M. Tech. Signal Processing and Digital Design Dhruv Sharma 2023

DRIVABLE AREA DETECTION USING YOLOV7

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

SIGNAL PROCESSING AND DIGITAL DESIGN

Submitted by:

DHRUV SHARMA

2K21/SPD/03

Under the supervision of

Prof. S. Indu
(Prof. ECE Dept.)

&

Dr. N. Jayanthi
(Asst. Prof. ECE Dept.)

DEPARTMENT OF ELECTRONICS AND COMM ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

JUNE, 2023

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Dhruv Sharma, Roll No. 2K21/SPD/03 student of MTech (Signal Processing and Digital Design), hereby declare that the project Dissertation titled "**Drivable Area Detection using YOLOv7**" which is submitted by me to the Department of Electronics and Communication Engineering, Delhi Technological university, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi — Dhruv Sharma

Date: — (2K21/SPD/03)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

We hereby certify that the project Dissertation titled "**Drivable Area Detection using YOLOv7**" which is submitted by Dhruv Sharma, Roll No. 2K21/SPD/03, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under our supervision. To the best of our knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

SUPERVISOR	SUPERVISOR
Prof. S. Indu	Dr. N. Jayanthi
Place: Delhi	Place: Delhi
Date:	Date:

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to express my gratitude towards all the people who have contributed their precious time and effort to help me without whom it would not have been possible for me to understand and complete the project. We express our profound gratitude to IIIT Hyderabad for their provision of dataset comprising of Indian road-scene videos diligently gathered by their esteemed research group, thus enabling us to embark upon this research work.

I would like to thank Prof. S. Indu and Dr. N. Jayanthi, DTU Delhi, Department of Electronics and Communication Engineering, my Project supervisors, for supporting, motivating and encouraging me throughout the period of this work was carried out. Their readiness for consultation at all times, their educative comments, their concern and assistance even with practical things have been invaluable.

Date:

Dhruv Sharma
MTech (SPDD)
Roll No. 2K21/SPD/03

ABSTRACT

Autonomous vehicle, also referred to as self-driving vehicles, represent a paradigm-shifting advancement in the realm of transportation, characterized by their ability to operate independently without any human intervention. Such vehicles require the usage of drivable area detection systems. In this dissertation, a YOLOv7 oriented framework using detection and segmentation was proposed to detect the safe-drivable area effectively and efficiently. The presence of numerous potholes on Indian roads is a serious problem because they increase the likelihood of accidents. Pothole detection is hence a must for drivable area detecting systems. Nonetheless, it can be challenging to tell the difference between the roadways and the potholes. HybridNets, being an advanced drivable area identification model, is limited to the tasks of lane-line segmentation and drivable region recognition, lacking the ability to detect potholes due to this limitation. To address this gap, the proposed approach suggests incorporating instance segmentation of road scenes using YOLOv7, which utilizes the E-ELAN layer in its backbone to facilitate the learning of more diverse and improved features. The utilization of various cardinalities and group convolutions in the E-ELAN layer promotes expansion and enhances the model's ability to acquire a wider range of features. Through extensive experimentation on our dataset, the proposed method achieves an impressive mean Average Precision (mAP) of 87%, outperforming the state-of-the-art HybridNets model which achieves an mAP of 67.8%. This demonstrates the effectiveness and superiority of the proposed approach in drivable area identification.

May, 2023 Dhruv Sharma

Delhi (India) (2K21/SPD/03)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

YOLO	You Only Look Once
FPS	Frames Per Second
CNN	Convolutional Neural Network
SPP	Spatial Pyramidal Parsing

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO AUTONOMOUS DRIVING

Autonomous vehicle, also referred to as self-driving vehicles, represent a paradigm-shifting advancement in the realm of transportation, characterized by their ability to operate independently without any human intervention. This conspicuously remarkable technological marvel epitomizes an unprecedented growth and transformative expansion within the transportation domain. It is amelioration of the human lifestyle and technology that has spurred a profound interest among the experts, researchers and industry enthusiasts on this clutter-breaking innovation of driverless vehicles [1]. Over the years, the interest among the researchers in this domain has been catapulted by its myriad of advantages including ameliorated road safety, decreased likelihood of road accidents due to human error, enhanced mobility, and reduced traffic congestion. The multifaceted merits of this domain continue to captivate researchers, driving them to explore and innovate in pursuit of a future driven by autonomous technologies.

Within the structure of a typical driverless vehicle lies the integration of an Advanced Driver Assistance System (ADAS). These intricately designed systems incorporate a plethora of cutting-edge technologies, including state-of-the-art sensors, advanced artificial intelligence algorithms, and sophisticated control systems. Such components synergistically collaborate to empower the vehicle with the ability to autonomously navigate and traverse roadways, all accomplished without any form of human intervention. Among the myriad of advanced systems, one particular system is the drivable area detection system, which leverages the power of deep learning methodologies to discern and evaluate the boundaries of the safe and drivable region. Through the utilization of intricate deep neural networks and sophisticated data processing techniques, this system epitomizes the synergy between cutting-edge technology and comprehensive analysis, thereby enabling the identification and delineation of areas conducive to safe vehicular navigation. In this dissertation, a drivable

area detection model is proposed based on YOLOv7 [2]. It infuses detection of vehicles on roads and segmentation of lane-lines, potholes and safe-drivable area.

1.2 LEVELS OF VEHICLE AUTOMATION

The concept of autonomous vehicles dates back several decades, with initial research and development efforts focused on exploring the feasibility of automated driving systems. Over the years, rapid advancements in computing power, machine learning, sensor technologies, and connectivity have propelled the development of autonomous vehicles to new heights. Today, major automotive companies, tech giants, and startups are investing substantial resources in the development and testing of autonomous vehicle technologies.

Over the past few decades, there has been a perpetual evolution in the operation mode of the automobiles. Hence, a hierarchy has been framed by the "Society of Automotive Engineers" (SAE) to categorize vehicles on the basis of their functioning and mode of operation, as summarized in figure 1. It comprises of six levels that vary from downright manual operations to accomplished fully automatic systems. While level 3 and 4 vehicles are available in the selected regions across the globe, the level 5 or the complete automation is yet to become a reality.

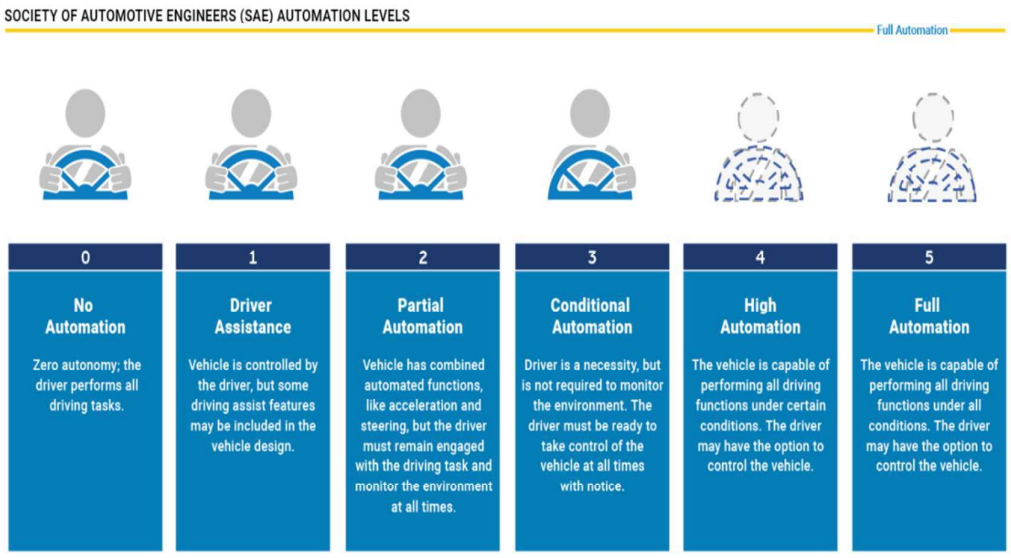


Fig. 1.1 Levels of Vehicle Automation

1.3 DRIVABLE AREA DETECTION SYSTEM

This dissertation proposes to implement a drivable area detection system using YOLOv7. The procedure of driveway area detection holds paramount importance within the realm of automated car systems. This pivotal detection system serves the purpose of assessing and determining the safe and drivable section of the road for a vehicle, by processing and analyzing the intricate details and attributes of the surrounding environment. Through the utilization of video frames and sensory inputs, this system harnesses the power of data-driven algorithms to decipher and delineate the optimal path for vehicular navigation, ensuring enhanced safety and efficiency in autonomous driving scenarios. The proposed model implements a typical detection system for drivable regions consists of vehicle detection, drivable region segmentation, lane-line segmentation, and pothole segmentation.

The vehicles are the obstacles on the road. For safe driving, it is essential that the autonomous car doesn't collide with the other vehicles, resulting into accidents. Hence, detection of the vehicles is a must. Potholes, or the uneasy concave depressions on roads, are highly undesirable due to numerous reasons. Potholes that are more profound can cause road accidents and damage to the vehicles. As described by Koch et al. [4], Potholes are optically characterized by their shape which is usually oval, with craggier and darker surfaces with respect to their circumambient roads. Such visually perceptible aspects could be manipulated using segmentation and shape extraction techniques to obtain numerical results that define a pothole mathematically. Segmentation of lane-lines are used to localize a vehicle and predict its lane. Tasks like automatic indication can only be informed if vehicle can recognize such lane-lines of varying widths. Finally, the drivable region is to be segmented to evaluate the safe path over which the vehicle can commute.

The primary aim of this dissertation is to implement a driveway area detection system using the YOLOv7 model. This model, characterized as an encoder-decoder network, amalgamates the principles of segmentation and object detection to facilitate the comprehensive detection of various aspects pertaining to driveways. Consequently, the model incorporates two distinct decoders: the vehicle detection head and the segmentation head. The overall workflow entails training the model weights utilizing driving scene images and subsequently evaluating the model's performance by testing it on video frames

sourced from a dedicated camera capturing the road during the driving process. Through this approach, the dissertation aims to demonstrate the effectiveness and applicability of the YOLOv7 model in detecting driveway areas in real-world scenarios [3]. Figure 2 shows the four tasks involved in the proposed drivable detection area system. Vehicles are represented through the bounding boxes. The lane-lines are represented by red line masks, potholes via pink masks and safe drivable region via orange mask.

