

Machine Learning Approaches to Intelligent Road Transport Systems

**Thesis submitted by
Srimanta Kundu**

DOCTOR OF PHILOSOPHY (Engineering)

**Department of Computer Science and Engineering,
Faculty Council of Engineering & Technology,
Jadavpur University
Kolkata, India
2024**

JADAVPUR UNIVERSITY
KOLKATA-700032, INDIA

INDEX. NO. 120/19/E

1. Title of the Thesis:

Machine learning Approaches to Intelligent Road Transport Systems

2. Name, Designation & Institution of the Supervisor:

(a) Prof. Ujjwal Maulik

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032, India

List of Publications

Papers in Journals and Transactions

1. S Kundu, U Maulik, R Sheshanarayana, and S Ghosh. "Vehicle Smoke Synthesis and Attention-Based Deep Approach for Vehicle Smoke Detection." IEEE Transactions on Industry Applications 59, no. 2 (2022): 2581-2589.
DOI: 10.1109/TIA.2022.3227532
2. U Maulik, and S Kundu. "Automatic Vehicle Pollution Detection using Feedback based Iterative Deep Learning." IEEE Transactions on Intelligent Transportation Systems 24, no. 5 (2023): 4804-4814.
DOI: 10.1109/TITS.2023.3239190
3. S Kundu, and U Maulik. "Passenger Surveillance Using Deep Learning in Post-COVID-19 Intelligent Transportation System." Transactions of the Indian National Academy of Engineering 7, no. 3 (2022): 927-941.
DOI: <https://doi.org/10.1007/s41403-022-00338-y>
4. S Kundu, and U Maulik. "Cloud deployment of game theoretic categorical clustering using Apache spark: an application to car recommendation." Machine Learning with Applications 6 (2021): 100100.
DOI: <https://doi.org/10.1016/j.mlwa.2021.100100>
5. S Kundu, U Maulik, and A Mukhopadhyay. "A game theory-based approach to fuzzy clustering for pixel classification in remote sensing imagery." Soft Computing 25, no. 7 (2021): 5121-5129.
DOI: <https://doi.org/10.1007/s00500-020-05514-2>
6. S Kundu, R De, S Bhattacharya, and U Maulik. "Automatic License Plate Recognition using Multi-scale Dual GAN." (In Process to communicate).

Indian Patent List

1. S Kundu, and U Maulik. "Method and device for determining safety condition of a place of interest remotely from vehicle infotainment based on IoT", Indian Patent. Application number: 202031033675 (Published)
2. U Maulik, and S, Kundu. "Device And Method for End-To-End Medical Safety in Social Shared Vehicles for Passengers and Drivers", Indian Patent. Application number: 202131013483 (Published)

Papers in Conference Proceedings

1. S Kundu, U B Maulik, A Bej, and U Maulik. "Deep Learning based Pollution Detection in Intelligent Transportation System." In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), pp. 292-297. IEEE, 2020.
DOI: 10.1109/ICCCA49541.2020.9250883
2. S Kundu, and U Maulik. "Vehicle pollution detection from images using deep learning." Intelligence Enabled Research: DoSIER 2019 (2020): 1-5. Springer, Singapore.
DOI: 10.1007/978-981-15-2021-1_1

PROFORMA - 1
STATEMENT OF ORIGINALITY

I, **SRIMANTA KUNDU** registered on **03.06.2019**, do hereby declare that this thesis entitled "**Machine Learning Approaches to Intelligent Road Transport Systems**" contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked the thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is 3%.

Srimanta Kundu

Signature of Candidate (Srimanta Kundu)

Date: 15/05/24

Professor
Computer Sc. & Engg. Department
Jadavpur University
Kolkata-700032

Ujjwal Maulik 15/5/24

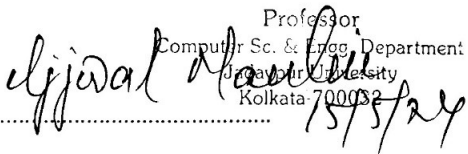
Certified by Supervisor (Prof. Ujjwal Maulik)
(Signature with date, seal)

PROFORMA – 2

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **Machine learning Approaches to Intelligent Road Transport Systems** submitted by Srimanta Kundu, who got his name registered on 3rd June, 2019 for the award of Ph.D. (Engineering) degree of Jadavpur University, is absolutely based upon his own work under the supervision of **Prof. Ujjwal Maulik, Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India**, and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

Professor
Computer Sc. & Engg. Department
Jadavpur University
Kolkata-700032



.....

Prof. Ujjwal Maulik
Signature of the Supervisor
and date with Office Seal

Dedication

To my family and supervisor

Acknowledgements

I would like to express my deepest appreciation and gratitude to my supervisor, Prof. Ujjwal Maulik, for his unwavering support, guidance, and mentorship throughout the journey of completing my PhD thesis. Without his invaluable expertise and constant encouragement, this work would not have been possible. After a significant break in my academic journey, when I made the decision to pursue a PhD degree, it was Prof. Ujjwal Maulik who greatly motivated and encouraged me to make this choice. I am really fortunate to have the opportunity to work under his supervision.

I wish to extend my sincere appreciation to Sayantari di and Saumik, for sharing their knowledge, guiding me in every research phase, and refining my ideas and methods. Their contribution to transform me, from a good developer to a complete research oriented academecianis has been immense. I know their patience level while reviewing my unprofessional writing in the initial days. I can still remember the constructive and critical feedbacks shared during a long online meeting after their hectic schedules. Their involvement and dedication have significantly enriched the quality of this thesis.

To my mother, father, sisters, brother-in-laws, and mother-in-law, all their values, ethics they taught me has helped to explore the absolute best in me. Their ethics and morale to life has guided me in this path as well. I would like to extend my heartfelt appreciation to them who believed in me throughout this journey. Their continous faith and support have been a constant source of inspiration and motivation. With a touch of fun, they persistently brought the fact to my attention "When can we expect you to wrap up your studies?"

Finally, I wish to express my deepest gratitude to my wife, Sayanshree. Her constant support has been the foundation of my success. She displayed incredible courage by shouldering the household responsibilities while I struggled with various family commitments. I'm deeply grateful for the extra responsibilities she took on, her patience during busy weekends and late-night study sessions, and her continuous assistance during crucial decision-making moments. This thesis is a tribute to her commitment and backing. In closing, I'd like to mention my seven-month-old son, Leo, who serves as a constant reminder of my responsibilities while drafting this thesis. I've made a promise to him that one day, I'll share the story of my journey; but for now, my focus will be on nurturing his growth and upcoming education. I must complete my PhD before that.

Srimanta Kundu.

(Srimanta Kundu)

Date: 15/05/24.

Contents

List of Figures	xvi
List of Tables	xix
Abstract	xxi
1 Introduction and Scope of the Thesis	1
1.1 Machine Learning for Road Transport	1
1.2 Fundamentals of Explored Deep Learning in the Research	2
1.2.1 Experimented Deep Learning database, and models	3
1.2.2 Attention and Multi-head Attention	5
1.2.3 Transfer Learning	6
1.2.4 Experimented Loss Functions	6
1.2.5 Explored Metrics for evaluation	7
1.3 Literature Survey on IRTS	8
1.4 Thesis overview and contributions	14
1.4.1 Synopsis and Contribution of Chapter 2	14
1.4.2 Synopsis and Contribution of Chapter 3	16
1.4.3 Synopsis and Contribution of Chapter 4	18
1.4.4 Synopsis and Contribution of Chapter 5	18
1.5 Summary	19
2 Smokey Vehicle Recognition	21
2.1 Overview of Smokey Vehicle Recognition	21
2.2 Vehicle smoke relation with the air pollution	21
2.3 Data Set Preparation	23
2.4 Methodology	27
2.5 Experimental Setup	29
2.6 Experimental Findings	31
2.7 Summary and Future Directions	39
3 Vehicle Smoke Embedding and Smokey Region Detection	41
3.1 An overview of smoke embedding and smokey region Detection	41
3.2 Smoke Generation	42
3.2.1 Statistics of the Dataset used in the experiment	43
3.2.2 Proposed Smoke Synthesis Algorithm	43
3.2.3 Validation of Smoke Synthesis	45
3.3 Vehicle Smoke detection	47
3.3.1 Proposed attention based object detector network	48
3.3.2 Experimental Setup	50

CONTENTS

3.3.3	Results	51
3.4	Summary and Scope	53
4	Automatic License Plate Recognition	55
4.1	An overview of Automatic License Plate Recognition	55
4.2	Proposed Methodology and Architecture	57
4.2.1	Generator Network	58
4.2.2	Discriminator Network	59
4.2.3	Defined Loss Function in GAN network	61
4.3	Experimental Setup and Result	64
4.3.1	Ground truth generation and data augmentation	64
4.3.2	License Plate experiment	64
4.3.3	Cross Dataset Evaluation	67
4.4	Summary and Future Prospects	69
5	Safety in Intelligent Road Transport System during Pandemic	71
5.1	Pandemic Era Road Transport Safety: An Overview	71
5.2	Passenger Surveillance during COVID period	72
5.3	Mask Data set Preparation	73
5.4	Experimental Setup	74
5.4.1	Model of Surveillance Camera Setup	74
5.4.2	Deep Model training Setup	75
5.5	Proposed Methodology	76
5.6	Results and Evaluations	80
5.7	Discussion and Future Prospects	85
6	Conclusions and Future Scope of Research	87
	Bibliography	91

List of Figures

2.1	Vehicle Pollution DATASET ₁	24
2.2	Night mode image generation	26
2.3	Transfer Learning Flow	29
2.4	Proposed TL Flow framework	29
2.5	Comparison of training loss	31
2.6	Comparison of training accuracy	32
2.7	Increase of training data	32
2.8	Testing accuracy of prediction	34
2.9	Confusion Matrix with confidence level ≥ 0.5	34
2.10	Confusion Matrix with confidence level ≥ 0.9	35
2.11	Average testing accuracy of models	35
2.12	Saliency Map	37
2.13	Successful identification in next iterations	37
2.14	Comparison of training time	38
2.15	Proposed testing framework	38
3.1	Smoke synthesis flow	42
3.2	Sample synthesized smoky images	43
3.3	Samples of realistic web-collected smokes	44
3.4	Ellipses fitting to define the boundary	44
3.5	Sample mask patterns	44
3.6	Flow chart of automatic vehicle smoke generation.	46
3.7	User feedback on smoke generation	46
3.8	The proposed architecture of deep network	48
3.9	Comparison of validation loss	50
4.1	The overall flow of proposed ALPR	57
4.2	Overall architecture of proposed BMDDNet	58
4.3	Generator with Multi-Scale Block	58
4.4	Multi-Scale Architecture	60
4.5	Patch Discriminator Network.	62
4.6	Image Discriminator Network	63
4.7	Sample images and ground truth produced by BMDDNet	66
4.8	Sample Visual Comparison on two data sets	67
4.9	Visual Comparison in bright light condition	67
4.10	Visual Comparison in low light condition	68
4.11	Sample Visuals on Media Lab data set	68
5.1	Comparison of data distribution	73
5.2	Sample images of our data set	73

LIST OF FIGURES

5.3	Surveillance camera (Gantry type) position	75
5.4	passengers' position while sitting inside vehicles	76
5.5	Proposed Transfer Learning (TL) based DNN	77
5.6	Testing flow of the proposed framework	78
5.7	Sample saliency-map images	78
5.8	Accuracy and loss curve during training and validation	79
5.9	Statistical comparison through Box-plot	80
5.10	Confidence level comparison with SSDMNv2	81
5.11	Confidence level comparison with ResNet-SVM	82
5.12	Sample visual comparison	82
5.13	Testing flow	83
5.14	Comparison of two best-performed existing techniques and proposed model	84
5.15	Sample cross data set test outcomes	85
5.16	The time comparison graph	85

List of Tables

1.1	Face-Mask detection Algorithms in the literature	14
2.1	Image augmentations	23
2.2	Day-Night image statistic	25
2.3	Synthetic night image count vs Accuracy	27
2.4	Comparison of test accuracy	36
2.5	Comparison of accuracy in cross data set	36
3.1	Image count statistics of different data sets	43
3.2	Different backbone layers of the proposed network.	49
3.3	Cross data set results	51
3.4	Overall real test results	52
4.1	Results over Open ALPR data set	65
4.2	Results over Open UFPR data set	65
5.1	Statistical comparison with data set of Prajna	74
5.2	Wilcoxon hypothesis comparison	79
5.3	Numerical Index (Average Weighted) comparison	81
5.4	Comparison of results for unusual conditions	82

Abstract

Various novel innovations and the integration of artificial intelligence (AI) components have significantly reshaped Intelligent Road Transport Systems (IRTS) in recent times. Among different challenges in IRTS, environmental issue, primarily originating from vehicle-emitted pollutants, pose a significant concern nowadays. Utilizing advanced devices like sensors, AI-cameras, and data analytics chips, these advanced systems monitor and evaluate air quality across road networks. Through the incorporation of real-time pollution detection mechanisms into transportation infrastructure, IRTS not only pinpoint pollutant sources but also play a important role in formulating impactful mitigation strategies. These systems empower authorities to swiftly address pollution incidents, enact traffic management interventions, and help to execute proper action in real-time. Through the collaboration of AI and transportation engineering, IRTS enhance our ability to create healthier and more sustainable urban environments by actively combating the adverse effects of pollution on both human health and the ecosystem.

Another important aspect in the revolutionized realm of traffic management is to apply law enforcement through their sophisticated Automatic License Plate Recognition (ALPR) capabilities. By deploying cutting-edge technologies such as machine learning (ML) solutions, and computer vision frameworks, IRTS can efficiently and accurately identify and process license plate information in real-time. This functionality proves invaluable for law enforcement agencies in monitoring traffic, managing parking, and ensuring public safety. The integration of ALPR into IRTS facilitates the automation of many tasks such as toll collection, parking management, handling anti-theft activity etc. This not only streamlines administrative processes but also enhances the overall efficiency and effectiveness of traffic control. The communication between IRTS and ALPR system represents a significant stride towards creating smarter and more secure transportation infrastructures.

In another domain, IRTS are playing a crucial role in handling inside-vehicle safety. These systems leverage advanced technologies such as GPS tracking, sensor deployment, and real-time data analysis to monitor and ensure the security of passengers using various AI-enabled processes. Even during the recent pandemic, IRTS has prompted the exploration of an entirely new research area focused on enhancing passenger life safety. Road transport played a significant role in the spread of COVID-like diseases. Monitoring the number of passengers automatically posed a considerable challenge to restraining such infectious diseases. Traffic agencies had to adapt rules governing passenger counts on roads to ensure a safer distance, providing a secure environment for both drivers and passengers. Amid this pandemic, it became crucial to enforce mandatory checks on mask-wearing status inside vehicles. Manual checking of the same status imposed significant

Abstract

hurdles. Numerous innovations in this area have profoundly transformed the landscape of IRTS during this crisis period.

While studying several IRTS innovations, object recognition and detection are two important and widely used techniques. Therefore, a nuanced comprehension and precise definition of these terms within the context of IRTS are imperative. Object detection, a facet of computer technology aligned with computer vision and image processing, involves identifying instances of semantic objects belonging to specific classes (such as humans, buildings, or cars) within digital images and videos. Object detection is crucial in IRTS as it enables real-time identification of vehicles, pedestrians, and obstacles, enhancing traffic management and safety. This technology facilitates efficient monitoring and control, optimizes traffic flow, and contributes to the overall effectiveness of IRTS innovations. Conversely, object recognition in the context of AI entails the capacity of AI implementations and systems to identify and categorize diverse entities. Recognition, in the same context, involves the identification of someone or something based on prior encounters or knowledge. The difference between object detection and object recognition becomes apparent in their distinct goals and outputs. Object detection not only involves the identification of objects but also the precise localization through bounding boxes, offering detailed information about their placement. In contrast, object recognition is primarily concerned with the higher-level understanding and categorization of objects within an image, obviating the necessity for precise localization.

The objective of this thesis is to introduce innovative concepts into the IRTS domain with the aim of enhancing system efficiency. The goal is to establish robust surveillance systems within road transport to provide a risk-free road environment. Implementing continuous automated monitoring and executing responsive actions significantly alleviates human effort as well. This represents an additional objective of the research endeavor. Another integral goal of the thesis is to automatically monitor on-road vehicle pollution and implement necessary measures, contributing to a pollution-free and eco-friendly environment. Ultimately, the thesis is determined by its commitment to ensure a safer road environment for all, aligning seamlessly with the objectives of smart transport.

Introduction and Scope of the Thesis

1.1 Machine Learning for Road Transport

Transport system is one of important the pillars of civilized society. The primary types of transport systems include roads, waterways, airports, and seaports. Among various communication processes, road transport has been the most frequently utilized way of communication from very ancient times. In present era, the road transport system has traversed through significant remodeling due to the development of advanced and cutting-edge technologies. This research field is inevitably complex, and the complexity arises from many sources. Transportation systems can involve factors like human activities, vehicles, conveyances, physical infrastructure, information technology etc., all interacting in complicated ways. However, the safety and security aspect has been considered the highest priority in road transport. With the rapid advancement of Artificial Intelligence (AI) technologies, the system has passed through notable evolution. The challenges associated with Intelligent Road Transport Systems (IRTS) incorporate various aspects, including technological, logistical, and societal considerations. Some of the key challenges include:

- **Traffic Management:** While IRTS aims to achieve smooth traffic flow, the transition from manual management to automatic maintained smart traffic and addressing several issues can be very complex. Determining vehicle identity, on-road safety handling, coordinating real-time data for effective traffic management etc. remains an ongoing challenge.
- **Environmental Impact:** Addressing the challenge of vehicle pollution requires a concerted effort to transition towards cleaner and more sustainable transportation technologies. The environmental impact of vehicular emissions highlights the urgency to implement eco-friendly practices and develop innovative solutions for a greener and healthier future.
- **Infrastructure Integration:** Implementing IRTS requires integrating new

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

technologies with existing transportation infrastructure. Renovating or upgrading legacy systems to accommodate advanced technologies can be costly and complex. Providing a low-cost solution is sometimes considered necessary.

- **Traffic and related object identification:** Major challenges in IRTS for object identification include dealing with complex environments, varied weather conditions, object occlusion, and ensuring real-time processing speed for accurate and timely decision-making.
- **Human Factors and Public Acceptance:** Automated systems may struggle to handle complex or unpredictable situations. Human judgment is essential for making decisions in scenarios that require subtle understanding, adaptability, and ethical considerations. Humans are better prepared to handle unforeseen circumstances, such as sudden changes in weather, road closures, emergencies, etc. At the same time, acceptance and understanding of IRTS among the citizens can be a hurdle. For example, wearing a mask while on the road during the pandemic time, obeying traffic rules, performing pollution certification on time, etc. need proper public awareness and adoption. Earning the trust and addressing concerns related to surveillance, data privacy, and the reliability of new technologies are crucial for extensive use of the same.

To develop a sustainable, effective, and socially acceptable IRTS, governments, industry stakeholders, researchers, and the general public must work together to address these challenges. In this context, the utilization of various fundamental ML frameworks, plays a vital role in persistently enhancing system efficiency.

1.2 Fundamentals of Explored Deep Learning in the Research

ML algorithms enable computers to ‘self-learn’ from historical instances and get better over time without having to be explicitly programmed. The fundamental goal of a ML system is to identify patterns in data and derive predictions by learning from existing data sets. In essence, machine learning algorithms and models acquire knowledge through experimental learning. While, one could think of DL algorithms [1] as an advanced and mathematically intricate development of ML algorithms. These days, the field is receiving a lot of attention. Recent advancements have produced outcomes that were previously deemed unattainable in terms of performance metrics.

Logically structured material is analyzed by DL algorithms in a manner akin to that of human conclusion-making. To achieve this, DL applications use a complex

1.2. FUNDAMENTALS OF EXPLORED DEEP LEARNING IN THE RESEARCH

layered structure within the algorithms which demands more processing resources compared to standard ML models. Being a subset of ML, DL is immensely used with very large data with a simpler framework to deploy any complex model [2, 3, 4]. DL along with Convolutional Neural Networks (CNNs) has been tremendously applied in several computer vision researches. The major advantage of DL is having the potential to support huge amounts of data during the learning. As an obvious impact, the accuracy will increase automatically by a considerable margin.

1.2.1 Experimented Deep Learning database, and models

In our research, we utilized several widely recognized deep learning models. We conducted multiple experiments using various popular computer vision databases. A brief description of those databases and the subsequent models has been provided in the following paragraphs.

ImageNet Database: ImageNet [5] is an image database categorized as per the WordNet hierarchy. Typically, the training data set was comprised of 1,000,000 images, with 50,000 for a validation set and 150,000 for a test set. This data set has been used by researchers in several articles and machine vision projects [6, 7].

Open ALPR Benchmark: Open ALPR Benchmark data set[8] has three different regions' vehicle number plates, i.e. Europe, Brazil and US. Total number of number plates present in this data set is 600.

UFPR: The 4,500 fully annotated photos (almost 30,000 License Plate (LP) characters) from 150 autos in actual driving situations are included in the UFPR data set¹ [9]. All the available images are of size $1,920 \times 1,080$ pixels.

Media Lab LPR: This LPR data set² has been created by National Technical University of Athens. It contains several daylight conditions for license plate images. The data set contains total 112 images in challenging lighting conditions. We have used this data set for cross data set inference only to evaluate the generalization of the proposed ALPR model. The mentioned model is not trained directly using this data set, instead it has been used to test the segmentation results generated using our model with the pre-trained weights of the Open ALPR data set experiment.

It is important to mention that ALPR, UFPR and Media Lab benchmark data sets do not contain the binarized ground truth segments. Thus, we had to generate the ground truth separately and used them for the training of the proposed ALPR network.

VGG16: VGG16 has won the ILSVR in 2014. This is one of the popular vision models till the date [10]. The main advantage of this VGG is that it uses a homogeneous topology with small size kernels. However, this network is very slow to train.

¹<https://web.inf.ufpr.br/vri/databases/ufpr-alpr/>

²<http://www.medialab.ntua.gr/research/LPRdatabase.html>

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

VGG19: Convolutional neural network VGG19 [10] was trained on over a million images from the ImageNet collection. With its 19 layers, the network is capable of classifying photos into 1,000 different object categories, including a keyboard, mouse, pencil, and numerous animals. Consequently, a vast array of image rich feature representations have been trained by the network. However, similar to VGG16, there are a lot of trainable parameters, which causes the training process to be extremely slow [11].

Inception-V3: Google had designed this popular DL-based Convolutional Architecture. It is an image recognition model that achieved more than 78.1% precision on the ImageNet. The training of Inception-V3 is performed using a data set of 1,000 classes from the original ImageNet data set. [12, 13]. Compared to the earlier Inception versions, Inception-V3 was more accurate, required less computing power, and had a very low error rate. However, this network mainly suffers due to the lack of homogeneity in its design.

MobileNet-V2: It is the enhanced version of MobileNet-V1 [14]. Having depth-wise Separable convolution layers is the main feature of this network. MobileNet-V2 [15] is having two main blocks, they are a residual block with one stride and another with two strides [15]. The residual structure is the important factor in this architecture³. Trainable parameters count are less with compared to other networks and its previous version. Also the model size is very less with compare to its previous version.

MobileNet-V3: For MobileNetV3, a combination of different layers has been used to make it more effective compared to MobileNetV2. MobileNetV3 backbones are slightly faster than its V2 counterparts [16].

InceptionResNet-V2: InceptionResNet-V2 is a CNN that achieves a new scientific development on the basis of accuracy on the ILSVRC⁴ image classification benchmark. This is a variation of the Inception-V3 [13]. It has been successfully deployed in many applications. Still, one major drawback of this model is that it takes a longer time to train.

Xception: Xception [17] stands for Extreme version of Inception. It has the main feature of modified depth-wise separable convolution in its network which makes it even better than Inception-V3. Xception outperforms VGGNet, ResNet, and Inception-V3. It uses a 2D filter followed by a 1D filter which makes itself efficient rather than learning with 3D filters. However the computation cost is very high [11].

AlexNet: This is the first CNN based architecture. Eight layers make up its architecture: three fully-connected layers and five convolutional layers. However, these are not the qualities that set AlexNet apart; instead, it makes use of a few CNN components that have given the network a unique aspect. These components

³<https://towardsdatascience.com/review-mobilenetv2-light-weight-model-image-classification-8febb490e61c>

⁴<http://www.image-net.org/challenges/LSVRC/>

1.2. FUNDAMENTALS OF EXPLORED DEEP LEARNING IN THE RESEARCH

are Overlapping Pooling, Multiple GPUs, and ReLU Non-linearity⁵.

ResNet: It stands for Residual Network: This architecture presented the idea of Residual Blocks as a solution to the vanishing/exploding gradient issue. The skip connections have been added to this network. By omitting some layers in between, the skip connection links activations of one layer to those of subsequent layers. In essence, this creates the residual block. These leftover blocks are stacked to create ResNet. Adding this kind of skip link has the benefit of allowing regularization to bypass any layers that degrade architecture performance. The CIFAR-10 data set's 100–1000 layers were the subject of experiments by the paper's authors.⁶

YoloV3: Total of 53 layers of YOLOv3 which employs Darknet Architecture, were trained using the ImageNet data set [5]. Upsampling and residual connections are used in the network. Three distinct scales have been used for the detection. Compared to earlier versions, it has a longer processing time but is more effective at recognizing smaller objects.

YoloV5: The CSP network and Focus structure serve as the foundation for the YOLOv5 network. PANet and the SPP block make up the neck. It uses GIoU-loss and has a YOLOv3 head. Although the previous iterations of YOLO were developed in C, YOLOv5 is written in Python. This facilitates integration and installation on IoT devices.

Vision Transformer (ViT): A Transformer-like design is used over image patches in the ViT, an image classification model. After dividing a picture into fixed-size patches and linearly embedding each one, along with adding position embedding, the series of vectors is put into a typical Transformer encoder. The conventional method of performing classification involves incorporating an additional learnable 'classification token' into the sequence.

1.2.2 Attention and Multi-head Attention

The attention mechanism, which was first proposed as an improvement for encoder-decoder RNNs used in sequence-to-sequence applications like machine translation, is the central concept of the Transformer model [18]. The entire input was compressed by the encoder into a single fixed-length vector to be given into the decoder in the initial sequence-to-sequence models for machine translation [19]. The idea behind attention is that it could be preferable for the decoder to examine the input sequence at each stage rather than reducing the input. Furthermore, one may assume that the decoder should focus on specific portions of the input sequence at specific decoding steps, as opposed to always viewing the same representation of the input. The attention method proposed by Bahdanau [18] offered a straightforward way for the decoder to dynamically attend to distinct input

⁵<https://towardsdatascience.com/alexnet-the-architecture-that-challenged-cnns-e406d5297951>

⁶<https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/>

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

segments at each stage of decoding.

On the other hand, multi-head attention is a key component in neural network architectures for natural language processing tasks. This has been introduced in the Transformer model by Vaswani et al.[20], it enables the model to focus on different parts of the input sequence simultaneously. By employing multiple attention heads, the model can capture various aspects of relationships within the data, enhancing its capacity to understand complex patterns and dependencies.

1.2.3 Transfer Learning

Transfer Learning (TL) [21] is a popular concept in ML which has been successfully applied in different application areas [22, 23]. It utilizes the knowledge acquired while resolving one problem and applies it to a different but related domain. In CNN based classification problems, usually, the output layer (i.e. the final layer used in the network) or few last layers need(s) to be replaced with new layers as per the problem domain. The main idea behind this approach to achieve higher accuracy with fewer data. It allows any DL-based framework to generate a robust feature set from the original input. With a smaller connected network we can achieve improved accuracy. The baseline performance might be improved due to this knowledge transfer along with a faster model development time. Better generalization can be attained with this approach.

We have utilized this framework in several experiments in the research. We have achieved higher accuracy with comparatively less data (compare to ImageNet [5]) in our experiment as well.

1.2.4 Experimented Loss Functions

Below are the definitions of three statistical loss functions which have been used in several ML experiments in the thesis.

Structural Similarity Index Measure (SSIM): SSIM is used for measuring the structural similarity between two images with equal dimension. This has been measured by using the Equation 1.1. Here, the covariance of x and y is σ_{xy} , the means of x and y are μ_x and μ_y , the variances of x and y are σ_x^2 and σ_y^2 respectively, and the two parameters for stabilizing the division with weak denominator are c_1 and c_2 .

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (1.1)$$

Mean Absolute Error (MAE): This is very basic error which denotes the average error calculating the difference between true and predicted target values using the Equation 1.2. Here $Target_i$ denotes the actual i^{th} target label while $Predicted_i$

1.2. FUNDAMENTALS OF EXPLORED DEEP LEARNING IN THE RESEARCH

denotes the corresponding predicted one.

$$MAE = \frac{1}{N} \sum_{i=1}^N |Target_i - Predicted_i| \quad (1.2)$$

Binary Cross Entropy (BCE): The BCE loss has been defined mathematically as follows (Equation 1.3). Here $Target_i$ carries the same meaning as stated in the last equation, while $Conf_i^{Train}$ is the confidence value of the prediction for the same i^{th} sample.

$$Loss_{BCE} = -\frac{1}{N_{Train}^{total}} \sum_{i=1}^{N_{Train}^{total}} Target_i \cdot \log Conf_i^{Train} + (1 - Target_i) \cdot \log (1 - Conf_i^{Train}) \quad (1.3)$$

1.2.5 Explored Metrics for evaluation

Several validation metrics have been utilized to compare the performance of the proposed ideas in difference experiments with various related methods.

