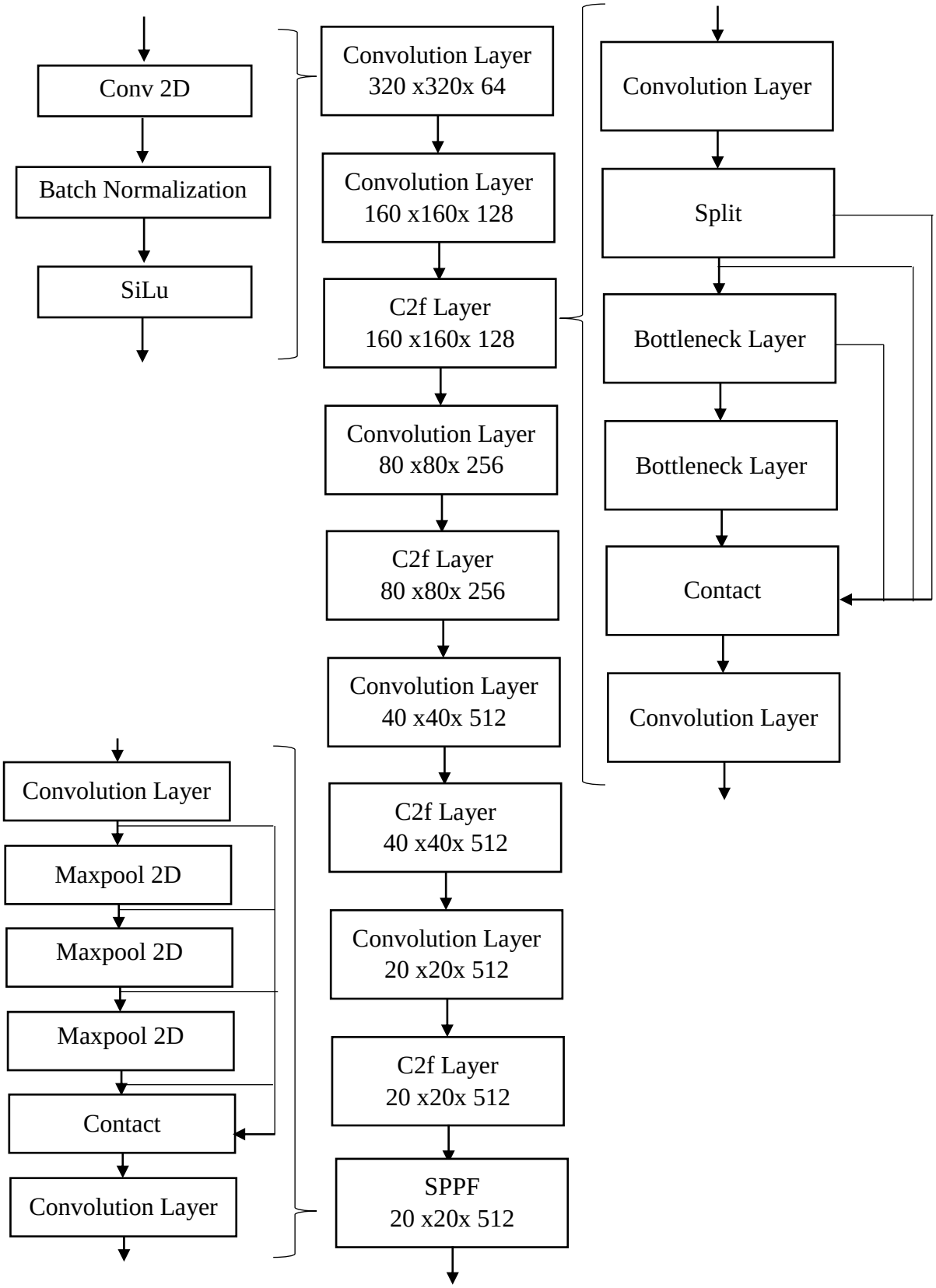
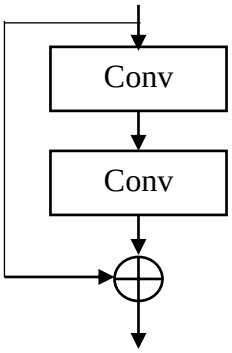


Figure 5. 1: Object tracking framework



(a) Top Level model



(b) Bottleneck layer

Figure 5. 2: YOLOv8 deep learning model

The layers used in the architecture of YOLOv8 are discussed here:

Convolution 2D layer: The Layer is an essential component of CNNs. Convolutional operations are carried out on the input data, such an image. The layer is made up of a group of kernels that move on top of the input and compute dot products between the filter weights and pixels of the input image. These convolutions are responsible for capturing the spatial patterns and feature representations that are present in the input image.

Batch Normalization is a method that is used to normalize the activations of a neural network's intermediate layers. Its goal is to stabilize and speed up the training process by decreasing internal covariate shift. This term refers to the change in the distribution of layer inputs that occurs as a result of the updating of network parameters during training. The input data are normalized using the process of batch normalization by first removing the mean and then dividing the result by the standard deviation of each mini-batch. It helps enhance gradient flow, enables greater learning rates, and offers better generalization by minimizing the model's sensitivity to the initial weight initialization.

Sigmoid Linear Unit (SiLU) activation is an activation function that is often used in deep learning models. The formula for computing it is as follows:

$$SiLU(x) = x * sigmoid(x)$$

This formula represents the element-wise product of the input and its sigmoid activation. The non-linear and smooth nature of SiLU is one of its primary features. It has been shown that it improves the learning and generalization capacities of neural networks by increasing information propagation, fostering greater gradient flow, and capturing more complicated connections within the data.

Maxpooling 2D is a downsampling process that is often employed in CNNs to lower the spatial dimensions of the input feature maps. It works on very tiny local areas, often 2x2 or 3x3, and it replaces the value in each region with the highest value that can be found in that region. In order to accomplish attaining spatial invariance, lowering the computational cost of following layers, and extracting the most important features from the input, maxpooling is a helpful technique.

Spatial Pyramid Pooling – Fast (SSPF): The SPP layer's primary function is to extract features at many scales without the need of any additional convolutional processes or pooling layers than those are already included in the network. This is achieved by arranging the input feature maps in a grid and carrying out pooling operations at many levels using a variety of kernel sizes. The model is able to successfully deal with objects of varying sizes and scales thanks to the pooling procedures, which often include maximum pooling. These processes gather information at varied receptive fields.

