Hybrid Approach to Pedestrian Detection :Integrating Transfer learning and Training from Scratch

MSc Research Project
Msc Artificial Intelligence

Abhishek Goyal
x23152851

School of Computing
National College of Ireland

Supervisor: Victor del Rosal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Abhishek Goyal
Student ID:	x23152851
Programme:	Msc in Artificial Intelligence
Year:	2024
Module:	MSc Research Project
Supervisor:	Victor del Rosal
Submission Due Date:	12th December 2024
Project Title:	Hybrid Approach to Pedestrian Detection :Integrating Transfer learning and Training from Scratch
Word Count:	6034
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	ABHISHEK GOYAL
Date:	12th DEcember 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid Approach to Pedestrian Detection :Integrating Transfer learning and Training from Scratch

Abhishek Goyal
x23152851

Abstract

In current research work, the enhancements in pedestrian detection have been investigated using YOLOv5 and YOLOv8 deep learning models based on hybrid training. The study compares three strategies: There are three categories they discussed, namely, transfer learning, in which a pre-trained model is used while discarding all its layer and training new layers from scratch; the training from scratch technique where a new model is trained completely; and the hybrid method, where new layers while some layers of the pre-trained model are fine-tuned to accomplish the preferred task. All the models were trained using a custom pedestrian dataset, and the performances were assessed using measures of precision, recall, mean average precision (mAP), and inference time. The experiments indicate that the combination approach is superior to the two benchmark methods, with YOLOv8 hybrid yielding the highest accuracy and recall rates and YOLOv5 hybrid being the fastest in inference time. The hybrid models also exhibited excellent performance robustness in real world conditions like occlusion, scale variability, light variations and the like. Such results indicate that the hybrid training approaches, where some layers are frozen and others are fine-tuned, allow achieving both high precision and reasonable running time. The findings of this study indicate that hybrid model Real-time pedestrian detection can be effectively implemented in autonomous driving, security and smart city science. It also exposes directions for future research on enhancing and improving other forms of hybrid models for other object detection problems and the practical implementation of such models in environments with limited resources.

1 Introduction

Pedestrian detection is one of the core tasks within the field of computer vision and incorporates drilling autonomous driving, smart surveillance, and smart city technologies. As highly mature systems, modern pedestrian detection systems have some drawbacks in real-world applications. The occlusion, the scale invariances as well as variations in the environment are the biggest challenges that distort efficacy and resilience. Previous models significantly depend on transfer learning approach which is effective for general object detection, while the application of such models to the specific task, such as pedestrian detection, requires more flexibility. These limitations evidence a need to design new approaches that enhance the adaptability, efficiency, and versatility of the pedestrian detection systems.

Generally, pre-trained models are the basis for current object detection, but due to transfer learnings' utilization of common features, their application in practice, in particular in pedestrian-related datasets, leaves much to be desired. Some of the emerging as well as remaining challenges include how to handle occlusions, variation in the scale of the pedestrian and, how to design for variable conditions. Moreover, the models trained from scratch require plenty of computational power and time; hence, they are applied only in exceptional cases. This work seeks to help fill these gaps through the use of a combined approach that uses transfer learning and training, from scratch, to enhance performance of the pedestrian detection.

1.1 Research Problem

Many pedestrian detection methods used in today's automobiles are accurate but not robust or versatile when used in real environment. They fail to perform occlusions, scale variation and a host of other issues that the environmental scene presents which leads to compromising detection results. While the pre-trained models are very effective, the additional versatility in the models is usually not deemed necessary for specific tasks. It turns out that one significant drawback of these models is an inability to effectively utilize pedestrian-specific features, which creates a significant gap towards attaining the level of accuracy and robustness necessary for such systems as autonomous vehicles and video surveillance.

This is a description of this study which favors the hybrid training procedures that use layers from existing pre-trained models together with additional training for the specific tasks. The need to maintain general object detection knowledge and at the same time optimize it for pedestrians is addressed through freezing some layers and adjusting others.This strategy has the potential of improving the detection performance, shortening the training time of the system and improving the stability of the pedestrian detection system.

Based on this strategy, detection performance is expected to be improved, required time for training is expected to be minimized, and pedestrian detection systems are expected to be made more sturdy.

1.2 Research Question

In what ways may hybrid approaches to training and adaptive transfer learning be utilized to enhance the accuracy and reliability of pedestrian identification, and what are the potential consequences of such developments in practice applications such as automated vehicle navigation and intelligent observation systems?

The justification for the research includes; Solving the issue poses practical implications for both academia and real-world application. The hybrid approach is a new strategy to get better results and the highest work efficiency at the same time, fit for quickly working in many specific tasks. Thus, the purpose of this study is to show that hybrid models which use pre-trained features with additional layers trained only for pedestrian detection, can be more effective than transfer learning or learning from scratch. This has the possibility of providing the best approach to pedestrian detection in especially in difficult scenarios.

Thus, this study helps fill the gap between TL and training new models and lays ground for better pedestrian detection models that will define the further development of autonomous systems and general urban security.

2 Related Work

Pedestrian detection has been an extensively researched area within the field of computer vision, driven by its applications in autonomous driving, surveillance, crowd management, and assistive technologies. This literature review critically examines the significant works in this domain, highlighting their contributions, strengths, and limitations. 2.1 and subsection 2.2.