Peak Signal to Noise Ratio (PSNR): It is defined as follows in dB. The maximum pixel value is shown by C , while the Mean Squared Error is indicated by MSE .

$$PSNR = 10 \log_{10} \left(\frac{C^2}{MSE} \right) \quad (1.4)$$

Accuracy, Precision, Recall and F-Measure (FM): FM is defined from the conventional definition of Recall and Precision while recall and precision have been calculated by Equations 1.6 and 1.7 respectively. Here in the notations, the symbols TP, FN, and FP stand for true positive, false negative, and false positive, in that order.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.5)$$

$$Recall = \frac{TP}{FN + TP} * 100 \quad (1.6)$$

$$Precision = \frac{TP}{FP + TP} * 100 \quad (1.7)$$

$$FMeasure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (1.8)$$

Specificity, Jaccard Similarity (JS): Below two equations define the Specificity

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

and JS respectively.

$$Specificity = \frac{TN}{TN + FP} \quad (1.9)$$

$$JS = \frac{TP}{TP + FP + FN} \quad (1.10)$$

Pseudo F-Measure (pFM): This metric is computed like F-Measure except that recall value is taken from skeletonized ground truth:

$$pFMeasure = \frac{2 * Recall_{skel} * Precision}{Recall_{skel} + Precision} \quad (1.11)$$

Negative Rate Measure (NRM): NRM can be computed as follows. For this metric, lower value indicates better binarization process.

$$NR_{FP} = \frac{FP}{FP + TN}; NR_{FN} = \frac{FN}{FN + TP}$$
$$NRM = \frac{NR_{FP} + NR_{FN}}{2} \quad (1.12)$$

OCR Score: For the Optical Character Recognition (OCR) task, we have used Google Tesseract baseline OCR library. The exact accuracy has been measured using the fuzzywuzzy⁷ library. fuzzywuzzy.ratio Application Program Interface (api) has been used to measure the similarity between the OCR retrieved data and the actual data. The baseline OCR accuracy measurement for the raw images and the ground truth images have been computed for every LPR data sets at the beginning.

1.3 Literature Survey on IRTS

In this thesis, we have explored three distinct surveillance aspects of IRTS: air pollution caused by vehicle emissions, automatic license plate recognition for vehicle identification, and various safety considerations in road transport. Each of these areas has seen numerous ongoing research initiatives conducted by researchers.

Survey on On-Road Air Pollution Surveillance: Regarding the previous related studies of vehicle pollution, in [24] the authors have presented a deep spatio-temporal based fusion network for vehicle emission forecast. In [25], Xu et al. have employed an LSTM based Auto-encoder model to forecast vehicle pollution. In [26], a vehicle monitoring based air pollution system has been developed using a wireless sensor network. A sensor and micro-controller based automated system in a vehicle have been proposed by Chandrasekaran et al. in [27]. Pal et al. have designed another sensor based system to share the individual personal car's

⁷<https://pypi.org/project/fuzzywuzzy/>

1.3. LITERATURE SURVEY ON IRTS

pollution status [28]. In [29] authors have presented a system to identify vehicle pollution using MQ-7 carbon monoxide sensor. However, in all these researches, the sensor installation itself will increase the overall costing. Cardoso et al. have filed a US patent for image based technique of several toxic gas absorption in vehicles [30]. Guo et al. have shown a different perspective with empirical evidence, how the electric vehicle sale got a jump due to the rising vehicle pollution [31]. However, we cannot ignore the fact that, in the third world countries it will take at least another decade to have 100% EV in the road. Several image processing approaches have been experimented by the researchers [32, 33, 34]. Tao et al. have proposed a volume Local Binary Pattern (LBP) based method for smoky vehicle detection in their paper [35]. In another work, they have proposed a TAMURA feature-based technique over video for smoke detection [33]. This TAMURA based approach lags to achieve higher accuracy in varying environment. Higher order Local Ternary Pattern (HLTP) has been explored for smoke detection by Yuan et al. in [36]. In [37], the combination of LBP and Support Vector Machine (SVM) has been utilized for identifying smoke. Their proposed framework might fail in case of overlapping smoke and non-smoke LBP. On the contrary, the single image-based DL approaches are very few primarily in this particular domain through vehicle smoke detection. The authors in the article [38] have presented a vision transformer-based deep learning approach for identifying smoky vehicles on the road. Ba et al. have developed an attention-based CNN network for the satellite smoke images [39]. The single image smoke detection technique has been presented through a convolution network in [40], which is known as DNCNN in short. Here, in this framework, the performance of the convolution layers is strongly dependent on its fully connected network part. Moreover, we have many other research solutions in the literature for smoke detection in different environments [41, 42] which can be utilized in our problem area as well.

In context of vehicle smoke localization, with advancements in computer vision, especially in the realm of object detection, researchers have started to address the vehicle pollution issue by coming up with efficient machine learning and deep learning models [38, 43, 33, 44, 45, 46]. Tao et al. proposed a method based on multi-feature fusion and ensembling to mitigate the need for a large human force to monitor traffic surveillance footage to recognize smoke-emitting vehicles [33]. In another study, aimed to identify smoke-emitting vehicles based on the diesel engine specifically, Wang et al. [44] presented an SDV-Net based framework which incorporates a two-stage approach toward smoky diesel vehicle detection, comprising of a detection stage and a fine-grained classification stage. The first stage detector was inspired by YOLOv3 [47], while SE-ResNet inspired the backbone of the second-stage classifier [48]. Harnessing the efficiency of pre-trained models, Kundu et al. applied TL paradigm to classify vehicle images captured

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

using on-road surveillance cameras into those emitting and not emitting smoke [45]. They experimented with three popular pre-trained deep convolutional neural network backbones– InceptionV3, MobileNetV2 and InceptionResNetV2. To ensure the effectiveness of the trained models, the authors also included image augmentation techniques in their pipeline. The authors achieved promising test results using a majority voting scheme through a robust TL framework. More recently, to improve the speed of smoke and smoke-emitting vehicle detection without compromising the accuracy, Wang et al. [46] proposed a lightweight architecture inspired by YOLOv5 [49]. Here, the authors incorporate MobileNetV3 [16] into YOLOv5's backbone to reduce the number of learnable weights in the model. They also include some advanced image augmentation methods such as 'cutout' [50] to improve their data set, reduce over-fitting, and enhance the robustness of their detection model. Wang et al. [51] have proposed a saliency-based object detection technique that has the advantage of faster inference with smaller number of annotated points. In our research, we have chosen YOLOv5 as the baseline model while deploying the transformer block within the same. The main advantages of this framework over the previous version are its small size and operational speed. Even, the accuracy of the corresponding lighter version (YOLOv5s) is better compared to other variations in different detection areas⁸. As this has been established that YOLOv3 and YOLOv5 worked really well in challenging object detection domains, we have decided to proceed with these two variations.

Employing deep-networks to identify vehicular pollution also relies on the visibility of smoke and the laws based on the same. In the process of internal combustion of diesel or gasoline engines, several harmful gases often leak out. Few components are invisible; however, more than 75% of those burning elements belong to NO_x (x=1, 2,..) compounds (with reddish brown color) which are visible to the eyes [52]. Besides these, particulate matter are of dark gray color, volatile organic compounds are of whitish shade⁹ [53, 54], and improper burning of engine oil produces black smoke. Different functional issues associated with the vehicle engine are producing different types of colors¹⁰. Undoubtedly, those materials affect the environment by degrading the air quality directly, and thus, it is possible to detect the pollutant vehicles by detecting visible smoke.

Proceeding with similar intentions, governments of different developed countries are adopting intervention strategies for controlling the massive air pollution by identifying and penalizing vehicles with smoke for controlling the massive air pollution. Transport agencies started enforcing laws based on smoky vehicle identification. In Dudley, air quality authority has planned to identify smoky vehicles¹¹

⁸<https://towardsdatascience.com/yolov5-compared-to-faster-rcnn-who-wins-a771cd6c9fb4>

⁹<https://www.ucsusa.org/resources/cars-trucks-buses-and-air-pollution>

¹⁰<https://www.universal-tyres.co.uk/news/exhaust-smoke-colour-warnings/>

¹¹<https://www.dudley.gov.uk/business/environmental-health/pollution-control/air-quality/vehicle-air-pollution-smoky-vehicles/>

1.3. LITERATURE SURVEY ON IRTS

which will improve the air quality directly. Even normal citizens can register to complain through the web interface¹². Oregon State has installed an automatic system to catch and take action against smoky vehicles¹³. Singapore government's environmental agency and Hongkong government have also adopted similar approaches to restrict air pollution^{14,15}. All these instances make it evident that from the perspective of governments and policy makers automated smoky vehicle identification is being considered as a sustainable way to enforce the laws related to vehicular air pollution.

Review of License Plate Recognition Studies: This ALPR domain is facing many challenges to make it a successful real product in the surveillance. Geometry of plate is a concern to fetch the size, position and angle along with the physical obstruction by other objects [55]. Different font, color are another challenge in the same task [56]. At the same time, distortion and damaged plate will hamper the accurate recognition. Varying lighting condition is another factor which can strongly effect this ALPR process. Besides these, camera hardware features which can affect the resolution of the image taken can directly impact the recognition [55]. Researchers are facing issues in different weather conditions like foggy, rainfall as well [57].

A notable amount of work has been done over the last decade on image processing techniques and deep learning methods in the field of license plate detection and recognition [58, 59, 60]. ALPR research studies mainly address the problem from two perspectives: one is detection-recognition solely based on real-world samples, while the other focuses on synthetic license plate generation [61]. The former one can be categorized into two segments: image processing based techniques [58, 59, 62] and machine learning based techniques [60, 63]. However, the plate generation through Generative adversarial network (GAN) in the literature is mainly used for generating huge number of synthetic training images for the main network to recognize the number plate in the experiment [61]. With the rapid advancement of deep learning through the last decade, ALPR researchers have focused much on implementing deep learning based methodologies for further improvement of the detection algorithms. Chen et al. [60] have showcased YOLO-based architecture with sliding window technique for Taiwan's car license plates. However, YOLO-based architectures face some limitations in detecting objects with high confidence which are close to camera-plane and are small in size. This is a genuine drawback, especially in case of ALPR. At the same time, as the YOLO-based networks work in overall detection process through feature encoding, therefore it fails to focus on the character-wise segmentation task, which can

¹²<https://www.gov.uk/report-smoky-vehicle>

¹³<https://www.oregon.gov/deq/FilterDocs/smokingvehicles.pdf>

¹⁴<https://www.nea.gov.sg/media/readers-letters/index/measure-to-enforce-against-smoky-vehicles>

¹⁵https://www.epd.gov.hk/epd/english/how_help/report_pollution/spotter_training.html

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

degrade the recognition accuracy in ALPR. Regarding the other deep networks, in [64], Bulan et al. have explored a localization method using Convolutional Neural Network (CNN) based architecture with Hidden Markov model based OCR for identification purpose. Wen et al. [65] have applied Support Vector Machine (SVM) to identify the characters of the plate with some variations in the ambiance. A CNN based approach considering sparse overlapping sub-regions of concerned images has been presented by Kurpiel et al. [66]. Using the YOLO-v2 architecture, Lin et al. [67] presented a three-stage license plate identification system based on Mask-RCNN that was utilized for a variety of shooting angles and oblique images. In early times, thresholding-based approaches [68] have been adopted by the researchers to segment the license plate characters to assist the recognizer. In [58], an adaptive thresholding strategy has been deployed. Dun et al. [59] have showcased another dual threshold strategy for Chinese number plate recognition. Connected component analysis is another important aspect in this context. Different filtering mechanisms like high-pass, band-pass, Gaussian and feature-based techniques have been applied as pre(or post)-processing to enhance the detection accuracy. Maximally Stable Extremal Regions (MSER) is another very useful approach which has been widely used in the similar and related application areas.

License plate generation is another task which has been explored by the researchers, mainly to generate training samples for the machine learning algorithms. Several deep learning techniques have been used for this purpose [61]. In last few years, GAN has been deployed in many application domains [69, 70] including ALPR. But mostly, in ALPR, it has been used for generating synthetic training image sets [71]. CycleGAN-based plate generation has been also used for the annotation purpose [72]. Overall, the divergent imaging conditions can affect the recognition task to a large extent while a robust segmentation technique can pull up the accuracy in those conditions. Considering this fact, the aim of the article is to propose a accurate segmentation algorithm which can enhance the final recognition accuracy. In this article, we have propose a GAN-based architecture which will be used for license plate binarization in order to get the license plate segmented into foreground and background; which ultimately provides superior performance in adverse conditions as well.

Exploration of Safety Aspects in Road Transport during Pandemic: During pandemic time, where on-road transmission of COVID like diseases posed a significant concern, researchers have tried their best to restrict such spread in different possible ways. With the prevalence of the COVID-19 pandemic and the implementation of social-distancing, road transportation rules have been modified and imposed with new criteria. In this article, we have outlined a different aspect of surveillance considering the post-COVID-19 circumstances. Here, we have focused on the two important ideas in the vehicular inspection process. Firstly,

we have put our attention into the safe distance norm inside the car. The government of different countries has enforced regulations to control the COVID-19 spread through transport. In several countries, passenger count is getting restricted where four-wheeler can only have three occupants, including the driver, in order to maintain social-distancing [73, 74]. For the next few years, this kind of rule should be there to control the contagious spread. In this article, we have incorporated a robust face detection technology for counting the number of passengers within a car. In this regard, from the road-side installed surveillance camera, video frames or snippets will be taken. Secondly, the use of face-mask is another mandatory criterion to prevent the spread via the transport system. Therefore, the identification of facial mask has become an important topic of research in very recent times. Researchers have started putting efforts in this particular area to detect mask over the human face which is very critical preventive step [75]. A feature-based technique that uses the images from mobile camera, has been developed by [76]. They have put focus on the feature extraction for the specific types of masks. K-Nearest Neighbor algorithm has been utilized for the detection purpose. [77] have utilized the well known PCA technique for the recognition of mask over human face. In [78], the authors have experimented to identify the presence of a mandatory medical-mask in the operation theater. Though, their application areas were quite novel, the performance of the proposed algorithms were not satisfactory due to the use of traditional machine learning based methods, and recently developed DL approaches provide better performance in the mentioned tasks in terms of accuracy. [79] have recently presented a Support Vector Machine (SVM) based approach over ResNet architecture for face mask detection. Though, the accuracy of the model is high, it has not been validated for the noisy image set. Moreover, the ResNet feature computation is bit complex due to the higher number of parameters. In [80], authors have achieved 95% recognition accuracy using the face-eye-based multi-granularity model. Single-shot detector (SSD) has been experimented in [81] for identifying the faces from the images and afterward, the MobileNet-V2 baseline architecture has been used for the classification task. Though, specifically the tiny face use cases, side-face, hand-covered use cases were not considered in this paper. Thus, the robustness of the proposed framework is low and it is difficult to use in real-world conditions. At the same time, several other CNN frameworks with different backbone architectures can be deployed for serving the same purpose. For example, AlexNet [82] can be used as a backbone of a low complexity process. As VGG [10] is known to be a robust feature extractor, VGG based models like VGGFace-ResNet [83] have also been developed for this purpose. Specialized models for face classification like FaceNet [84] can be modified to detect face mask in images. VGGNet [10], AlexNet [82] have similar drawback of having huge numbers of parameters, 138M and 62M re-

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

spectively. In this respect, Inception-V3 is comparatively a lighter network which can be utilized in TL framework. Table 1.1 has outlined a comparative sketch of the state-of-the-art approaches in this domain.

Table 1.1: An overview of different related DL based state-of-the-art algorithms: data set used, the face positions and conditions like front or side, hand-covered, image quality on which they operate, face detection strategy, cross data set evaluation status

Algorithm	Data set used	Special Attention				Face de- tection	Cross data set evalu- ation
		Front Face	Side Face	Hand Cov- ered	Low Reso- lution		
VGGFace [83]	LFW [85], Youtube Face DB [86]	✓	✓	×	×	×	×
FaceNet [84]	LFW [85], Youtube Face DB [86]	✓	✓	×	×	picasa type	×
SSDMNV2 [81]	Mikolaj (Kaggle), Prajna (github) [87], RT-MDD [88]	✓	×	×	×	single shot	×
ResNet-SVM [79]	RMFD [80], Prajna (github) [87], LFW [85]	✓	×	×	×	×	×

1.4 Thesis overview and contributions

The research objective of this thesis is applying machine learning methods in the domain of intelligent road transport. Intelligent road transport covers a huge research area where different innovative and technical strategies are deployed to automate several manual activities. Here, the research has been focused mainly in the surveillance perspective which ultimately provide a safe environment in the road transport. The research question addressed in this thesis are:

1. How to identify polluting vehicle in the road?
2. How to recognize License plate number so that surveillance can be automatically served?
3. What measures can be implemented to safeguard passenger well-being amid the ongoing pandemic, similar to recent times?

1.4.1 Smokey Vehicle Recognition

In the literature, existing image based smoky vehicle identification processes suffer from having very minimal training images availability. Even when a automated system deployed into the road transport, it will have limited training data to start with. This fact indirectly affects the accuracy in the outcome.

1.4. THESIS OVERVIEW AND CONTRIBUTIONS

In Chapter 2 of this thesis, A new, highly general method has been proposed that may be applied to any supervised learning framework. Our suggested approach works particularly well in cases where the supervised system initially has a small amount of labeled data. Usually, in this scenario, traditional methods fail to perform while the proposed framework demonstrated better efficiency compared to those traditional approaches. Most notably, numerous trials have shown that the suggested approach's confidence level has significantly grown in addition to its accuracy. We presented a framework which is based on the visibility and color of the vehicle smoke. Regarding the color aspect of vehicle emission, mostly the carbon particles are present in various forms which are responsible for black smoke. Bluish or gray color specifies the burning of engine oil which is having some toxic particles. Milky white reflects the burning of coolant with pollutant components¹⁶. In a nutshell, visible smoke from the vehicle is firmly indicating the presence of impure molecules which can be harmful to the environment¹⁷. Vehicle pollution detection is mostly performed from sensor data [26, 27, 28]. While sensor-based detection methods are accurate and standardized, from the surveillance perspective these tools have some implementation issues. The sensor systems are installed either in pollution testing centers or the Original Equipment Manufacturer (OEM) needs to attach them with individual vehicle. From the manufacturer's end, this results in rising cost of cars. Sensor installation by car owners, on the other hand, is not only very costly, but these are also dependent on the intentions and awareness of the car users. A very important information that should be considered at this point, is that around 75% of the pollutant gases, generated from the vehicle engine cause visible smoke. For enforcing the law, thus, a low-cost, automated detection method based on road surveillance images would be much more helpful and effective for the administrative authorities. To address this issue, in this paper we have proposed a cost-effective iterative process to identify vehicle pollution from on-road images using the DL approach. Considering the research studies and laws enforcement of different countries based on smokey vehicles, we can claim that our proposed solution provides an economical iterative process to identify vehicle pollution from on-road images smoothly and promptly, using this DL approach.

In our proposed research, initially, we have developed an image data set with notable variations for detecting vehicle pollution. In an earlier work [89], we have built a prototype, where a small image data set was used. Here we have enriched the data substantially for building effective DL models. We have focused on different daylight conditions along with the quality of the surveillance camera images while building this data set. On the contrary for night mode image generation, we have used the Style-GAN [90]. Other image augmentation operations viz. blur-

¹⁶<https://www.carthrottle.com/post/different-engine-smoke-colours-and-what-they-all-mean/>

¹⁷<https://www.tmr.qld.gov.au/Community-and-environment/Environmental-management/How-you-can-make-a-difference/Smoky-vehicles>

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

ring, rotation, flip, generation of rainy and foggy images have been used to include large variations in the image data set. Based on the images captured from surveillance cameras, the framework will automatically identify the pollutant vehicles. We have utilized seven CNN based DL models, i.e Inception-V3, MobileNet-V2, MobileNet-V3, InceptionResNet-V2, VGG16, VGG19 and XceptionNet as baseline weights to feed to our proposed network. Transfer Learning (TL) concept has been exploited to build the final models which will be used for the prediction. The main advantage of TL [91] is rapid progress and improved performance in executing the ML tasks. In our experiment as well, we have exploited it with the baseline ImageNet [5] data set. Our framework has started with a very small number of images for training the models. Subsequent feedback from the test images will help the model to improve its performance. To deal with the false alarm, we have given importance to the confidence level of prediction for each of the models. Instead of relying upon a single model, we have exploited the advantage of multiple models using majority voting consensus to take the final decision. On-road rule breakers will be spotted in real-time. Our main goal is to investigate vehicle pollution identification with a higher trust level. An additional motivation for this proposed approach is to deal with very low labeled training data which is a very common challenge in several real-life scenarios. A significant increase of accuracy has proven the effectiveness of the developed model. The following are the planned work's primary contributions: A brand-new, highly adaptable method has been put out that can be used with any supervised learning-based framework. Our suggested approach works particularly well in cases where the supervised system initially has a small amount of labeled data. Usually, in this scenario, traditional methods fail to perform while the proposed framework demonstrated better efficiency compared to those traditional approaches. In particular, numerous trials have shown that the suggested approach's confidence level has significantly grown in addition to its accuracy. The majority voting scheme's implementation improves the approach's inherent reliability. For this purpose, we have prepared two vehicle pollution data sets with huge variations in environment. The data sets are publicly available in <https://github.com/srimantacse/VehicleSmokeDataset>. We have applied our approach to several DL methods and also compared the same with related techniques to demonstrate its superiority.

1.4.2 Vehicle smoke embedding and smokey region Detection

To properly identify the smoky vehicle, we need to identify the smoke region and correlate with the corresponding vehicle. This step boils down to a object detection problem where smoke and smoky vehicle is the intended objects. For applying any object detector, we need to have a robust marked data set which is very rare in the literature. This situation hampers the detector training model.

1.4. THESIS OVERVIEW AND CONTRIBUTIONS

On the other hand, in recent times different YoLo detectors are very popular and getting industrialized as well. Researchers are actively enhancing and optimizing this framework through their innovative efforts to achieve greater efficiency.

In Chapter 3, we propose a novel realistic vehicle smoke synthesis algorithm that can accommodate several generation conditions like color, shape, etc. Afterwards, we propose a lambda implemented attention based detector module development for the vehicle smoke detection. This is the first attempt, as far as we are aware, to use lambda-attention for the smoke detection task. The supremacy of the proposed technique over the SOTA methods are showcased with cross data set testing approach using real-world data. For this identification process, we need a good number of smoky vehicle images for Deep Neural Network (DNN) model building. However, a significant research gap is present in all these reported works as there is a very limited publicly available vehicular pollution image data sets available. Therefore, we break down the problem in two levels: we first propose an algorithm to synthesize a robust data set containing smoky vehicle images, which will assist researchers in training deep models. In this chapter, we have described a robust method for synthesizing smoke on top of any on-road vehicle image. This will give rise to a ‘pair’ of images in smoky and non-smoky class, and there are no such paired data sets in the literature till the date; this synthetic generation helps us to train deep models with large data sets. At the same time, we also focus on designing a robust deep network design so that we can very accurately identify on-road vehicle smoke. An attention-based [20] backbone network in the YOLOv5 has been designed for the smoke detection phase utilizing a lambda-implemented transformer layer. Extensive analyses have been performed on different data sets along with cross data set evaluation to demonstrate the supremacy of the proposed framework. Considering the other techniques for identifying vehicle pollution, we do not need to install costly hardware setups like a laser spectral instrument, sensors etc. To achieve high accuracy, it would be better if those devices could be attached to the individual car which indirectly increases the infrastructure cost. Moreover, this external device installation is also dependent on the intentions and awareness of the car users. On the other hand, our primary objective is to identify on-road smoky vehicles on a busy road where it is not possible to manually detect the lawbreakers among thousands of cars passing within a minute. Here, with minimal infrastructure cost, this problem can be solved through our proposed approach. Moreover, our proposed approach can detect multiple smoky vehicles simultaneously in a single frame which is not possible using several costly available solutions. This first level of screening will help the traffic authority to reduce manual effort drastically. Moreover, regarding the dataset scarcity problem for the DL model training we are utilizing the proposed smoke synthesis method. The main contributions of this proposed work are as follows–

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

- We proposed a novel realistic vehicle smoke synthesis algorithm that can accommodate several generation conditions like color, shape, etc.
- We proposed a lambda implemented attention based detector module development for the vehicle smoke detection. To the best of our knowledge, this is the first attempt to use lambda-attention for the smoke detection task.
- The supremacy of the proposed technique over the SOTA methods have been showcased with cross data set testing approach using real-world data.

1.4.3 Automatic License Plate Recognition

The complexity of surrounding environments make the task more complicated. ALPR performance also critically depends on the quality of the image which deteriorates based on ambient conditions like fog, mist, rain, snow etc. Further, other challenges such as partial occlusion, shadow etc. affect the overall performance of ALPR significantly.

In Chapter 4, we have provided a GAN based method for automatic identification of number plate. To the best of knowledge, this is the first approach in ALPR that attempts to generate binarized image from original number plate using GAN architecture. The major contribution of this work is to design the GAN-based technique which can deal with the number plate binarization even in the presence of different image degradation. To achieve this task, we have developed our own ground truth segmentation data set for LP as the same is not available in open-source platforms. The dual discriminator concept is introduced here to deal with the local and global information of an image simultaneously. The proposed generator minimizes structural similarity-based loss along with cross-entropy loss designed precisely for binarization task. Our method performs satisfactorily even for document binarization problem.

1.4.4 Safety in Intelligent Road Transport System during Pandemic

Another road safety involves restricting passenger air transmitted diseases spread which was very vital during the pandemic period. Researchers have put their effort to discover medicines to prevent such spread. However, how to automatically restrict during journey, mainly inside the vehicle was not surveilled during this time.

In Chapter 5, we have initially proposed a image based passenger mask identification framework which was very relevant in the pandemic context. A robust data set with huge variations and real-life use cases have been prepared. Special attention has been given to the side-face, hand-covered and low-resolution images which will be essential for the traffic surveillance. The data set is publicly avail-

able at <https://github.com/srimantacse/MaskSurveillance>. An efficient deep learning model to identify face mask over the human face with the adoption of very tiny face detection has been adopted. We have shown that the proposed model outperforms the other state-of-the-art techniques with the extensive comparison as well as statistical hypothesis testing. As per the authors' knowledge, this was the first work to find out the in-vehicle mask-wearing status for the surveillance. This sort of traffic monitoring has been very essential in the current COVID-19 related pandemic context.

For this task, we have initially prepared a robust data set with significant variations for the mask prediction over a human face image. Different essential conditions like noise-induced image, day-night mode, angle of faces, the direction of heads, etc. have been considered to enrich the overall database. We have also put our effort to deal with several other use cases like, side-face, hand-covered, tiny face with varying illumination, imaging height, resolution etc. A separate on-road image set has been accumulated for the real-time evaluation of the model. Concurrently, a network model has been proposed and optimized with different variations of these image data inputs during the training which ultimately provides more accurate outcomes. The Transfer Learning (TL) process has been exploited in our framework. We have explored the popular DNN Inception-V3 in the experiment. Using TL mechanism, the model has been re-trained through the proposed network with our data set. The result has demonstrated the efficiency of the framework. We have compared our approach with the recent state-of-the-art DL techniques like SSDMNv2, ResNet-SVM, AlexNet, VGGFace-ResNet, FaceNet using several numerical metrics viz. Accuracy, Precision, Recall, F1-Score, Specificity and Jaccard score (JS). Visual comparison and other related results also prove the superiority of the proposed approach. Along with these, we have showcased the efficiency of the proposed method through cross data set evaluation steps with the Real-Time-Medical-Mask-Detection (RTMDD) data set experimented in [81]; this supports that the method is not bottleneck by the over-fitting problem. The research benefits of the proposed framework mainly include real-time traffic surveillance. Preventing the spread of COVID-19 like contagious diseases through the transport system is one important motive of this research.

1.5 Summary

The process of IRTS enhancement through different approaches using ML activities have been discussed in the consecutive chapters. Initially vehicle smoke detection, pollution monitoring has been discussed. Afterwards, to create full automation in the complaint lodging process, license plate identification is a crucial step. We have innovated a unique GAN mechanism to serve this purpose. Later on,

CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

a pandemic time safety context and measures in IRTS have been elaborated in the final chapters. In accordance with recent advancements in research, these mentioned innovations will enhance the IRTS significantly. Human efforts will be drastically reduced. Along with the Government agencies, the common people will surely gain substantial benefits from these applied research activities. These innovations will strongly contribute to the successful realization of our overarching goal to establish an intelligent and smarter road transportation system.

Smokey Vehicle Recognition

2.1 Overview of Smokey Vehicle Recognition

The integration of innovative solutions in IRTS not only focus to establish a smarter system but also becomes imperative in addressing the health risks posed by air pollution, ultimately ensuring a safer and sustainable future. Air pollution is highly hazardous for health and may lead to premature death [92]. This type of pollution in urban and peri-urban areas, caused mostly through vehicles, affects the environment significantly [93]. Several challenges remain in IRTS, like identifying the pollutant cars, enhancing the technology to reduce emission, monitoring highly polluted areas, enforcing traffic rules etc. Real-time detection is another challenge so that Government Traffic Authority can take necessary measures from on-road surveillance to reduce this catastrophic effect. Therefore, several DL approaches have been used immensely in IRTS in recent times [94, 95, 96]. In different perspective, some of the methods [97, 98] have been developed to identify and monitor the air pollution as well.