Fig. 2. Proposed drivable area detection output classes.

1.4 CHALLENGES

There are several challenges and limitations of the existing deep learning methods for detection of safe-drivable area. In real-time scenarios, where vehicles are operated at high speeds, latency becomes a critical parameter that cannot be compromised. The trade-off between latency and performance metrics is evident: when latency is improved, the performance metrics tend to decline significantly, and vice versa. Object detection and segmentation techniques individually exhibit impressive performance metrics, but when combined, they can adversely impact the model's speed in terms of FPS, thereby compromising latency. Encoder-Decoder architectures offer improved latency at the cost of a noticeable decline in performance metrics. Consequently, one of the key challenges in this field is to establish an efficient deep learning model that strikes a balance between latency and performance metrics, effectively addressing the trade-off dilemma.

The detection of potholes on roads using deep learning methods poses a significant challenge in drivable-region detection algorithms and systems. Distinguishing these concave-sculpted depressions from the road surface is particularly challenging, especially during nighttime and unfavorable weather conditions. Currently, there is a lack of fusion systems that incorporate the detection of both drivable regions and potholes present on the roads. This gap in the existing research hinders the development of comprehensive solutions for drivable-region detection that effectively address the issue of pothole detection.

1.5 SCOPE OF WORK

Significant objectives and contributions of this work are:

- To extend the drivable area detection system to segment potholes as well, something that the SOTA systems are devoid of.
- Proposing to use the YOLOv7, an encoder-decoder architecture mainly used for detection and also extending it in order to the carry out the segmentation processes.
- To improve upon the performance metrics of the test results without reducing the inference speed and vice-versa. This is achieved through trainable bag of freebies.
- To assess the performance of the proposed model and compare it with SOTA method HybridNets. Notably, HybridNets is constrained to three specific tasks and does not encompass the critical task of pothole segmentation.

1.6 DISSERTATION ORGANIZATION

The content of the dissertation is organized into six chapters:

- Chapter I INTRODUCTION TO AUTONOMOUS DRIVING
- Chapter II LITERATURE SURVEY
- Chapter III BACKGROUND TECHNIQUES
- Chapter IV PROPOSED METHODOLOGY
- Chapter V EXPERIMENTAL RESULTS
- Chapter VI CONCLUSION AND FUTURE SCOPE

Chapter I – Includes the introduction to autonomous driving system and overview about drivable area detection tasks.

Chapter II – This chapter is literature survey, which gives an insight about the research papers published based on object detection, segmentation and multi-task approaches.

Chapter III – This chapter gives an insight into the background techniques that are being used in the implementation of the proposed work.

Chapter IV – This chapter covers the methodology that includes detection-segmentation architecture of YOLOv7 and system architecture along with the model losses.

Chapter V – This chapter includes the experimental results. The results also involve performance comparison between YOLOv7 and HybridNets.

Chapter VI – This includes the conclusion about the research work and future scope.

CHAPTER 2

LITERATURE SURVEY

In the field of accurate detection of drivable areas, significant research and advancements have been accomplished over the past few decades. Arising from the realm of conventional geometry-based modeling and energy-focused optimization schemes, the detection methods have undergone a transformation, embracing intricate methodologies rooted in deep learning. The approach of geometry-oriented modeling entails a comprehensive solution that encompasses the detection of driveway edges and the subsequent fitting of lines onto them [4]. In order to facilitate the detection of the road edges accurately, filter-oriented strategies have been employed. Aspects such as texture and colour may also be integrated to contribute to the aforementioned objectives. Such strategies include incorporation of gradient-oriented methods such as Gaussian [5] and Gabor filtering [6]. While Continuous Random Fields (CRF) represents a method that leverages energy optimization to effectively identify the lanes [7], it is the Hough transform techniques that have been employed to refine the alignment of lines over the detected edges [8]. The emergence of deep learning techniques led to ousting of such techniques with some of their usage being restricted to the pre-processing part. Based on the application, intent, and requirement, such deep learning methods can be categorized into three categories described in the following.

2.1 OBJECT DETECTION METHODS

Deep learning models specializing in bounding-box-based object detection on two-dimensional image signals have can be particularly useful in predicting obstacles, such as vehicles, on roadways. Several methods have been proposed over the years to improve upon the process of object detection. Deep learning algorithms, notably Convolutional Neural Networks (CNNs) [9], have demonstrated remarkable capabilities in achieving precise object identification. However, they are often associated with much prolonged inference times The Region-Convolutional Neural Network (R-CNN) generates throughput comprising sets of bounding boxes corresponding to obstacles, along with the corresponding class identification outputs [10]. RCNN relies on a selective

search algorithm as its foundation, which aids in predicting objects within designated regions. To accomplish this, RCNN employs linear Support Vector Machines (SVM) as its classification method. The approach exhibits limitations in terms of real-time detection speed primarily stemming from redundant feature computations. Subsequent advancements were introduced through the enhancements made via Fast R-CNN [14] and Faster R-CNN [15]. In contrast to RCNN, Fast R-CNN adopts the approach of utilizing the entire image as input for feature extraction, rather than processing each proposed region individually. However, it still faces challenges in achieving efficient detection speed, primarily due to limitations imposed by proposal detection. Conversely, Faster R-CNN eliminates the need for the selective search algorithm by enabling the network to learn region proposals, thereby positioning itself as one of the pioneering near-real-time deep learning object detectors. Although Faster R-CNN successfully overcomes the speed limitations of Fast R-CNN, it still encounters computational redundancy in subsequent detection stages.

In its quest to capture smaller objects, the Feature Pyramidal Network (FPN) employs a strategy involving the utilization of straightforward image pyramids. By rescaling the image to various sizes before inputting it into the main network, FPN enhances its ability to capture diverse object scales. The final prediction generated by FPN is derived from a fusion of outputs from different methods, resulting in a comprehensive and inclusive detection capability across a wide range of scales [10]. The Single Shot Detector (SSD) exhibits a commendable frame rate ranging from 22 to 59 frames per second (FPS) during inference, thanks to its efficient design. Leveraging multi-resolution techniques, SSD carries out object detection effectively. Notably, SSD employs a single progressive pass to seamlessly perform localization and categorization steps. However, its performance deteriorates when dealing with smaller objects, presenting a challenge in accurately detecting them [14]. Perceiving object detection as a regression problem, the You Only Look Once (YOLO) approach [15] involves the partitioning of ground truth into grids. Consequently, the implementation of this methodology necessitates only a single forward propagation. YOLO is designed to perform single stage detection. Notably, YOLO achieves the simultaneous prediction of bounding boxes across all classes within an image, resulting in an impressive frame rate of 45 frames per second. However, YOLO experiences a decline in localization accuracy despite the rise in its inference speed, particularly for certain small objects. YOLOv2 [16]

and its subsequent iteration, YOLOv3 [17], demonstrated advancements in both object detection accuracy and speed, yet there remains considerable room for further enhancement.

2.2 SEGMENTATION METHODS

Numerous segmentation techniques encompassing both instance and semantic segmentation have been introduced to facilitate the segmentation and masking of lane lines and drivable road regions. When it comes to identifying potholes, the maintenance team can opt for the bounding box method, which provides a general area of interest. Alternatively, if a more precise representation of the pothole shape is necessary while driving, segmentation methods can be employed to accurately delineate its boundaries on the road. Further several segmentation methodologies have also been put into use to detect the driveway area. Different segmentation methods sing deep learning have been proposed over the years. The Fully Convolutional Network (FCN) is employed for conducting semantic segmentation tasks [18]. It utilizes locally connected layers, resulting in fewer model parameters and thus enhancing the network's computational efficiency. The FCN architecture incorporates downsampling and upsampling paths, with the former focusing on feature extraction and the latter on the precise localization of segmentation masks. By employing this approach, FCN achieves accurate semantic segmentation while maintaining a faster net- work operation. Arjapure et al. [19] used VGG annotator tool to annotate the images containing potholes. They further utilize Mask-RCNN in order to identify potholes under the Regions of Interest (ROI). Subsequently, the area occupied by the pothole is calculated based on the generated ROI.

UNet is also renowned for its semantic segmentation capabilities [20], undertakes the task of categorizing each pixel within the ground truth. This attribute empowers UNet with remarkable border differentiation capabilities. Its unique architecture resembles a U shape, leveraging both contractive and expansive components to ensure input and output have matching dimensions. LaneNet [21] adopts a sophisticated approach by combining binary segmentation with clustering-oriented embedding to achieve robust lane-line detection. Its primary focus lies in segmenting lane-lines accurately. Operating on an encoder-decoder architecture, LaneNet utilizes the decoder to generate pixel-wise predictions efficiently. Notably, LaneNet is designed specifically

for lane-line segmentation and does not cater to multi-class segmentation tasks. Spatial-CNN (SCNN) exhibits the remarkable ability to capture translational and rotational relationships among ground-truth pixels, resulting in significant enhancements in the segmentation process [22]. This advancement is achieved through the utilization of a slice-by-slice convolution method, which facilitates the exchange of information within individual pixels. In the SCNN model, rows and columns of feature maps are treated as layers, allowing for a sequential flow of pixel details. This unique approach proves particularly advantageous in capturing long, continuous shapes with precision and accuracy.