5.3 SIMULATION RESULTS

This section provides an overview of the simulation results obtained from the Moving Object detection model on diverse videos. The research methodology involved applying an object detection model to the videos, followed by the implementation of a tracking model. The proposed detection and tracking models were then evaluated and compared to existing state-of-the-art models using different videos. Detailed information regarding the videos used in the study and a comprehensive analysis of the performance of the proposed object detection and tracking models will be presented in subsequent sections.

5.3.1 Datasets

In the first stage of this research work, Various video sequence images are considered for object detection and continuous video footage is applied for object tracking. The Video-1 is considered from the CDNET [94], which stands for change detection dataset, is a benchmark that is used for object identification, instance segmentation, and several other computer vision applications. However, the CDNET dataset that was used initially did not have a category that was dedicated to the identification of pedestrians. The data collection largely concentrates on 80 different types of common object categories, including things like people, animals, automobiles, and goods found in the home, among other things. From this huge data set, pedestrians' footage is considered as Video-1 and vehicles on highway is considered as Video-2. The Video-3 and 4 is considered from real video footages. The Video-3 is the M6 motorway which is the longest and most significant route in Britain, spanning from the Midlands all the way up to the west coast. It is also the nation's longest motorway. Each and every day, tens of thousands of cars make use of it. This video-3 shows the traffic on the M6 between the Knutsford Services junction and junction 19. Since the camera is pointing south, the roadway heading south is on the left, while the carriageway heading north is on the right. In terms of automobiles, this video has close to two hundred huge articulated vehicles. In addition to that, there are eight Eddie Stobart trucks. In addition to the heavy goods vehicles, there are around 30 coaches, several thousand automobiles and vans, a small number of motorcycles, an ice cream truck, and an ambulance. Similarly, Video-4 is the traffic jam located in Dhaka, Bangladesh. This video footage is live traffic jam captured during morning time. The image samples which are considered for proposed object detection tracking models are depicted in Figure 5.3.

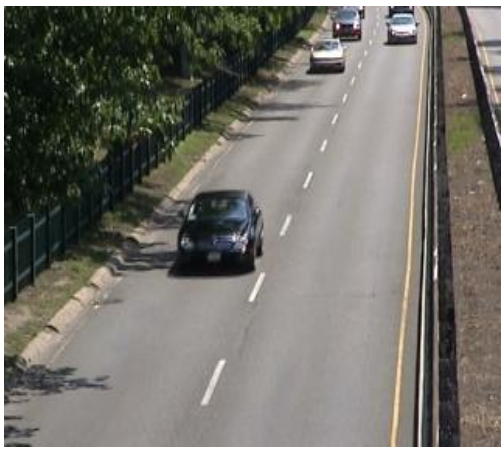
(i)

(ii)

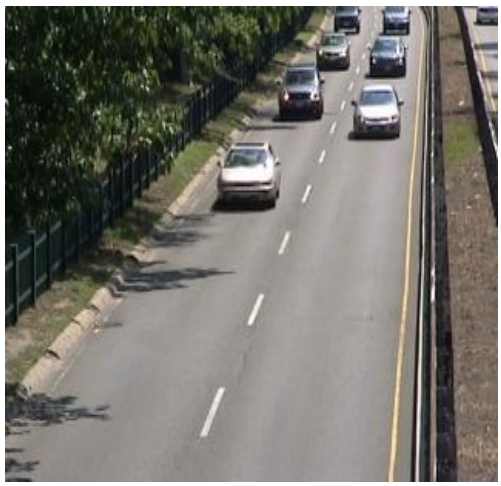
(iii)

(a) Video-1

(i)



(ii)



(iii)

(b) Video-2

(i)

(ii)

(iii)

(c) Video-3

(i)

(ii)

(iii)

(d) Video-4

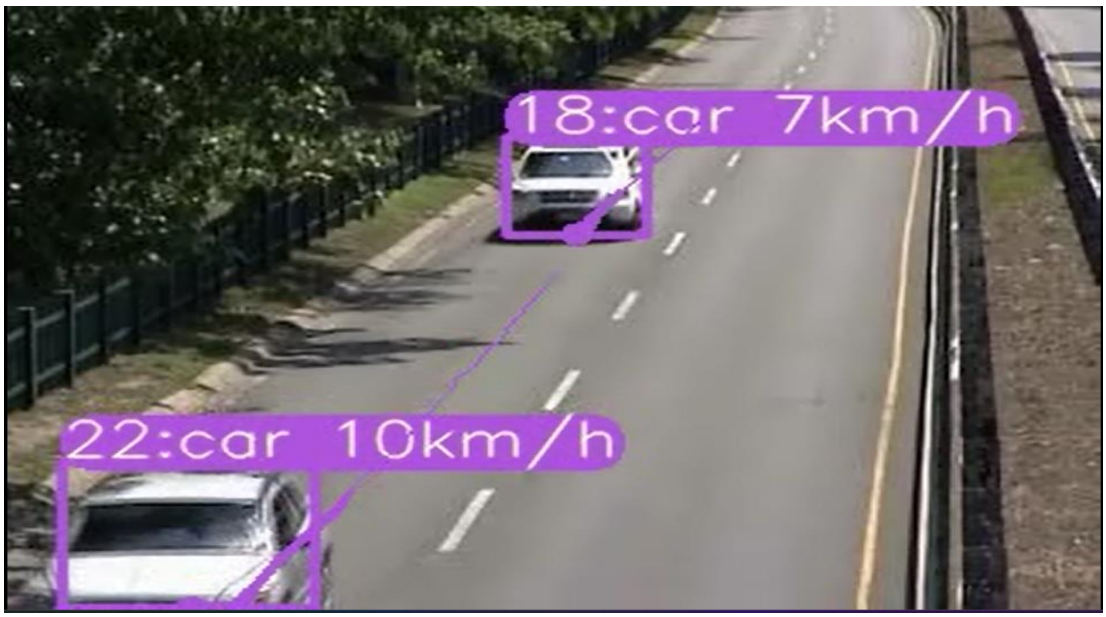
Figure 5. 3: Sample images in different videos