2.2 Vehicle smoke relation with the air pollution

The main two types of engines that have been used by the vehicle-OEM are diesel (and bio-diesel) and gasoline (also known as petrol engine in some geographical locations). During the internal combustion of those engines, more than 200 toxic gases can possibly come out [52]. The harmful toxic gases and compounds are listed in Table I of the supplementary which are coming out from the exhaust pipe of running vehicles [53, 54].

Color perspective: The smoke generated from the engine fuel burning consists of many different compounds. Adulteration caused by improper mixing of materials like methanol, kerosene, acetone and sec-butyl acetate etc. increases the environmental contamination by producing polluting gas during the fuel burning [99]. 75% of those compounds are actually Nitrogen oxide (NO_x) compounds which have colors and are visible [52]. Three major visible pollutants from mo-

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

tor vehicles are NO_x (Reddish brown), Particulate Matter (PM) (Dark Grey) and Volatile Organic Compounds (VOCs) (Smog type white)¹. Different functional issues associated with vehicle engine are producing different types of color². For example, the reason behind the black or gray smoke is incomplete fuel combustion, which can be originated by defective injection system, engine overheating or a choked air filter. White smoke is produced by improper fuel burning, which can happen due to low compression, faulty timing etc. On the other hand, black smoke is generated by engine oil burning due to high oil levels. But the materials coming out due to these mentioned reasons will affect the environment undeniably. In Table I of the supplementary, the visibility of different components are mentioned. Therefore, it could be concluded that a major percentage of the pollutant gases have specific colors, and visible smoke guarantees that the vehicle is causing air pollution.

Law enforcement in different countries by Environmental agencies: Several laws have been already enforced by different countries where based on the visible smoke vehicles are identified and penalized. A properly maintained vehicle engine should not emit smoke. Excessive smokes, normally from ill-maintained vehicles, are polluting the air. Therefore, reporting smoky vehicles become a challenging task nowadays through which we can restrict the air pollution at a great extent. In Dudley, a town of England, air quality management has decided to identify smoky vehicles³. Anyone can report a smokey vehicle and lodge complain through their official web interface⁴. In the state of Oregon (US) has already deployed a system to identify smoke and take proper action⁵. Singapore Government's National Environmental Agency has also adopted stringent rules to identify smoky vehicles⁶. Very interestingly, Hong Kong government recruits efficient people for spotting emission visually⁷. It establishes the fact that most of the law enforcing agencies take their decision based on smoky vehicle identification, considering those as a reason for air pollution. However, this type of manual spotting vehicles over a very busy road is really a challenging task.

Therefore, from all theses above instances, we can surely conclude that automated detection of pollution-causing vehicles from visible smoke emission is an essential solution needed for air pollution control. Traffic surveillance cameras,

¹<https://www.ucsusa.org/resources/cars-trucks-buses-and-air-pollution>

²<https://www.universal-tyres.co.uk/news/exhaust-smoke-colour-warnings/>

³<https://www.dudley.gov.uk/business/environmental-health/pollution-control/air-quality/vehicle-air-pollution-smoky-vehicles/>

⁴<https://www.gov.uk/report-smoky-vehicle>

⁵<https://www.oregon.gov/deq/FilterDocs/smokingvehicles.pdf>

⁶<https://www.nea.gov.sg/media/readers-letters/index/measure-to-enforce-against-smoky-vehicles>

⁷https://www.epd.gov.hk/epd/english/how_help/report_pollution/spotter_training.html

2.3. DATA SET PREPARATION

Table 2.1: Different image augmentations used in the experiment for generating the data set

Operations	Generation details	Purpose
Blurring	Kernel size up to (15x15)	Training & testing
Rotation	Angle (-15° to $+15^\circ$)	Training & testing
Flip	Mode: Horizontal	Training & Testing
Night Mode	Style Transfer GAN	Training & testing
Fog	Used imgaug library	Cross data set evaluation
Rain	Used imgaug library	Cross data set evaluation

which are already installed in most of the major cities can act as a helpful tool in this regard. From on-road surveillance frames, we can identify these vehicles automatically within reasonable time and take proper action. Thus, we proceed to develop a very low cost solution where from the surveillance images, we can very easily identify the pollution rule breakers.

2.3 Data Set Preparation

In literature, there are many data sets available for air pollution detection. The majority of those have taken care of the data which contains the percentage of toxic gases present in the air. However, in a busy urban road, amidst thousand of running vehicles per hour, it would not be easy to detect the probable pollution sources using those data sets.

To the best of the authors' knowledge, this is the first attempt to work in this particular domain through the DL approach. We have specifically captured different types of vehicle smoke images which strongly indicate to have pollutant particles present in those vehicles. Our main objective is to prepare a robust image base with considerable variations. We have collected instances from google image and captured real images from New town road, Kolkata. For introducing variations in the data set, we have used image augmentation operations for the transformation as shown in Table 2.1. During this augmentation, true labels are preserved, which means after the generation process, the same labels have been put to the corresponding output samples. This is done to ensure that irrespective of the data augmentation, a polluting (non-polluting) vehicle should be identified as hazardous (non-hazardous) in all conditions.

We have intentionally introduced the noise effect in the images through blurring. By including this noise level, during the testing, the image quality will not affect the accuracy of the proposed models. The kernel size of the blur operation has been set maximum to (15×15) . Beyond this range, the image quality has been



Figure 2.1: Sample images of the DATASET₁ with different variations. Each row presents a particular type of images. a) These are the images without pollution. b) This row shows images with pollution. c) This is the row containing images with rotation in the range $(-15^\circ$ to $+15^\circ)$. d) Images captured in night time. e) Blurred images are shown in this row. f) Rainy image set has been presented; and g) foggy images are shown in the last row.

2.3. DATA SET PREPARATION

degraded so much and object detection is becoming almost impossible. Another important aspect of the surveillance camera is the angle of rotations. The position of the cameras will impact the outcome of the models. So the range of angle will be vital during the training. The range has been fixed to $\pm 15^\circ$ for significant variations in the data set. Importantly, as the angle range is maximum $\pm 15^\circ$, the chance of label change is negligible due to border region cropping. Along with these, the horizontal flip has been added for better generalization. These operations we have performed using the opencv⁸ functionalities. On the other hand, for generating the rainy and foggy images, we have used the imageaug⁹ library. For rain generation, the default range of speed (0.04, 0.2) and drop size (0.01, 0.02) have been used. The severity parameter for fog generation has been varied from 1 to 5 which is the allowable range in the library.

Here, the data set contains a large variety of images that are collected in various conditions. Overall, in the image base, there are 3000 images. We have mainly focused on having day and night mode images in the collection. This will be making the approach more realistic. For the night mode image synthesis, we have adopted the Style-GAN [90] which is a novel Generative Adversarial Network (GAN), shown in Figure 2.2. We have used 5 different styler images (Figure 2.2b) that represent the different types of the night scene. The original day mode images (Figure 2.2a) have been fed into the GAN network. The trained Style-GAN will produce the night images (Figure 2.2c) after applying the saved weight values in the given input. Figure 2.2 has shown just a single styler with seven different input image examples. In the data set, there is a total of 500 night mode images present as shown in Table 2.2. All the colored images are resized to (224×224) resolution. The diversity in the data set produces better generalization of the CNN-based models. Figure 2.1 has shown some sample images in the data set in different categories.

Table 2.2: Image statistic of the overall data set from day-night perspective

Day Images		Night Images	
Pollution	Non-pollution	Pollution	Non-pollution
1500	1000	275	225

Regarding the authenticity of the style transfer generation process, it is important to mention that this synthesis through style transfer for model training is becoming a very commonly used approach nowadays. Paolo et al. have used style transfer for agar plate segmentation [100]. In [101], authors have used this

⁸<https://opencv.org/>

⁹https://imageaug.readthedocs.io/en/latest/source/api_augmenters_weather.html



Figure 2.2: Night mode image generation using Style GAN: a) Set of input images. b) Sample styler image and Style GAN network. c) Set of corresponding transformed images.

neural network technique to generate a large number of micro-Doppler signatures which will help to recognize the human activity. Rafael et al. have utilized Style Transfer for augmenting Parkinson's Disease EMG signals [102]. In many applications, researchers have very rigorously utilized this technique for data validation [103, 104, 105, 106]. We have used this Style GAN for the night mode image generation as we have ample day mode images in our data set, while the samples for night time vehicle pollution count is very less. For supporting the use of Style-GAN, we have run a separate set of experiments as ablation. We have initially run the TL-Xcp model without any night mode during the training, and we have tested with 120 sample real night mode images (70 pollution and 50 non-pollution). While we are increasing the synthetically generated images during training in consecutive experiment sets, the real testing accuracy increases (overall 10% increase through the injection of 300 synthetic night-mode images). Thus, the results reported in Table 2.3 validate that the generation mode is useful in the real-world cases as well.

2.4. METHODOLOGY

Table 2.3: Synthetic night image count and statistics of testing accuracy for **real** night images (Best accuracy has been considered)

Day images	Synthetic night images	Accuracy (10 epochs)	Accuracy (20 epochs)
500	0	55%	60%
500	100	61%	66%
500	200	63%	66%
500	300	65%	70%

2.4 Methodology

In this section, we discussed our proposed feedback-based framework which will build a more confident (compared to the previous cycle) model at the end of each cycle throughout the iterative run. In different ML fields where the labeled training data (l_d : count of such samples) set is very limited, this method will be very effective, unlike the traditional training testing paradigm. This framework, in consecutive iterations, will inject more data points for training purposes from the prediction of multiple models, built in the previous step. We have exploited seven DL methods with a majority voting scheme, to provide reliability in the approach. During this process, each testing image has been validated with several models. The prediction status along with the confidence value has been utilized for the final decision making. We have not considered those test images which fall below the threshold of confidence termed as θ^* . This parameter basically denotes the allowable confidence level of the prediction during testing. Based on the domain criticality, the value has to be set. For example, in the case of a sensitive medical experiment, this value should be very high, whereas, for the domain with less severity, we can select a lower threshold. We have set the threshold parameter to 0.7 to perform an initial check on vehicles on a heavy traffic road. As a matter of fact, based on the value of this parameter, the number of injected images during training will be varied in every iteration. The higher the value of θ^* , the more authentic but less number of images will be fed (which is termed as m_{pd}) in the training set. If the majority of algorithms predicted the same class label with the higher confidence value, we have included the test image in the training for the next iterations. Another parameter ϵ , as mentioned in Algorithm 1, signifies the maximum allowed number of training samples for any experiment. In addition, the proposed framework may also take into account a maximum loop count as a termination condition. Here, we have iterated for four and three iterations for DATASET₁ and DATASET₂ respectively. Algorithm 1 has shown detailed steps of

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

the proposed framework.

We have started with only 600 labeled training data points (l_d) and 2400 unlabeled points (u_d) (for DATASET₁) in the first iteration. In consecutive iterations, the training size is gradually increased using the Algorithm 1. We have run till the fourth iteration for the DATASET₁. We have built seven DL models in every training phase. For a particular testing image, the predicted class labels and the confidence values of all these models have been used in the voting scheme. If it satisfies the criteria, we included the same in the next training cycle. To evaluate the performance of the framework, we will consider the output of the last iteration always. We have shown the accuracy value, confidence level and confusion matrix in each iteration for the validation and will take the final iteration's score only for evaluation purposes. We have deployed this framework on on-road traffic images to identify pollutant vehicles. To note, in any surveillance domain where we have limited labeled data to start with, this framework will be quite appropriate.

Algorithm 1 Iterative Deep Training

Input: Labeled data points(count: l_d), unlabeled data points(count: u_d) where $u_d \gg l_d$

Output: Iterative model (s)

```
1: /* Define the Parameters before execution */
2: #define itrmax Maximum iteration count
3: #define  $\epsilon$  Maximum allowable count for ( $l_d + m_{pd}$  (model-predicted))
4: #define  $\theta^*$  Threshold of confidence
5: procedure ITERATIVEIMAGETRAIN(dataset, loopCount)
6:   loopCount  $\leftarrow$  loopCount + 1
7:    $m_{pd} \leftarrow 0$ 
8:   newTrainDataset  $\leftarrow$  dataset
9:   if loopCount == itrmax or  $l_d + m_{pd} \geq \epsilon$  then
10:     Stop the process
11:   else
12:     Train all the models with the newTrainDataset
13:     for each testImage in the TestImage set do
14:       Test the testImage by all the models
15:       if passCount  $\geq$  majorityAlgoCount and all the confidenceValue  $\geq \theta^*$ 
16:         newTrainDataset.append(testImage)
17:          $m_{pd} + +$ 
18:   iterativeImageTrain (newTrainDataset, loopCount)
```

2.5 Experimental Setup

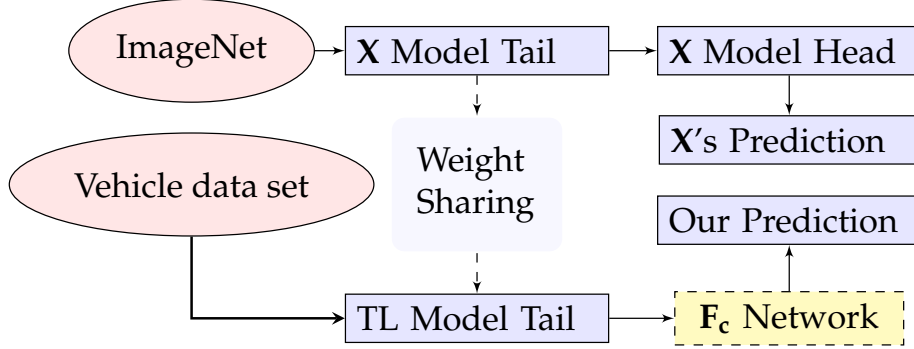


Figure 2.3: Overall block diagram of a TL (Transfer Learning) process using a baseline model X and the placement of F_c .

We have performed our experiment in Google colab¹⁰ provided GPU system. The baseline operating system was Linux 4.19.104+. The RAM size was 12.7 GB. The experiment has been performed iteratively in colab with our image-based pollution data set. We have incorporated seven popular CNN-based models, viz. Inception-V3, MobileNet-V2, MobileNet-V3 InceptionResNet-V2, VGG16, VGG19 and Xception for the evaluation and comparison. Exploiting the TL concept, the models have been fit to our problem.

Figure 2.3 has shown a block diagram of the overall process that utilized tail part of the mentioned baseline models (denoted as X). As shown in Figure 2.3, in the TL phase, we have used the pre-trained weights of the existing model at the beginning. We have selected some of the most popular deep models trained

¹⁰<https://colab.research.google.com/>

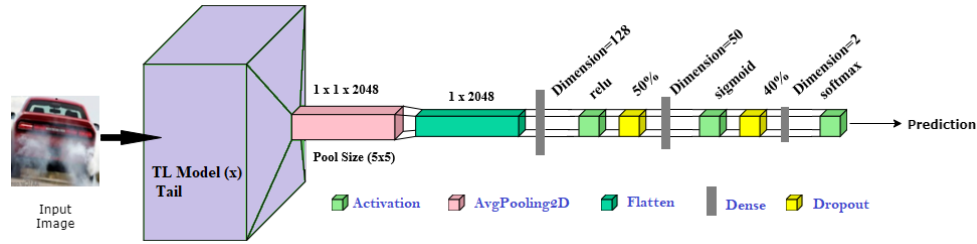


Figure 2.4: The detailed network architecture of the proposed framework. The TL (X) box can be any one of the networks used in our experiment.

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

on ImageNet data set. As these deep models are pre-trained on hundreds¹¹ of different classes, they are able to extract robust representations from the input images that will be helpful for our task as well. Our research is mostly dealing with vehicle and its surroundings, and there are various classes in imagenet data set related to vehicles. Therefore, this can be considered as the related task of those several Deep tasks (which built on top of imagenet). Also to note, “ConvNet features are more generic in early layers and more original-data-set-specific in later layers”¹². Now starting with different pre-trained weights, we have fine tuned the network with our sub task which is only dealing with non-pollutant vehicle and pollutant vehicle in our network. The last few layers mostly will be updated for our task specific problem. We have appended a fully connected (F_c) network followed by our prediction classes on top of the tail of a pre-trained model (X) and trained the F_c network using our data set. In Figure 2.4, we have put the detailed architecture diagram with different DL layers. The tail part of the TL Model (X) has been utilized to get the robust image representation through the shared weight values. The framework starts with an average pooling followed by a flatten. Subsequently, 3 sets of consecutive dense and different activation layers have been used with an intermediate dropout. These dropout layers made the network more generic which help to overcome the over-fitting problem. The pool size in the average pooling (2D) step has been fixed to (5×5) with the same stride. The padding mode is set to valid in this pooling operation. Before using the fully connected network we have flattened the nodes. Thereafter, the first dense layer with the dimension of output space 128 has been placed. We have used *relu* as the activation which converts all the negative values to zero. The next two dense layers with the output dimensions 50 and 2 (to match with the class count) consecutively have been placed. The adjacent corresponding activation layers are *sigmoid* and *softmax*. The initial dropout percentage has been kept as 50%. Subsequently, it is reduced to 40% as shown in Figure 2.4. *Adam* optimizer has been deployed during the training with *binary-cross-entropy* as the loss function. The optimizer parameters β_1 , β_2 have been set to the default values 0.9, 0.999 with epsilon as 10^{-7} . The loss Loss_{BCE} has been defined in the Equation 2.1. Here, $\text{con}(I_i)$ denotes the confidence level of prediction for i^{th} image I_i . O_i is the original class label for the corresponding i^{th} image. I_{total} is the total number of training images.

$$\text{Loss}_{\text{BCE}} = -\frac{1}{I_{\text{total}}} \sum_{i=1}^{I_{\text{total}}} (O_i \cdot \log \text{con}(I_i) + (1 - O_i) \cdot \log (1 - \text{con}(I_i))) \quad (2.1)$$

The learning rate has been experimentally set to a low value of 0.0001. We have

¹¹<https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a>

¹²<https://cs231n.github.io/transfer-learning/>

iterated the network for 20 epochs and the batch size has been kept to 40 during the training for each of the models. To note, the spatial input image size to each DL network has been kept to 224x224 for both the data sets.

The naming convention for the abbreviated TL models is like, TL-x denotes 'TL with x'; for example, TL-IncV3 indicates TL with Inception-V3 model.

2.6 Experimental Findings

In our experiment, in each iteration, every model has been trained for 20 epochs. During training, we used a disjoint validation set of size equals to 10% of the training set size. The training loss has been decreased gradually. This loss values for each 7 models have been plotted in Figure 2.5 for the mentioned iterations. From the figure it is clear that TL-VGG16 and TL-MNet3 are having the maximum and minimum average loss respectively between the 7 models used in our experiment. Similarly the accuracy of the all 7 models have gradually increased with iterations while TL-MNet3 has the maximum average accuracy compare to the others as shown in Figure 2.6.

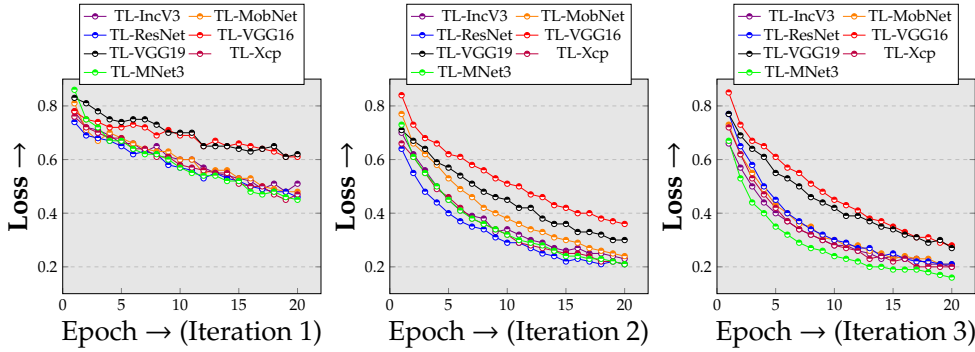


Figure 2.5: Training loss comparison for different deep models for first three iterations.

The result has demonstrated the superiority of the proposed approach in every iteration. As plotted in Figure 2.7, the initial training image count was kept to 600 (for DATASET₁) which was 20% of the overall image set. Gradually with the iterations, our proposed approach has injected image samples for training based on the models' outcome. In the second, third and fourth iterations, $mp_d = 889, 625$ and 148 images have been added in the training set respectively. While increasing the training samples in each iteration, only high confident images whose confidence are greater than 0.7 (θ^*) have been considered. This, in turn, increases the prediction accuracy thereby improve the overall efficiency of the model. Similar observation has been shown for the DATASET₂ which is placed in the supplementary.

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

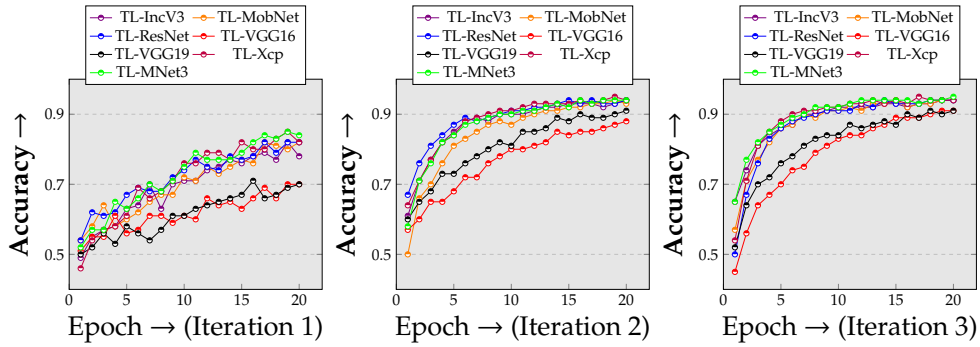


Figure 2.6: Training accuracy comparison for different deep models for first three iterations.

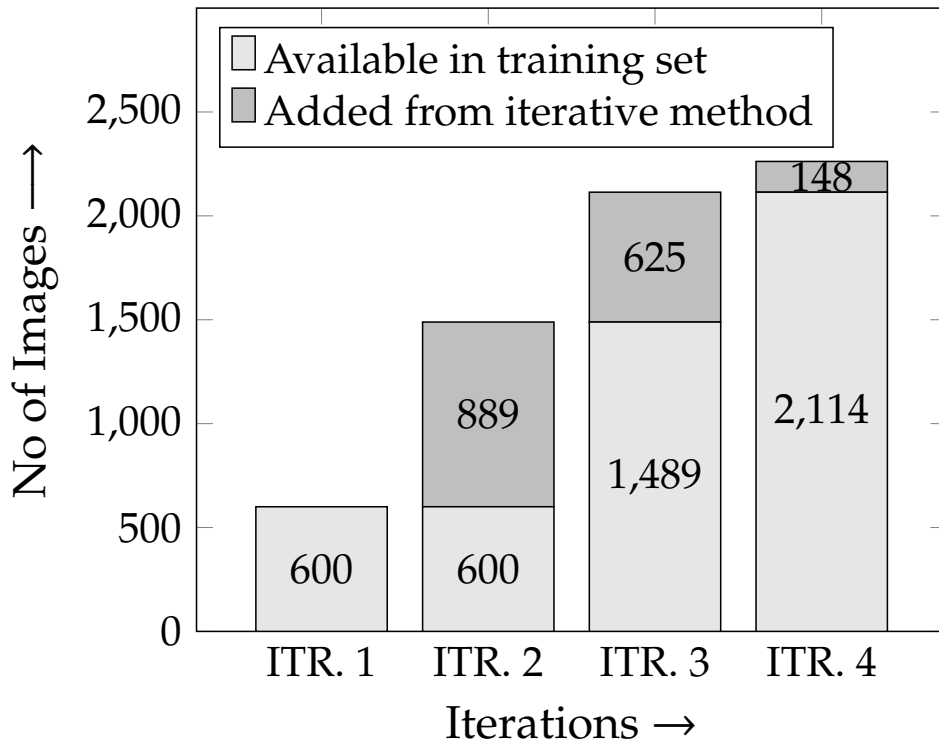


Figure 2.7: Increase of images in training set in consecutive iterations based on the testing output with higher confidence. Originally, we have 600 labeled data samples in DATASET₁.

2.6. EXPERIMENTAL FINDINGS

Any coordinate (a, θ) in Figure 2.8 denotes the $a\%$ of correctly predicted images with θ confidence level. Here, the domain of $\theta \in (0, 1)$. The objective is to increase the value of confidence level and testing accuracy simultaneously. We have considered those images for the next training iteration where the threshold value is greater than θ^* as pointed in Figure 2.8. As can be seen from Figure 2.8, in iteration 1, the prediction accuracy with confidence parameter $\theta \geq 0.5$ is 71.83%, but for $\theta \geq 0.9$, the accuracy drastically falls down to 7%. It indicates the fact that at the initial stage, the models are not confident enough to provide their decisions for sufficient number of test images. However, in iteration 2, the positive prediction accuracy substantially increased to 42.66% for $\theta \geq 0.9$. The similar improvement has also been observed in consecutive iterations. In iteration 3, the accuracy for $\theta \geq 0.9$ is 56.45%. So, this iterative approach, not only increased the accuracy percentage, but the trust factor of the models has also been increased which is a very important aspect in any deep learning model. The \uparrow in the Figure 2.8 has shown the nature of confidence increment of our proposed approach. In the experiment we have chosen $\theta^* = 0.7$ to select the testing images with high trust factor to enrich the training data set. Eventually, the curve becomes straighter with the increased iterations which are clearly observed in the same figure. This demonstrates the fact that the proposed technique increases the confidence and accuracy level simultaneously. This aspect cannot be achieved in traditional training testing paradigms. In every iteration we have also plotted the confusion matrix based on the output produced by our proposed method. Figure 2.9 and 2.10 have shown two sample confusion matrix sets with three sub-images each for every iteration of our experiment at point $\theta = 0.5$ and $\theta = 0.9$ respectively. In Figure 2.9, we can see that the percentage wise True-Positive and True-Negative counts are increasing and the same representation has been observed in Figure 2.8. In first iteration, the accuracy was 72% and in next two iterations it has increased to 77% and 82% respectively. Similar observation is applicable for the Figure 2.10 as well at $\theta = 0.9$. Figure 2.11 has shown the percentage of testing accuracy with different iterations as bar chart. It can be noted that the testing accuracy percentage has been increased significantly with 3 iterations only. For example, TL-IncV3 has reached to 86% accuracy starting from 82%. In the same way, other models also have gained the similar accuracy growth. The average testing accuracy has been jumped to 85.1% from 78.4%.

We have compared our model with other state-of-the-art techniques as well. In Table 2.4, we have shown the comparison with two feature based approaches and with one Deep network. HLTP has achieved 71.1% accuracy while the LBP technique has reached to 69.3% over the testing data images. The DNCNN accuracy has reached to 79.9%, while our proposed technique has achieved 82.1% accuracy over the same test set. SmokeNet, slic-CNN and kpca-CNN lags in this

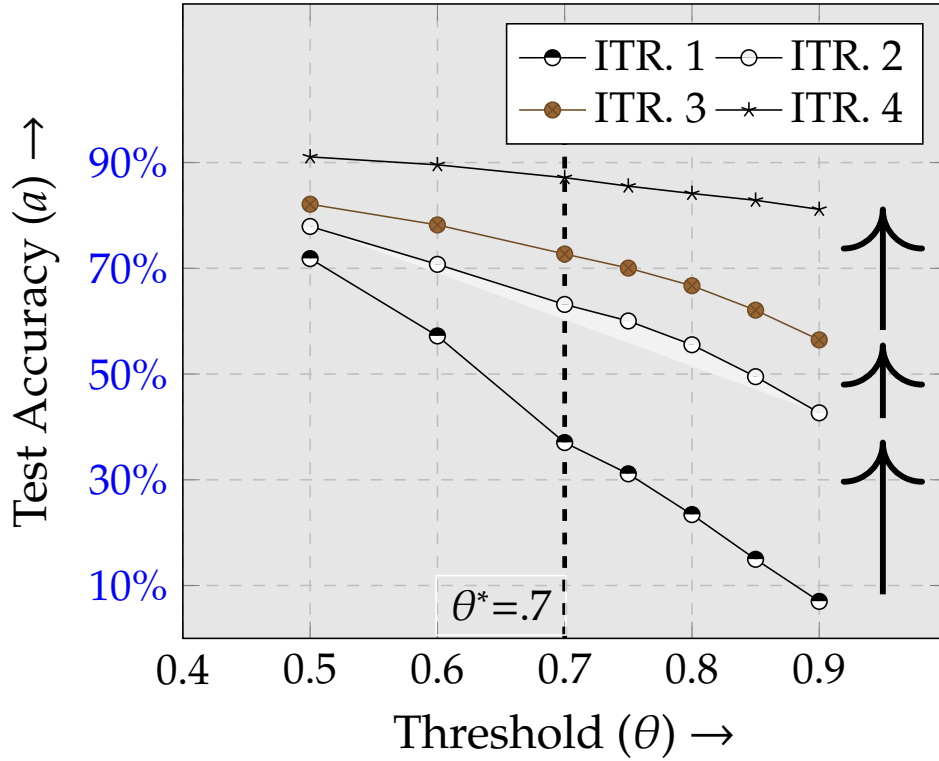


Figure 2.8: Testing accuracy of prediction in consecutive iterations for different confidence levels.