SegNet [23], an additional model employed for semantic segmentation, incorporates an encoder, decoder, and a pixel categorization layer. Notably, the novelty of SegNet lies in its decoder, which distinguishes itself by skillfully upsampling both low-resolution and full ground-truth feature maps. Remarkably, the upsampling process in SegNet is characterized by a non-linear nature, enabling the model to effectively reconstruct detailed segmentation maps with higher resolution. This distinctive feature of SegNet contributes to its superior performance in accurately delineating object boundaries and capturing intricate semantic information within images. ENet [24], a cutting-edge framework, adopts a compact encoder-decoder architecture that effectively reduces the number of parameters and enhances processing speed. This architecture is designed with a specific focus on efficiency. ENet leverages the Parametric Rectified Linear Unit (PReLU) as its activation function. Additionally, ENet employs dilated convolutions. Notably, ENet optimizes processing costs through early downsampling, which strategically optimizes the initial stages of the network, leading to improved overall performance and faster inference times. In the research work referenced as [25], an innovative hybrid encoder-decoder network is introduced, incorporating both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) units. This advanced architecture is purposefully devised to efficiently handle data exhibiting temporal sequence characteristics. Within this framework, the CNN component undertakes the processing of input data, taking into account its inherent temporal properties. Simultaneously, the LSTM component intelligently learns and captures significant attributes from the data. The resultant learned features are subsequently employed for accurate lane forecasting, showcasing the network's ability to predict and anticipate lane behavior. Through the integration of CNN and LSTM, this hybrid network

exemplifies substantial potential in effectively addressing dynamic scenarios and enabling informed decision-making processes.

2.3 MULTI-TASK APPROACH

Drivable area detection is a multi-task application that requires detection of the road-obstacles such as vehicles., and segmentation of lane-lines, potholes and safe-drivable region. This necessitates the usage of both detection as well as segmentation processes in such a system. Consequently, there is a need to combine detection and segmentation methods to effectively determine the safe drivable region. However, integrating different models for fusion purposes poses challenges in terms of inference speed and latency. The performance of the model is compromised, making it unsuitable for real-time applications where timely coordination, aligned with the vehicle's speed, is of utmost importance. In order to address this limitation, researchers have proposed simpler architectures that integrate these networks into streamlined encoder-decoder-based models. By utilizing these simplified architectures, the computational efficiency is improved while still maintaining the ability to perform accurate driveway area detection and segmentation. Mask R-CNN is an encoder-decoder-oriented network [26] that is capable of performing segmentation of images along with the detection of objects. It is an extension of Faster-RCNN.

The Drivable-area, Lane line and Traffic detection-Network (DLT-Net) is a comprehensive framework designed to address the three crucial tasks of drivable-region detection, lane line detection, and traffic detection. The architecture of DLT-Net heavily relies on the contextual information provided by specially designed tensors, enabling efficient coordination between the three task-specific decoders [27]. The encoder component generates feature maps of various sizes, with the final feature map being fed into the dedicated decoders for each task. The context tensors, strategically positioned between the encoder and the decoders, contribute to improving the discriminative capability among the tasks. However, DLT-Net does possess certain limitations. It struggles to yield optimal results in scenarios involving intense reflection while driving. Additionally, it demonstrates limitations in accurately predicting recurring divider lines as lane-lines. These areas present opportunities for further refinement and improvement in the performance of DLT-Net. The Dynamic Fusion Module-RGB Fusion Thermal

Network (DFM-RTFNet) [28] is an advanced model that incorporates a feature fusion technique through the dynamic integration of the DFM module into the RTFNet architecture [40]. This methodology involves generating a dynamic kernel based on a two-dimensional feature and applying it to another feature to achieve fusion. By leveraging the properties of spatial variance and content dependency, distinct kernels can be selectively applied to different regions of an image, optimizing the fusion process and enhancing network efficiency. The RTF-Net, which forms the foundation of DFM-RTFNet, consists of two encoders dedicated to RGB and Thermal inputs, drawing inspiration from the renowned ResNet architecture. Subsequently, a series of five decoder layers is employed to process the fused features. Finally, a softmax layer is applied to generate detection and segmentation results, marking the conclusive stage of the network's operation.

YOLOP, an innovative deep learning model for driveway-area detection [29], has been developed by the authors specifically for real-time scenarios on the Jeston Tx2 embedded device. This model aims to excel in multiple tasks, including traffic object detection, drivable-region segmentation, and lane-line detection. Built upon the foundation of YOLOv4, the model utilizes Adam as the optimizer during training [29]. The encoder module is composed of two main architectures: the backbone architecture, based on the Cross Stage Partial Dark Network (CSPDarkNet), and the neck architecture, utilizing the Spatial Pyramid Pooling (SPP) module. The input to the heads is obtained from the bottom layer of the neck. To restore the output feature map to the appropriate size, triple-upsampling is employed in the heads. This approach, inspired by the trainable Bag of Freebies (BoF), contributes to achieving high-precision throughput and faster inference. HybridNets stand apart from models like YOLOP and others due to its unique architecture, which features a shared encoder and only two decoders—one dedicated to vehicle detection and the other to lane-driveway area segmentation—rather than three separate decoders . The encoder component has been carefully designed to incorporate model scaling, a critical aspect for efficient hardware system implementation. Model scaling is achieved through the utilization of Bidirectional Feature Pyramid Networks (BiFPN), which enables efficient information flow across different scales. The backbone of the model is EfficientNet-B3, which plays a vital role in optimizing the network by evaluating scaling parameters, ensuring the stability of the network design. The encoder also incorporates a neck network based on EfficientDet, which receives the feature maps

from the backbone [30]. The Bidirectional Feature Pyramid Networks (BiFPN) combine attributes at various resolutions, leveraging cross-scale connections for each node through bidirectional paths. This approach empowers the model to effectively handle complex datasets. The segmentation head of the model focuses on examining three categories: background, driveway-area, and lane-line, enabling comprehensive analysis and detection.

CHAPTER 3

BACKGROUND TECHNIQUES

3.1 ENCODER-DECODER ARCHITECTURES

Encoder-Decoder architectures are simple structures that perform feature extraction by generation of feature maps in the first part and detection/segmentation in the next part. In other words, the proposed architecture consists of an encoder component designed to process input sequences of varying lengths, and a decoder component that functions as a conditional language model. The encoder receives the input sequence and performs encoding operations to capture its underlying features. On the other hand, the decoder utilizes the encoded input as well as the preceding context of the target sequence to predict the next token in the target sequence. This dynamic interaction between the encoder and decoder enables the model to effectively generate accurate predictions for subsequent tokens based on the contextual information. YOLOv7 is also an encoder-decoder architecture, with its backbone serving as the encoder and head as its decoder.

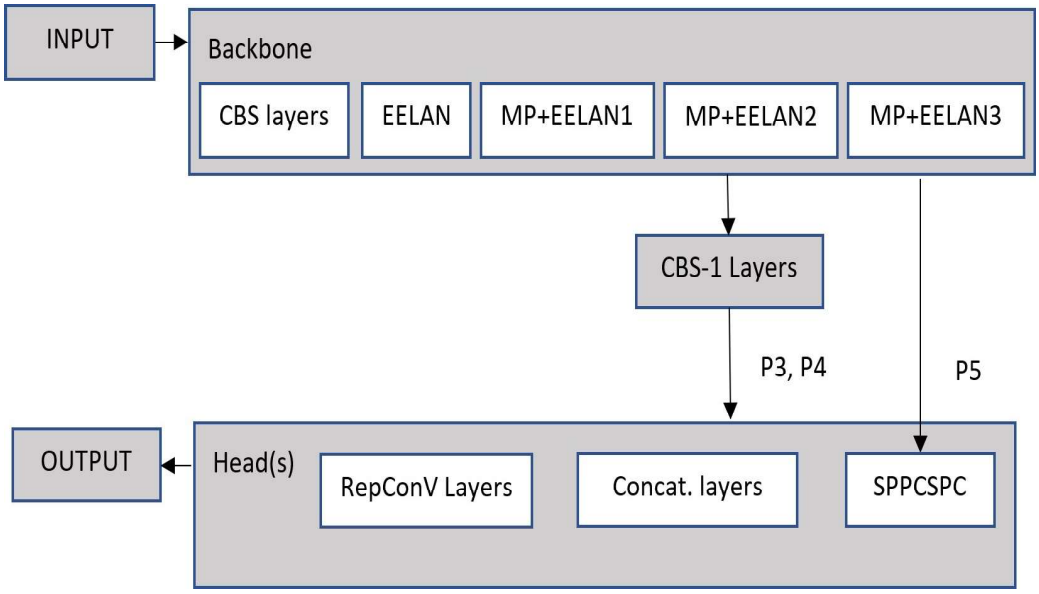


Fig. 3.1 An Overview of YOLOv7 Architecture

3.2 AN OVERVIEW OF HYBRIDNETS

The HybridNets model, similar to the proposed methodology, adopts an encoder-decoder network architecture. It integrates both semantic segmentation and object detection techniques, allowing for the comprehensive detection of various elements within the scene. By incorporating these two decoders, namely vehicle detection and driveway-lane segmentation heads, the model is capable of simultaneously addressing multiple aspects of interest. This comprehensive approach enables the HybridNets model to effectively analyze and interpret the complex visual information captured on the road. EfficientNet-B3, an influential component of the architecture, assumes the role of the encoder-backbone. Its primary function entails the optimization of the network through the meticulous assessment of scaling parameters [31]. This pivotal contribution serves to bolster the stability of the network design while concurrently mitigating the computational burden. The neck network effectively receives the feature maps derived from the backbone, thereby serving as an intermediary connecting element [32]. To facilitate the seamless integration of attributes at distinct resolutions, the neck network incorporates EfficientDet-inspired BiFPN (Bidirectional Feature Pyramid Network) components. This innovative approach not only enables the fusion of multi-level features but also contributes to the overall efficiency and effectiveness of the segmentation process. Each grid of the multi-scale fusion feature maps from the Neck network will be assigned nine prior anchors with different aspect ratios. K-means clustering is utilized [33] to establish the anchor boxes.