5.3.2 Object Tracking

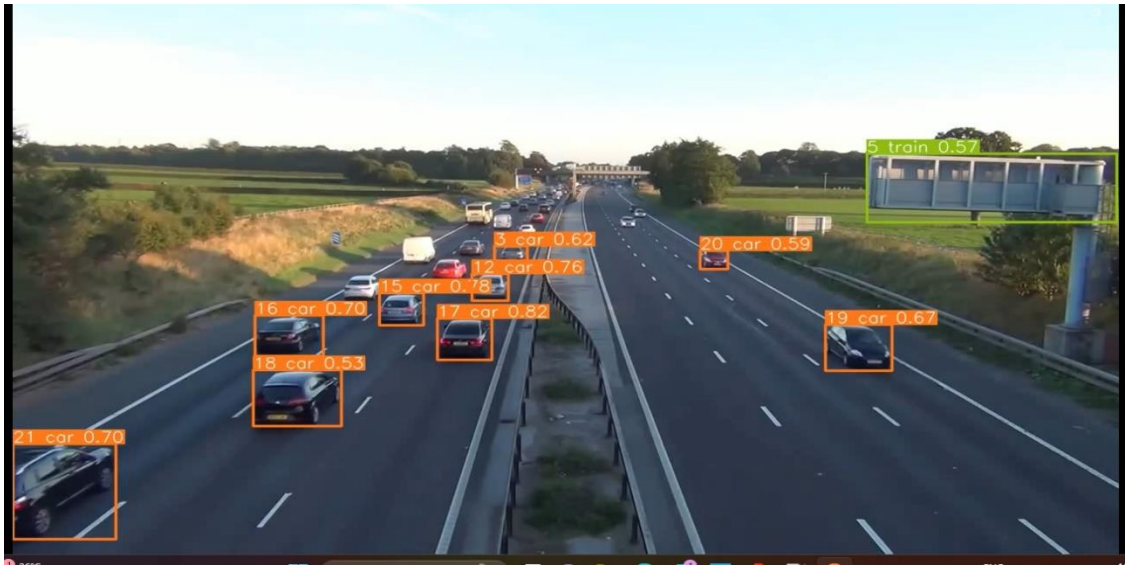
The proposed object tracking module proved reliable performance by properly tracking objects over many frames. Even when there were occlusions or momentary disappearances of objects, the model was able to quickly connect the bounding boxes of identified objects across frames, hence ensuring consistency in the tracking process.

Sample frames from the simulation are shown here in order to provide a graphical illustration of the findings obtained from the object recognition and tracking processes. The frames show the items that have been correctly recognised and tracked by using proposed model, together with the bounding boxes and class labels that belong to those objects. The visual results make it abundantly clear that the model is capable of dealing with complicated sceneries, occlusions, and object interactions. The visualization results of the proposed object tracking results are depicted in Figure 5.4.

(a) Object tracking on Video-1



(b) Object tracking on Video-2



(c) Object tracking on Video-3

(d) Object tracking on Video-4

Figure 5. 4: Object Tracking results on different videos

The model obtained an accuracy of 69% for person-1, 68% for person-2, 63% for Car and 52% for truck, which indicates that the bounding boxes that were predicted and those that were ground-truth had a significant overlap. Real-time tracking capability was also proven by the model, which was able to achieve a frame rate of thirty frames per second (FPS) on a regular CPU. This suggests that proposed object tracking model is successfully used in real-time applications, such as video surveillance or autonomous cars, where it is essential to have low latency and high tracking accuracy.

Table 5. 1: Comparative Analysis of proposed Object tracking model

Model Name	**Time**	**Accuracy**			
		Person-1	**Person-2**	**Car**	**Truck**
Deepocsort [100]	96.472	69	68	63	45
Strongsort [101]	309.511	69	66	55	38
Bytetrack [76]	37.519	71	Not detected	Not detected	Not detected
Botsort [102]	98.854	69	68	63	44
Proposed Ocsort	**50.519**	**69**	**68**	**63**	**52**

The table 5.1 offers insights into various models utilized for car detection and tracking, along with corresponding inference times and accuracy rates for specific car classes (Person-1, Person-2, Car and Truck). One of the models listed in the table is Deepocsort, which demonstrates an inference time of 96.472 seconds. It achieves

reasonable accuracy rates, with 69% for detecting person-1, 68% for person-2, 63% for Car and 45% for Truck. These results indicate that Deepocsort is effective in detecting and tracking persons, although there is room for improvement in accuracy. Bytetrack, another model in the table, shows a faster inference time of 37.519 seconds. It achieves an accuracy rate of 71% for detecting person-1, but unfortunately, it fails to detect person-2 ,car and truck in the provided Videos. This suggests that Bytetrack may struggle with certain car classes or may require further refinement for more accurate results. Figure 5.5 indicates the comparative time analysis chart and Figure5.6 represents the comparative accuracy analysis chart.