	P	NP		P	NP		P	NP		P	NP
P	20 %	20 %	P	21 %	20 %	P	25 %	16 %	P	32 %	7 %
NP	8 %	52 %	NP	3 %	56 %	NP	2 %	57 %	NP	2 %	57 %
(a)			(b)			(c)			(d)		

Figure 2.9: Sample Confusion Matrix for the proposed method with confidence level ≥ 0.5 . P=Pollution; NP=Non-Pollution. (a) Iteration 1, (b) Iteration 2, (c) Iteration 3, (d) Iteration 4

comparison as well. Even in terms of F1-Score, as shown in the same Table 2.4, it is understandable that our proposed technique has dominated significantly. Besides these, In Table 2.5 we have reported Cross data set evaluation with the synthetically generated rainy images and foggy images. Here also, for the rainy image set, HLTP, LBP, DNCNN, SmokeNet, slic-CNN and kpca-CNN have achieved 66.5%,

2.6. EXPERIMENTAL FINDINGS

	P	NP		P	NP		P	NP		P	NP
P	3 %	38 %	P	16 %	24 %	P	17 %	23 %	P	35 %	12 %
NP	55 %	4 %	NP	33 %	27 %	NP	20 %	40 %	NP	8 %	45 %
(a)			(b)			(c)			(d)		

Figure 2.10: Sample Confusion Matrix for the proposed method with confidence level ≥ 0.9 . P=Pollution; NP=Non-Pollution. (a) Iteration 1, (b) Iteration 2, (c) Iteration 3

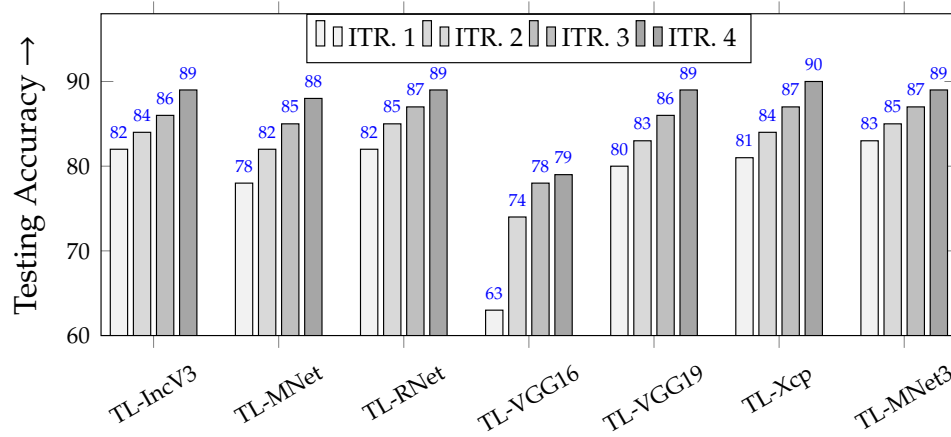


Figure 2.11: Average testing accuracy of different models used in our experiment in different iterations.

63.3%, 74.8%, 75%, 70.5% and 76.5% accuracy while we have reached to 79.5% testing accuracy. The similar results have been observed for the foggy image set as well which are reported in the same Table 2.5.

Figure 2.12 demonstrated how TL-Xcp model was able to learn from training images. In this regard, two popular visualization saliency techniques viz. GradCAM [107] and SmoothGrad [108] with heat map have been considered. Six samples images along with their saliency map produced by GradCAM and SmoothGrad are shown in Figure 2.12 (a), (b) and (c) respectively. The resulting saliency map depicts that the model has learnt mainly the vehicle and its close surrounding parts from the training images. For images with pollution, the smoke part along with the relevant neighboring have been highlighted.

Figure 2.13 has shown some sample images where in the first iteration, the framework has failed to predict correctly while in the next iteration onwards, the images are predicted with more than 90% confidence level. This exhibits the

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

Table 2.4: Comparison of test accuracy for the existing image-based smoke detection methodologies

Algorithm type	Algorithm	Test Accuracy	F1-Score
Feature based	LBP (riu2) [37]	69.3%	67.1%
	TAMURA[33]	67.1%	70.1%
	HLTP[36]	71.1%	68.1%
DL based	DNCNN[40]	79.9%	80.1%
	SmokeNet[39]	81.0%	79.9%
	slic-CNN[41]	78.0%	78.3%
	kpca-CNN[42]	75.1%	75.0%
	Proposed (ITR. 3)	82.1%	85.1%
	Proposed (ITR. 4)	91.0%	88.8%

Table 2.5: Comparison of accuracy in Cross data set evaluation over bad weather images.

Algorithm type	Algorithm	Rainy Image	Foggy Image
Feature based	LBP (riu2)	63.3%	49.1%
	TAMURA	68.5%	58.5%
	HLTP	66.5%	54.1%
DL based	DNCNN	74.8%	53.5%
	SmokeNet	75.0%	78.3%
	slic-CNN	70.5%	45.8%
	kpca-CNN	76.5%	64.8%
	Proposed (ITR. 3)	79.5%	71.3%
	Proposed (ITR. 4)	84.1%	77.3%

effectiveness of the proposed approach. Moreover, the reliability of the decision making process improves due to the use of majority voting scheme. In Figure 2.15, we have outlined one sample flow of the testing process that can be directly deployed in the real surveillance scenario. From the surveillance video, the 'Frame

2.6. EXPERIMENTAL FINDINGS

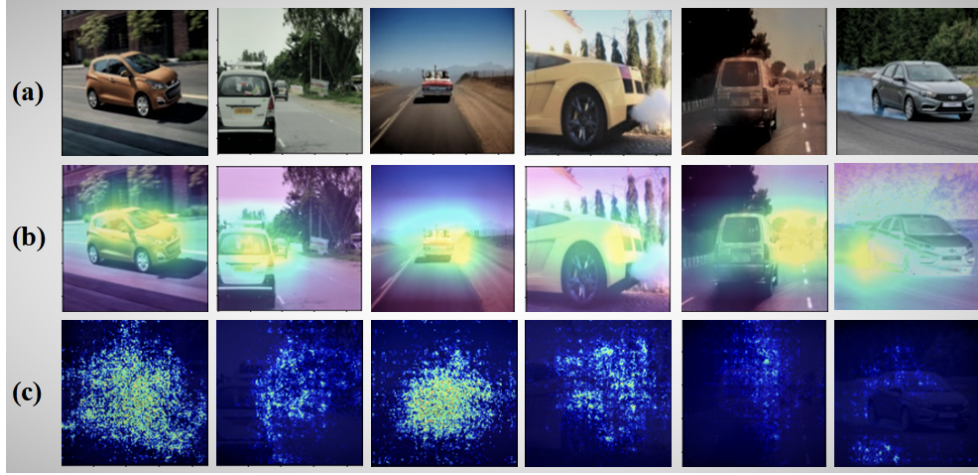


Figure 2.12: Saliency Map generation: a) Original images b) Grad-CAM c) Gradient-Smooth.



Figure 2.13: Sample images where the proposed model failed in first iteration but succeed in next with $\geq 90\%$ confidence.

Extractor Module' will process individual frame and feed it to the next 'Vehicle Detector Module' as shown in Figure 2.15. Here, we can detect the vehicles through any object detector model like YOLO, SSD etc. Afterwards, our proposed approach will classify and determine whether it is polluting or not. The simultaneous tracking of the number plate is another important aspect to make the overall surveillance process fully automatic. However it is beyond our scope of research.

Ablation study: We have run those seven DL methods without the iterative approach for both the data sets. For DATASET₁, Inception-V3, MobileNet-V2, InceptionResNet-V2, VGG16, XceptionNet, VGG19 and MobileNet-V3 have achieved 83.1%, 80.37%, 82.3%, 67%, 82%, 79% and 80% respectively. Here we have taken the average of one K-Fold run where the K has been set to 5. Result for DATASET₂ without iterative process has been presented in the section 4 of the sup-

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

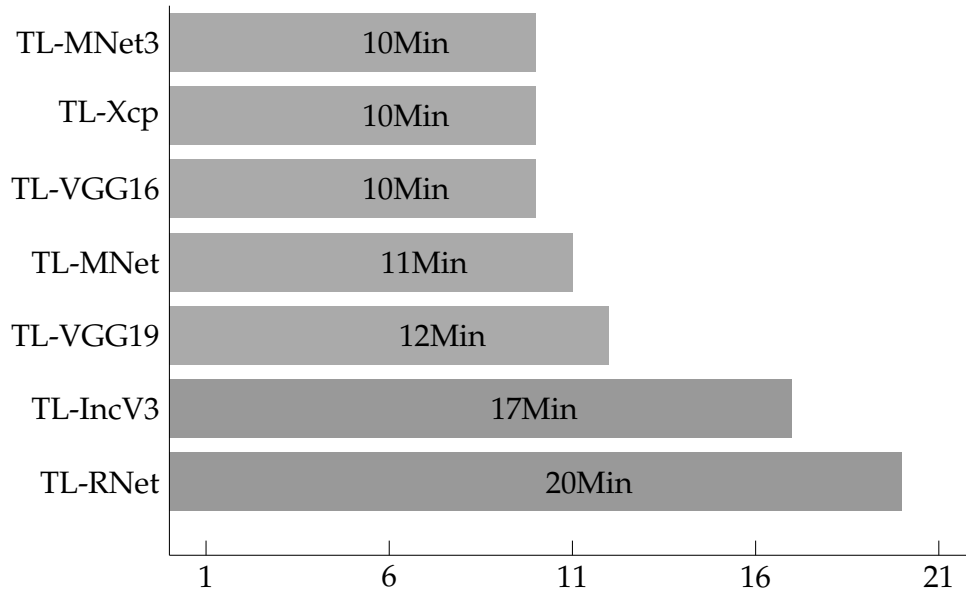


Figure 2.14: Training time taken by different TL-based models used in the experiment for 20 epochs in iteration 3 with 2114 images.

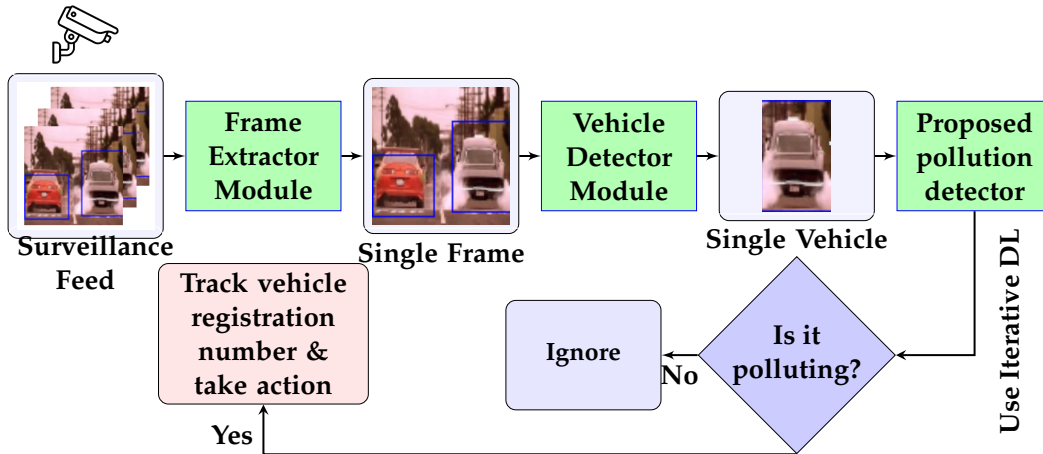


Figure 2.15: Proposed testing framework for on-road vehicle pollution surveillance for the first level of screening.

plementary. We have observed that our framework outperforms the traditional training in every cases in terms of accuracy.

Time Analysis: We have performed our experiment in the Google Colab platform. Figure 2.14 has shown the the over all timing for individual network to

2.7. SUMMARY AND FUTURE DIRECTIONS

complete the training in Iteration 3 where the training image count was 2114 and the batch size has been kept to 40. At the same time the average testing time for one image are like 0.04 Sec, 0.03 Sec, 0.06 Sec, 0.03 Sec, 0.03 Sec, 0.02 Sec and 0.04 Sec respectively for TL-IncV3, TL-MNet, TL-RNet, TL-VGG16, TL-VGG19, TL-MNet3 and TL-Xcp. As we are considering all the models to take a single decision, therefore the overall timing is 0.3 Sec per image including the decision making logic. Moreover if we can deploy parallel computing framework where the decision will be made simultaneously we can make it faster using the following timing calculation. $T_{test} = \max(t_1, t_2, ..t_M) + t_d$. Here, t_i denotes the prediction time for i^{th} model and M is the number of models. t_d is the time to execute the final decision which is approximately 0.001 Sec here. In our experiment, T_{test} will be 0.06 second which is quite faster compare to the manual checking. In video surveillance, we can support 16fps (frames per second) processing with the current setup in colab as mentioned earlier. If high end hardware has been used then timing can be improved further. Therefore our model can be integrated with any current spec (speed camera)¹³ type system for pollution surveillance.

2.7 Summary and Future Directions

We have developed an efficient feedback based vehicle pollution tracking strategy by exploiting the advantages of several Deep Learning models. The performance of our method has been compared with other state-of-the-art techniques over the various real life images which demonstrates the superiority of the same. The accuracy in the Cross data set evaluation has supported the robustness of the process as well. Our proposed approach is generic in nature and can be applied to any similar deep learning based model. This framework combines the decision of several models with higher confidence level. It can be used effectively in the Intelligent Transport System to identify the on-road pollutant vehicles.

However, there are some limitations associated with the framework. It would be really difficult to handle the scenarios of invisible smoke and dense fog. During the image generation through style-GAN, the variation of generated night mode images is dependent on the number of styler references. Also, specifically for the images with pollutant vehicles, the smoke becomes invisible in case of very dark styler image. In case of motion blur and occlusion of frames, the detection of number plate and vehicle will be tough; this can indirectly hamper the automatic complain lodging process. Also regarding the camera shooting parameter, decent resolution is expected to see the smoke properly. As a scope of future research, one can work to enhance the database from different perspectives and can apply better Deep Learning based tracking mechanism for the surveillance. Color based

¹³[https://en.wikipedia.org/wiki/SPECS_\(speed_camera\)](https://en.wikipedia.org/wiki/SPECS_(speed_camera))

CHAPTER 2. SMOKEY VEHICLE RECOGNITION

pollution classification is another area where we can apply a shallow network to determine the possible pollution type. At the same time, for efficient government tracking purposes, full automation of monitoring requires simultaneous identification of the smoky region on the road and respective concerned vehicle along with. In the next chapter we have mainly focused on this task.

3

Vehicle Smoke Embedding and Smokey Region Detection

3.1 An overview of smoke embedding and smokey region Detection

IRTS is evolving fast through different innovations in AI and Computer vision in recent times. Researchers are trying to minimize human involvement with significant efficiency in several activities in this domain. AI based IRTS includes pedestrian management, traffic load based signaling mechanism, vehicle speed control, license plate recognition, and essentially, pollution monitoring [45, 109]. Vehicle pollution identification and automatic complain lodging is one of the important aspect of smart city building as vehicle pollution is a major source of air pollution. Due to this fact, we are aiming to build 100% Electric Vehicles (EV) in every road of the world. However, in the developing countries, considering the infrastructure, it will take another decade to mitigate the goal. Besides this, there has some other disadvantages of EV as well, like high cost of the battery and its associated components, short range cover due to their smaller battery capacity, limited charging stations available in the current infrastructure and importantly EV also have environmental impact while manufacturing the battery itself. Therefore, researchers need to focus on vehicle pollution identification with recent cutting edge innovations. We can restrict vehicle pollution through two ways like proactive and reactive mechanisms. The first category includes improvising the existing fuel burning process to reduce the emission, preparing better vehicle engine etc. where in the second process we need to focus on identifying pollutant vehicles [45], monitoring highly polluted areas, enforcing traffic rules etc. Accurate detection of vehicle smoke is deemed necessary so that traffic authorities can take necessary measures from on-road surveillance to reduce the emission. Deep network-based approaches have been used immensely in IRTS in recent times [110, 96], while few approaches are there to identify and restrict air pollution [97, 98].

Here, we've divided our tasks into two segments. Initially, our focus was on

CHAPTER 3. VEHICLE SMOKE EMBEDDING AND SMOKEY REGION DETECTION

creating a comprehensive data set for the identification of vehicle smoke. Subsequently, utilizing this data set, we successfully identified smoky vehicles along with their respective smoke regions.

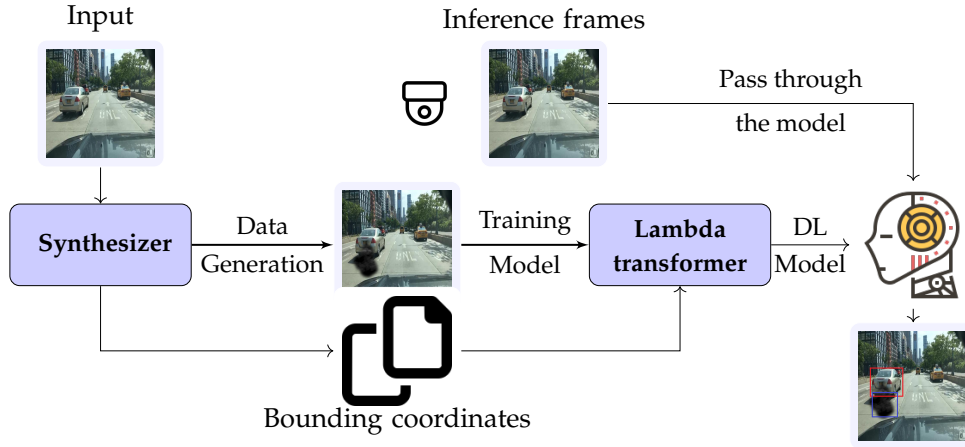


Figure 3.1: The flow of the proposed framework; the synthesizer of smoke module and lambda-based transformer module for detection modules are highlighted blue

3.2 Smoke Generation

Plenty of data sets of on-road automobile images are available on the web. Many researchers have collected on-road vehicle photos for different purposes. Some of them either focus on the road surveillance [111], while others have been used for object tracking. However, publicly available vehicle pollution data sets are not readily available in the literature. This non-availability of the data set creates an imbalance during the training for identifying pollutant vehicles. Even if we install an on-road vehicle surveillance camera in a particular place, the ratio of pollutant and non-pollutant vehicles will be very less, causing a drastic class imbalance in collected images.

To eliminate this problem, we have proposed a novel smoke generation process to help to overcome the data set imbalance issue by generating a sufficient number of images. In this data set creation process, we have collected real smoke instances, along with the publicly available images collected from the web. To note, our proposed synthesis algorithm can be easily applied to any available on-road vehicle data set to produce the paired pollutant vehicle data set. For the experiment purpose, we have used three publicly available data sets as baseline; Berkeley Deep Drive (BDD) [112], Stanford Cars [113], and Boxy [114]. Figure 3.2 has shown a few samples (in terms of color and density) of such synthesis on the BDD data set.

3.2. SMOKE GENERATION

Our proposed method can help researchers to build an unbiased model during the training phase. Algorithm 2 shows the detailed outline for the vehicle smoke generation steps. We will elaborate on the detailed methodology in upcoming subsections.



Figure 3.2: Sample generated images using our proposed technique. Light smoke images which are very common in the traffic while dark smokes which are rare but easily detectable; 1) whitish light smoke 2) grayish dark smoke 3) grayish light smoke 4) bluish light smoke 5) very dark black smoke.

3.2.1 Statistics of the Dataset used in the experiment

We utilized a subset of the three previously mentioned on-road vehicle data sets in our experiment. Smoke has been embedded into the rear-end of the on-road vehicles to generate a relatively large data set for bench-marking our detection model. Table 3.1 outlines the count-wise statistics during the training, validation and testing phase.

Table 3.1: Image count statistics of three different baseline data sets which have been modified to smoky vehicle data sets for our experiment

Dataset	Total	Training	Validation	Test
BDD [112]	2436	1500	200	736
Stanford Cars [113]	1861	1161	200	500
Boxy [114]	2330	1680	150	500

3.2.2 Proposed Smoke Synthesis Algorithm

In this section, we have presented a novel smoke generation algorithm that generates and overlays synthetic yet realistic smoke on the vehicle exhaust. We have focused on many aspects like the position of the smoke, color, shape, density and similarity with real smoke etc. while embedding the same on any vehicle image. Regarding the color of the smoke, there are three primary colors of the vehicle smoke based on the chemical gas components from the exhaust pipe have been observed. NO_x compounds are generally of Reddish brown, Particulate Matter (PM) which shows dark gray variation and Volatile Organic Compounds (VOCs)

CHAPTER 3. VEHICLE SMOKE EMBEDDING AND SMOKEY REGION DETECTION



Figure 3.3: Samples of realistic smoke samples scraped from the internet and considered in the smoke synthesis process.

(Smog type white)¹. These colors are mainly produced due to the functional issues linked to the engine². Different smoke patterns were taken from the real instances (see Figure 3.3). We have created a variations of these patterns with image augmentations to make the generation process more generic.



Figure 3.4: Combining and fitting multiple ellipses onto natural smoke in the boundary region.



Figure 3.5: Sample mask patterns used in the experiment for embedding; First two are Elliptical Pattern while third one is a perturbed circle and the last pattern is a funnel shaped towards left

On the other hand, the shape of the smoke seems to be random to naked eyes. However, smoke from the exhaust of a vehicle is known to reasonably resemble a collection of ellipses [115, 116, 117]. On small scale, we can decompose multiple elliptical shapes to generate a big shape as well. This shape strongly depends on the vehicle motion's direction as well. For synthesizing the shape of the smoke, small regions from the boundaries of different ellipses were combined end-to-end

¹<https://www.ucsusa.org/resources/cars-trucks-buses-and-air-pollution>

²<https://www.universal-tyres.co.uk/news/exhaust-smoke-colour-warnings/>

to form a cloud-shaped mask for the smoke. This assumption of the smoke being composed of small parts of ellipses was validated by manually fitting elliptical curves on realistic smoke patterns (as shown in Figure 3.4). The ‘Turtle’ module was used for generating cloud-shaped masks for the smoke. In addition to the elliptical shape, we took into account the arbitrary outline of the smoke with the help of Bézier curves. We have selected any one of these shape patterns randomly and pre-processed it before embedding into the given vehicle image. Sample shape patterns have been depicted in Figure 3.5. Another essential point regarding the generated embedding is that in the border region the smoke gets gradually thinner and disappeared and the background becomes prominent. To mitigate this issue, we have applied a Gaussian filter in the selected smoke pattern. The smoke pattern has its original high density in the center and the density gradually lowers in the border region. The algorithms to generate these mask patterns are explained in the Supplementary Material (See Algorithm 1 and 2 in the supplementary). After that, the mask is inverted by subtracting each of its entries from one. A Gaussian filter is applied to the inverted binary mask to make it smooth/continuous in the border area mainly. Note that the kernel/radius parameter (σ) controls the extent of blur induced into the mask by the filter. The resulting continuous mask, with pixels lying between zero and one, is then used to filter out the required portion of the actual smoke based on the continuous pixel values of the mask. The pre-processed smoke is finally overlaid on the input vehicle image based on the exhaust coordinates. The exhaust coordinates for all the vehicles in a given image are obtained using a YOLOv3 [47] model trained to detect the rear-end of vehicles. Note that the above algorithm can easily be extended to handle smoke embedding into multiple vehicles in an image by looping over all the rear-ends detected by the same trained model. Figure 3.6 has shown the overall flow of the synthesis process.

We use the Gaussian blur to process the smoke’s mask. Specifically, the blur function is applied on the mask to make the border smooth and move towards transparency radially outwards. The Gaussian function is of the following form:

$$G_{\sigma}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ is the specified kernel size. Note that the equation is applied to each of the pixels (x, y) in the smoke’s mask.

3.2.3 Validation of Smoke Synthesis

We have validated our generation through user feedback for randomly selected 40 images. 9 images are of original smoke and remaining are synthesized smoke images. We have taken random review of those images through a Fake/ Real

CHAPTER 3. VEHICLE SMOKE EMBEDDING AND SMOKEY REGION DETECTION

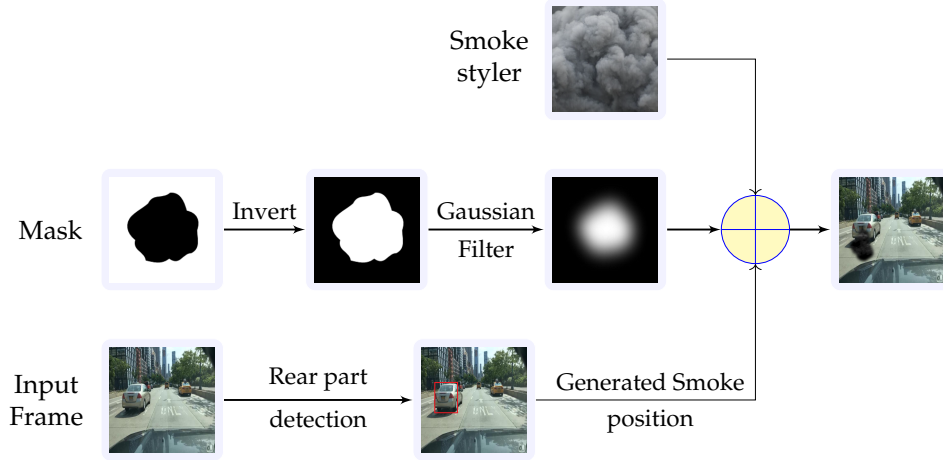


Figure 3.6: Flow chart of automatic vehicle smoke generation.

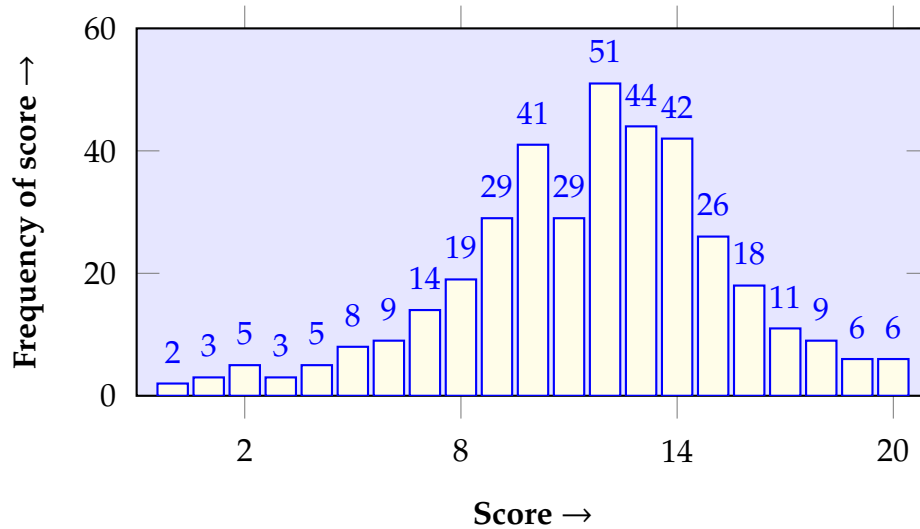


Figure 3.7: User feedback status collected online on randomly selected 40 images. Every right choice awards 1 mark while wrong provides 0.

online voting. Based on Total 360 voting count, we have drawn the histogram of points scored (Each correct answer gives 1 mark while wrong answer gives 0). From figure 3.7, it is clear that maximum candidates have scored 12 with mean 11.6 which proves that they are failed to identify the generated images as fake. This fact indirectly establishes that the generation process has been good with compare to real image set.