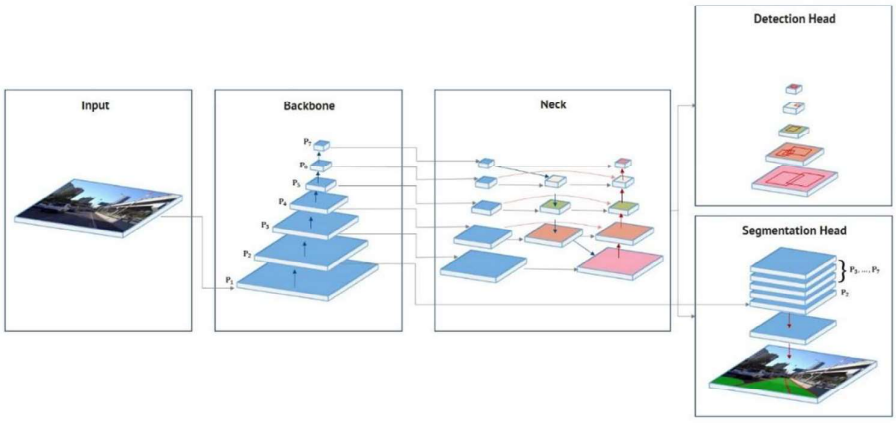


Fig. 3.2 HybridNets Architecture [30]

3.3 BOUNDING-BOX REGRESSION

Bounding box regression is a sophisticated technique employed to estimate the precise spatial location of an object within an image. This technique involves training a model to predict the coordinates that define the bounding box encompassing the object of interest. By leveraging this approach, the model can effectively localize and outline the object's boundaries. Despite its relative simplicity in implementation, bounding box regression serves as an advantageous starting point in object detection tasks, facilitating accurate localization and subsequent analysis of objects within an image.

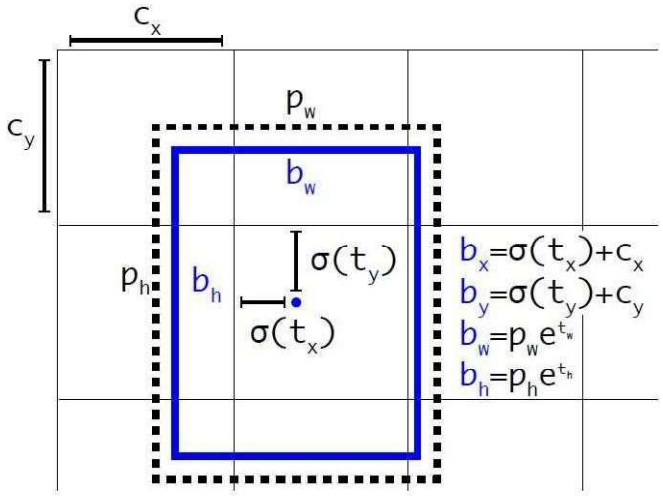


Fig 3.3 Bounding Box Detection [YOLO]

In figure 3.3, *cx* and *cy* represent the top upper left corner of the image while *pw* and *ph* represent width and height.

3.4 SEGMENTATION

Image segmentation is a sophisticated technique employed in the realm of computer vision to partition a digital image into discrete and meaningful subgroups known as image segments. This process serves the purpose of simplifying the complexity of the image, thereby facilitating subsequent processing or analysis on each individual segment. From a technical standpoint, segmentation entails the intricate task of assigning labels to individual pixels, thereby discerning and identifying objects, people, or other significant elements present within the image. In order to optimize computational

resources and improve the efficiency of object detection, a prevalent approach involves the utilization of an image segmentation algorithm to identify and extract objects of interest within the image. By employing this strategy, the subsequent object detector can focus exclusively on the predefined bounding boxes derived from the segmentation algorithm's output. This targeted approach eliminates the need for the detector to process the entire image, resulting in enhanced accuracy and reduced inference time.

Fig 3.4 Segmentation Example of Road and Pothole

CHAPTER 4

PROPOSED METHODOLOGY

4.1 PROPOSED SYSTEM

The proposed system is summarized in figure 4.1. The proposed methodology employs a combination of segmentation and object detection techniques in order to accurately detect and delineate the safe drivable area. By leveraging the capabilities of

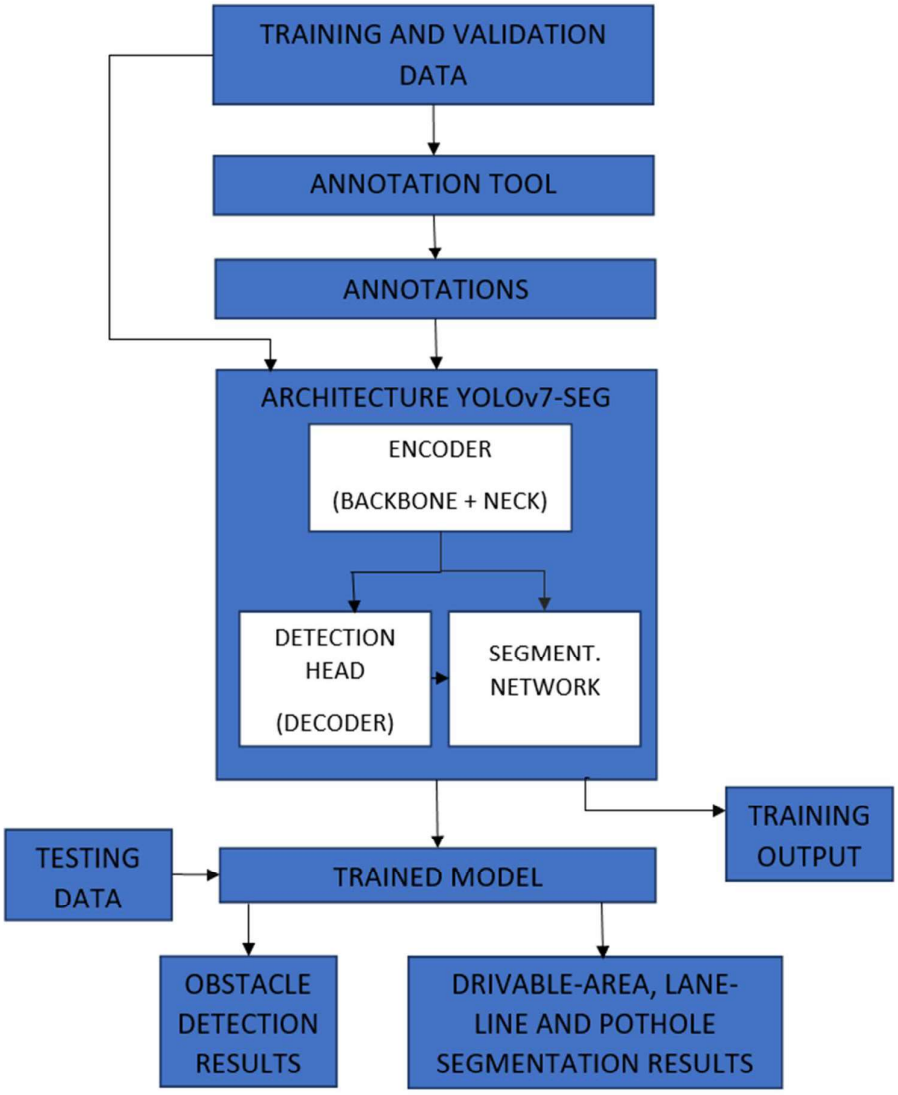


Fig. 4.1 Proposed System

both segmentation and object detection, the methodology aims to enhance the precision and reliability of the drivable area detection process. Road-scene images are firstly annotated manually. For vehicles, bounding boxes are drawn around them, while polygon-based annotations are used for segmentation-oriented classes. Prior to the commencement of the training process, the road images undergo an annotation procedure wherein detailed annotations are generated and formatted in a text (txt) file. Notably, these annotations comprise numerical values represented as floating-point numbers, ranging between 0 and 1, capturing the essential attributes of the road scene. Subsequently, these annotated images are inputted into the training module, wherein the robust YOLOv7 model assumes a pivotal role. There are three segmentation labels that we have used, that is, the drivable region, lane lines and the potholes and two object detection labels, cars and trucks. Instance segmentation has been proposed in this model that is inspired from YOLOv5 [34] and YOLOv6 [35]. Based on it, an anchor-free instance segmentation method has been developed. The testing data is finally fed to the trained model to obtain the detection and segmentation results from their heads, respectively.

4.2YOLOV7 ARCHITECTURE

4.3 At its core, YOLOv7 adopts an intricate encoder-decoder architecture, which encompasses two essential components: the backbone and the neck network. The backbone serves as a feature extraction network, responsible for generating feature maps of distinct resolutions. Beginning with P1, representing a downsampled resolution of half the original size, the feature maps progressively evolve up to P5, with a downsampled resolution of 1/32 of the original size. This intricate process spans across approximately 50 layers within the network. The backbone integrates an array of sophisticated techniques and components, strategically incorporated to enhance its performance and overall efficacy. The backbone consists of the following modules.

- Extended Efficient Layer Aggregation Network – E-ELAN ameliorates the feature learning capabilities of the backbone by leveraging the ”Expand, Shuffle and Merge” cardinalities and group convolutional techniques. This approach significantly improves upon the diversity of aspect learning while still maintaining the gradient path without compromising it.
- MaxPool Layer – The module incorporates a pooling technique where the input tensor is pooled using various kernel sizes without deducting the resolution of the

input(s), which signifies stride being unity. This is accomplished using a Max-Pooling (MP) layer. The module then concatenates the input tensor with the pooled outputs, generating multi-scale region aspects through this pooling and concatenation operation.

- Convolution Batch Normalization – In the realm of deep learning, the merging of convolution and batch normalization layers into a single convolutional layer is a popular technique. The merits of this method have been well established, including ameliorated efficiency and deducted computational cost. The integration of the batch normalization step into the convolution operation enables the model to learn more informative representations of the data, while simultaneously normalizing the inputs to each layer.