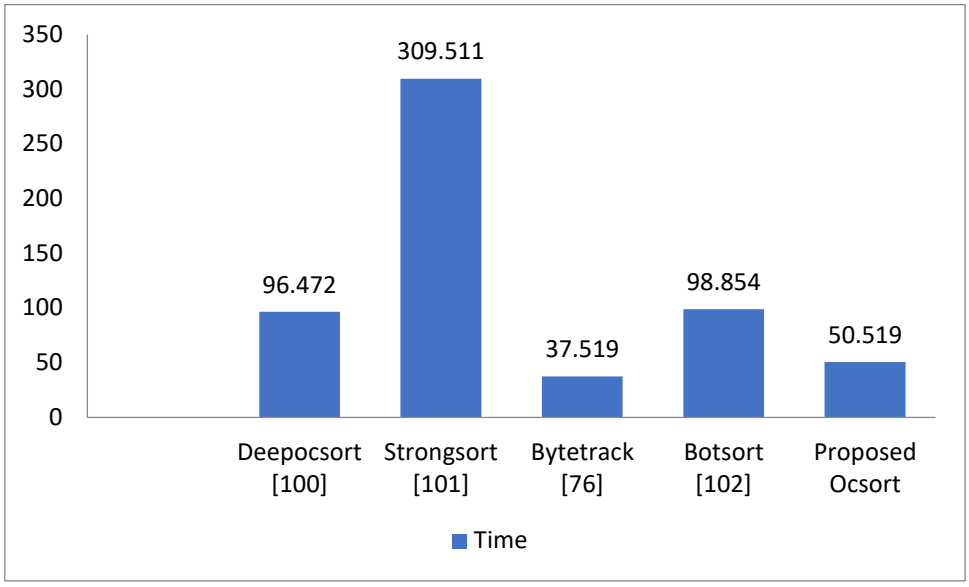


Figure 5.5: Comparative Time Analysis Chart

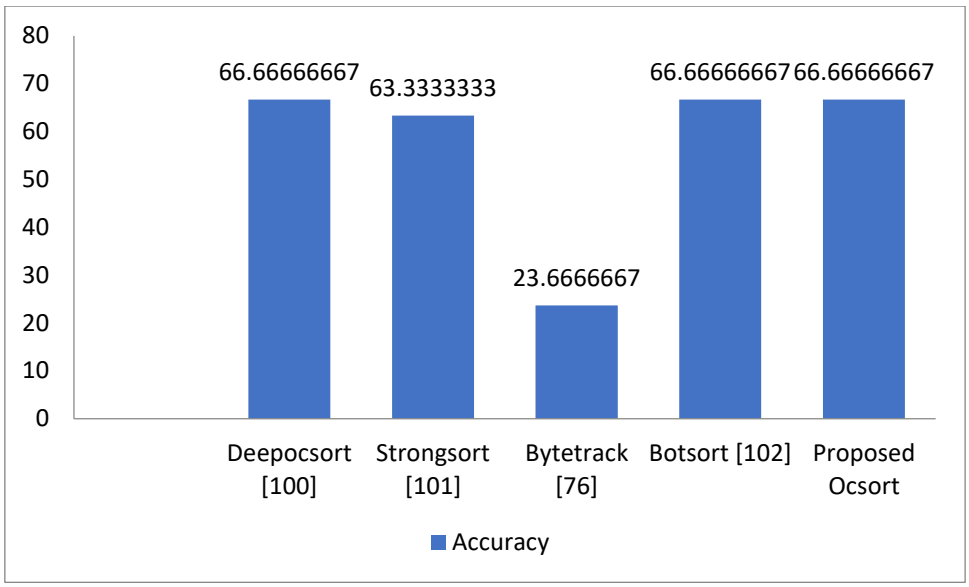


Figure 5. 6: Comparative Accuracy Analysis Chart

The proposed Ocsort model demonstrates an inference time of 50.519 seconds. It achieves accuracy rates similar to Deepocsort, with 69% for person-1, 68% for person-2, 63% for Car and 52% for Truck. These results indicate that the proposed Ocsort model performs comparably to Deepocsort in terms of accuracy, while potentially offering a slightly faster inference time. It is important to note that the accuracy values reported in the table pertain specifically to the mentioned car classes and may not represent the overall detection accuracy for all car classes.

5.4 SUMMARY

The investigation into video object tracking utilizing the OC Sort model has shown a domain of considerable importance within the continuously developing field of computer vision. The current chapter extensively explored the fundamental principles, methodology, and actual implementations of OC Sort, hence offering significant perspectives on its potential as a reliable tracking solution in diverse fields. The primary insight that has been derived from this exploration is the pivotal significance of object tracking within the realm of video analytics. Considering the escalating quantity of video data becoming produced globally, it has become imperative to possess the capability to effectively monitor and track objects over several frames. The OC Sort algorithm, because to its seamless integration of object detection and

tracking capabilities, represents a significant advancement in this field. OC Sort offers a sophisticated resolution to a multifaceted issue by effectively connecting the identification of objects with their subsequent tracking. The chapter underscored the need of utilizing labeled object tracks for training data across the entire process, starting from data collection and extending to model implementation. The process of refining a pre-trained object detector, such as Fast RCNN, was examined as a strategy to enhance object recognition within the OC-SORT model. Furthermore, the significance of using the SORT algorithm was emphasized as an essential element in achieving real-time object tracking.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

The domain of computer vision has witnessed remarkable advancements in recent years, enabling machines to perceive and interpret visual information with increasing accuracy. However, a persistent and challenging problem in this domain revolves around video object detection and tracking under poor lighting conditions. In many practical applications such as surveillance, autonomous vehicles, and industrial automation, the quality of visual data is often compromised by insufficient or uneven illumination. This dimly lit and unpredictable environment poses substantial hurdles for traditional computer vision systems, which primarily rely on well-lit scenarios. The need for effective solutions to address the challenges posed by poor lighting conditions has never been more pronounced. This thesis embarks on a journey to explore innovative methodologies for video object detection and tracking in such challenging environments, aiming to enhance the reliability and efficiency of computer vision systems in real-world applications.

This thesis has thoroughly explored the complex domains of video object detection and tracking under poor lighting conditions. Through the introduction of innovative methodologies, including the proposed LLIE approach and the incorporation of the Fast RCNN with ResNet model for object detection, this research has showcased the potential of advanced computer vision techniques to address the challenges posed by challenging lighting environments. This potential has been further validated with the introduction of the OC Sort model for video object tracking. The application of these techniques has the capacity to revolutionize surveillance systems, autonomous vehicles, and numerous other fields, ensuring reliable and efficient object detection and tracking under adverse lighting conditions.

In Chapter 3, an innovative approach to LLIE was presented, involving a unique methodology for uneven illumination correction. Techniques like exponential transformation, hyperbolic tangent profiles, LIP, and logarithmic scaling functions

were employed to demonstrate a substantial improvement in image quality. The results, as indicated by metrics such as Lightness Order Error, PSNR Mean Squared Error, and Average Implementation Time, demonstrated the efficacy of the approach in addressing illumination challenges in imaging.

In Chapter 4, video object detection is explored using the Fast RCNN model integrated with the ResNet architecture. This chapter presents the methodology and outcomes of research in video object detection, with the potential and effectiveness of the selected model being demonstrated. The proposed model, the Fast RCNN with ResNet, takes center stage in this chapter. The model's architecture is detailed, explaining how video frames are processed, objects are identified, and contributions are made to the advancement of video object detection technology. A comprehensive understanding of the model's structure, features, and suitability for the task at hand is aimed to be provided. The practical outcomes of implementing the model are presented. The discussion encompasses the datasets used in the simulations, the model's performance in detecting objects within video streams, and the metrics used to evaluate its effectiveness. Through these results, the potential applications of the model in real-world scenarios are highlighted, offering insights into its strengths and areas for further improvement.

In Chapter 5, video object tracking is discussed using the OC Sort model. This chapter presents the methodology and outcomes of research in video object tracking, with the potential and effectiveness of the selected model being demonstrated. The proposed model, OC Sort, takes center stage in this chapter. The model's architecture is detailed, explaining how it tracks objects within video streams and contributes to the advancement of video object tracking technology. A comprehensive understanding of the model's structure, features, and its suitability for the task at hand is aimed to be provided. Through these results, the potential applications of the model in real-world scenarios are highlighted, offering insights into its strengths and areas for further improvement. In particular, the model demonstrates strong performance in tasks requiring precise object localization, making it suitable for various applications. However, to ensure robust deployment across diverse environments, future work will

focus on enhancing generalization capabilities, improving computational efficiency, and optimizing the model for real-time inference on resource-constrained devices.