3.3. VEHICLE SMOKE DETECTION

Algorithm 2 Vehicle Smoke Embedding

Input: Image with vehicle (I_v), smoke template (s), and mask pattern (M_p)

Output: Image with smoke-embedded vehicle (O_v)

```

#define  $\sigma$  (kernel size)
#define N (number of points on ellipse)
#define e (number of sharp edges for the circle)
#define bw (border thickness; default is 0)
if pattern type is elliptical then
    mask_pattern  $\leftarrow$  generate_ellipse(N) //See Algorithm 1 in Appendix
else if pattern type is perturbed then
    mask_pattern  $\leftarrow$  random_shape(e) //See Algorithm 2 in Appendix
mask_pattern  $\leftarrow$  invert(mask_pattern)

if is_high_contrast(s, threshold = 0.075) then
    // double the kernel size
     $\sigma \leftarrow 2\sigma$ 
( $x_{crop}$ ,  $y_{crop}$ )  $\leftarrow$  dimension of the mask_pattern
if  $\sigma > 25$  then
    // For larger kernel size decrease the effect of border visibility through increasing the
    border; set the crop window on the same basis
    bw  $\leftarrow 2 \times \sigma$ 
     $x_{crop} += 2 \times bw$ 
     $y_{crop} += 2 \times bw$ 
    //create a fake border around the mask pattern
    mask_pattern  $\leftarrow$  adjust_border(mask_pattern, bw)

mask_pattern  $\leftarrow G_\sigma$ (mask_pattern)
where  $G_\sigma(M_p) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{M_p^2}{2\sigma^2}}$ 
rear_end_coor  $\leftarrow$  get_exhaust_area_vehicle( $I_v$ )
//use opencv.paste
 $O_v \leftarrow paste_{smoke}(x, s, mask\_pattern, rear\_end\_coor)$ 
return  $O_v$ 

```

3.3 Vehicle Smoke detection

In the second phase of our experiment, we utilized the synthetically generated data sets to train our deep CNN-based detector network. For identifying the vehicle smoke and its originator, we need to have a robust detector module as well. Here, we have designed and developed a novel attention-based network for this purpose. In this section, the proposed architecture for smoke and the corresponding smoke-emitting vehicle detection has been described in detail. The

CHAPTER 3. VEHICLE SMOKE EMBEDDING AND SMOKEY REGION DETECTION

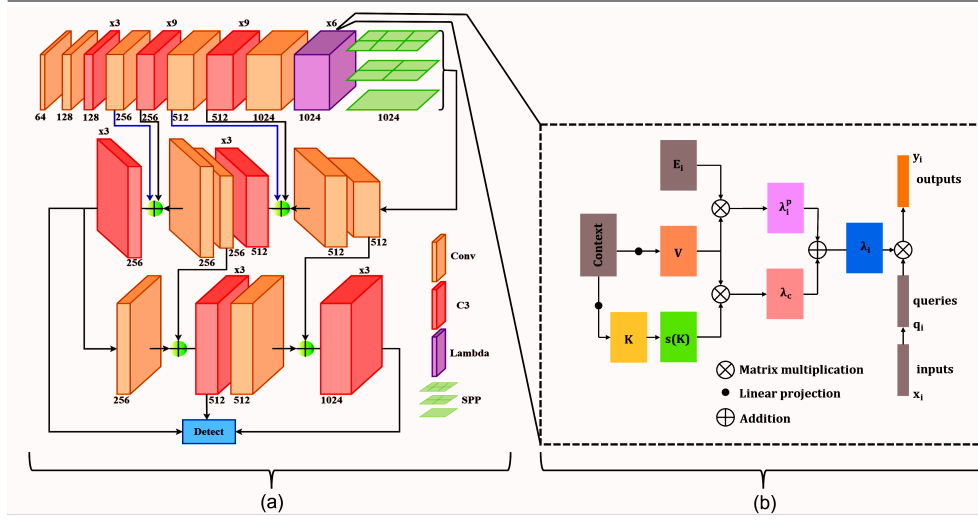


Figure 3.8: The proposed architecture of deep network. (a) First row depicts the **Backbone** layer, second row depicts **Head** while final row depicts the **Detection** layer. (b) Lambda implemented attention layer has been shown here.

main focus is to enhance the backbone of the YOLOv5 network by introducing of the attention layer. We have implemented this attention module with the lambda concept for generating the attention map, which is computationally less expensive. The overall experimental setup and the results for all these data sets have been elaborated in this section subsequently.

3.3.1 Proposed attention based object detector network

The backbone of our proposed architecture is inspired by CSPNet [118] and lambda implemented attention network [119]. Feature Extractor (FE)/ Backbone layer has been enriched with the attention-based module to provide a robust input to the Detection layer through the Head. There are few basic components present in this FE layer, as shown sequentially in Table 3.2 along with the depth of repetition. The CSP-inspired network has been integrated with convolution blocks (Conv), and three stacked convolution layers known as C3. This C3 layer helps to reduce the duplication in gradient information prevalent in the usual DenseNet [120]. At the end of FE, to perform the initial aggregation of extracted 2D features effectively, a Spatial Pyramid Pooling (SPP) [121] layer is incorporated. The SPP layer ensures that there is no loss of information due to cropping/ warping of input images by allowing for features of different sizes to be aggregated. It is worth noting that the outputs of the SPP layer has ensured to have a constant length. In addition, SPP uses a fixed number of 'spatial bins' (average/max pooling layers) which pool the individual channels of the features to form a vector representation of a fixed

3.3. VEHICLE SMOKE DETECTION

length. This is advantageous in the sense that the aggregated features are uniform throughout images of different sizes and hence, do not require any pre-processing in this regard before being passed through the next layers of the architecture. The proposed model's head has been modified by introducing extra deep connections compared to the standard PANet [122] used in YOLOv5. Here, we improved PANet's architecture for the task of smoke and smoke-emitting vehicle detection by adding residual connections starting from convolution blocks, as shown in blue lines in Figure 3.8a. This enhancement has been made to ensure an efficient flow of extracted features, especially the low-layer features, for making better use of localized information. Finally, the standard YOLOv5 detector was used to aggregate the feature maps from the last three C3 layers to predict the class and the bounding box of all the targets in the image. The complete architecture, including a closeup of the lambda network, is shown in Figure 3.8. Table 3.2 has shown the component-wise backbone structure with layer information. Note that, a single convolution block was constructed as a sequential list consisting of a convolution layer with no bias, a batch normalization layer, and an activation function (sigmoid in our case). All the convolution blocks in the backbone were initialized with a kernel size of (3×3) and the stride has been fixed to 2. Additionally, the different kernel sizes have been used to construct the SPP layer like 5, 9, and 13.

Table 3.2: Different backbone layers of the proposed network.

Sequence	Repetition	Layer	Output channels
1		Conv (Focus)	64
2		Conv	128
3	3 times	C3	128
4		Conv	256
5	9 times	C3	256
6		Conv	512
7	9times	C3	512
8		Conv	1024
9	6 times	Lambda	1024
10		SPP	1024

The lambda-implemented attention block has been introduced into our backbone of the proposed network. The lambda network captures long-range interactions between the input image and the context of the smoke/smoke-emitting vehicle, i.e., information of pixels surrounding the pixels of smoke/smoke-emitting vehicles. It is particularly efficient due to its ability to showcase attention without calculating the computationally expensive attention maps. The detailed structure is shown in Figure 3.8b. The lambda concept has been developed on top of the attention module, where the contextual information gets priority. In the lambda layer, the context is linearly projected into keys (K) and values (V). Then, the keys

CHAPTER 3. VEHICLE SMOKE EMBEDDING AND SMOKEY REGION DETECTION

are normalized by applying the softmax function (s) across them, termed as $s(K)$. A dot product is then performed between $s(K)$ and V to yield content lambda (λ^c), which encodes the information to transform each of the queries (q_i , where i is the query position index) solely based on the context. Simultaneously, a positional lambda (λ_i^p), corresponding to the i^{th} query position is calculated using the positional embedding (E_i) by taking its dot product with V . The positional lambdas are meant to capture the information to transform queries based on their positional information. Finally, the i^{th} positional output (y_i) is deduced by applying a dot product between q_i and the corresponding contextual lambda (λ_i), which is the sum of λ^c and λ_i^p . Each of the lambda-layer outputs can be mathematically represented as:

$$y_i = \underbrace{(s(K)^T V)}_{\lambda^c} + \underbrace{E_i^T V}_{\lambda_i^p} q_i$$

3.3.2 Experimental Setup

We have developed and deployed the above mentioned network on two different platforms and executed various experiments there. Google Colab³ provided Tensor Processing Unit (TPU) which has Linux (Ubuntu-18.04) as the baseline operating system with 12.7 GB Graphics RAM. On the other hand, another GPU Quadro RTX 6000 with 24 GB RAM (With CUDA 11.6) has been used to execute most of the experiments for real data sets. The base line operating system in this GPU machine is Windows 10. The three data sets mentioned above were modified to include smoke based on the proposed algorithm and were used for training and evaluating our models.

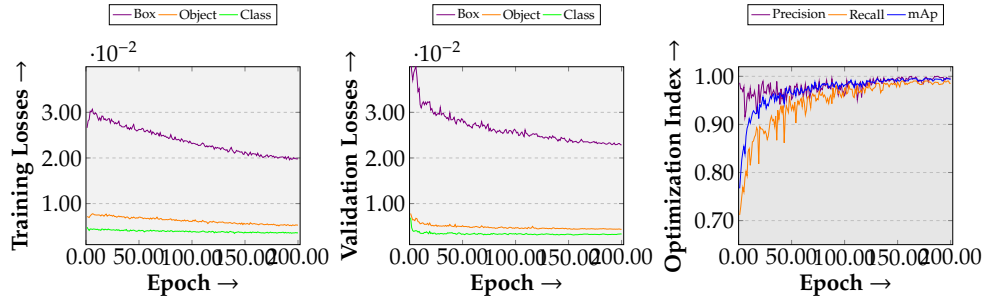


Figure 3.9: Validation loss and accuracy comparison for different detection models for BDD data set during training phase.

We trained and built all the models using all three data sets under different experiments for the same number of epochs, i.e., 200. We have ensured that within

³<https://colab.research.google.com/>

3.3. VEHICLE SMOKE DETECTION

Table 3.3: Results, using Cross data set inference, in the form of precision, recall, mAP@0.5 and mAP@[0.5, 0.95], all in percentage (%), obtained using different models on the test set (complete data set) corresponding to the mentioned data sets. Metric scores in **bold** indicate best relative performance.

Train set (image count)	Test set (image count)	Model	Precision	Recall	mAP@0.5	mAP@[0.5, 0.95]
BDD (2436)	Stanford Cars (1861)	YOLOv3	92.9	80.2	87.7	60.3
		YOLOv5s	95.0	80.7	89.9	63.5
		Proposed (Lambda)	96.1	82.3	90.7	59.6
	Boxy (2330)	YOLOv3	96.3	96.1	98.3	69.3
		YOLOv5s	97.0	95.9	98.2	68.2
		Proposed (Lambda)	97.0	96.3	98.4	70.7
Stanford Cars (1861)	BDD (2436)	YOLOv3	83.4	71.7	77.0	51.2
		YOLOv5s	81.1	74.2	77.9	48.9
		Proposed (Lambda)	83.5	73.1	78.6	51.3
	Boxy (2330)	YOLOv3	87.7	70.7	76.3	36.8
		YOLOv5s	84.7	82.1	86.0	42.7
		Proposed (Lambda)	89.5	78.3	83.8	43.5
Boxy (2330)	BDD (2436)	YOLOv3	80.5	54.1	62.5	36.9
		YOLOv5s	93.3	62.5	75.2	49.0
		Proposed (Lambda)	93.7	75.6	85.7	60.1
	Stanford Cars (1861)	YOLOv3	45.3	38.2	33.0	10.2
		YOLOv5s	60.4	52.9	54.3	20.9
		Proposed (Lambda)	70.0	65.1	69.5	34.6

these epochs, the models got stabilized with respect to different losses (box loss, object loss and class loss) and the index parameters have been saturated and not improving for the validation set. Figure 3.9 has supported this fact in this regard. In the same figure, we have shown one sample validation loss decreases and index values (Precision, Recall and mAP) increase for the BDD data sets. Similar nature has been observed for the other experiments as well. We have set the batch size to 8 for all the training tasks. The stochastic gradient descent (SGD) optimizer was used for training, with the initial and final learning rates set to 10^{-2} and 10^{-3} , respectively. We have taken the input image size of 256×256 . We incorporated image augmentation techniques before the training, such as RGB to HSV transformation, translation (fraction: 0.1), scaling (up to 0.5), and horizontal left-right flip. To ensure that our model has not overfitted during the training, we have applied cross data set approaches in different experiments. From the results of those experiments, we have proved that our proposed model has outperformed others in terms of different index parameters.

3.3.3 Results

For establishing the efficiency of our proposed model, we have utilized three index parameters, i.e. precision, recall and mAP, over all the data sets. The separated testing images have been used for this purpose, as mentioned in Table 3.1. We have compared our method with two recent well-known object detectors, i.e. YOLOv3 and YOLOv5. The overall test results for all the models on different data sets

CHAPTER 3. VEHICLE SMOKE EMBEDDING AND SMOKEY REGION DETECTION

Table 3.4: Overall real test results in terms of three indexes, all in percentage (%), obtained using different models on the test set (880 images). The training has been done on Boxy data set. Metric scores in bold-blue indicate best performance.

Model	Precision	Recall	mAP@0.5	mAP@[0.5, 0.95]
YOLOv3 [47]	87.8	83.1	82.9	48.4
YOLOv5s [49]	90.3	84.6	86.7	52.1
Proposed (Lambda)	91.2	86.9	88.2	54.1

is summarized in Table 3.4. Here the testing has been performed on real data sets while the training has been done on Boxy data set. As evident from the numbers, the proposed architecture incorporating the Lambda network performs better than YOLOv3 and YOLOv5s in terms of precision, recall, mAP@0.5 and mAP@[0.5, 0.95]. For all the indexes, our proposed method outperformed the other techniques. We have visually compared our results with the other state-of-the-art techniques as well.

Cross data set Evaluation: We incorporated the Cross data set learning-based evaluation technique to showcase our model’s ability on generalization on the unseen test sets. In this process, we trained the models on a single data set and tested their performance using the rest of the data sets considered in our study. For example, as shown in Table 3.3, in the first group, we trained YOLOv3, YOLOv5s, and the proposed architecture using the BDD data set. Then, we tested the individual model’s performance using the Stanford Cars and Boxy data sets, independently. In terms of superiority, the proposed architecture has achieved the best results in most cases. As an example, consider Stanford Cars to be the training set with 1861 images (2nd group in Table 3.3). Thereafter, we have tested these three trained models using the BDD and Boxy data sets, separately. In terms of the evaluation metrics considered, the proposed attention-based model surpasses YOLOv3 and YOLOv5s. Similarly, in the first group with the BDD training data set, the proposed model achieved a precision of 96.1%, recall of 82.3%, mAP@0.5 of 90.7%, and mAP@[0.5, 0.95] of 59.6% for Stanford Cars, which are higher than the other two methods. On the other hand, for the same trained model with Boxy testing data set, the proposed model achieved a precision of 97%, recall of 96.3%, mAP@0.5 of 98.4%, and mAP@[0.5, 0.95] of 70.7%. A similar observation has been noted for the other experiments as well, as shown in Table 3.3

Performance for normal scenes without vehicle smoke: We have performed a separate testing experiment to prove that our proposed method is not biased toward vehicle smoke. For this, we have taken the Stanford vehicle data set. Total of 7019 out of 7600 images, we have not detected any object through the built DL model.

Run time analysis: We have performed our experiment in the mentioned GPU platform. The detection inference time t_d is approximately 0.013 Sec for a particular image frame (during the testing) which is quite faster compared to the manual checking. In video surveillance, we can support more than 75 fps (frames per second)($\approx 1/t_d$) processing with the current experimental setup. Hence the proposed DL model can be fused with the speed camera for smoke surveillance.

Failure cases: We report a couple of failure cases: 1. Identification of the smoky vehicle will be tricky due to the occlusion of objects. 2. We are facing issues while taking care of the shadow of the vehicles sometimes, which leads to the wrong classification.

3.4 Summary and Scope

In this chapter, We have shown a vehicle smoke synthesis technique first, followed by a novel attention-based detection technique. The performance of the proposed smoke detection method was compared with the state-of-the-art detection techniques, which proved the superiority of our proposed method. Through extensive ablation studies, we established that the proposed attention module had enhanced the overall backbone of the network. The deployment of the proposed lambda implemented attention based model also established its efficiency in the real-world test cases. Our proposed technique can be used very effectively in IRTS to accurately identify vehicle smoke. At the same time, different efficient approaches can be deployed so that the pollution identification process will have better accuracy and confidence in its decision. Automatic decision taking against such pollution rule breaking cases is another crucial aspect in IRTS. For this purpose the system needs to identify the license plate number along with any surveillance task. In the next chapter, we are going to present a novel approach for identifying license plate from on-road vehicles.

4

Automatic License Plate Recognition

4.1 An overview of Automatic License Plate Recognition

In Traffic Surveillance System (TSS), there are several challenges to identify the registration number from the license plate of vehicles. For the surveillance purpose, Automatic License Plate Recognition (ALPR) is a mandatory step to execute irrespective of the main objectives like pollutant vehicle identification, accident detection or spotting other unsocial activities performed by any vehicle. ALPR uses character identification from the plates to obtain vehicle registration number received through traffic installed camera or gantry camera. In recent times, several companies and government agencies are fascinated to improve their systems of traffic surveillance which justifies the need of developing a precise and efficient ALPR technique in uncontrolled conditions.

One of the major steps in ALPR is to segment the characters from the number plate by removing the background. Real time foreground and background separation from a video or image is an essential task in several computer vision applications. However, the complexity of surrounding environments make the task more complicated. ALPR performance also critically depends on the quality of the image which deteriorates based on ambient conditions like fog, mist, rain, snow etc. Further, other challenges such as partial occlusion, shadow etc. affect the overall performance of ALPR significantly. In literature, the recognition part mostly rely on classical image processing approaches to segment the characters of a number plate. Recently deep learning based techniques have become very popular in this regard.

In this chapter, we have showcased a Generative Adversarial Network (GAN) based approach to extract the binarized characters from the original number plate for simplifying the recognition activity. The main objective of the study is to achieve higher recognition accuracy in several adverse conditions through this GAN based segmentation. Different conditions have been considered during the training. To generate good quality segmented images from the real traffic scenario, we have provided the dual GAN approach. We have focused on novel generator

CHAPTER 4. AUTOMATIC LICENSE PLATE RECOGNITION

and discriminator architectures to come up with a model which significantly outperform the classical image processing based segmentation algorithms for ALPR. In our Bi-directional ConvLSTM and MultiScale (which has been termed as BM) based generator, the network mainly concentrates on the segmentation part with skip connections and the multi scale view of the given input. On the other hand Dual Discriminators (DD) have been incorporated to deal with the local and global information together. Together we name the network as BMDDNet. We have experimented our method in three real life license plate data sets, viz. Media Lab¹, Open ALPR Benchmark [8] and UFPR² [9]. At the same time we have experimented the same methods on three document binarization data sets, viz. DIBCO 17³, and DIBCO 18⁴ and Palm Leaf⁵ [123] for demonstrating its efficiency over similar image processing applications. Noisy environment, varying background and lighting condition are major challenges in the ALPR task. We have tried to overcome these issues through our robust data set with varying environment while deploying the same in the model building process. Moreover, We have focused two different perspectives in our approach to analyze the plate from global perspective to find out the generic pattern and analyze local information as well for finding the stroke details for better performance. Also, generating binarized image in the intermediate step has provided better OCR.

To the best of knowledge, this is the first approach in ALPR that attempts to generate binarized image from original number plate using GAN architecture. The major contribution of this work is to design the GAN-based technique which can deal with the number plate binarization even in the presence of different image degradation. Proposed GAN network has provided better binarized foreground text compared to other state-of-the-art popular GAN networks. To achieve this task, we have developed our own ground truth segmentation data set for LP as the same is not available in open-source platforms. The dual discriminator concept is introduced here to deal with the local and global information of an image simultaneously. We have also shown the results with single discriminator which is not performing well compared to our network. The proposed generator minimizes structural similarity-based loss along with cross-entropy loss designed precisely for binarization task. Our method performs satisfactorily even for document binarization problem which indirectly shows the generality of the proposed framework. The document binarization related experiments and results have been placed in the appendix section separately.

¹<http://www.medialab.ntua.gr/research/LPRdatabase.html>

²<https://web.inf.ufpr.br/vri/databases/ufpr-alpr/>

³<https://vc.ee.duth.gr/dibco2017/>

⁴<https://vc.ee.duth.gr/h-dibco2018/>

⁵http://amadi.univ-lr.fr/ICFHR2016_Contest/index.php/download-123

4.2. PROPOSED METHODOLOGY AND ARCHITECTURE

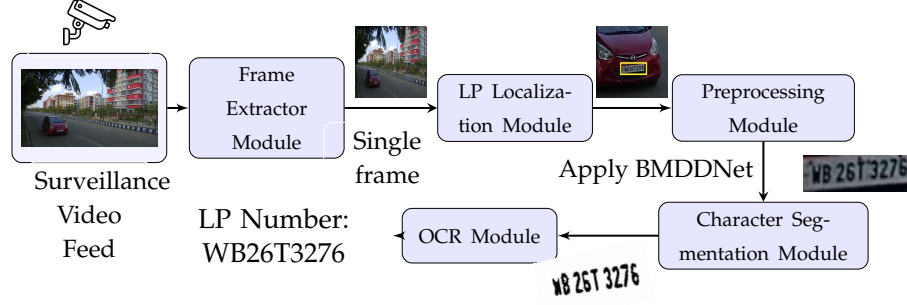


Figure 4.1: The overall flow of license plate number recognition used in our experiment.

4.2 Proposed Methodology and Architecture

The end-to-end ALPR module, as shown in Figure 4.1, takes a surveillance footage of the vehicle in question and outputs the license plate number of that vehicle. The first step of the task is to extract individual frame from the video through the Frame Extractor Module. This frame is passed to the LP localization Module which uses the yoloV3 algorithm to identify the location of the number plate in the image. Subsequently, this region of interest is cropped and de-noising is performed on the same. The Pre-processing Module aims to perform some enhancement on the cropped license plate image, before passing the result as input for the proposed BMDDNet in the Character Segmentation Module. The BMDDNet is used to binarize and segment the number plate and the binarized image is passed through the Google OCR module to ultimately obtain the license plate number. The proposed BMDDNet is a GAN based model having several sub-parts with the specific goal to achieve an overall high performance in the segmentation task in order to obtain an overall better recognition accuracy.

In our proposed BMDDNet, we have used three networks – a generator (G) and two discriminators, one Patch Discriminator (D_p) and another Image Discriminator (D_I). Further details of the generator and discriminator networks are provided in Subsections 4.2.1 and 4.2.2. Figure 4.2 has outlined the overall block diagram of the proposed architecture. Here the generator will produce binarized number plates which will be so close to the ground truth binarized plates to outwit both these discriminators at the end of the training phase. Based on the loss values of the generator and discriminators, the weights of the overall network will be fine-tuned. It is important to note that D_p takes image patches obtained from the Patch Breaker (PB) module as input while D_I takes the whole image as input as shown in the Figure 4.2.

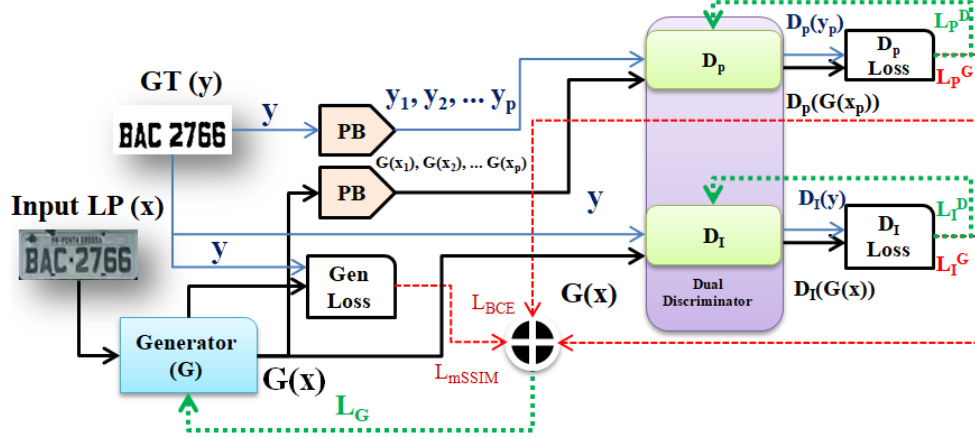


Figure 4.2: Overall architecture of proposed BMDDNet with the flow of different training losses

4.2.1 Generator Network

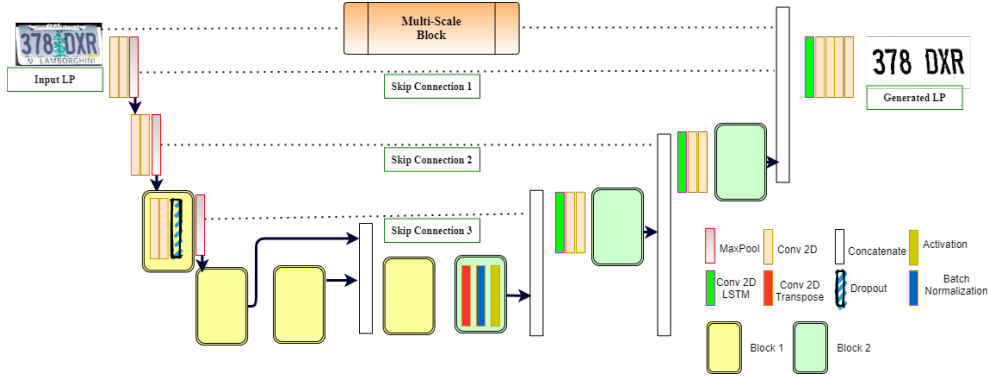


Figure 4.3: Generator with Multi-Scale Block and several Skip Connections. The Multi-Scale Block has been further demonstrated in Figure 4.4.

In the proposed model, the generator accepts the images of license plate of vehicles as input and generates corresponding binarized images on which Optical Character Recognition (OCR) have been performed to identify the registration number. For the generator network, we have proposed a densely connected Bi-Directional ConvLSTM with Multi-Scale structure (BM). This densely connected part has few skip connections which enhanced the overall network by providing an alternative path for the gradient flow. These skipped paths are often beneficial for the model convergence. At the same, time Multi-Scale architecture will help the model to analyze a given input in different scales. In different resolutions,

4.2. PROPOSED METHODOLOGY AND ARCHITECTURE

we can focus on distinct features which will ultimately provide better results in constrained surroundings. Figure 4.3 has shown the overall diagram of the generator.

Bi-Directional ConvLSTM with Skip Connections: The main motive of the generator is to perform segmentation task over the given input for classifying the foreground and background properly. As shown in Figure 4.3, few skip connections (shown by dotted line) have been placed in the network. Channel-wise concatenation (shown by white box) has been performed in every skipped path with a convolutional blocks in the last layer. Features learnt in every block are rolled forward, whereby a diverse set of features can be learnt by a block based on collective knowledge acquired by all the previous blocks. This helps to overcome the problem of a network learning redundant features. Binarization of text images [124] needs error free localization which is very tough in case of down sampled images. This skip connection helps to acquire different resolution features from different convolution layers. At the same time, Bi-Directional ConvLSTM provides a non-linearity while combining the skipped features which eventually contributes more precise segmented output. Furthermore, we accelerate the convergence speed of the network by employing Batch Normalization (shown by blue box) after the up-convolution filters. It has been discovered in some applications of CNN that computing features at multiple scales leads to enhanced performance in dense pixel prediction problems [125]. This yields a sophisticated method to fuse local and global features for predictions.

Multi-Scale Architecture: Multi-scale network has two basic module blocks, one is coarse block and another is fine scale network as shown in the Figure 4.4. The input has been provided in both the module blocks separately. In the mid-way, the output of the coarse network has been concatenated to one of the convoluted layers of the fine scale network and passed forward. In the coarse network, the size of the convolution filters are set to (11×11) , (9×9) and (7×7) respectively. On the contrary, size of the filters in the fine scale network is lower compared to the coarse one. The filter sizes have been kept as (7×7) , (5×5) and (3×3) respectively in the fine scale network. In between, in both the blocks, max pooling (2×2) and up-sampling have been applied to incorporate invariance keeping the resolution of the features same with the input. The coarse feed has been concatenated in the 2nd convolution layer of the fine scale.

4.2.2 Discriminator Network

Traditionally, GANs consist of a single discriminator which is used to distinguish fake images (generated by generator) from real images (original ground truth in the output domain). The generator targets to generate ground truth-like images in the output domain to fool the discriminator. In the BMDDNet, we have proposed

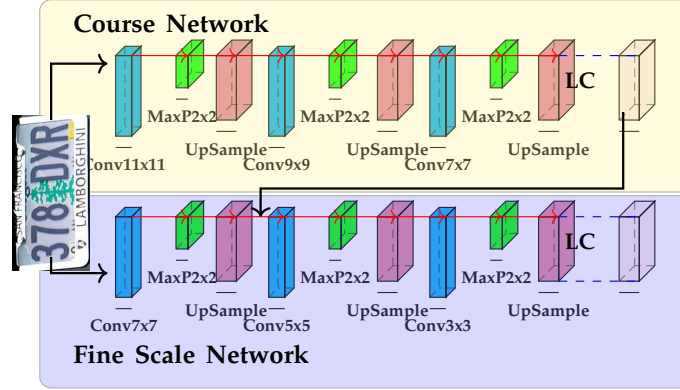


Figure 4.4: Multi-Scale Architecture: The yellow block and the blue block contain the coarse network and the fine network respectively. LC denotes Linear Combination.

to use two discriminators instead of one – a Patch Discriminator and an Image Discriminator. The main intuition behind this is to analyze the images at a local as well as global level for a reliable segmentation. In both the discriminators, a sequence of convolution with normalization layers have been incorporated. Usually, the more the number of convolutions applied, the receptive field increases. This is because each convolution operation is to sum up the previous information, therefore the current layer holds more global information than the previous one. Thus, we have kept the minimum depth of the proposed discriminators to be 15. Below are the descriptions for the patch and image discriminators used in our experiment.