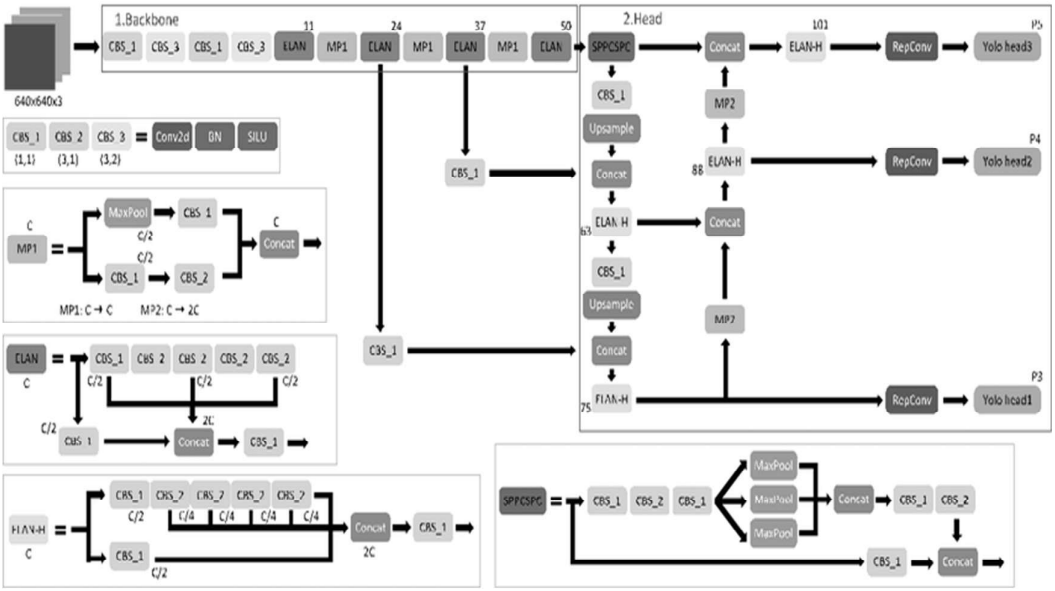


Fig. 4.2 YOLOv7 detailed Architecture

The YOLOv7 model uses a compound scaling method to multiple alternative models with different inference speeds of variable scales. This constitutes of amendment of scale factors including depth, width, resolution, and stage in an attempt to prepare models of distinct sizes. By scaling depth and width factors, models can achieve optimal performance while retaining the characteristics of the original representation. In order to implement trainable BoF, the planned re-parameterized convolution technique with a module-level ensemble is used in YOLOv7, utilizing the RepConv method that combines 3x3 and 1x1 convolutions with an identity connection. However, in YOLOv7, RepConvN

skips the identity connection to preserve residuals and concatenation. Additionally, re-parameterization is also applied to CBN for superior results.

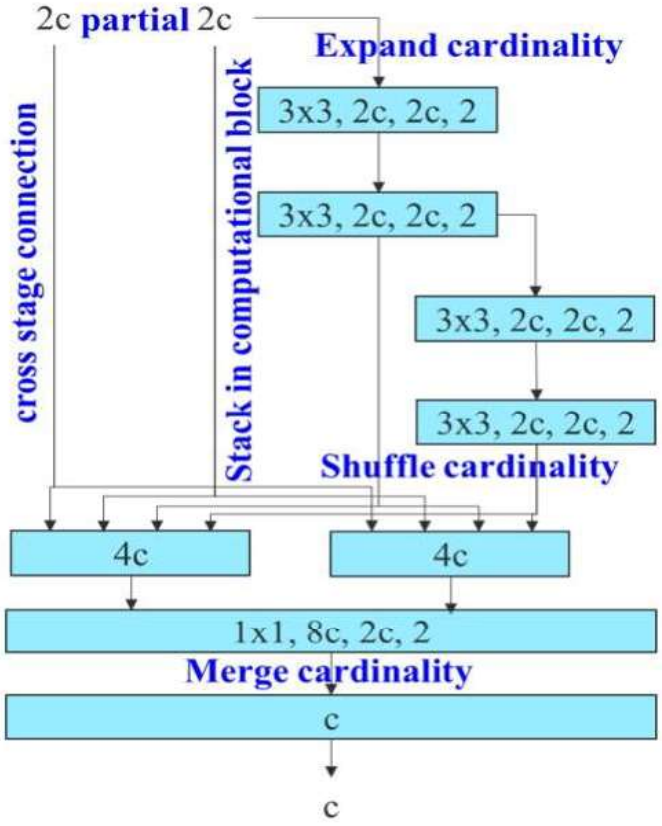


Fig 4.3 E-ELAN Module [3]

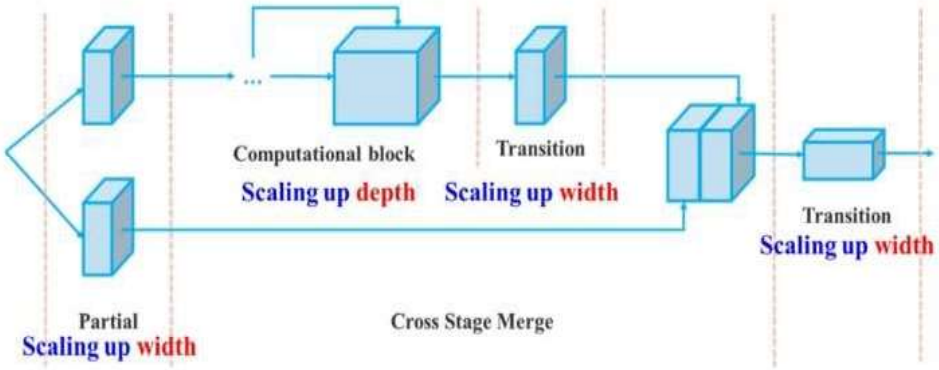


Fig 4.4 Compound Scaling [3]

The SPP (Spatial Pyramid Pooling) module plays a pivotal role in the network architecture by employing pooling operations with various kernel sizes, without reducing

the input resolution (stride == 1). Maxpooling layers are utilized for this purpose. The module then combines the original input tensor with the pooled outputs through concatenation. This process generates multiscale region features, leveraging the pooling and concatenation operations.

On the other hand, the CSPNet (Cross Stage Partial Network) is designed to address the issue of redundant gradient information by dividing the feature map of the base layer into two parts. One part undergoes a block operation, such as a dense block or a resblock, along with a transition layer. The other part is combined with the transmitted feature map for the subsequent stage. This approach effectively reduces the computational complexity, memory usage, and duplicate gradient information. As a result, the network benefits from improved inference speed and accuracy.

The YOLOv7 model implements a novel two-headed approach to enhance its ability to segment objects and locate them accurately. The lead head is responsible for the final output, while the auxiliary head assists in training by utilizing shallow and superficial weights, as well as an assistant loss to provide guidance. The model incorporates deep supervision techniques, such as the label assigner, which generates soft and coarse labels to optimize learning and allow the lead head to capture residual information.

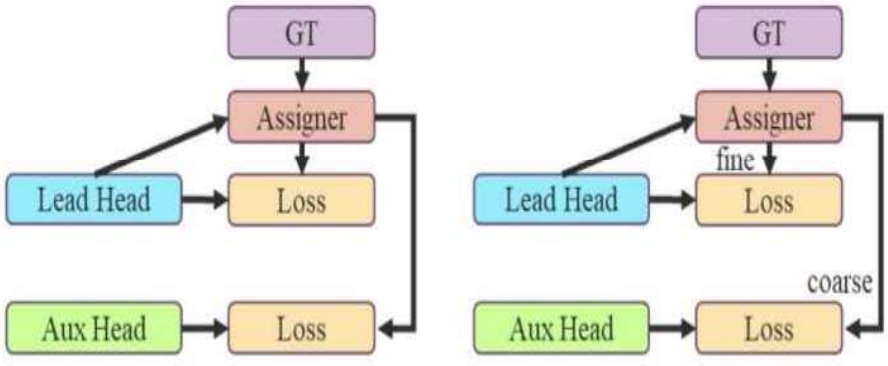


Fig 4.5 Label Assignment in YOLOv7 [3]

```
# yolov7 backbone
backbone:
  # [from, number, module, args]
  [[-1, 1, Conv, [32, 3, 1]],  # 0

   [-1, 1, Conv, [64, 3, 2]],  # 1-P1/2
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [128, 3, 2]],  # 3-P2/4
   [-1, 1, Conv, [64, 1, 1]],
   [-2, 1, Conv, [64, 1, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [[-1, -3, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [256, 1, 1]],  # 11
   [-1, 1, MP, []],
   [-1, 1, Conv, [128, 1, 1]],
   [-3, 1, Conv, [128, 1, 1]],
   [-1, 1, Conv, [128, 3, 2]],
   [[-1, -3], 1, Concat, [1]],  # 16-P3/8
   [-1, 1, Conv, [128, 1, 1]],
   [-2, 1, Conv, [128, 1, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [[-1, -3, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [512, 1, 1]],  # 24
   [-1, 1, MP, []],
   [-1, 1, Conv, [256, 1, 1]],
   [-3, 1, Conv, [256, 1, 1]],
   [-1, 1, Conv, [256, 3, 2]],
   [[-1, -3], 1, Concat, [1]],  # 29-P4/16
   [-1, 1, Conv, [256, 1, 1]],
   [-2, 1, Conv, [256, 1, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [[-1, -3, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [1024, 1, 1]],  # 37
   [-1, 1, MP, []],
   [-1, 1, Conv, [512, 1, 1]],
   [-3, 1, Conv, [512, 1, 1]],
   [-1, 1, Conv, [512, 3, 2]],
   [[-1, -3], 1, Concat, [1]],  # 42-P5/32
   [-1, 1, Conv, [256, 1, 1]],
   [-2, 1, Conv, [256, 1, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [[-1, -3, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [1024, 1, 1]],  # 50
  ]
```