The proposed method integrates the Fast-RCNN and OCSort algorithms into the YOLOv8 framework. The Fast-RCNN model utilizes the ResNet152 pre-trained model to extract relevant features from the input frame, improving the search process through the selective search algorithm. These extracted features are subsequently input into RoI pooling and fully connected layers to achieve precise object detection and bounding box assignment. Concurrently, the OCSort model is employed to ensure effective object tracking, even in cases of occlusions, maintaining consistent tracking performance across the video sequence. The developed object identification model exhibited an average accuracy of 93% for detecting persons, 88% for cars, and 43% for trucks. The tracking module yielded an average object tracking accuracy of 69%.

6.2 FUTURE SCOPE

The research undertaken in this thesis presents opportunities for further investigation and advancement in the domains of computer vision and image processing. Some of the promising directions include:

1. **Advanced Deep Learning Models:** The potential for even more accurate and efficient image enhancement, object detection, and tracking lies in continuously evolving deep learning architectures, which can be explored in future research.

2. **Real-time Applications:** Practical applications, such as surveillance, autonomous vehicles, and robotics, necessitate the development of real-time implementations for the proposed image enhancement, object detection, and tracking methods.

3. **Multi-Modal Integration:** Enhanced robustness and accuracy in object detection and tracking systems in challenging environments can be achieved

by combining data from various sensors, such as LiDAR, radar, and thermal imaging, with visual information.

4. **Semi-Supervised and Unsupervised Learning:** Future research can focus on the investigation of semi-supervised and unsupervised learning approaches for object detection and tracking, reducing the reliance on labeled data and increasing adaptability.

5. **Hardware Acceleration:** Optimizing algorithms for specific hardware, such as GPUs and TPUs, can lead to the significant enhancement of efficiency and speed in image enhancement, object detection, and tracking in resource-constrained environments.

6. **Cross-Domain Applications:** Valuable innovations and interdisciplinary collaborations can be achieved by expanding the application domains beyond surveillance to fields like medical imaging, wildlife monitoring, and industrial automation.

7. **Ethical and Privacy Considerations:** Addressing ethical and privacy concerns is vital as these technologies are integrated into various aspects of society. Future research should focus on the development of safeguards and guidelines for responsible use.

LIST OF PUBLICATIONS

[1] Kumar, Chinthakindi Kiran, Gaurav Sethi, and Kirti Rawal. "Adapting to the Dark: A Novel Adaptive Low Light Illumination Correction Algorithm for Video Sequences in Wireless Communications."

[2] Kumar, Chinthakindi Kiran, Gaurav Sethi, and Kirti Rawal. "An Effective Approach for Object Detection Using Deep Convolutional Networks." *2022 International Conference on Breakthrough in Heuristics And Reciprocation of Advanced Technologies (BHARAT)*. IEEE, 2022.

[3] Kumar, Chinthakindi Kiran, Gaurav Sethi, and Kirti Rawal. "Deep Network Architectures for Object Detection and Tracking: A Review." *Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2022* (2023): 117-128.

[4] Kumar, Chinthakindi Kiran, Gaurav sethi and Kirti Rawal. "A Brief Study on Object Detection and Tracking." *Journal of Physics: Conference Series*. Vol. 2327. No. 1. IOP Publishing, 2022.

[5] Kumar, Chinthakindi Kiran, Gaurav Sethi, and Kirti Rawal. "42 Motion detection and tracking of surveillance videos under distorted environments." *Intelligent Circuits and Systems* (2021): 267.

[6] Kumar, Chinthakindi Kiran, Gaurav Sethi, and Kirti Rawal. "Advancements in Object Detection and Tracking Techniques: A Comprehensive Review in Variable Illumination Conditions." *Intelligent Circuits and Systems for SDG 3–Good Health and well-being*: 628-635.

REFERENCES

[1] . Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., Lim, J., Yang, M., & Lyu, S. (2015). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.*, 193, 102907.

[2] . Lee, D. (2020). CNN-based single object detection and tracking in videos and its application to drone detection. *Multimedia Tools and Applications*, 80, 34237 - 34248.

[3] . Pérez-Hernández, F., Tabik, S., Lamas, A., Olmos, R., Fujita, H., & Herrera, F. (2020). Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. *Knowl. Based Syst.*, 194, 105590.

[4] . Sreenu, G., & Saleem Durai, M.A. (2019). Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6.

[5] . Fernández-Sanjurjo, M., Bosquet, B., Mucientes, M., & Brea, V.M. (2019). Real-time visual detection and tracking system for traffic monitoring. *Eng. Appl. Artif. Intell.*, 85, 410-420.

[6] . Wang, W., Wu, X., Yuan, X., & Gao, Z. (2020). An Experiment-Based Review of Low-Light Image Enhancement Methods. *IEEE Access*, 8, 87884-87917.

[7] . Zhang, Q., Huang, N., Yao, L., Zhang, D., Shan, C., & Han, J. (2019). RGB-T Salient Object Detection via Fusing Multi-Level CNN Features. *IEEE Transactions on Image Processing*, 29, 3321-3335.

[8] . Wang, J., Song, K., Bao, Y., Huang, L., & Yan, Y. (2021). CGFNet: Cross-Guided Fusion Network for RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32, 2949-2961.

[9] . Avola, D., Cinque, L., Diko, A., Fagioli, A., Foresti, G.L., Mecca, A., Pannone, D., & Piciarelli, C. (2021). MS-Faster R-CNN: Multi-Stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images. *Remote. Sens.*, 13, 1670.

[10] . Mhalla, A., Chateau, T., & Amara, N.E. (2019). Spatio-temporal object detection by deep learning: Video-interlacing to improve multi-object tracking. *Image Vis. Comput.*, 88, 120-131.

[11] . Ji, Y., Zhang, H., Zhang, Z., & Liu, M. (2021). CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Inf. Sci.*, 546, 835-857.

[12] . Wang, Y., Choi, J., Zhang, K., Huang, Q., Chen, Y., Lee, M., & Kuo, C.J. (2020). Video object tracking and segmentation with box annotation. Signal Process. *Image Commun.*, 85, 115858.

[13] . Sun, L., Zhao, S., Li, G., & Liu, B. (2019). High accuracy object detection via bounding box regression network. *Frontiers of Optoelectronics*, 12, 324 - 331.