Patch Discriminator: A Patch model tends to analyze the features extracted from the image patches which ensures faithful generation of structural information at local level. Patch discriminator (D_p) is used in this work as we can focus on smaller regions of the output separately to identify these low level intricacies rather than making predictions on the entire output. The number plate images possess a lot of local characteristics such as the text strokes of the characters, key points in the license number sequence and patterns for different countries or states in a country, which can be exploited to reduce distortions in the generator output images. Moreover, no pooling layer has been used as it is often difficult to obtain such precise localized details when the images are scaled down [125]. The architecture diagram of patch discriminator network is provided below in Figure 4.5. It consists of five 4×4 convolutional layers with a stride of 2 in each direction with padding, each layer followed by a Batch Normalization layer and Leaky ReLU as activation. Finally, a 4×4 convolution is applied followed by sigmoid to obtain the output patch predictions. For given generator G , the loss \mathcal{L}_p^D of

4.2. PROPOSED METHODOLOGY AND ARCHITECTURE

the patch discriminator D_p is defined by the below equation. Based on this loss discriminator D_p will be trained.

$$\begin{aligned}\mathcal{L}_p^D = & \mathbb{E}_{\mathbf{y}_p \sim p_{y_p}(\mathbf{y}_p)} [\log(D_p(\mathbf{y}_p))] \\ & + \mathbb{E}_{\mathbf{x}_p \sim p_{x_p}(\mathbf{x}_p)} [\log(1 - D_p(G(\mathbf{x}_p)))]\end{aligned}\quad (4.1)$$

where $p_{y_p}(\mathbf{y}_p)$ is the distribution of the ground truth patches and $p_{x_p}(\mathbf{x}_p)$ is the distribution of the input patches. $D_p(\mathbf{y}_p)$ and $D_p(G(\mathbf{x}_p))$ are the patch discriminator's outputs over the Ground truth \mathbf{y}_p and generated patch $G(\mathbf{x}_p)$. Instead of the overall image to classify as fake or real, we have utilized individual patch information for the decision.

Image Discriminator: The Image Discriminator (D_I), as shown in Figure 4.6, is used to make decisions based on a lot of global information present in the overall image. This Discriminator uses a convoluted network with a fully connected component at the end to capture more detailed global information from the image such as the overall shape of the license number pattern and the individual digits in it, background texture, etc. Thus, more number of layers are required to compute features at a higher level. Moreover, It has been found that performing features at different scales improves performance. Pooling is therefore a crucial part of this network. The Image Discriminator comprises five 4×4 convolutional layers with a stride of 2 in each direction with padding, followed by Batch Normalization and Leaky ReLU as activation function. Pooling (max) is performed after the third and fifth layers since computing features at multiple scales is observed to be effective in capturing global information. The fully connected network is used to store further higher level details, before finally adding an output sigmoid layer for prediction. For given generator G , the loss \mathcal{L}_I^D of the patch discriminator D_I is defined as follows.

$$\begin{aligned}\mathcal{L}_I^D = & \mathbb{E}_{\mathbf{y} \sim p_y(\mathbf{y})} [\log(D_I(\mathbf{y}))] \\ & + \mathbb{E}_{\mathbf{x} \sim p_x(\mathbf{x})} [\log(1 - D_I(G(\mathbf{x})))]\end{aligned}\quad (4.2)$$

where $p_y(\mathbf{y})$ is the distribution of the ground truth images and $p_x(\mathbf{x})$ is the distribution of the input images. $D_I(\mathbf{y})$ and $D_I(G(\mathbf{x}))$ are the image discriminator's output over the Ground truth \mathbf{y} and generated image $G(\mathbf{x})$. Both the discriminators use a BCE loss, since the discriminators output a simple binary prediction of real/fake.

4.2.3 Defined Loss Function in GAN network

In the experiment, we have used two pixel-wise losses in the generator side along with the adversarial loss. Binary Cross Entropy (BCE) and Mean Structural Similarity Index Measure (mSSIM) losses have been incorporated in the generator to

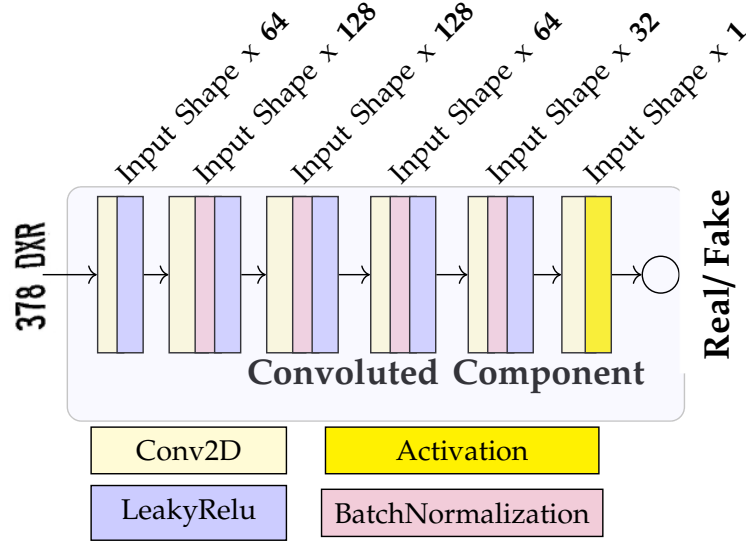


Figure 4.5: Patch Discriminator Network.

preserve the structural information of the segmented characters. As we have to separate the background pixels from the foreground, therefore BCE loss (\mathcal{L}_{BCE}) is appropriate in this context. During the computation of this loss, we have performed the pixel-wise loss computation as follows.

$$\mathcal{L}_{BCE} = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N \{y_{(i,j)} \log(G(x)_{(i,j)}) + (1 - y_{(i,j)}) \log(1 - G(x)_{(i,j)})\} \quad (4.3)$$

where the generated image has dimension $M \times N$, and at $(i, j)^{th}$ pixel location, the generated image has intensity $G(x)_{(i,j)}$. The actual label at the same pixel location in the ground truth segmented image is $y_{(i,j)}$. On the other hand, for improving the structural details of the generated binarized image, mSSIM-based loss would be very useful. This loss is denoted by \mathcal{L}_{mSSIM} which will be minimized to maximize the patch-wise SSIM (using Equation 4.4) between the reconstructed output and the corresponding ground truth binarized image using the following expression:

$$\mathcal{L}_{mSSIM} = 1 - \mathbb{E} \left[\frac{1}{N_p} \sum_{i=1}^{N_p} SSIM(|G(x_i)|, |y_i|) \right] \quad (4.4)$$

4.2. PROPOSED METHODOLOGY AND ARCHITECTURE

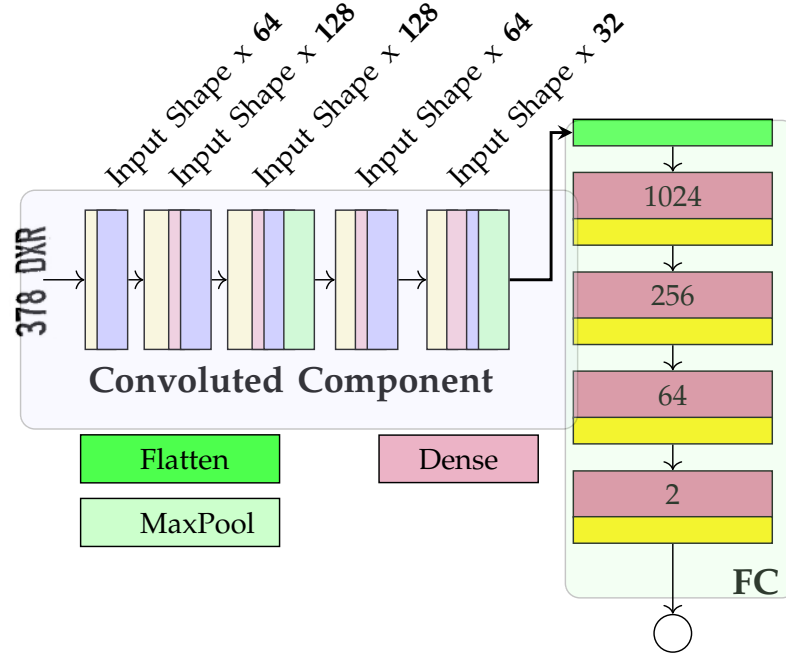


Figure 4.6: Image Discriminator Network. FC denotes fully connected network. Other color representations are same as Figure 4.5

where N_p denotes the number of patches in the image. Since the SSIM is used to compare the perceptual similarity between two images, we use ground truth binarized image and reconstructed binarized image of the license plate in the aforementioned formulation to ensure faithful generation.

The overall Generator loss is denoted by \mathcal{L}_G which is calculated as

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{mSSIM} + \lambda_3 \mathcal{L}_I^G + \lambda_4 \mathcal{L}_p^G \quad (4.5)$$

$$\begin{aligned} \text{where, } \mathcal{L}_I^G &= \mathbb{E}_{\mathbf{x} \sim p_x(\mathbf{x})} [\log(D_I(G(\mathbf{x})))] \\ \mathcal{L}_p^G &= \mathbb{E}_{\mathbf{x}_p \sim p_{x_p}(\mathbf{x}_p)} [\log(D_p(G(\mathbf{x}_p)))] \end{aligned}$$

In this experiment, λ_1 and λ_2 have been set to 100, λ_3 and λ_4 have been set to 0.5 and 1 respectively. The four λ values are chosen experimentally. However, the BCE and mSSIM losses are directly effecting the generation process that is why we have always set the higher values of λ_1 and λ_2 compared to λ_3 and λ_4 , i.e. λ_1, λ_2

>> λ_3, λ_4

4.3 Experimental Setup and Result

We have executed our experiment in Google Colab⁶ provided Tensor Processing Unit (TPU) which has Linux (Ubuntu-18.04) as baseline operating system with 12.7 GB RAM. In this section, we have briefed the ground truth generation and data augmentation techniques followed by the detailed results. A short description of the data sets with the metrics used in the experiments have been discussed in the Sections A and B of the appendix materials.

4.3.1 Ground truth generation and data augmentation

We must mention here that in case of Open ALPR benchmark and UFPR data sets the unavailability of binarized ground truths were a major problem for training and performance evaluation. For overcoming this difficulty, we have manually marked the ground truth characters for these two data sets for enhancing the algorithm performance. To prepare the data set, one annotator has marked the segments whereas other two annotators have verified the marked ground truth sequentially. Total 1250 number of images were processed for ground truth, with 645 from Open ALPR and 605 from UFPR data set. To the best of our knowledge, there is no open-source segmented data set is available to assist license plate recognition task.

Moreover, we have put intentional distortions like blurring, random cropping, color jitters etc., while augmenting the training set. This would help us to deal with several challenging scenarios where the binarization process is very tough to execute.

4.3.2 License Plate experiment

Open ALPR Benchmark and UFPR data sets have been used for the training purpose in the experiment. These two sets have been enriched with degraded images through augmentation like blurring, random cropping, color jitters etc. The kernel size has been kept to 15×15 in blur operation. For both the data sets, the actual number plate areas are extracted initially using the given coordinates. Afterwards, using the proposed BMDDNet network we have trained the system using the input images along with the augmentation. The method has been iterated for 100 iterations during the training. The input size of the network is kept to (128×128) and (256×128) for these data sets respectively. In both the training cases, the average loss has been gradually decreased with epochs. In the testing

⁶<https://colab.research.google.com/>

4.3. EXPERIMENTAL SETUP AND RESULT

Table 4.1: Qualitative results on the Metrics over Open ALPR data set. PSNR is measured in dB; while SSIM, Precision, F-Score and pF-Score are unit-less and normalized to (0-1). OCR accuracy has been presented in %. Larger value of all the indexes are indicating better algorithm.

Type	Model Used	PSNR	SSIM	Precision	F-Score	pF-Score	OCR Accuracy
Image Processing	Otsu[126]	06.77	0.44	0.75	0.83	0.540	11.68
	Bradley-Roth[127]	08.33	0.60	0.86	0.89	0.663	27.25
	Nick[128]	08.34	0.59	0.89	0.90	0.632	24.64
	Niblack[129]	06.77	0.52	0.75	0.84	0.622	22.57
	Feng[130]	07.97	0.61	0.81	0.88	0.671	29.35
	Sauvola[68]	08.34	0.59	0.91	0.90	0.563	20.10
	Wolf[131]	08.71	0.63	0.87	0.90	0.654	28.88
Deep Learning	Pix2Pix[132]	11.68	0.75	0.94	0.95	0.823	54.54
	Selectional auto-encoder[133]	11.58	0.74	0.96	0.95	0.819	47.69
	DD-GAN[134]	11.24	0.72	0.97	0.95	0.750	43.36
	BCDUnet[120]	13.63	0.830	0.96	0.97	0.901	66.78
	BMSDNet (Single D)	09.56	0.65	0.93	0.92	0.632	38.64
	BMDDNet	13.98	0.843	0.97	0.97	0.907	69.64

Table 4.2: Qualitative results on the Metrics over UFPR data set. PSNR is measured in dB; while SSIM, Precision, F-Score and pF-Score are unit-less and normalized to (0-1). OCR accuracy has been presented in %. Larger value of all the indexes are indicating better algorithm.

Type	Model Used	PSNR	SSIM	Precision	F-Score	pF-Score	OCR Accuracy
Image processing	Otsu[126]	6.229	0.480	0.715	0.762	0.385	13.98
	Bradley-Roth[127]	7.361	0.553	0.900	0.877	0.480	17.86
	Nick[128]	7.347	0.575	0.934	0.885	0.403	15.80
	Niblack[129]	5.012	0.289	0.662	0.757	0.453	6.82
	Feng[130]	7.275	0.523	0.839	0.859	0.542	23.46
	Sauvola[68]	7.272	0.601	0.957	0.888	0.278	10.46
	Wolf[131]	7.812	0.592	0.916	0.886	0.491	27.01
Deep Learning	Pix2Pix[132]	9.950	0.645	0.913	0.918	0.680	13.25
	Selectional Auto-encoder[133]	9.710	0.658	0.976	0.928	0.626	12.98
	DD-GAN[134]	9.592	0.706	0.981	0.930	0.663	13.06
	BCDUnet[120]	12.890	0.797	0.976	0.964	0.884	60.68
	BMSDNet (Single D)	5.203	0.408	0.677	0.705	0.325	11.20
	BMDDNet	15.203	0.849	0.980	0.976	0.924	71.86

phase, separate image set has been deployed which are not considered during the training. Over the model's binarized output, we have applied post processing for clearing the noises distributed over the regions. Connected component has been applied for removing the unwanted foreground. At the same time row and column wise vertical and horizontal projection count has been considered



Figure 4.7: Sample images and ground truth produced by the proposed algorithm with different environment condition: 1) blur. 2) Tricky background patch and distortion. 3) Water bubble effect. 4) Extra lighting condition.

to remove the outlier part from the foreground text section. In case of outlier in binarized number plate, if we take the row-wise count, the foreground pixel counts ideally should be less than a certain percentage of the column length and vice versa. Applying this logic, we have reduced some unwanted pixel noises in the boundary region. Table 4.1 has shown several metrics (described with mathematical formulae in appendix section) for seven image processing based binarized methods and six deep learning based approaches over Open ALPR data set. In all the cases, BMDDNet has provided better index values. PSNR (13.98%), SSIM (0.84%), Precision (0.97%), F-Score (0.97%), pFScore (0.91%) have supported the efficiency of the network. Even the OCR accuracy 69.64% has reached to the highest value compared to the other methods while the base line raw image and the corresponding GT OCR accuracies are 28.11% and 80.07% respectively. This OCR accuracy has been improved significantly with the help of our segmentation approach even in the presence of adverse imaging conditions. At the same time, Figure 4.8a has visually established how the proposed network is able to produce almost accurate binarized image which is quite similar to the original GT image. The input image that has been shown in Figure 4.8a contains significant amount of blur. As shown in Figure 4.8a(d2), the registration number (**WB03TN7601**) has been generated perfectly while other methods failed to segment the characters correctly.

Similar type of results have been observed on the UFPR data set too in Table 4.2. PSNR value reaches to 15.20 which is the highest measure compare to the other methods. Other index values also prove the supremacy over the other techniques. The average OCR accuracy in the testing data set hits to 71.86% which is very close to the baseline accuracy 73.09% with the GT images while the raw image OCR accuracy is 20.57%. This indirectly reinforces the needs of the binarization in

4.3. EXPERIMENTAL SETUP AND RESULT

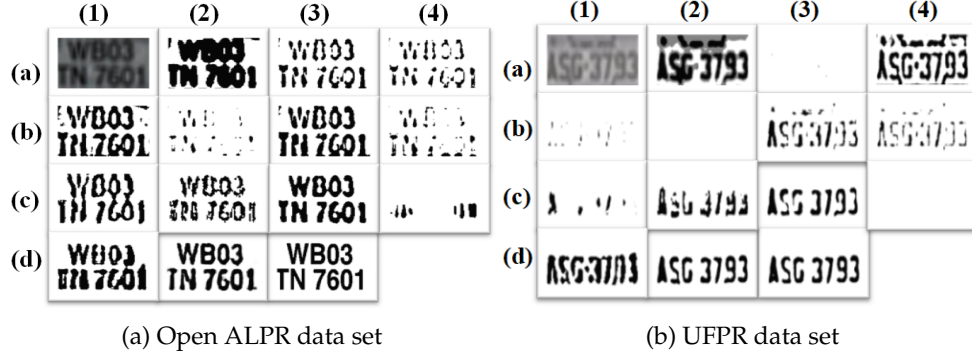


Figure 4.8: Sample Visual Comparison on two data sets: (a1) original, (a2) Otsu, (a3) Nick, (a4) Niblack, (b1) Bradley-Roth, (b2) Sauvola, (b3) Wolf, (b4) Feng, (c1) Selectional auto-encoder, (c2) DDGAN, (c3) BCDUNet, (c4) BMSDNet (Single D), (d1) Pix2Pix, (d2) **BMDDNet**, (d3) Ground Truth.

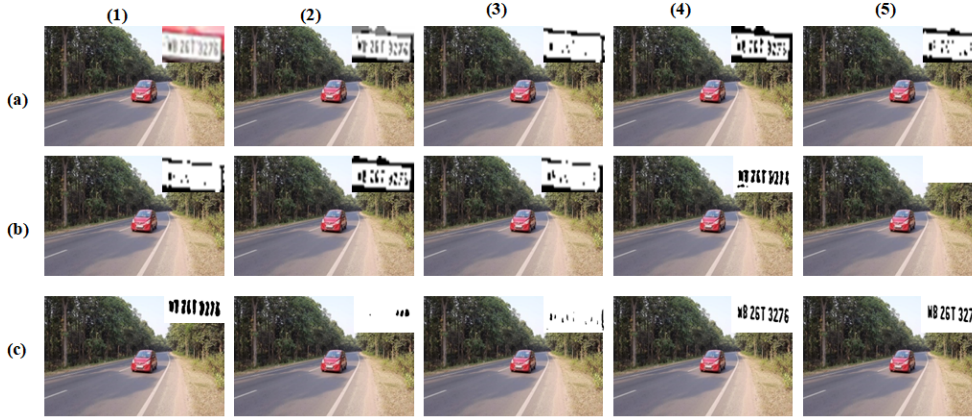


Figure 4.9: Visual Comparison in bright light condition (Original Number- **WB26T 3276**): (a1) Cropped Plate, (a2) Otsu, (a3) Nick, (a4) Niblack, (a5) Bradley-Roth, (b1) Sauvola, (b2) Feng, (b3) Wolf, (b4) Pix2Pix, (b5) BCDUNet, (c1) BMSDNet (Single D), (c2) DDGAN, (c3) Selectional Auto-Encoder, (c4) **BMDDNet** c5. Ground Truth.

ALPR. Figure 4.8b(d2) has shown the binarized output generated using proposed method for another blurred number plate (**ASG3793**). The segmented output is close to GT in terms of characters' thickness, stroke etc. while the outputs generated by other methods exhibit flaws.

4.3.3 Cross Dataset Evaluation

The Media lab ALPR data repository has been used for zero shot learning purpose. The samples used in this data set for the evaluation steps are captured under

CHAPTER 4. AUTOMATIC LICENSE PLATE RECOGNITION

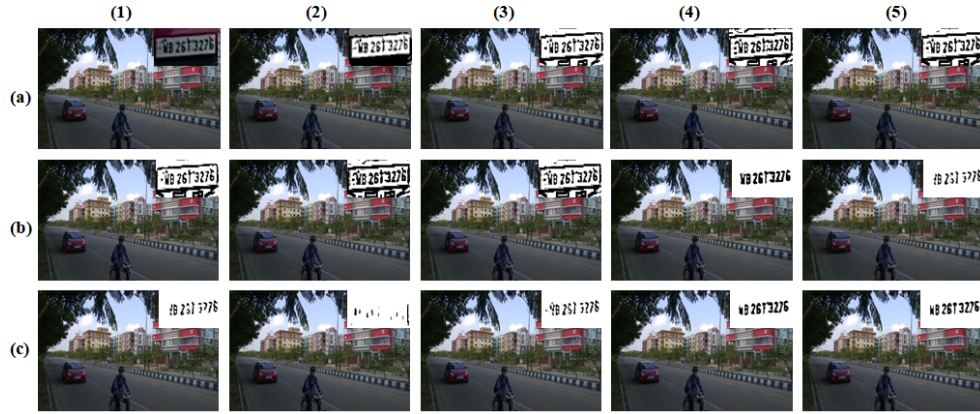


Figure 4.10: Visual Comparison in low light condition (Original Number- **WB26T 3276**): (a1) Cropped Plate, (a2) Otsu, (a3) Nick, (a4) Niblack, (a5) Bradley-Roth, (b1) Sauvola, (b2) Feng, (b3) Wolf, (b4) Pix2Pix, (b5) BCDUNet, (c1) BMSDNet (Single D), (c2) DDGAN, (c3) Selectional Auto-Encoder, (c4) **BMDDNet**, (c5) Ground Truth.



Figure 4.11: Sample Visuals on Media Lab data set: a) Night flash. b) Tricky shadows. c) Difficult night samples.

4.4. SUMMARY AND FUTURE PROSPECTS

different constrained conditions like night flash, tricky shadow, brighter lighting condition etc. Figure 4.11 has displayed few sample output of such constrained conditions. The segmentation output for each of these input images are shown in the offsets of the respective images. This evaluation has proven the model's efficiency in unknown environment even with degraded image quality.

Figure 4.9 has shown one real life image test result of a running vehicle over the road in bright day condition. Similarly, Figure 4.10 has shown one low light condition where the image has been captured for the same vehicle. BMDDNet has segmented the license plate accurately, compared to the GT image. Our proposed network also outperformed other related algorithms in both environment conditions evidently.

Inference time of the proposed model: Our inference framework is mainly comprising of two parts; generator and the OCR module. The inference time of the binarized approach is approximately 1ms while the google OCR is taking 2ms. So altogether it is taking 3ms. Therefore in any surveillance system we can process approximately 160 fps frame processing which can be easily incorporated to yellow-vulture camera.

4.4 Summary and Future Prospects

The importance of a robust segmentation algorithm in case of ALPR is often underestimated, and thus it is comparably a less explored area of research. With growing complications of YOLO-based frameworks for recognition in the wild, especially in case of ALPR tasks, the necessity of the character-wise segmentation is becoming more prominent. As this could be an extremely effective tool to improve performance of conventional ALPR algorithms, we explore this problem with a novel GAN-based architecture. In this chapter, we have developed a dual discriminators based GAN network to assist ALPR. We presented a state-of-the-art GAN-based methodology for image-to-image translation to generate binarized LP images using real LP images for end-to-end ALPR technique. With extensive manual effort to create a substantial number of ground truth data and exhaustive data augmentation, we train our model for ALPR task. We have given attention to shape our proposed algorithm for different imaging conditions with drastically divergent variations. The superiority of this method over related recognition strategies along with document binarization have been elaborately demonstrated. The results in terms of several domain specific numerical metrics and visual observations have been showcased to prove the method's efficiency over other approaches present in literature. To the best of our knowledge, no GAN-based technique has been used for license plate segmentation task yet. We observed that the multi-scale dual discriminator architecture preserves the structural information in the final

CHAPTER 4. AUTOMATIC LICENSE PLATE RECOGNITION

binarized image successfully.

Therefore, safety concern in the transport is very high priority in recent days. Through technology researchers are trying to minimize the risks in IRTS. In the next chapter we will discuss more safety concerns that is associated in Intelligent Car.

5

Safety in Intelligent Road Transport System during Pandemic

5.1 Pandemic Era Road Transport Safety: An Overview

The current pandemic situation will impact every industry in the world. The automotive industry will not be an exception. All the domains in this industry like car design, manufacturing, marketing strategy, maintenance, ride safety, etc. will need to address the upcoming restrictions. The vehicle monitoring system will have a domain shift too from its existing version. Generally, Government and the Traffic Management Authority install surveillance systems to increase on-road safety and security. Those systems inspect pollution, speed of a vehicle, anomalies in driving behavior and other traffic rule violations. ‘Yellow-Vulture’ cameras have become very popular in UK’s road¹ which captures several offenses such as smoking, drinking, etc. inside the car. But in the near future, some additional constraints related to COVID-19, like social-distancing, mask-wearing, etc., need to be compulsorily considered. Therefore, it is essential to address issues related to contagious diseases in IRTS as well.

In the last few years, several research works have been reported to address issues related to the surveillance system in the vehicular field. [135] showed a smart camera-based system to detect anomalies in IRTS. In the paper, [136] have presented one intelligent traffic management system. [137] proposed an efficient vehicle detection technique for traffic surveillance data in real-time. In the paper, [138], recognized vehicles on the basis of generation, make and model which will be used for surveillance. In [139], the authors proposed a unique network model for multi-UAV surveillance. Drone-based traffic management is also becoming popular in the last few years [140, 141]. But in most cases, video processing is commonly used in traffic management systems [142, 143]. [144] have developed a video surveillance system for road safety with pre-event detection

¹<https://www.express.co.uk/life-style/cars/1070006/speed-camera-mobile-phone-driving-UK-fine>

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

feature. Different vehicle detection and counting techniques have been applied by the researchers for the last decades [145, 146]. Detection of several driving anomalies has been deployed for traffic management in many articles [147, 148]. Seat-belt violation has been captured efficiently by [149]. In recent pandemic state, researchers also have put their effort to identify and enforce social distancing norms [150, 151, 152]. In [153], authors have shown different possibilities of implementing social distancing in public transport. The spread of COVID through transport is becoming a huge challenge to the world now. IRTS should impose such relevant norms to break the chain of COVID-transmission.

5.2 Passenger Surveillance during COVID period

In this paper, we have mainly focused to prevent the spread of such contagious diseases through enforcing rules in IRTS. The key intention here is to identify the passengers' mask-wearing status inside a vehicle. We have deployed a Deep Learning (DL) based approach to deal with the mask status of any passenger seating inside the vehicle. At the same time, the passenger count is another important criterion to maintain during the pandemic time; this indirectly takes care of the social distancing aspect. Though face detection is a well-known problem in computer vision, detection of tiny degraded faces inside a vehicle for passenger counting is not a widely explored problem. A near real-time integrated process flow for the passenger counting and mask detection from image frames has never been explored to the best of our knowledge. In this work, we not only design a robust framework for counting the number of passengers inside a vehicle but also designed a face mask detection algorithm using transfer learning with very high accuracy.

A robust data set with huge variations and real-life use cases have been prepared. Special attention has been given to the side-face, hand-covered and low-resolution images which will be essential for the traffic surveillance. The data set is publicly available at <https://github.com/srimantacse/MaskSurveillance>. An efficient deep learning model to identify face mask over the human face with the adoption of very tiny face detection has been implemented. We have shown that the proposed model outperforms the other state-of-the-art techniques with the extensive comparison as well as statistical hypothesis testing. As per the authors' knowledge, this is the first work to find out the mask-wearing status in inside-vehicle surveillance. This sort of traffic monitoring will be very essential in the current COVID-19 related pandemic context.

5.3. MASK DATA SET PREPARATION

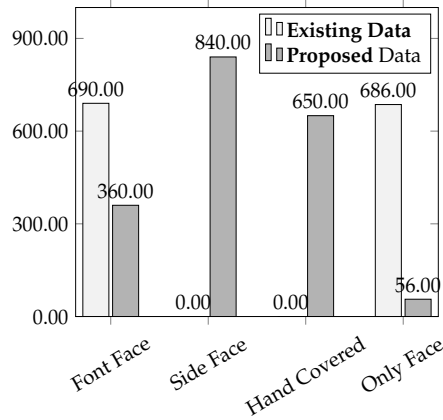


Figure 5.1: Data distribution comparison between data set of Prajna [87] and our enhanced data set

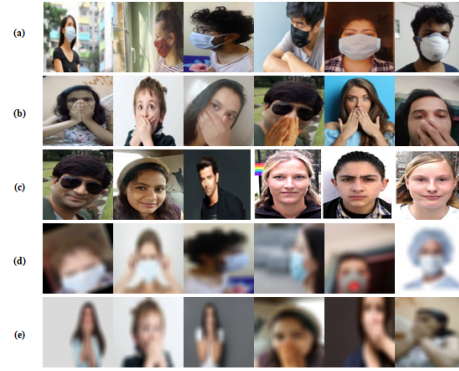


Figure 5.2: Sample images of our data set: a) Masked. b) Hand on top of Mouth. c) Non Masked. d) Blurred Masked. e) Blurred Hand on top of Mouth

5.3 Mask Data set Preparation

In post-COVID-19 road transportation, the mask should be a mandatory item that has to be used all the time while going outside mainly. But some people are deliberately guarding their faces by hand that should be properly identified and penalized. We have focused mainly on two aspects while preparing the data set. Firstly, we planned to prepare a substantial data set for the generic mask detection in the human face. Secondly, we have focused on the on-road inside car image set which will be very much important for the surveillance purpose.