Fig 4.6 YOLOv7 Backbone layers

```
head:
  [[-1, 1, SPPCSPC, [512]], # 51
   [-1, 1, Conv, [256, 1, 1]],
   [-1, 1, nn.Upsample, [None, 2, 'nearest']],
   [37, 1, Conv, [256, 1, 1]], # route backbone P4
   [[-1, -2], 1, Concat, [1]],
   [-1, 1, Conv, [256, 1, 1]],
   [-2, 1, Conv, [256, 1, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [[-1, -2, -3, -4, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [256, 1, 1]], # 63
   [-1, 1, Conv, [128, 1, 1]],
   [-1, 1, nn.Upsample, [None, 2, 'nearest']],
   [24, 1, Conv, [128, 1, 1]], # route backbone P3
   [[-1, -2], 1, Concat, [1]],
   [-1, 1, Conv, [128, 1, 1]],
   [-2, 1, Conv, [128, 1, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [-1, 1, Conv, [64, 3, 1]],
   [[-1, -2, -3, -4, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [128, 1, 1]], # 75
   [-1, 1, MP, []],
   [-1, 1, Conv, [128, 1, 1]],
   [-3, 1, Conv, [128, 1, 1]],
   [-1, 1, Conv, [128, 3, 2]],
   [[-1, -3, 63], 1, Concat, [1]],
   [-1, 1, Conv, [256, 1, 1]],
   [-2, 1, Conv, [256, 1, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [-1, 1, Conv, [128, 3, 1]],
   [[-1, -2, -3, -4, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [256, 1, 1]], # 88
   [-1, 1, MP, []],
   [-1, 1, Conv, [256, 1, 1]],
   [-3, 1, Conv, [256, 1, 1]],
   [-1, 1, Conv, [256, 3, 2]],
   [[-1, -3, 51], 1, Concat, [1]],
   [-1, 1, Conv, [512, 1, 1]],
   [-2, 1, Conv, [512, 1, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [-1, 1, Conv, [256, 3, 1]],
   [[-1, -2, -3, -4, -5, -6], 1, Concat, [1]],
   [-1, 1, Conv, [512, 1, 1]], # 101
   [75, 1, RepConv, [256, 3, 1]],
   [88, 1, RepConv, [512, 3, 1]],
   [101, 1, RepConv, [1024, 3, 1]],
   [[102,103,104], 1, Detect, [nc, anchors]],   # Detect(P3, P4, P5)
  ]
```

Fig 4.7 YOLOv7 Head layers

4.3 SEGMENTATION

The model architecture utilizes an instance segmentation method on YOLOv7 that is inspired by YOLOv5 [18] and YOLOv6 [19], which is an anchor-free approach. The YOLOv7 segmentation architecture draws inspiration from a compact, fully connected neural network known as ProtoNet. This influential concept serves as the foundation for the object detection head integrated within the YOLOv7 framework. By combining the ProtoNet principles with the object detection capabilities, YOLOv7 achieves its distinctive segmentation architecture. It consists of three convolutional layers. It operates by generating prototype masks that are integral to the segmentation model. This approach shares similarities with the concept of FCNs employed in semantic segmentation tasks. By leveraging this network, the segmentation model is equipped with the ability to generate prototype masks, enabling effective and precise segmentation of the desired objects.

4.4 LOSS FUNCTION

The comprehensive loss function encompasses a range of individual losses to address different aspects of the network's objectives. These include the segmentation loss, which measures the dissimilarity between predicted and ground truth segmentation masks. The bounding box loss quantifies the disparity between predicted and actual bounding box coordinates. The detection loss evaluates the dissimilarity between predicted and ground truth object detections. Lastly, the classification loss captures the discrepancy between predicted and true class labels. By incorporating these diverse loss components, the network can effectively optimize its performance across various aspects of the task at hand.

$$loss = Segment.loss + Bounding\ Box\ loss + Detection\ loss + Class.loss \quad (4.1)$$

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 DATASET PREPARATION

The dataset utilized for evaluating the proposed detection and segmentation task was sourced from a renowned research group based at IIIT Hyderabad, India. Specifically, a video dataset was selected, and a total of 1000 frames were extracted from it for further analysis. These frames were meticulously annotated in a text (txt) format to ensure compatibility with the YOLOv7 model. The annotations encompassed five key classes: cars, trucks, potholes, drivable regions, and lane lines. For the precise annotation of cars and trucks, rectangular bounding boxes were employed, while polygons were used to accurately delineate the boundaries of potholes, lane lines, and drivable regions. Furthermore, binary masks were generated specifically for the segmented lane lines and the overall road area, which were indispensable for the subsequent analysis and evaluation of the HybridNets model.

5.2 PERFORMANCE METRICS

A diverse range of performance metrics has been employed to assess the effectiveness of the model. Intersection over Union (IoU) is a metric that quantifies the extent of overlap between the segmented areas by calculating the ratio of their common intersection to their union. This metric has been utilized to evaluate the segmentation aspect of the model, as referenced in [36]. For the vehicle detection component, precision, recall, F1-score, and mean Average Precision (mAP) have been adopted as evaluation measures. Precision refers to the proportion of correctly identified positive detections out of the total detections made. Recall, on the other hand, represents the proportion of actual positive instances that are correctly identified. F1-score is a combined metric that balances both precision and recall. Lastly, mAP calculates the average precision across different object categories, providing an overall assessment of the detection performance.

$$Recall(R) = TP/(TP + FN) \tag{5.1}$$

$$Precision(P) = TP/(TP + FP) \quad (5.2)$$

$$F1 - score = 2RP/(R + P) \quad (5.3)$$

mAP is calculated from the recall-precision curve. The PR curve helps in evaluating the trade-off between the two parameters and evaluate the best result in terms of mAP. [37]

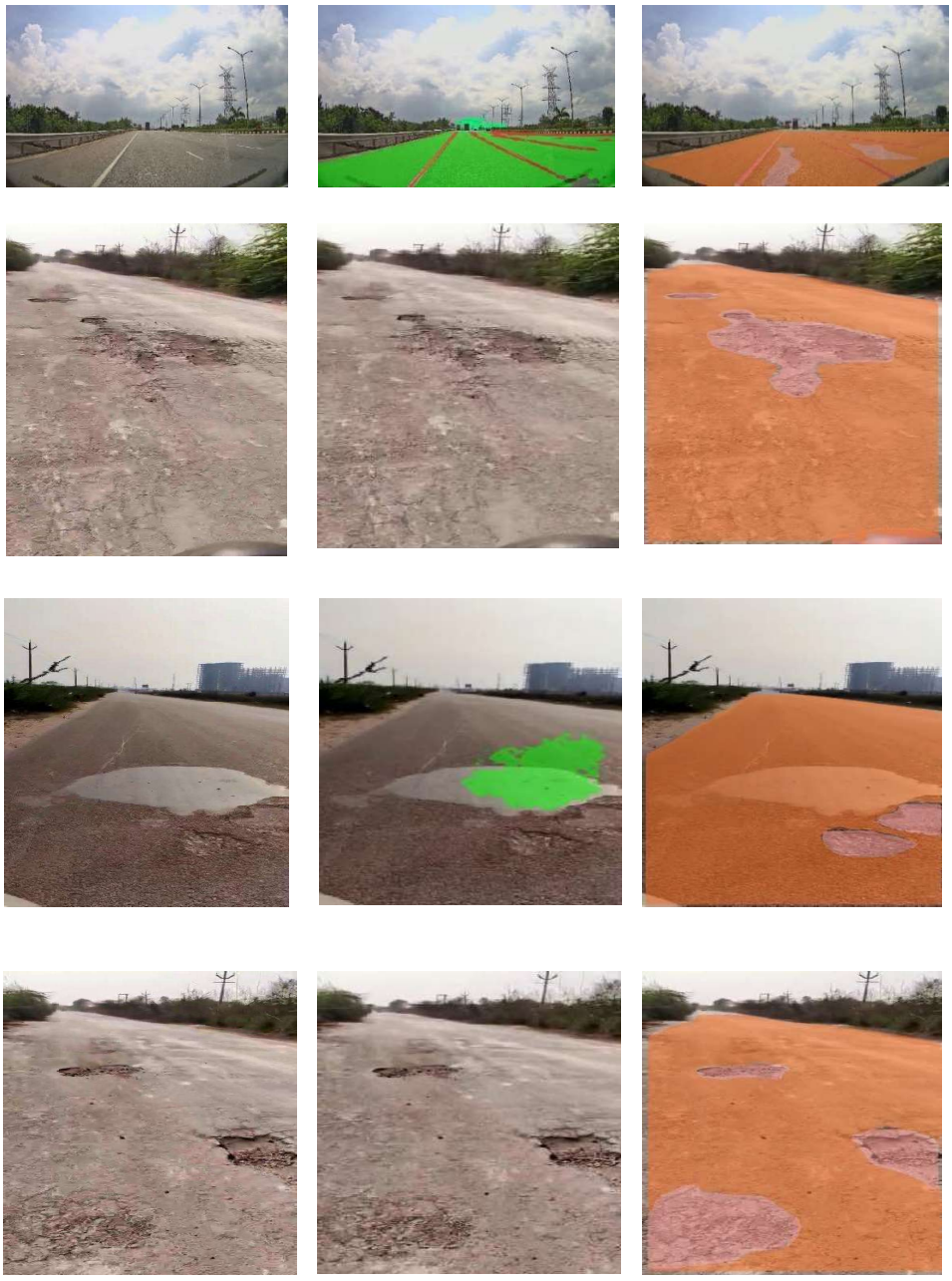
5.3 SYSTEM REQUIREMENTS

The implementation of the proposed methodology was conducted on the Linux Ubuntu platform, specifically version 22.04.01. To facilitate the training, validation, and testing processes of the model, the NVIDIA RTX A5000 GPU was employed, leveraging its computational power. The model itself was developed using the PyTorch 1.10 library, which provided a robust framework for deep learning tasks and facilitated efficient execution of the proposed algorithms.

5.4 RESULTS AND DISCUSSION

A subset of 1000 frames was extracted from a video dataset for further analysis. Out of these frames, 900 were utilized for training the model, while the remaining 100 frames were reserved for validation purposes. The training process involved multiple iterations, with variations in epoch numbers and batch sizes, in order to optimize the model's performance. After thorough experimentation, it was determined that utilizing a batch size of 8 over 100 epochs yielded the most favorable results.