[14] . Xu, Y., Zhou, X., Chen, S., & Li, F. (2019). Deep learning for multiple object tracking: a survey. *IET Comput. Vis.*, 13, 355-368.

[15] . Hazra, S., Mandal, S., Saha, B., & Khatua, S. (2022). UMTSS: a unifocal motion tracking surveillance system for multi-object tracking in videos. *Multimedia Tools and Applications*, 82, 12401-12422.

[16] . Wawrzyniak, N., Hyla, T., & Popik, A. (2019). Vessel Detection and Tracking Method Based on Video Surveillance. *Sensors (Basel, Switzerland*), 19.

[17] . Wang, L.X., Shu, X., Zhang, W., & Chen, Y. (2019). Design and Optimization of Evaluation Metrics in Object Detection and Tracking for Low-Altitude Aerial Video. *ICBDS*.

[18] . Avşar, E., & Avşar, Y.Ö. (2022). Moving vehicle detection and tracking at roundabouts using deep learning with trajectory union. *Multimedia Tools and Applications*, 81, 6653 - 6680.

[19] . Maltezos, E., Douklias, A., Dadoukis, A., Misichroni, F., Karagiannidis, L., Antonopoulos, M., Voulgary, K., Ouzounoglou, E., & Amditis, A.J. (2021). The INUS Platform: A Modular Solution for Object Detection and Tracking from UAVs and Terrestrial Surveillance Assets. *Comput.*, 9, 12.

[20] . Sun, S., Akhtar, N., Song, X., Song, H., Mian, A.S., & Shah, M. (2020). Simultaneous Detection and Tracking with Motion Modelling for Multiple Object Tracking. *European Conference on Computer Vision.*

[21] . Rakotoniaina, Z.A., Chelbi, N.E., Gingras, D., & Faulconnier, F. (2023). LIV-DeepSORT: Optimized DeepSORT for Multiple Object Tracking in Autonomous Vehicles Using Camera and LiDAR Data Fusion. 2023 IEEE Intelligent Vehicles Symposium (IV), 1-7.

[22] . Yang, D., Zhang, S., Wang, S., Yu, Q., Su, Z., & Zhang, D. (2022). Real-time illumination adjustment for video deflectometers. *Structural Control and Health Monitoring*, 29.

[23] . Liao, B., Hu, J., & Gilmore, R.O. (2021). Optical flow estimation combining with illumination adjustment and edge refinement in livestock UAV videos. *Comput. Electron. Agric.*, 180, 105910.

[24] . Gómez, P., Semmler, M., Schützenberger, A., Bohr, C., & Döllinger, M. (2019). Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network. *Medical & Biological Engineering & Computing*, 57, 1451 - 1463.

[25] . Zhang, Q., Nie, Y., Zhu, L., Xiao, C., & Zheng, W. (2019). Enhancing Underexposed Photos Using Perceptually Bidirectional Similarity. *IEEE Transactions on Multimedia*, 23, 189-202.

[26] . Li, C., Guo, C., Han, L., Jiang, J., Cheng, M., Gu, J., & Loy, C.C. (2021). Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 9396-9416.

[27] . Guo, Y., Lu, Y., & Liu, R.W. (2021). Lightweight deep network-enabled real-time low-visibility enhancement for promoting vessel detection in maritime video surveillance. *Journal of Navigation*, 75, 230 - 250.

[28] . Huang, J., Fu, X., Xiao, Z., Zhao, F., & Xiong, Z. (2023). Low-Light Stereo Image Enhancement. *IEEE Transactions on Multimedia*, 25, 2978-2992.

[29] . He, Y., Yang, P., Qin, T., & Zhang, N. (2023). End-Edge Coordinated Joint Encoding and Neural Enhancement for Low-Light Video Analytics. GLOBECOM 2023 - 2023 *IEEE Global Communications Conference*, 7363-7368.

[30] . Guo, Y., Lu, Y., Liu, R.W., Yang, M., & Chui, K.T. (2020). Low-Light Image Enhancement With Regularized Illumination Optimization and Deep Noise Suppression. *IEEE Access,* 8, 145297-145315.

[31] . Li, J., Feng, X., & Hua, Z. (2021). Low-Light Image Enhancement via Progressive-Recursive Network. *IEEE Transactions on Circuits and Systems for Video Technology,* 31, 4227-4240.

[32] . Liu, M., Cui, Y., Ren, W., Zhou, J., & Knoll, A.C. (2025). LIEDNet: A Lightweight Network for Low-light Enhancement and Deblurring. *IEEE Transactions on Circuits and Systems for Video Technology.*

[33] . Ayoub, A., El-shafai, W., El-Samie, F.E., Hamad, E.K., & El-Rabaie, S. (2025). Video and image quality enhancement using an enhanced lower bound on transmission map dehazing technique. *Multim. Syst.*, 31, 158.

[34] . Jiang, J., Xu, Y., Cao, B., & Xiao, C. (2025). Darkseg: a lightweight and edge-optimized network for nighttime semantic segmentation. *Signal, Image and Video Processing.*

[35] . Yue, H., He, C., Wang, L., Yu, B., Yin, X., Zhou, Z., & Yang, J. (2025). Staggered HDR video reconstruction with a real-world benchmark dataset for night scenes. *Displays,* 88, 103029.

[36] . Zhu, H., Wei, H., Li, B., Yuan, X., & Kehtarnavaz, N. (2020). A Review of Video Object Detection: Datasets, Metrics and Methods. *Applied Sciences,* 10, 7834.

[37] . Han, M., Wang, Y., Chang, X., & Qiao, Y. (2020). Mining Inter-Video Proposal Relations for Video Object Detection. *European Conference on Computer Vision*.

[38] . Dasiopoulou, S., Mezaris, V., Kompatsiaris, Y., Papastathis, V., & Strintzis, M.G. (2005). Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology,* 15, 1210-1224.

[39] . Fan, L., Zhang, T., & Du, W. (2021). Optical-flow-based framework to boost video object detection performance with object enhancement. *Expert Syst. Appl.*, 170, 114544.

[40] . Jiang, Z., Liu, Y., Yang, C., Liu, J., Gao, P., Zhang, Q., Xiang, S., & Pan, C. (2019). Learning Where to Focus for Efficient Video Object Detection. *European Conference on Computer Vision.*

[41] . Xu, Z., Hrustic, E., & Vivet, D. (2020). CenterNet Heatmap Propagation for Real-Time Video Object Detection. *European Conference on Computer Vision.*

[42] . Lu, S., Wang, B., Wang, H., Chen, L., Ma, L., & Zhang, X. (2019). A real-time object detection algorithm for video. *Comput. Electr. Eng.,* 77, 398-408.