Several perspectives are required for accurate mask detection. The result of the human face and mask identification strongly depends on the image resolution and dots per inch (dpi). These two qualities will be affected if the distance from the camera increases. Even the position and size of the face will be another vital point in this context. So, while designing the data set, we have intentionally introduced the blurring effect for increasing data variation. The side face images with masked and non-masked variations have been added as well to predict them correctly irrespective of face direction. Different angles of the face need to be considered here. We have contemplated all these in the data set for mask detection.

To construct a substantial database, we took an available data set of [87] as a baseline which has 690 masked and 686 non-masked faces. In the mentioned data set, all the face images were taken from the front side and the mask has been superimposed on the non-masked images to prepare the same. The data set has

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

been enriched by adding another 1906 images with the existing one. The major contribution is that different side face images have been added and used during the training. At the same time, some images of the hand located on top of the mouth have been additionally included which is obviously a restricted use-case in the post-COVID-19 situation. A total of 650 such images are there in the data set. Again there is another important aspect, in the testing phase, if the image quality of the individual face is not good then it may hamper the testing accuracy. The chance of getting noisy images is higher in case of gantry camera input while capturing the vehicle in motion. Keeping this limitation in mind, we have introduced blurred images in the data set. Total 400 masked and 130 non-masked images are converted to blurred with the kernel size (20×20). Table 5.1 has shown the category-wise distribution for the blurred image set. Considering the feasible angle of head position in the real scenario, we have augmented those images with two rotation angles 10° and 20° to increase the variation. The total number of images in the data set is now 3282. Figure 5.1 has shown the comparative statistics of images in the overall database. Figure 5.2 shows some sample images of the enriched data set in different conditions.

Table 5.1: Statistical comparison between data set of Prajna [87] and our enhanced data set for the blurred image set

	Blurred Masked (400)		Blurred Non-Masked (130)		
Data set	Front-Face	Side-Face	Hand-Covered	Only Face	Total
Prajna data set [87]	0	0	0	0	0
Proposed data set	100	300	130	0	530

5.4 Experimental Setup

We have deployed our framework in google colab² system where the training of the proposed DNN network and testing of several use cases have been performed separately. We have tuned different DL parameters of the proposed network and used them after proper optimization. The google colab environment which has been chosen for our experiment has a baseline operating system (OS) Linux-4.19.104 (Ubuntu-18.04) with 12.7 GB RAM. We have deployed our experiment using tensorflow and keras environment along with the required APIs.

5.4.1 Model of Surveillance Camera Setup

In our experiment, we have mostly considered low-height surveillance cameras like gantry-setup. The tentative height of this camera should be around 6.5ft

²<https://colab.research.google.com/>

5.4. EXPERIMENTAL SETUP

above ground level. It will not be elevated so high which may prevent getting the proper view of the inside vehicle. As shown in Figure 5.3, the region of interest should be focused on the side windows and the windshield of a vehicle from a certain height. We have collected such images that will actually replicate the real gantry camera scenario for the test simulation.

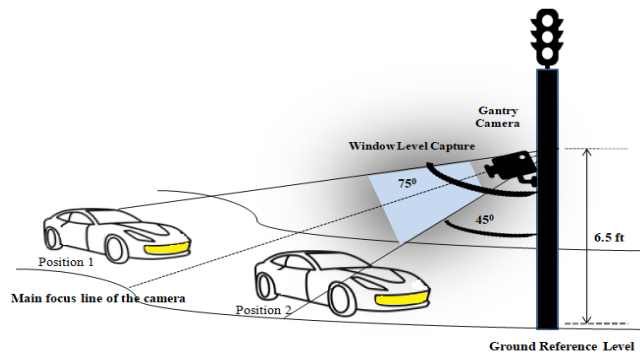


Figure 5.3: Surveillance camera (Gantry type) position and focus to the model car's window and windshield. Sample angle range has been showcased for windshield.

Figure 5.4 has shown such images which have been used in our experiment. This kind of mask surveillance will be the part of traffic 'Yellow-Vulture'³ shortly. The height of the surveillance camera has to be set experimentally based on the dimension of several vehicles. Mostly for the Light Motor Vehicle (LMV) the optimum height range is 4ft to 7ft. In our experiment, we have mostly con-

sidered LMV and the height from the ground level has been considered to be 6.5ft. For the windshield images (a maximum distance of 25ft to a minimum distance of 6.5ft from the camera), the angle range is +75° to +45°. For the side window images, a short distance range (a maximum distance of 10ft to a minimum distance of 4.5ft from the camera) will be considered where the angle range lies approximately in between +55° to +35°. Figure 5.3 has shown this sample calculation for the front window screen only. On the other hand, for SUVs or big buses, this height should be higher. To examine all the use cases we need to consider a multi-camera setup installed in different heights. Inside vehicle camera is another solution for this surveillance. However, this will increase the cost and create an extra burden for individual installation and necessary changes.

5.4.2 Deep Model training Setup

We have utilized the pre-trained Inception-V3 weights at the beginning and appended fully-connected network components with average pooling followed by flatten. Subsequently, 3 sets of consecutive dense and different activation layers have been used with an intermediate dropout. The pool size in the average pooling step was kept as (5 × 5). The three activation layers used in the network are

³<https://www.pendlelease.co.uk/news/new-uk-speed-cameras>

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC



Figure 5.4: On-road images with proper view of the passengers sitting inside vehicles

relu, sigmoid and softmax consecutively. The initial dropout percentage has been kept as 50%. Subsequently, it is reduced to 40%. Figure 5.5 has shown the overall training architecture of the Deep Neural Network (DNN) used in the experiment. 70% of the images from the newly built data set have been used for training and the remaining images have been kept for testing. Again, 10% of the training data has been used as the validation set during the training. *Adam* optimizer has been used during the training with a learning rate η . We have tested the range of η from 0.000001 to 0.01 at 9 discrete levels [0.000001, 0.000005, 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01] to select the optimal value. The experimentally-found optimum value of η where the proposed model has achieved the best result is 0.0001. The learning rate η has been uniformly distributed in the decay parameter of the optimizer over the epochs which has been set to 30. To note, for all the other algorithms too, we have run for the same number of iterations and compared the results. ‘binary-cross-entropy’ has been utilized for this binary classification task as the loss function. The two dropout layers have been used for better generalization in the network which will reduce the chance of over-fitting.

5.5 Proposed Methodology

In the proposed approach, we have used the enhanced data set as mentioned in Section 5.3 to train the deep model. The tiny face detector model [154] has been adopted to find out the face images from the given input. In our experiment, the TL concept has been deployed to build the deep network for mask detection. The baseline Inception-V3 model has been built on imagenet data set. This data set has few variations of mask (normal-mask, gas-mask, ski-mask and oxygen-mask) images as well. Therefore, our mask identifying target problem domain, D_t , is a subset of the Inception-v3’s problem domain, D_s ($D_t \subset D_s$). Subsequently, the learning task T_t is also a subset of the corresponding source domain’s learning task T_s ($T_t \subset T_s$). From a thousand class problem, we have boiled down to two class problem using this TL framework. After removing the last output layer of the base

5.5. PROPOSED METHODOLOGY

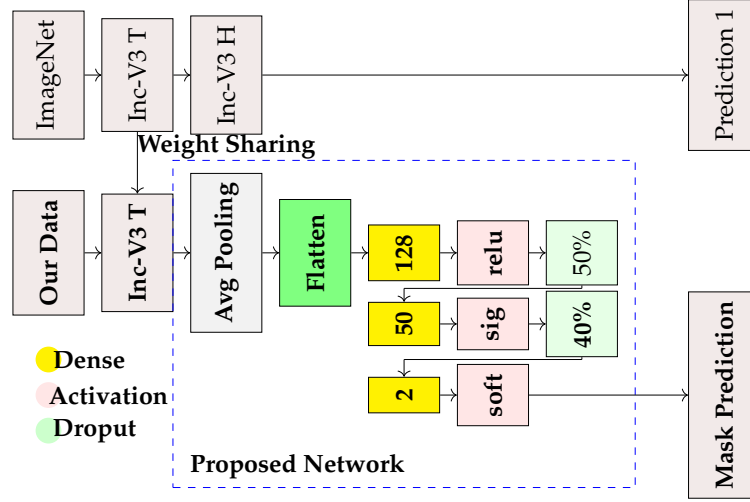


Figure 5.5: Proposed Transfer Learning (TL) based DNN, used in our mask prediction task.

model, we have introduced unique two-class layers (masked and non-masked) at the end in *Mask Prediction* node as shown in Figure 5.5. Algorithm 1 has shown the detailed structure of the process. Two input parameters, the vehicle image I with size $M \times N \times 3$ and the threshold parameter δ for the passenger limit will be taken by the algorithm for providing the proper on-road surveillance. The δ parameter will be set based upon the vehicle type and capacity. The social-distancing norm will be strongly dependent on this parameter. For example, for four-wheeler LMV, the value of δ should be set to less than or equal to (\leq) 3 including the driver. Let's assume, there are N_{Faces} faces present in the given input $I_{M \times N \times 3}$. The function \mathcal{R} has performed the resizing operation over every face image to make them a fixed size of $(m \times m \times 3)$. Here, in the experiment, we have kept $m=n=224$. For the i^{th} face, the corresponding feature set $feature_i(f_1^i, f_2^i, \dots, f_k^i)$ with cardinality k has been computed using \mathcal{E} . The generated Inception-V3 based TL model (\mathcal{G}) has considered this feature set to provide the confidence level of prediction which has been noted as $Conf_i^{Test}$. Subsequently, the mask-wearing status of i^{th} face has been processed from $Conf_i^{Test}$ using the function $Stat$. Based on the status and face count, the rule violation decision will be taken. During the training time, following the similar process, $Conf_i^{Train}$ has been generated for the i^{th} face. The $Target_i$ is the corresponding original class label. For computing the loss (binary-cross-entropy) during the training phase, we have used the below Equation 1.3. N_{total}^{Train} denotes the total number of images used in the training phase of the experiment.

In the testing phase, we have used the tiny face detector model to extract the individual face images with the proper bounding box. As per the flow shown

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

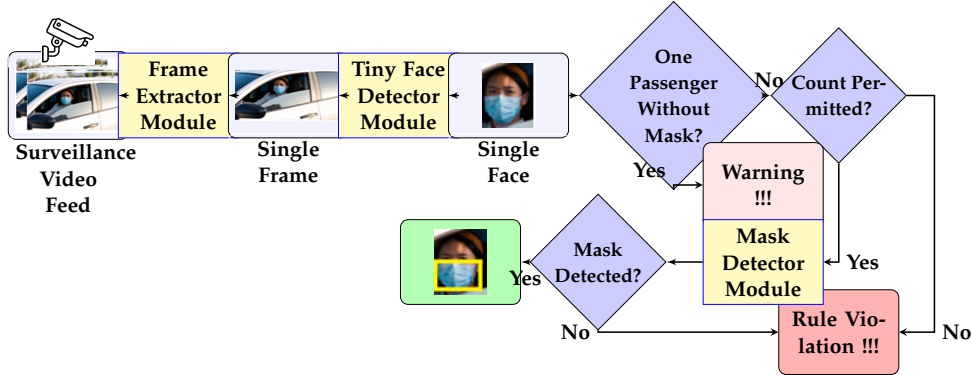


Figure 5.6: Testing flow of the proposed framework for passenger surveillance to find mask-wearing status inside vehicle.

in Figure 5.6, the frame extractor module will extract images from the video feed taken from the surveillance camera. Based on the face count, we can make a decision about the rule violation. The threshold value would be taken as prior information to the system. For a special case, only the driver without a mask inside a car would have been warned for ignoring the safety measure. At the final step of the testing phase, using the proposed mask detection model we can detect the mask from the face images.

The learning has been demonstrated through the visualization of the saliency map for individual training images. Figure 5.7 has demonstrated six sample saliency images with the respective original images. These sample images indicate that the model has learned the overall facial part in case of normal images and the face part without the covered zone in case of masked images. As shown in Figure 5.8 the training accuracy has been gradually increased and the loss has been decreased with the epoch while training. Moreover, for statistically demonstrating the efficiency of the generated model we have run the repeated K-Fold validation process separately. The repetition parameters we have set to 5 and the K has been set to 5 as well in K-Fold. For displaying the outcome, we have drawn the boxplot of the corresponding validation accuracy for different folds in different

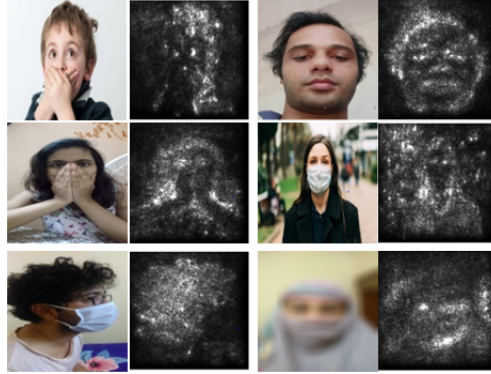


Figure 5.7: Sample saliency-map images generated for our proposed model

run.

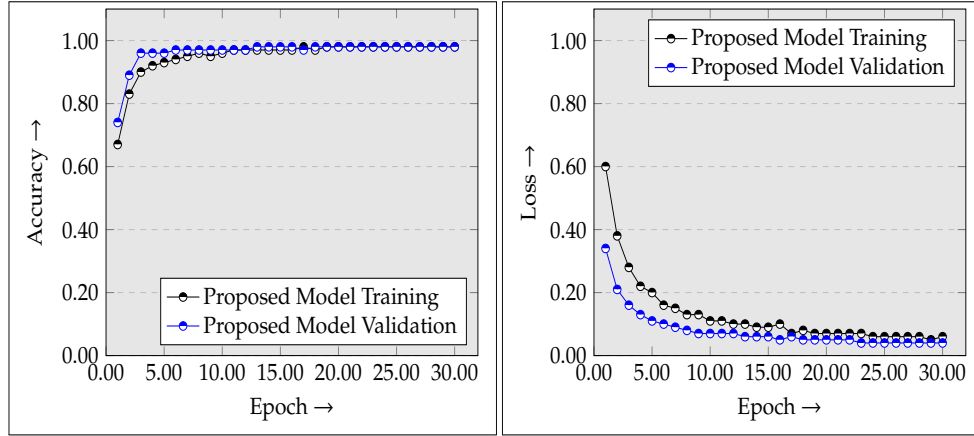


Figure 5.8: Accuracy and loss curve during the training and validation (10% of the training data) process

Table 5.2: Wilcoxon hypothesis comparison for the different models on the K-Fold validation with mean and standard deviation of the achieved by the different algorithms

Method	Wilcoxon p-Value (vs Proposed)	Mean	Standard deviation
AlexNet	0.000012	0.907	0.018
VGGFace	0.000012	0.827	0.019
FaceNet	0.000012	0.874	0.038
SSDMNV2	0.000012	0.934	0.011
Res-SVM	0.016599	0.964	0.017
Proposed	NA	0.974	0.006

the Table 5.2. Here we can see, there is not such similarities in between the proposed models and other models used for the comparison in the experiment. As per the hypothesis, the obtained p-values, less than 0.05, are supposed to be statistically significant. The mean and standard deviation of the different models have been noted in the same Table 5.2. This clearly shows the primacy of the proposed model.

Figure 5.9 has shown that our model exhibits the minimal span of the box in the boxplot while the other models have shown varying accuracy in different runs. To hypothetically demonstrate the supremacy, we have computed the pairwise (State-of-the-art technique vs proposed model) p-value of the Wilcoxon rank [155] which has been shown in

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

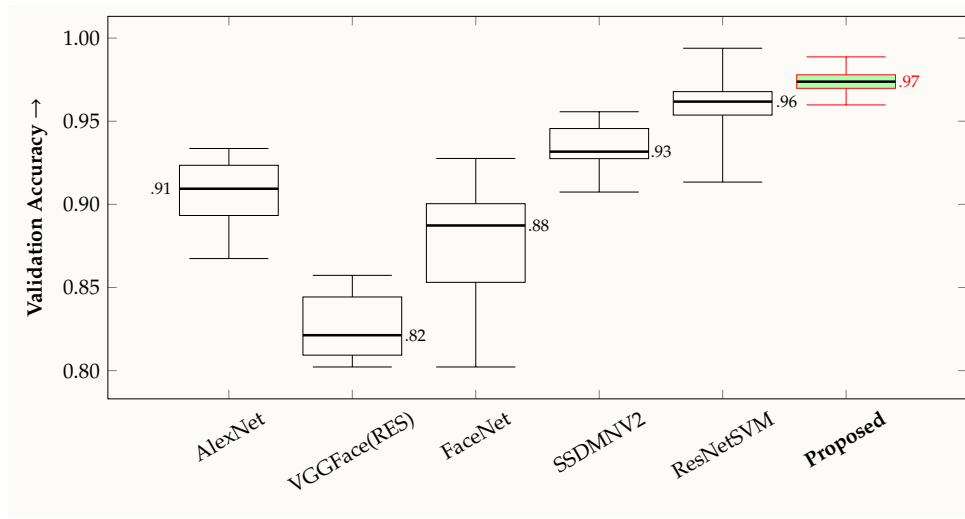


Figure 5.9: Statistical comparison through Box-plot of *Repeated K-Fold's* ($K=5$, *Repeat=5*) validation accuracy achieved by different models during training. Our proposed method has been outline in red color with green highlight. The median accuracy line with corresponding value has been presented with every plot.

5.6 Results and Evaluations

To demonstrate the superiority of the proposed approach we have tested on multiple variant images as well as on-road images of the vehicle. The primary objective of the method is to count the number of passengers inside the car, and at the same time, we need to check whether they wear a mask or not during the journey. The testing car database contains images from different angles and heights taken by the surveillance camera, night mode pictures, mirror-reflected face images, glass-covered faces etc. These images are mainly used for testing purposes. For the evaluation process, we have utilized several numerical indices viz. Accuracy, Precision, Recall, F1-Score, Specificity and Jaccard Score (JS) using Equations 1.5, 1.7, 1.7, 1.6, 1.9 and 1.10 respectively along with the visual comparisons.

Table 5.3 has shown the different indexes used for the evaluation. We have computed those indexes for all the mentioned algorithms. From Table 5.3, it is clear that in terms of accuracy over the testing data set, our method has outperformed other techniques. While the proposed network has achieved 98.5% accuracy, the best accuracy among other techniques is 95.8% obtained by ResNet-SVM. A similar observation is applicable for the other indexes as well. The proposed model distinctly outperforms the existing algorithms in various aspects.

For comparing the confidence level of the prediction, we have chosen the best two models i.e. SSDMNv2 and ResNet-SVM as per Table 5.3. Here we have used

5.6. RESULTS AND EVALUATIONS

Table 5.3: Numerical Index (Average Weighted) comparison with the state-of-the-art techniques and other variations over 1066 [Masked (634) & Non-Masked(432)] testing images

Model Used	Accuracy	Precision	Recall	F1Score	Specificity	JS
SSDMNV2	0.935	0.910	0.854	0.891	0.940	0.880
ResNet-SVM	0.958	0.964	0.964	0.965	0.951	0.939
AlexNet	0.905	0.913	0.904	0.893	0.924	0.805
VGGFace(ResNet)	0.905	0.904	0.906	0.897	0.941	0.765
FaceNet	0.825	0.857	0.827	0.817	0.845	0.784
Proposed Net	0.985	0.990	0.983	0.990	0.986	0.970

several unusual and critical environment conditions like low-quality images, side-face images and hand-covered images for the comparison. Table 5.4 has clearly demonstrated the superiority of the proposed approach in all those conditions. 400 blurred, 50 side-face and 200 hand-covered images have been experimented. SSDMNV2 and ResNet-SVM have correctly identified 174 and 351 blurred images respectively, while our proposed model has predicted 389 successfully. Similar types of results have been observed while testing side-face and hand-covered category too as shown in Table 5.4. Figure 5.10 and 5.11 have shown the confidence level of prediction where the state-of-the-art techniques predicted correctly. For example, as shown in Figure 5.10, among SSDMNV2's correct prediction of 174 images, the confidence value of the prediction of our proposed model is higher for 71% of such images. For side-face and hand-covered images, our model has beaten SSDMNV2 by 100% and 97% margin in terms of confidence level. The same observation is applicable for ResNet-SVM model as well for the same scenarios that have been depicted in Figure 5.11. Figure 5.12 has shown few sample test images where the recommended approach correctly predicted but other existing models failed.

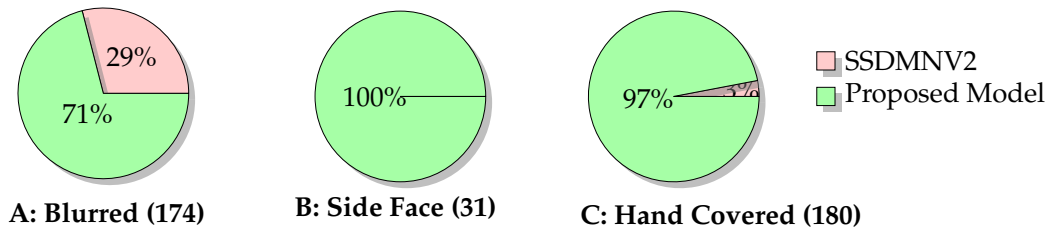


Figure 5.10: Confidence level comparison with SSDMNV2 where SSDMNV2 predicted correctly as mentioned in Table 5.4

Figure 5.13 has shown one sample testing flow with four passengers inside a car and all of them have used masks while traveling. The confidence level of

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

Table 5.4: Special test set comparison with state-of-the-art techniques over different unusual conditions of mask images

Image Type	Model Used	Correct Prediction
Blurred (400)	SSDMNV2	174 (43.50%)
	ResNet-SVM	351 (87.75%)
	Proposed Net	389 (97.25%)
Side Face (50)	SSDMNV2	31 (62.00%)
	ResNet-SVM	46 (92.00%)
	Proposed Net	49 (98.00%)
Hand-Covered (200)	SSDMNV2	180 (90.00%)
	ResNet-SVM	196 (98.00%)
	Proposed Net	198 (99.00%)

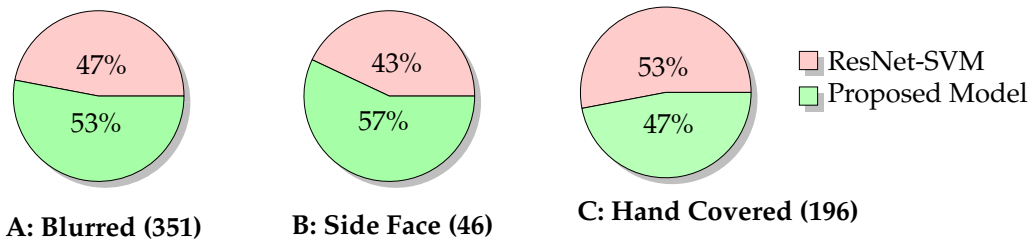


Figure 5.11: Confidence level comparison with ResNet-SVM where ResNet-SVM predicted correctly as mentioned in Table 5.4



Figure 5.12: Sample test images where the proposed model predicted correctly but the existing models failed



Figure 5.13: Testing flow: a) Input image. b) Face detected image. c) Mask detected images with confidence level

prediction has appeared with the bounding box. The back seat passengers' tiny faces also have been captured and predicted the mask-wearing status successfully. Figure 5.14 has shown the step-wise results for these best-performed algorithms in various road conditions with the confidence level of successful predictions. Those conditions have been mentioned explicitly in the first column of the same Figure 5.14.

Cross data set evaluation

In this cross data set evaluation process, we have used the data set that has been experimented by Nagrath et. al. in the paper [81]. We have tested on 10,212 images in this data set with 5520 non-masked and 4692 masked images respectively. Our model has successfully predicted 5511 non-masked and 4336 masked images. Collectively 9847 (96.4%) successful prediction is a remarkable figure without deploying training while SSDMNv2 has achieved 92.6% accuracy after the training phase. Figure 5.15 has shown some sample evaluation output over the mentioned data set.

Testing Time analysis

The face detection module has taken almost uniform time (around 3 sec) for any number of human faces. However, the mask detection process has been conducted sequentially for every face image. It has linearly grown with the number of faces present in a single image. The average time taken for this process is around 0.5 sec for one human face. The time graph has been shown in Figure 5.16. So, generally, the surveillance framework will take around 4 seconds to complete the process

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

Condition Type	Original Image	Detection of Passenger Faces	Detection of Mask by Proposed Model	Proposed Model	SSDMNV2	RESNET-SVM
1. Blurred Image With Mask				Pass 99.80%	Fail	Pass 99.20%
2. Side View Image With Mask				Pass 90.46%	Pass 85.92	Pass 96.44%
3. Night Mode Image With Mask				Pass 99.60%	Pass 74.45	Pass 83.79%
4. Very Bright Image With Mask				Pass 98.50% 99.30%	Fail - Fail	Pass 92.79% 79.86%
5. Mirrored Image With Mask				Pass 99.73%	Fail	Pass 99.08%
6. Reflected Tiny Image With Mask				Pass 99.5% 99.865%	Fail - Pass 99.55	Pass 85.17% 82.26%
7. Image With No Mask				Pass 99.65%	Pass 98.50%	Pass 99.10
8. Side Face With No Mask				Pass 98.90%	Pass 99.91%	Pass 97.90%
9. Sunlight reflection on windscreen				Pass 99.26%	Fail	Pass 78.02%
10. Night mode with Poor illumination				Pass 98.43%	Fail	Fail
11. Passengers at the Back seat				Pass 99.18%	Pass 90.10%	Pass 83.64%
12. Very Bright light over window				Pass 99.66%	Pass 81.88%	Fail

Figure 5.14: Results of two best-performed state-of-the-art techniques and proposed model over unusual and challenging on-road conditions with inside-vehicle passengers for mask surveillance along with confidence levels of individual prediction

5.7. DISCUSSION AND FUTURE PROSPECTS



Figure 5.15: Sample cross data set test outcomes: a) Masked Faces. b) Non-Masked faces.

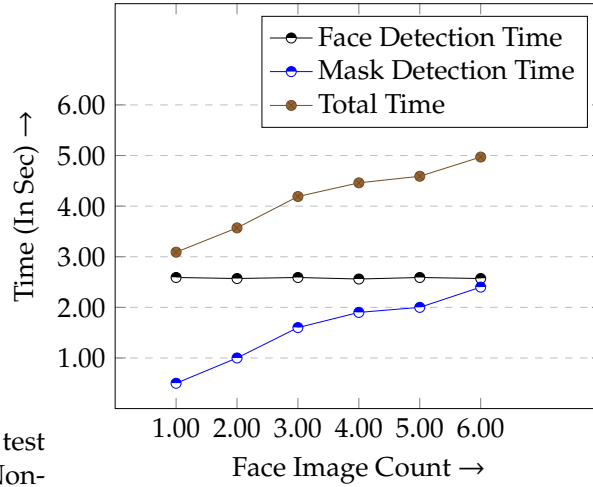


Figure 5.16: The time graph of the proposed model with varying number of faces

for a standard vehicle with 3 passengers (maintaining social-distancing norm). The timing measurement has been performed on the same google colab setup as mentioned in the subsection 5.4. With Graphics Processing Unit (GPU) or Tensor Processing Unit (TPU) system, the timing graph can be improved further.

5.7 Discussion and Future Prospects

The advantages of Transfer Learning have been exploited to develop new traffic surveillance which is extremely useful during COVID-19 like pandemic situation. Our proposed network has outperformed all the other state-of-the-art techniques in terms of several metrics. The zero-shot evaluation has reached 96.4% accuracy as well which supports the potency of the technique. This framework will also be applicable for similar contagious diseases where the mask will be compulsory to use. Considering the post-COVID-19 urgency, different traffic rules need to be imposed in the transport system. The surveillance strategy should also be modified accordingly. We have demonstrated an efficient way to identify the on-road rule-breakers. The developed deep model has achieved high accuracy in different conditions. As a scope of future research, one can work to enhance the training data set from a different perspective related to mask-wearing. Designing proper network to deal with the hazy and rainy images is another future scope. Researchers can apply better deep learning based tracking mechanism for the surveillance. Hyper-parameter optimization is another area to work on. For

CHAPTER 5. SAFETY IN INTELLIGENT ROAD TRANSPORT SYSTEM DURING PANDEMIC

handling the night-mode on-road images more accurately, we can generate a large number of such samples using Generative Adversarial Networks. For managing varying heights of several types of vehicles, one can use a multi-camera setup. Also, the multi-view of a specific vehicle will help to provide higher confidence in the outcome. Real field testing with this type of multi-camera setup will be another challenging task. Currently, the authors are working in these directions.

Conclusions and Future Scope of Research

IRTS represent a groundbreaking advancement in the realm of transportation, seamlessly merging cutting-edge technology with traditional infrastructure to enhance efficiency, safety, and sustainability. By integrating real-time data analytics, communication technologies, and automation, this system has the potential to revolutionize the way we perceive and interact with transportation networks. The implementation of sensors, cameras, and communication devices allows for the continuous monitoring of transport systems in various surveillance aspects. This smart technologies not only improves the overall travel experience for commuters but also helps Government agencies to deploy and maintain laws for safety measurement. IRTS are not only about technology but also about sustainability. By highlighting different health issues and other safety things, this system aligns with the global push towards environmentally conscious transportation solutions. In essence, the smart transport system holds the promise of assisting a new era of intelligent, safe, and sustainable transportation. As we continue to embrace technological innovations, this system serves as a beacon guiding us towards a future where roads are not merely paths for vehicles, but dynamic ecosystems that adapt and evolve for the benefit of all road users and the planet.