To evaluate the model's effectiveness, a selection of video frames from the same dataset was randomly chosen for testing purposes. In addition to our proposed detection model, we also trained the state-of-the-art drivable area detection model, HybridNets, using the same dataset. The outputs generated by both models, including the ground truth and the combined driveway area detection output, are visually presented in Figure 5.1 for comparative analysis. The precision recall curve, F!-confidence curve, Precision-Confidence curve and Recall-Confidence curve are shown in the subsequent figures 5.2, 5.3, 5.4 and 5.5.



a) b) c)

Fig. 5.1 Model Results a) Ground Truth b) HybridNets c) YOLOv7. HybridNets is incapable of performing pothole detection. Also, it is only capable of identifying smooth roads, as in case1 and fails to detect roads that are improper or slightly damaged in rest of the cases. YOLOv7 detects the roads in either of the cases.

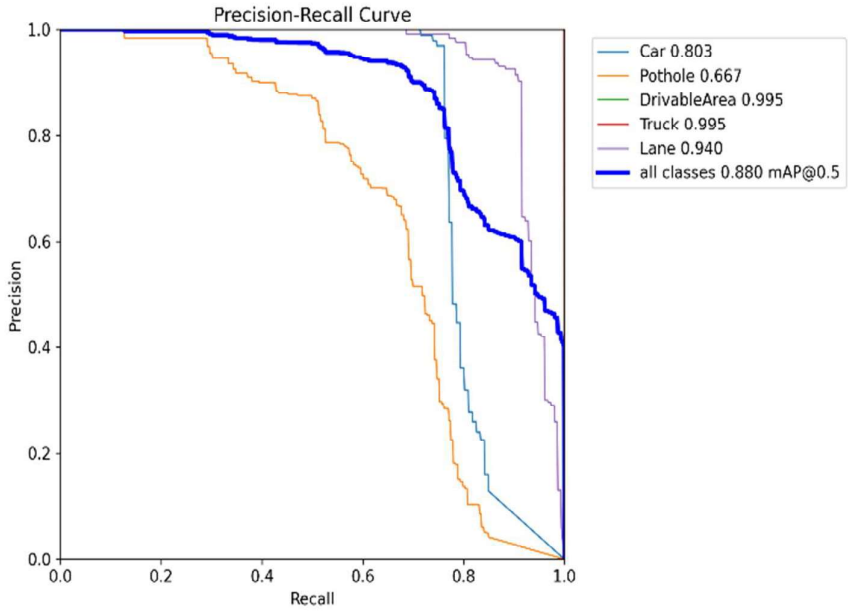


Fig 5.2 Precision-Recall Curve of YOLOv7

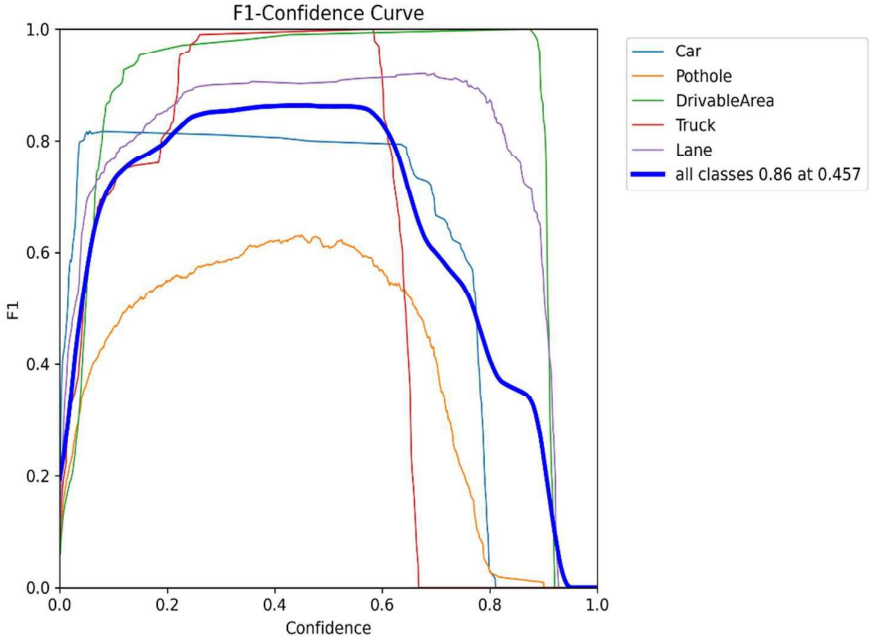


Fig. 5.3 F1-Confidence Curve of YOLOv7

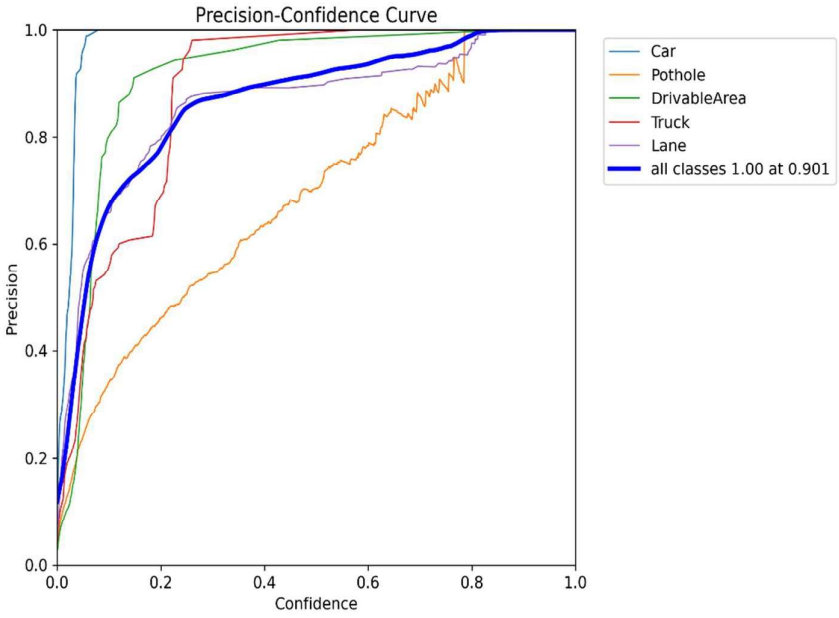


Fig. 5.4 Precision-Confidence Curve of YOLOv7

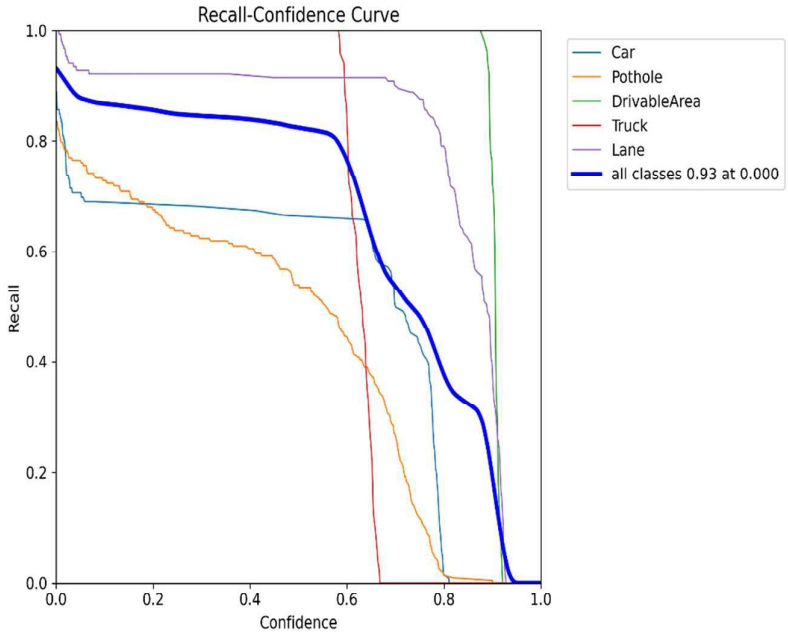


Fig. 5.5 Recall-Confidence Curve of YOLOv7

Fig 5.6 Confusion Matrix

Figure 5.6 shows the resulting confusion matrix of the proposed model. The outcomes of vehicle detection for both the proposed method and HybridNets have been succinctly presented in Table 1. The comparative evaluation of these models has been conducted based on key performance metrics such as mean Average Precision (mAP), Recall, and F1-score. The findings indicate that YOLOv7 exhibits superior performance over HybridNets, even when accounting for the inclusion of pothole detection, with a significant margin of 20% in terms of mAP.

Table I. Comparison between HybridNets and YOLOv7-seg

Model	Speed (FPS)	mAP@ 50%	Recall (%)	F1-Score
HybridNets	55	67.8	87.8	78.6
Proposed Model	160	88	93	87