[43] . Yao, C., Fang, C., Shen, X., Wan, Y., & Yang, M. (2020). Video Object Detection via Object-Level Temporal Aggregation. *European Conference on Computer Vision.*

[44] . Yuan, Z., Song, X., Bai, L., Wang, Z., & Ouyang, W. (2021). Temporal-Channel Transformer for 3D Lidar-Based Video Object Detection for Autonomous Driving. *IEEE Transactions on Circuits and Systems for Video Technology,* 32, 2068-2078.

[45] . Wu, Y., Zhang, H., Li, Y., Yang, Y., & Yuan, D. (2020). Video Object Detection Guided by Object Blur Evaluation. *IEEE Access,* 8, 208554-208565.

[46] . Jha, S., Seo, C., Yang, E., & Joshi, G.P. (2020). Real time object detection and trackingsystem for video surveillance system. *Multimedia Tools and Applications,* 80, 3981 - 3996.

[47] . Subudhi, B.N., Nanda, P.K., & Ghosh, A. (2011). A Change Information Based Fast Algorithm for Video Object Detection and Tracking. *IEEE Transactions on Circuits and Systems for Video Technology,* 21, 993-1004.

[48] . Wang, W., Shen, J., & Shao, L. (2017). Video Salient Object Detection via Fully Convolutional Networks. *IEEE Transactions on Image Processing,* 27, 38-49.

[49] . Brazil, G., Pons-Moll, G., Liu, X., & Schiele, B. (2020). Kinematic 3D Object Detection in Monocular Video. *European Conference on Computer Vision.*

[50] . Perreault, H., Bilodeau, G., Saunier, N., & H'eritier, M. (2021). FFAVOD: Feature Fusion Architecture for Video Object Detection. *Pattern Recognit. Lett.,* 151, 294-301.

[51] . Han, L., Wang, P., Yin, Z., Wang, F., & Li, H. (2022). Class-Aware Feature Aggregation Network for Video Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32, 8165-8178.

[52] . Lyu, Y., Yang, M.Y., Vosselman, G., & Xia, G. (2021). Video object detection with a convolutional regression tracker. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176, 139-150.

[53] . Yazdi, M., & Bouwmans, T. (2018). New trends on moving object detection in video images captured by a moving camera: *A survey. Comput. Sci. Rev.*, 28, 157-177.

[54] . Yin, J., Shen, J., Gao, X., Crandall, D.J., & Yang, R. (2021). Graph Neural Network and Spatiotemporal Transformer Attention for 3D Video Object Detection from Point Clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 1-1.

[55] . Ren, S., Han, C., Yang, X., Han, G., & He, S. (2020). TENet: Triple Excitation Network for Video Salient Object Detection. *ArXiv, abs*/2007.09943.

[56] . Oreifej, O., Li, X., & Shah, M. (2013). Simultaneous Video Stabilization and Moving Object Detection in Turbulence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 450-462.

[57] . Chen, M., Chen, M., Han, X., Zhang, H., Lin, G., & Kamruzzaman, M.M. (2019). Quality-guided key frames selection from video stream based on object detection. *J. Vis. Commun. Image Represent.*, 65.

[58] . Tu, Z., Guo, Z., Xie, W., Yan, M., Veltkamp, R.C., Li, B., & Yuan, J. (2017). Fusing disparate object signatures for salient object detection in video. *Pattern Recognit.*, 72, 285-299.

[59] . Cong, R., Song, W., Lei, J., Yue, G., Zhao, Y., & Kwong, S. (2022). PSNet: Parallel Symmetric Network for Video Salient Object Detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7, 402-414.

[60] . Murugan, A.S., SuganyaDevi, K., Sivaranjani, A., & Srinivasan, P. (2018). A study on various methods used for video summarization and moving object detection for video surveillance applications. *Multimedia Tools and Applications*, 77, 23273-23290.

[61] . Tatana, M. M., Tsoeu, M. S., & Maswanganyi, R. C. (2025). Low-Light Image and Video Enhancement for More Robust Computer Vision Tasks: A Review. *Journal of Imaging*, *11*(4), 125.

[62] . Jia, Z., Wang, C., Wang, Y., Gao, X., Li, B., Yin, L., & Chen, H. (2025). *Recent Research Progress of Graph Neural Networks in Computer Vision. Electronics*, 14(9), 1742.

[63] . Du, L., Ma, C., Chen, J., Zheng, H., Nie, X., & Gao, Z. (2025). Survey on deep learning-based weakly supervised salient object detection. *Expert Systems with Applications*, 281, 127497.

[64] . Jamali, M., Davidsson, P., Khoshkangini, R., Ljungqvist, M.G., & Mihailescu, R. (2025). Context in object detection: a systematic literature review. *ArXiv, abs*/2503.23249

[65] . Ciaparrone, G., Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2019). Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381, 61-88.

[66] . Zhao, M., Li, S., Xuan, S., Kou, L., Gong, S., & Zhou, Z. (2022). SatSOT: A Benchmark Dataset for Satellite Video Single Object Tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-11.

[67] . Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Harshit, Huang, M., Liu, J., Xu, Y., Liao, C., Yuan, L., & Ling, H. (2020). LaSOT: A High-quality Large-scale Single Object Tracking Benchmark. *International Journal of Computer Vision*, 129, 439 - 461.

[68] . Dike, H.U., & Zhou, Y. (2021). A Robust Quadruplet and Faster Region-Based CNN for UAV Video-Based Multiple Object Tracking in Crowded Environment. *Electronics*, 10, 795.

[69] . Micheal, A.A., & Vani, K. (2022). Deep Learning-Based Multi-class Multiple Object Tracking in UAV Video. *Journal of the Indian Society of Remote Sensing*, 50, 2543-2552.

[70] . Fischer, T., Pang, J., Huang, T.E., Qiu, L., Chen, H., Darrell, T., & Yu, F. (2022). QDTrack: Quasi-Dense Similarity Learning for Appearance-Only Multiple Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 15380-15393.

[71] . Wu, D., Song, H., & Fan, C. (2022). Object Tracking in Satellite Videos Based on Improved Kernel Correlation Filter Assisted by Road Information. *Remote. Sens*., 14, 4215.