However, different challenges such as restricting pollution, automatic number plate recognition, addressing several safety aspects while in the road, traffic rule enforcement and taking actions, etc. demand an extensive infrastructure upgrades. Dealing with these issues will be crucial for ensuring the a safe and technologically advanced road transport. Interoperability and Standardization of rules from different angle, Infrastructure Cost and Retrofitting, Public Acceptance and Education, Limited Funding and Budget Constraints, different challenges we need to address to provide a seamless transport system. Addressing these challenges will be crucial for the successful and widespread adoption of IRTS, ensuring that the potential benefits in terms of safety, efficiency, and sustainability are realized.

We seek to respond to the following research questions in this thesis:

1. How to identify polluting vehicle in the busy road?

CHAPTER 6. CONCLUSIONS AND FUTURE SCOPE OF RESEARCH

2. How to recognize License plate number so that surveillance can be automatically served?
3. How to provide passenger safety during COVID like pandemic time?

In order to address the research questions, it is important to adopt recent computer vision based technologies and cutting edge AI algorithms. We have tried our best to address those mentioned research questions in different chapters.

In Chapter 2, We have developed an efficient feedback based vehicle pollution tracking strategy by exploiting the advantages of several Deep Learning models. The performance of our method has been compared with other state-of-the-art techniques over the various real life images which demonstrates the superiority of the same. The accuracy in the zero-shot evaluation has supported the robustness of the process as well. Our proposed approach is generic in nature and can be applied to any similar deep learning based model. This framework combines the decision of several models with higher confidence level. It can be used effectively in the IRTS to identify the on-road pollutant vehicles.

In Chapter 3, We have shown a vehicle smoke synthesis technique first, followed by a novel attention-based detection technique. The performance of the proposed smoke detection method was compared with the state-of-the-art detection techniques, which proved the superiority of our proposed method. Through extensive ablation studies, we established that the proposed attention module had enhanced the overall backbone of the network. The deployment of the proposed lambda implemented attention based model also established its efficiency in the real-world test cases. Our proposed technique can be used very effectively in IRTS to accurately identify vehicle smoke. At the same time, different efficient approaches can be deployed so that the pollution identification process will have better accuracy and confidence in its decision.

In Chapter 4, the importance of a robust segmentation algorithm in case of ALPR, which is often under-estimated, has been demonstrated. With growing complications of YOLO-based frameworks for recognition in the wild, especially in case of ALPR tasks, the necessity of the character-wise segmentation is becoming more prominent. As this segmentation could be an extremely effective tool to improve performance of conventional ALPR algorithms, we have explored this problem with a novel GAN-based architecture. In this chapter, we have developed a dual discriminator based GAN network to assist ALPR. We presented a state-of-the-art GAN-based methodology for image-to-image translation to generate binarized LP images using real LP images for end-to-end ALPR technique. With extensive manual effort to create a substantial number of ground truth data and exhaustive data augmentation, we train our model for ALPR task. We have given attention to shape our proposed algorithm for different imaging conditions with drastically divergent variations. The superiority of this method over related

recognition strategies along with document binarization have been elaborately demonstrated. The results in terms of several domain specific numerical metrics and visual observations have been showcased to prove the method's efficiency over other approaches present in literature. To the best of our knowledge, no GAN-based technique has been used for license plate segmentation task yet. We observed that the multi-scale dual discriminator architecture preserves the structural information in the final binarized image successfully.

In Chapter 5, The advantages of Transfer Learning have been exploited to develop new traffic surveillance which is extremely useful during COVID-19 like pandemic situation. Our proposed network has outperformed all the other state-of-the-art techniques in terms of several metrics. The cross data set evaluation has reached 96.4% accuracy as well which supports the potency of the technique. This framework will also be applicable for similar contagious diseases where the mask will be compulsory to use. Considering the post-COVID-19 urgency, different traffic rules need to be imposed in the transport system. The surveillance strategy should also be modified accordingly. We have demonstrated an efficient way to identify the on-road rule-breakers. The developed deep model has achieved high accuracy in different conditions. As a scope of future research, one can work to enhance the training data set from a different perspective related to mask-wearing. Designing proper network to deal with the hazy and rainy images is another future scope. Researchers can apply better deep learning based tracking mechanism for the surveillance. Hyper-parameter optimization is another area to work on. For handling the night-mode on-road images more accurately, we can generate a large number of such samples using Generative Adversarial Networks. For managing varying heights of several types of vehicles, one can use a multi-camera setup. Also, the multi-view of a specific vehicle will help to provide higher confidence in the outcome. Real field testing with this type of multi-camera setup will be another challenging task. Currently, the authors are working in these directions.

However, in general, there are some common hurdles associated with every proposed ideas in computer vision based researches. It would be really difficult to handle the scenarios of invisible smoke and dense fog in case of pollution identification tasks. Handling night mode images, dark shade images, images with high brightness is a common pain point. In case of motion blur and occlusion of frames, the detection of any road transport entity will be tough; this can indirectly hamper the process outcome. Also, regarding the camera shooting parameters, camera angle, position, resolution, etc. need to be properly handled while collecting the image data set.

As a scope of future research of pollution identification task, one can work to enhance the database from different perspectives and can apply better Deep Learning based tracking mechanism for the surveillance. Color based pollution

CHAPTER 6. CONCLUSIONS AND FUTURE SCOPE OF RESEARCH

classification is another area where we can apply a shallow network to determine the possible pollution type. As a scope of future research, one can work to enhance the synthesis approach from several other perspectives by considering more realistic shapes or deformable object models. Smoke content estimation in a mixed smoke scenario is another area where the proposed system needs extension. Differentiating vapor from white pollutant smoke is a challenging task that needs to be addressed in future. Considering additional features like weather conditions, time of the day etc. while dataset preparation will result into more precise synthesis for scenarios like vapory smoke, especially in colder weather conditions. In addition, researchers can focus on reducing the size of the proposed network. Ensemble methods can also be utilized for better decision-making with higher confidence.

For the license plate identification task, network complexity of the dual GAN network can be reduced. At the same time instead of two separate sub tasks, segmentation and identification, we can combine the same into a single network as well. The future of Automatic License Plate Recognition holds significant potential, with advancements expected in accuracy, integration with other technologies, and diverse applications across various sectors. As the technology evolves, it will be crucial to address privacy concerns, establish clear regulations, and ensure responsible deployment to harness its full benefits.

In addressing pandemic time safety considerations, there is a concerted effort to augment the training dataset by focusing on mask-wearing from a distinct perspective. Researchers are actively exploring the implementation of advanced deep learning-based tracking mechanisms within surveillance systems. Additionally, there is a focus on hyper-parameter optimization to refine the efficiency of these systems. To enhance the accuracy of handling night-mode on-road images, the generation of an extensive dataset using Generative Adversarial Networks (GANs) is being pursued. Managing the diverse heights of various vehicle types is being approached through the utilization of a multi-camera setup. This setup not only accommodates varying vehicle heights but also enhances confidence in the outcomes through a multi-view perspective. Real-world field testing with this specialized multi-camera configuration poses a noteworthy challenge, and ongoing efforts are dedicated to exploring these avenues. The authors are actively engaged in advancing research in these directions. At the same time, the future of point-of-interest safety while driving is poised for innovation, leveraging emerging technologies to create a more responsive and secure driving experience. Integrating advanced safety measures around specific locations through a combination of data analytics, connectivity, and AI-driven solutions is likely to play a pivotal role in shaping the next generation of road safety initiatives.

Bibliography

- [1] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," *Machine learning*, pp. 3–23, 1983.
- [2] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [4] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.
- [6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [7] J. Deng, O. Russakovsky, J. Krause, M. S. Bernstein, A. Berg, and L. Fei-Fei, "Scalable multi-label annotation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 3099–3102.
- [8] M. Hill, . [Online]. Available: <https://github.com/openalpr/benchmarks>
- [9] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "A robust real-time automatic license plate recognition based on the yolo detector," in *2018 international joint conference on neural networks (ijcnn)*. IEEE, 2018, pp. 1–10.
- [10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [11] A. Khan, A. Sohail, U. Zahoor, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial intelligence review*, vol. 53, no. 8, pp. 5455–5516, 2020.
- [12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

BIBLIOGRAPHY

- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [16] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.
- [17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 16, 2009.
- [22] J. J. Bird, J. Kobylarz, D. R. Faria, A. Ekárt, and E. P. Ribeiro, "Cross-domain mlp and cnn transfer learning for biological signal processing: eeg and emg," *IEEE Access*, vol. 8, pp. 54 789–54 801, 2020.
- [23] W. Pan, E. Xiang, N. Liu, and Q. Yang, "Transfer learning in collaborative filtering for sparsity reduction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, 2010.

-
- [24] Z. Xu, Y. Cao, and Y. Kang, "Deep spatiotemporal residual early-late fusion network for city region vehicle emission pollution prediction," *Neurocomputing*, vol. 355, pp. 183–199, 2019.
- [25] Z. Xu, Y. Kang, Y. Cao, and L. Yue, "Residual autoencoder-lstm for city region vehicle emission pollution prediction," in *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. IEEE, 2018, pp. 811–816.
- [26] E. Suganya and S. Vijayashaarathi, "Smart vehicle monitoring system for air pollution detection using wsn," in *2016 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2016, pp. 0719–0722.
- [27] S. S. Chandrasekaran, S. Muthukumar, and S. Rajendran, "Automated control system for air pollution detection in vehicles," in *2013 4th International Conference on Intelligent Systems, Modelling and Simulation*. IEEE, 2013, pp. 49–51.
- [28] S. Pal, A. Ghosh, and V. Sethi, "Vehicle air pollution monitoring using iots," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 400–401.
- [29] A. Bhatnagar, V. Sharma, and G. Raj, "Iot based car pollution detection using aws," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2018, pp. 306–311.
- [30] G. C. Cardoso and L. K. Mestha, "Image-based determination of co and co2 concentrations in vehicle exhaust gas emissions," Oct. 7 2014, uS Patent 8,854,223.
- [31] J. Guo, X. Zhang, F. Gu, H. Zhang, and Y. Fan, "Does air pollution stimulate electric vehicle sales? empirical evidence from twenty major cities in china," *Journal of Cleaner Production*, vol. 249, p. 119372, 2020.
- [32] P. Pyykönen, P. Peussa, M. Kutila, and K.-W. Fong, "Multi-camera-based smoke detection and traffic pollution analysis system," in *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2016, pp. 233–238.
- [33] H. Tao and X. Lu, "Smoky vehicle detection based on multi-scale block tamura features," *Signal, Image and Video Processing*, vol. 12, no. 6, pp. 1061–1068, 2018.
- [34] Y. Cao and X. Lu, "Learning spatial-temporal representation for smoke vehicle detection," *Multimedia Tools and Applications*, vol. 78, no. 19, pp. 27 871–27 889, 2019.
-

BIBLIOGRAPHY

- [35] H. Tao and X. Lu, "Smoke vehicle detection based on robust codebook model and robust volume local binary count patterns," *Image and Vision Computing*, vol. 86, pp. 17–27, 2019.
- [36] F. Yuan, J. Shi, X. Xia, Y. Fang, Z. Fang, and T. Mei, "High-order local ternary patterns with locality preserving projection for smoke detection and image classification," *Information Sciences*, vol. 372, pp. 225–240, 2016.
- [37] A. U. Russo, K. Deb, S. C. Tista, and A. Islam, "Smoke detection method based on lbp and svm from surveillance camera," in *2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2)*. IEEE, 2018, pp. 1–4.
- [38] L. Yuan, S. Tong, and X. Lu, "Smoky vehicle detection based on improved vision transformer," in *The 5th International Conference on Computer Science and Application Engineering*, 2021, pp. 1–5.
- [39] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo, "Smokenet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention," *Remote Sensing*, vol. 11, no. 14, p. 1702, 2019.
- [40] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A deep normalization and convolutional neural network for image smoke detection," *Ieee Access*, vol. 5, pp. 18 429–18 438, 2017.
- [41] D. Sheng, J. Deng, and J. Xiang, "Automatic smoke detection based on slic-dbscan enhanced convolutional neural network," *IEEE Access*, vol. 9, pp. 63 933–63 942, 2021.
- [42] X. Sun, L. Sun, and Y. Huang, "Forest fire smoke recognition based on convolutional neural network," *Journal of Forestry Research*, vol. 32, no. 5, pp. 1921–1927, 2021.
- [43] J. Zhou, S. Qian, Z. Yan, J. Zhao, and H. Wen, "Esa-net: A network with efficient spatial attention for smoky vehicle detection," in *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2021, pp. 1–6.
- [44] X. Wang, Y. Kang, and Y. Cao, "Sdv-net: A two-stage convolutional neural network for smoky diesel vehicle detection," in *2019 Chinese Control Conference (CCC)*, 2019, pp. 8611–8616.
- [45] S. Kundu, U. B. Maulik, A. Bej, and U. Maulik, "Deep learning based pollution detection in intelligent transportation system," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 2020, pp. 292–297.

-
- [46] C. Wang, H. Wang, F. Yu, and W. Xia, "A high-precision fast smoky vehicle detection method based on improved yolov5 network," in *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, 2021, pp. 255–259.
- [47] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [49] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, and K. Michael, "ultralytics/yolov5: v6. 1-tensorrt, tensorflow edge tpu and openvino export and inference," *Zenodo*, Feb, vol. 22, 2022.
- [50] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
- [51] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2017.
- [52] L. Sassykova, Y. Aubakirov, S. Sendilvelan, Z. K. Tashmukhambetova, M. Faizullaeva, K. Bhaskar, A. Batyrbayeva, R. Ryskaliyeva, B. Tyussyupova, A. Zhakupova *et al.*, "The main components of vehicle exhaust gases and their effective catalytic neutralization," *Oriental Journal of Chemistry*, vol. 35, no. 1, p. 110, 2019.
- [53] İ. A. Reşitoğlu, K. Altinişik, and A. Keskin, "The pollutant emissions from diesel-engine vehicles and exhaust aftertreatment systems," *Clean Technologies and Environmental Policy*, vol. 17, no. 1, pp. 15–27, 2015.
- [54] A. Joshi, "Review of vehicle engine efficiency and emissions," *SAE International Journal of Advances and Current Practices in Mobility*, vol. 1, no. 2019-01-0314, pp. 734–761, 2019.
- [55] P. Mukhija, P. K. Dahiya, and P. Priyanka, "Challenges in automatic license plate recognition system: An indian scenario," in *2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT)*. IEEE, 2021, pp. 255–259.
- [56] J. Shashirangana, H. Padmasiri, D. Meedeniya, and C. Perera, "Automated license plate recognition: a survey on methods and techniques," *IEEE Access*, vol. 9, pp. 11 203–11 225, 2020.
-

BIBLIOGRAPHY

- [57] M. Rahmani, M. Sabaghian, S. M. Moghadami, M. M. Talaie, M. Naghibi, and M. A. Keyvanrad, "Ir-lpr: A large scale iranian license plate recognition dataset," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2022, pp. 053–058.
- [58] A. M. Al-Ghaili, S. Mashohor, A. R. Ramli, and A. Ismail, "Vertical-edge-based car-license-plate detection method," *IEEE transactions on vehicular technology*, vol. 62, no. 1, pp. 26–38, 2012.
- [59] J. Dun, S. Zhang, X. Ye, and Y. Zhang, "Chinese license plate localization in multi-lane with complex background based on concomitant colors," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 3, pp. 51–61, 2015.
- [60] R.-C. Chen *et al.*, "Automatic license plate recognition via sliding-window darknet-yolo deep learning," *Image and Vision Computing*, vol. 87, pp. 47–56, 2019.
- [61] B.-G. Han, J. T. Lee, K.-T. Lim, and D.-H. Choi, "License plate image generation using generative adversarial networks for end-to-end license plate character recognition from a small set of real images," *Applied Sciences*, vol. 10, no. 8, p. 2780, 2020.
- [62] V. Tadic, M. Popovic, and P. Odry, "Fuzzified gabor filter for license plate detection," *Engineering Applications of Artificial Intelligence*, vol. 48, pp. 40–58, 2016.
- [63] P. Kaur, Y. Kumar, S. Ahmed, A. Alhumam, R. Singla, and M. F. Ijaz, "Automatic license plate recognition system for vehicles using a cnn," *CMC-Computers, Materials & Continua*, vol. 71, no. 1, pp. 35–50, 2022.
- [64] O. Bulan, V. Kozitsky, P. Ramesh, and M. Shreve, "Segmentation-and annotation-free license plate recognition with deep localization and failure identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2351–2363, 2017.
- [65] Y. Wen, Y. Lu, J. Yan, Z. Zhou, K. M. von Deneen, and P. Shi, "An algorithm for license plate recognition applied to intelligent transportation system," *IEEE Transactions on intelligent transportation systems*, vol. 12, no. 3, pp. 830–845, 2011.
- [66] F. D. Kurpiel, R. Minetto, and B. T. Nassu, "Convolutional neural networks for license plate detection in images," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3395–3399.

-
- [67] C.-H. Lin and Y. Li, "A license plate recognition system for severe tilt angles using mask R-CNN," in *2019 International Conference on Advanced Mechatronic Systems (ICAMechS)*. IEEE, 2019, pp. 229–234.
- [68] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.
- [69] Z. Pu, D. Cabrera, C. Li, and J. V. de Oliveira, "Vgan: Generalizing mse gan and wgan-gp for robot fault diagnosis," *IEEE Intelligent Systems*, 2022.
- [70] R. Ding, B. Chen, G. Guo, and X. Yang, "Adversarial path sampling for recommender systems," *IEEE Intelligent Systems*, vol. 36, no. 6, pp. 23–31, 2020.
- [71] X. Wang, Z. Man, M. You, and C. Shen, "Adversarial generation of training examples: applications to moving vehicle license plate recognition," *arXiv preprint arXiv:1707.03124*, 2017.
- [72] L. Zhang, P. Wang, H. Li, Z. Li, C. Shen, and Y. Zhang, "A robust attentional framework for license plate recognition in the wild," *arXiv preprint arXiv:2006.03919*, 2020.
- [73] Online, <https://gulfnews.com/world/gulf/qatar/covid-19-qatar-issues-restrictions-on-number-of-passengers-in-one-car-1.1590517292042>, May 26, 2020, [Online;].
- [74] —, <https://auto.hindustantimes.com/auto/news/driving-during-lockdown-3-0-here-is-faqs-answered-41589167300775.html>, May 12, 2020, [Online;].
- [75] A. M. Rahmani and S. Y. H. Mirmahaleh, "Coronavirus disease (covid-19) prevention and treatment methods and effective parameters: A systematic literature review," *Sustainable cities and society*, p. 102568, 2020.
- [76] Y. Chen, M. Hu, C. Hua, G. Zhai, J. Zhang, Q. Li, and S. X. Yang, "Face mask assistant: Detection of face mask service stage based on mobile phone," *IEEE Sensors Journal*, vol. 21, no. 9, pp. 11 084–11 093, 2021.
- [77] M. S. Ejaz, M. R. Islam, M. Sifatullah, and A. Sarker, "Implementation of principal component analysis on masked and non-masked face recognition," in *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*. IEEE, 2019, pp. 1–5.
- [78] A. Nieto-Rodríguez, M. Mucientes, and V. M. Brea, "System for medical mask detection in the operating room through facial attributes," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2015, pp. 138–145.
-

BIBLIOGRAPHY

- [79] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic," *Measurement*, vol. 167, p. 108288, 2021.
- [80] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei *et al.*, "Masked face recognition dataset and application," *arXiv preprint arXiv:2003.09093*, 2020.
- [81] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, "SS-DMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2," *Sustainable cities and society*, p. 102692, 2020.
- [82] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [83] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.
- [84] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
- [85] M. Kawulok, E. Celebi, and B. Smolka, *Advances in face detection and facial image analysis*. Springer, 2016.
- [86] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*. IEEE, 2011, pp. 529–534.
- [87] P. Bhandary, <https://github.com/prajnasb/observations/tree/master/experiments/data>, 2020.
- [88] Wang, W. Z, H. G, X. B, H. Z, W. Q, C. H, and H, "Real-Time-Medical-Mask-Detection," <https://github.com/TheSSJ2612/Real-Time-Medical-Mask-Detection/>, 2020, [Online; accessed 14-Nov-2020].
- [89] S. Kundu and U. Maulik, "Vehicle pollution detection from images using deep learning," in *Intelligence Enabled Research*. Springer, 2020, pp. 1–5.
- [90] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, "Gated-gan: Adversarial gated networks for multi-collection style transfer," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 546–560, 2018.

-
- [91] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [92] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, no. 7569, pp. 367–371, 2015.
- [93] W. H. Organization, *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization, 2016.
- [94] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Transactions on Intelligent transportation systems*, 2019.
- [95] N. Kumar, S. S. Rahman, and N. Dhakad, "Fuzzy inference enabled deep reinforcement learning-based traffic light control for intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [96] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [97] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, no. 22, pp. 22 408–22 417, 2016.
- [98] M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal, and C. Sekar, "Air quality modelling using long short-term memory (lstm) over nct-delhi, india," *Air Quality, Atmosphere & Health*, vol. 12, no. 8, pp. 899–908, 2019.
- [99] A. P. F. dos Santos, K. K. da Silva, G. A. Borges, and L. A. d'Avila, "Fuel quality monitoring by color detection," in *Color Detection*. IntechOpen, 2019.
- [100] P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, and F. Scarselli, "Image generation by gan and style transfer for agar plate image segmentation," *Computer methods and programs in biomedicine*, vol. 184, p. 105268, 2020.
- [101] S. Vishwakarma, W. Li, C. Tang, K. Woodbridge, R. Adve, and K. Chetty, "Neural style transfer enhanced training support for human activity recognition," *arXiv preprint arXiv:2107.12821*, 2021.
- [102] R. Anicet Zanini and E. Luna Colombini, "Parkinson's disease emg data augmentation and simulation with dcgans and style transfer," *Sensors*, vol. 20, no. 9, p. 2605, 2020.
-

BIBLIOGRAPHY

- [103] C. Ma, Z. Ji, and M. Gao, "Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 128–136.
- [104] Y. Chi, L. Bi, J. Kim, D. Feng, and A. Kumar, "Controlled synthesis of dermoscopic images via a new color labeled generative style transfer network to enhance melanoma segmentation," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 2591–2594.
- [105] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*. IEEE, 2019, pp. 953–956.
- [106] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *INTERSPEECH*, 2019, pp. 2828–2832.
- [107] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [108] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.
- [109] X.-G. Luo, H.-B. Zhang, Z.-L. Zhang, Y. Yu, and K. Li, "A new framework of intelligent public transportation system based on the internet of things," *IEEE Access*, vol. 7, pp. 55 290–55 304, 2019.
- [110] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *arXiv preprint arXiv:2005.00935*, 2020.
- [111] G. T. S. Ho, Y. P. Tsang, C. H. Wu, W. H. Wong, and K. L. Choy, "A computer vision-based roadside occupation surveillance system for intelligent transport in smart cities," *Sensors*, vol. 19, no. 8, p. 1796, 2019.
- [112] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.
- [113] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

-
- [114] K. Behrendt, "Boxy vehicle detection in large images," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
- [115] M. Zdravkovich, "Smoke observations of the formation of a kármán vortex street," *Journal of Fluid Mechanics*, vol. 37, no. 3, pp. 491–496, 1969.
- [116] A. F. Ghoniem, X. Zhanga, O. Knioa, H. R. Baum, and R. G. Rehm, "Dispersion and deposition of smoke plumes generated in massive fires," *Journal of Hazardous Materials*, vol. 33, no. 2, pp. 275–293, 1993.
- [117] G. R. Blackman and J. W. Marvin, "Automatic temporal analysis of smoke/dust clouds," in *Image Understanding Systems II*, vol. 205. International Society for Optics and Photonics, 1980, pp. 180–184.
- [118] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.
- [119] I. Bello, "Lambdanetworks: Modeling long-range interactions without attention," *arXiv preprint arXiv:2102.08602*, 2021.
- [120] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [121] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [122] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.
- [123] J.-C. Burie, M. Coustaty, S. Hadi, M. W. A. Kesiman, J.-M. Ogier, E. Paulus, K. Sok, I. M. G. Sunarya, and D. Valy, "Icfhr2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 596–601.
- [124] H. Mittal and M. Saraswat, "An optimum multi-level image thresholding segmentation using non-local means 2d histogram and exponential kbest gravitational search algorithm," *Engineering Applications of Artificial Intelligence*, vol. 71, pp. 226–235, 2018.
-

BIBLIOGRAPHY

- [125] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 99–104.
- [126] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [127] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of graphics tools*, vol. 12, no. 2, pp. 13–21, 2007.
- [128] K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of niblack inspired binarization methods for ancient documents," in *Document Recognition and Retrieval XVI*, vol. 7247. International Society for Optics and Photonics, 2009, p. 72470U.
- [129] F. Patin, "An introduction to digital image processing," *online*: <http://www.programmersheaven.com/articles/patin/ImageProc.pdf>, 2003.
- [130] M.-L. Feng and Y.-P. Tan, "Contrast adaptive binarization of low quality document images," *IEICE Electronics Express*, vol. 1, no. 16, pp. 501–506, 2004.
- [131] C. Wolf and J.-M. Jolion, "Extraction and recognition of artificial text in multimedia documents," *Formal Pattern Analysis & Applications*, vol. 6, no. 4, pp. 309–326, 2004.
- [132] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [133] J. Calvo-Zaragoza and A.-J. Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, vol. 86, pp. 37–47, 2019.
- [134] R. De, A. Chakraborty, and R. Sarkar, "Document image binarization using dual discriminator generative adversarial networks," *IEEE Signal Processing Letters*, vol. 27, pp. 1090–1094, 2020.
- [135] R. Baran, T. Rusc, and P. Fornalski, "A smart camera for the surveillance of vehicles in intelligent transportation systems," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10 471–10 493, 2016.
- [136] F. Mehboob, M. Abbas, A. Rauf, S. A. Khan, and R. Jiang, "Video surveillance-based intelligent traffic management in smart cities," in *Intelligent Video Surveillance*. IntechOpen, 2019, p. 19.

-
- [137] F. Zhang, C. Li, and F. Yang, "Vehicle detection in urban traffic surveillance images based on convolutional neural networks with feature concatenation," *Sensors*, vol. 19, no. 3, p. 594, Jan 19, 2019.
- [138] M. A. Manzoor, Y. Morgan, and A. Bais, "Real-time vehicle make and model recognition system," *Machine Learning and Knowledge Extraction*, vol. 1, no. 2, pp. 611–629, Jun, 2019.
- [139] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-UAV," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 82–89, Apr 13, 2018.
- [140] B. S. Ali, "Traffic management for drones flying in the city," *International Journal of Critical Infrastructure Protection*, vol. 26, p. 100310, Sep 1, 2019.
- [141] H. Y. Ong and M. J. Kochenderfer, "Markov decision process-based distributed conflict resolution for drone air traffic management," *Journal of Guidance, Control, and Dynamics*, pp. 69–80, Jan, 2017.
- [142] Y. Xia, X. Shi, G. Song, Q. Geng, and Y. Liu, "Towards improving quality of video-based vehicle counting method for traffic flow estimation," *Signal Processing*, vol. 120, pp. 672–81, Mar 1, 2016.
- [143] J.-W. Hsieh, S.-H. Yu, Y.-S. Chen, and W.-F. Hu, "Automatic traffic surveillance system for vehicle tracking and classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 175–187, May 30, 2006.
- [144] A. Pramanik, S. Sarkar, and J. Maiti, "A real-time video surveillance system for traffic pre-events detection," *Accident Analysis & Prevention*, vol. 154, p. 106019, 2021.
- [145] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system," *Multimedia tools and applications*, vol. 76, no. 4, pp. 5817–5832, Feb 01, 2017.
- [146] W. Balid, H. Tafish, and H. H. Refai, "Intelligent vehicle counting and classification sensor for real-time traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 6, pp. 1784–1794, Sep 06, 2017.
- [147] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1198–1209, Sep 12, 2016.
- [148] S. Hua, M. Kapoor, and D. C. Anastasiu, "Vehicle tracking and speed estimation from traffic videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 153–160.
-

BIBLIOGRAPHY

- [149] A. Elihos, B. Alkan, B. Balci, and Y. Artan, "Comparison of image classification and object detection for passenger seat belt violation detection using nir & rgb surveillance camera images," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.
- [150] N. S. Punna, S. K. Sonbhadra, and S. Agarwal, "Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deepsort techniques," *arXiv preprint arXiv:2005.01385*, 2020.
- [151] D. Yang, E. Yurtsever, V. Renganathan, K. A. Redmill, and Ü. Özgüner, "A vision-based social distancing and critical density detection system for covid-19," *arXiv preprint arXiv:2007.03578*, pp. 24–25, 2020.
- [152] S. Saponara, A. Elhanashi, and A. Gagliardi, "Implementing a real-time, ai-based, people detection and social distancing measuring system for covid-19," *Journal of Real-Time Image Processing*, pp. 1–11, 2021.
- [153] D. Hörcher, R. Singh, and D. J. Graham, "Social distancing in public transport: mobilising new technologies for demand management under the covid-19 crisis," *Transportation*, pp. 1–30, 2021.
- [154] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 951–959.
- [155] F. Wilcoxon, "Individual comparisons by ranking methods. biom. bull., 1, 80–83," 1945.

Primanta Kunda

Signature of Candidate