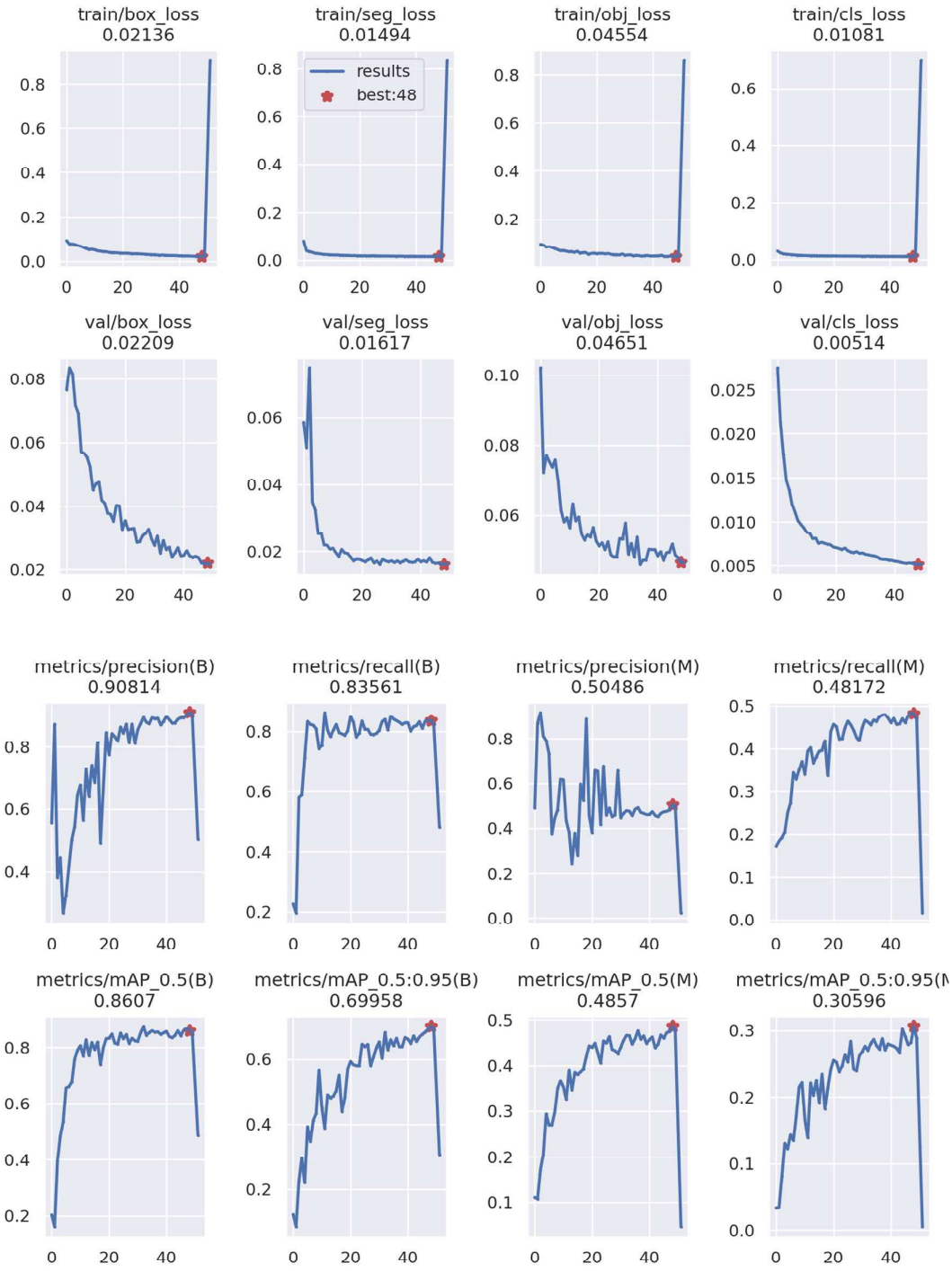


Fig 5.7 Performance metrics plots.

Additionally, YOLOv7 showcases a substantially higher processing speed of 160 frames per second (FPS) in contrast to the 55 FPS offered by HybridNets, thus excelling in both swiftness and accuracy. This notable achievement can be attributed to the

enhanced bag of freebies integrated into the proposed method, enabling a harmonious enhancement in both speed and precision without compromising either aspect. Fig 5.7 shows the performance metrics plots for YOLOv7.

The superior performance of YOLOv7 can be attributed to the utilization of E-ELAN, a technique that harnesses diverse cardinalities and group convolutions to achieve a comprehensive expansion. This expansion results in the acquisition of a wider range of varied and refined features, ultimately enhancing the model's capability. Furthermore, the proposed method implements a dynamic label assignment process in its head, leveraging soft labels that combine both ground truth and prediction outcomes. By incorporating this strategic approach, the model becomes adept at capturing and learning residual information from the image, thereby amplifying its capacity to extract meaningful insights and optimize overall performance.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

Autonomous vehicle, also referred to as self-driving vehicles, represent a paradigm-shifting advancement in the realm of transportation, characterized by their ability to operate independently without any human intervention. Such vehicles require the usage of drivable area detection systems. In this dissertation, a YOLOv7 oriented framework using detection and segmentation was proposed to detect the safe-drivable area effectively and efficiently. The presence of numerous potholes on Indian roads is a serious problem because they increase the likelihood of accidents. Pothole detection is hence a must for drivable area detecting systems. Nonetheless, it can be challenging to tell the difference between the roadways and the potholes.

HybridNets, being an advanced drivable area identification model, is limited to the tasks of lane-line segmentation and drivable region recognition, lacking the ability to detect potholes due to this limitation. To address this gap, the proposed approach suggests incorporating instance segmentation of road scenes using YOLOv7, which utilizes the E-ELAN layer in its backbone to facilitate the learning of more diverse and improved features. The utilization of various cardinalities and group convolutions in the E-ELAN layer promotes expansion and enhances the model's ability to acquire a wider range of features.

Furthermore, the proposed method employs a dynamic label assignment process in the head of the model, allowing for the learning of residual information from the image. Additionally, the introduction of a trainable bag of freebies contributes to improved performance metrics without requiring excessive training costs.

Through extensive experimentation on our dataset, the proposed method achieves an impressive mean Average Precision (mAP) of 87%, outperforming the state-of-the-art HybridNets model which achieves an mAP of 67.8%. This demonstrates the effectiveness and superiority of the proposed approach in drivable area identification.

In the future scope, the detection system is to be extended to include vehicle tracking, to make the system more advanced.

REFERENCES

[1] R. Chai, A. Tsourdos, A. Savvaris, S. Chai, Y. Xia, and C. P. Chen, "Multiobjective overtaking maneuver planning for autonomous ground vehicles," IEEE transactions on cybernetics, vol. 51, no. 8, pp. 4035– 4049, 2020.

[2] T. Liu, Q. hai Liao, L. Gan, F. Ma, J. Cheng, X. Xie, Z. Wang, Y. Chen, Y. Zhu, S. Zhang, et al., "The role of the hercules autonomous vehicle during the covid-19 pandemic: An autonomous logistic vehicle for contactless goods transportation," IEEE Robotics & Automation Magazine, vol. 28, no. 1, pp. 48–58, 2021.

[3] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022.

[4] S. Zhou, Y. Jiang, J. Xi, J. Gong, G. Xiong, and H. Chen, "A novel lane detection based on geometrical model and gabor filter," in 2010 IEEE Intelligent Vehicles Symposium, pp. 59–64, IEEE, 2010.

[5] M. Aly, "Real time detection of lane markers in urban streets," in 2008 IEEE intelligent vehicles symposium, pp. 7–12, IEEE, 2008.

[6] J. C. McCall and M. M. Trivedi, "Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation," IEEE transactions on intelligent transportation systems, vol. 7, no. 1, pp. 20–37, 2006.

[7] J. Hur, S.-N. Kang, and S.-W. Seo, "Multi-lane detection in urban driving environments using conditional random fields," in 2013 IEEE Intelligent vehicles symposium (IV), pp. 1297–1302, IEEE, 2013.

[8] A. Borkar, M. Hayes, and M. T. Smith, "Polar randomized hough transform for lane detection using loose constraints of parallel lines," in 2011 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1037–1040, IEEE, 2011.

[9] L. O. Chua and T. Roska, "The cnn paradigm," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 40, no. 3, pp. 147–156, 1993.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587, 2014.

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

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

[13] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125, 2017.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37, Springer, 2016.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once:Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, 2016.

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

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

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440, 2015.

[19] S. Arjapure and D. Kalbande, "Deep learning model for pothole detection and area computation," in 2021 International Conference on Communication information and Computing Technology (ICCICT), pp. 1–6, IEEE, 2021.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241, Springer, 2015.

[21] Z. Wang, W. Ren, and Q. Qiu, "Lanenet: Real-time lane detection networks for autonomous driving," arXiv preprint arXiv:1807.01726, 2018.

[22] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial cnn for traffic scene understanding," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE. transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[24] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint arXiv:1606.02147, 2016.

[25] Q. Zou, H. Jiang, Q. Dai, Y. Yue, L. Chen, and Q. Wang, “Robust lane detection from continuous driving scenes using deep neural networks,” IEEE transactions on vehicular technology, vol. 69, no. 1, pp. 41–54, 2019.

[26] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[27] Y. Qian, J. M. Dolan, and M. Yang, “Dlt-net: Joint detection of drivable areas, lane lines, and traffic objects,” IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 11, pp. 4670–4679, 2019.

[28] H. Wang, R. Fan, Y. Sun, and M. Liu, “Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms,” IEEE transactions on cybernetics, vol. 52, no. 10, pp. 10750–10760, 2021.

[29] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, “Yolop: You only look once for panoptic driving perception,” Machine Intelligence Research, vol. 19, no. 6, pp. 550–562, 2022.

[30] D. Vu, B. Ngo, and H. Phan, “Hybridnets: End-to-end perception network,” arXiv preprint arXiv:2203.09035, 2022.

[31] M. Tan, Q.V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” arXiv:1905.11946, 2020.

[32] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10778-10787

[33] J.B. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, 281-297, 1967.

[34] Jocher, Glenn, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang et al. "ultralytics/yolov5: V7. 0-YOLOv5 SOTA realtime instance segmentation." Zenodo, 2022.

[35] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W. and Li, Y., 2022. "YOLOv6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

[36] Rahman, M. A., & Wang, Y., "Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation". Lecture Notes in Computer Science, 234–244, 2016.

[37] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2017, pp. 5108–5115.

LIST OF PUBLICATIONS

[1] D. Sharma, S. Indu, N. Jayanthi, "A Review on Recent Encoder-Decoder Architectures for Drivable Region Detection and Segmentation", 5th IEEE International Conference on Energy, Power and Environment (ICEPE), 2023.

[2] D. Sharma, S. Indu, N. Jayanthi, "Lane-Line Segmentation Using YOLOv7", 2nd IEEE conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), 2023.

PAPER NAME

end1_merged.pdf

WORD COUNT	CHARACTER COUNT
8803 Words	**53367 Characters**
PAGE COUNT	FILE SIZE
50 Pages	**2.8MB**
SUBMISSION DATE	REPORT DATE
May 31, 2023 5:00 PM GMT+5:30	**May 31, 2023 5:01 PM GMT+5:30**

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- Crossref database
- 12% Submitted Works database
- 5% Publications database
- Crossref Posted Content database

Excluded from Similarity Report

- Bibliographic material