[72] . Liu, Q., Li, X., He, Z., Fan, N., Yuan, D., & Wang, H. (2019). Learning Deep Multi-Level Similarity for Thermal Infrared Object Tracking. *IEEE Transactions on Multimedia,* 23, 2114-2126.

[73] . Wang, X., Hu, Y.H., Radwin, R.G., & Lee, J.D. (2019). Temporal Frame Sub-Sampling for Video Object Tracking. *Journal of Signal Processing Systems,* 92, 569 - 581.

[74] . Noor, S., Waqas, M., Saleem, M.I., & Minhas, H.N. (2021). Automatic Object Tracking and Segmentation Using Unsupervised SiamMask. *IEEE Access,* 9, 106550-106559.

[75] . Xia, R., Chen, Y., & Ren, B. (2022). Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter. *J. King Saud Univ. Comput. Inf. Sci.*, 34, 6008-6018.

[76] . Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2021). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *ArXiv, abs/*2110.06864.

[77] . Jiang, M., Li, R., Liu, Q., Shi, Y., & Tlelo-Cuautle, E. (2021). High speed long-term visual object tracking algorithm for real robot systems. *Neurocomputing*, 434, 268-284.

[78] . Chen, Y., Wang, J., Xia, R., Zhang, Q., Cao, Z., & Yang, K. (2019). RETRACTED ARTICLE: The visual object tracking algorithm research based on adaptive combination kernel. *Journal of Ambient Intelligence and Humanized Computing,* 10, 4855 - 4867.

[79] . Harley, A.W., Fang, Z., & Fragkiadaki, K. (2022). Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. *European Conference on Computer Vision.*

[80] . Wu, R., Wen, X., Yuan, L., & Xu, H. (2022). DASFTOT: Dual attention spatiotemporal fused transformer for object tracking. Knowl. *Based Syst*., 256, 109897.

[81] . Lu, X., Ma, C., Ni, B., & Yang, X. (2021). Adaptive Region Proposal With Channel Regularization for Robust Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 31, 1268-1282.

[82] . Xu, Z., Zhang, W., Tan, X., Yang, W., Huang, H., Wen, S., Ding, E., & Huang, L. (2020). Segment as Points for Efficient Online Multi-Object Tracking and Segmentation. *ArXiv, abs/*2007.01550.

[83] . Zhang, J., Sun, J., Wang, J., Li, Z., & Chen, X. (2022). An object tracking framework with recapture based on correlation filters and Siamese networks. *Comput. Electr. Eng.*, 98, 107730.

[84] . Xuan, S., Li, S., Han, M., Wan, X., & Xia, G. (2020). Object Tracking in Satellite Videos by Improved Correlation Filters With Motion Estimations. *IEEE Transactions on Geoscience and Remote Sensing*, 58, 1074-1086.

[85] . Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., & Lu, H. (2022). Towards Grand Unification of Object Tracking. *European Conference on Computer Vision.*

[86] . Kim, D.Y., Vo, B., Vo, B., & Jeon, M. (2016). A labeled random finite set online multi-object tracker for video data. *Pattern Recognit.*, 90, 377-389.

[87] . Du, B., Cai, S., & Wu, C. (2019). Object Tracking in Satellite Videos Based on a Multiframe Optical Flow Tracker. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12, 3043-3055.

[88] . Zhou, X., Koltun, V., & Krähenbühl, P. (2020). Tracking Objects as Points. *ArXiv, abs/*2004.01177.

[89] . Keh, J.J., Cruz, M.D., Rivera, M., Jose, J.A., Sybingco, E., Dadios, E.P., Madria, W., & Miguel, A. (2020). AutoTrack: Interactive Visual Object Tracking for Efficient Object Annotations. 2020 *IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 1-4.

[90] . Ren, S. (2025). Multi-Camera Association Tracking Algorithm for Pedestrian Target Based on Difference Image. *Systems and Soft Computing*, 200282.

[91] . Shajeena, J., Shiny, R. M., Bini Palas, P., Mary Vespa, M., Stanley, B. F., & Jeen Retna Kumar, R. (2025). Siamese Deep Q-Learning Based Online

Correlation Filter Adaptation for Visual Object Tracking in Complex Scenarios. *Circuits, Systems, and Signal Processing*, 1-44.

[92] . Zebarjadi, M., Organ, A. J., Zachs, D. P., & Lim, H. H. (2025). Hybrid KLT-LSTM Tracking for Robust Organ Motion Monitoring in 2D Ultrasound-Guided End-Organ Therapies. *IEEE Transactions on Biomedical Engineering*.

[93] . Al-Ameen, Z. (2019). Nighttime image enhancement using a new illumination boost algorithm. *IET Image Process.*, 13, 1314-1320.

[94] . Wang, Y., Jodoin, P., Porikli, F.M., Konrad, J., Benezeth, Y., & Ishwar, P. (2014). CDnet 2014: An Expanded Change Detection Benchmark Dataset. 2014 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 393-400.

[95] . Momeni Pour, A., Seyedarabi, H., Abbasi Jahromi, S.H., & Javadzadeh, A. (2020). Automatic Detection and Monitoring of Diabetic Retinopathy Using Efficient Convolutional Neural Networks and Contrast Limited Adaptive Histogram Equalization. *IEEE Access*, 8, 136668-136673.

[96] . Fu, G., Duan, L., & Xiao, C. (2019). A Hybrid L2 −LP Variational Model For Single Low-Light Image Enhancement With Bright Channel Prior. 2019 *IEEE International Conference on Image Processing (ICIP)*, 1925-1929.

[97] . Shi, Y., Wu, X., & Zhu, M. (2019). Low-light Image Enhancement Algorithm Based on Retinex and Generative Adversarial Network. *ArXiv, abs*/1906.06027.

[98] . Ren, Y., Ying, Z., Li, T.H., & Li, G. (2019). LECARM: Low-Light Image Enhancement Using the Camera Response Model. *IEEE Transactions on Circuits and Systems for Video Technology*, 29, 968-981.

[99] . Al-Hashim, M.A., & Al-Ameen, Z. (2020). Retinex-Based Multiphase Algorithm for Low-Light Image Enhancement. *Traitement du Signal*, 37, 733-743.

[100] . Pramanik, A., Pal, S.K., Maiti, J., & Mitra, P. (2022). Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6, 171-181.

[101] . Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., & Meng, H. (2022). StrongSORT: Make DeepSORT Great Again. *IEEE Transactions on Multimedia*, 25, 8725-8737.

[102] . Aharon, N., Orfaig, R., & Bobrovsky, B. (2022). BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *ArXiv, abs*/2206.14651.