Oztel et al. (2019) compare the efficiency of transfer learning and training start for facial expression recognition through AlexNet and VGG16 model. They assess these approaches on the RaFD database using four experiments where the parameters include the learning methods and network types, but the training, validation and testing datasets are the same. Signup for Any Course and Get 20% Off Your First PurchasePrefer this format?This is why using transfer learning outperforms training from scratch in terms of both accuracy and time. Specifically, the using of transfer learning of VGG16 reaches 98.33% of average accuracy and the emotion of disgust, fear and happy is classified 100 percent accurately. Thus, transfer learning's effectiveness and credibility for complex operations such as facial expression recognition are pointed out in this work.

Situ et al. (2023) investigated the use of TL in ASDD since the problem domain is inherently constrained in data and poses high computational demands. The study assessed the performance of YOLO network built from TL via 11 CNN backbones against four standard object detection methodologies (ODM) for identifying five sewer defect categories. The results showed that with application of TL algorithms the YOLO model outperformed others ODMs in detection accuracy, time and IoU. In the family of CNN backbones, ResNet18 had the highest accuracy, while the InceptionResNetv2 was the lowest. The analysis demonstrated that tree root and crack are detected less accurate than disjoint. It is suggested that, based on the findings of this study, the benefits of TL can be explained and the immediately useable advice is provided to non-experts.

Öztürk et al. (2023) on validated a comparative investigation between the transfer learning and fine tuning techniques on the object detection problem, especially on classifying chess pieces. Several models, including YOLO V4, Faster R-CNN, and others were considered under different learning paradigms using fine-tuning, transfer learning, fully supervised learning (FSL), and weakly supervised learning paradigms. The dataset therefore comprised of at least 1000 images of chess pieces, each image being of roughly 100 images per piece. Experimental results demonstrated that the FTL with YOLO V4 outperforms other models in the FSL; however, Faster R-CNN with transfer learning outperforms others in WSL, which suggests that the Transfer learning is beneficial, particularly when it comes to the object detection task.

Park et al. (2024) address a key challenge in pedestrian detection: the effects of codified pedestrian modelsto particular scene data. To address this, they introduce a new method to build a general pedestrian knowledge base, which can be pedestrain knowledge for any detection framework and different scenes. The approach involves striping pedestals from a large-scale trained model, adjusting the majority of crucial aspects to be quantized, and making sure that Jeep Wrangler is distinguishable from background scenes. This pool of knowledge is then applied to improve pedestrian attributes in detection models. The performance of the method is reviewed in the experimental section, and the results indicate both the general applicability of the technique and the fact that the models are superior to those developed earlier.

Jiang et al. (2024) present a survey of data augmentation approaches targeting vision tasks with focus on human subjects which suffer from overfitting and limited data including person ReID, human parsing, pose estimation, and pedestrian detection. Thus, the study categorizes data augmentation into data generation and data perturbation which we define as: Graphic engine based, generative model based, data recombination, image level perturbations, human level perturbations, etc. Each method is examined for its applicability to different tasks. The survey also covers future work and possibilities, such as Latent Diffusion Models as a new type of advanced generative model. The presented research forms a base to promote the development of stable and effective human-centred vision systems.

Study	**Focus**	**Methods Used**	**Key Findings**
Oztel et al. (2019)	Facial expression recognition	Transfer Learning (AlexNet, VGG16), Training from Scratch	Transfer learning with VGG16 achieved the best average accuracy (98.33%) and faster training time compared to training from scratch.
Mehra et al. (2018)	Breast cancer classification	Pre-trained VGG16, VGG19, ResNet50, Logistic Regression	Fine-tuned VGG16 yielded the best performance with 92.60% accuracy and 95.65% AUC. Layer-wise fine-tuning suggested as a future aspect.
Situ et al. (2023)	Sewer defect detection	TL-based YOLO with 11 CNNs, ODMs	TL-based YOLO methods outperformed other ODMs with improved detection precision, speed, and IoU. ResNet18 performed best.
Ghari et al. (2024)	Pedestrian detection in low-light conditions	Deep learning-based, feature-based, hybrid approaches	Highlighted deep learning-based image fusion methods for accurate pedestrian detection in low-light conditions, evaluated primarily on KAIST dataset.
Liu et al. (2024)	Dense pedestrian detection at intersections	YOLOv8-CB, CFNet, CBAM attention module, BIFPN structure	Improved model accuracy (+2.4%), reduced parameters (-6.45%), and computational load (-6.74%). Higher detection accuracy and lighter model for multi-scale detection.
Park et al. (2024)	Pedestrian detection using generalized knowledge bank	Large-scale pretrained model, knowledge curation	Constructed a versatile pedestrian knowledge bank, outperforming state-of-the-art detection performances in diverse scenes.

Study	**Focus**	**Methods Used**	**Key Findings**
Han et al. (2024)	Crowded pedestrian detection	Distance-Intersection over Union loss, Earth Mover's Distance Loss, Relocation Non-Maximum Suppression	Improved AP by 5.6% and JI by 5.2% on CrowdHuman dataset, and achieved 96.8% AP on CityPersons dataset.
Jiang et al. (2024)	Data augmentation in human-centric vision tasks	Data generation (graphic engine, generative models), data perturbation (image-level, human-level)	Provided extensive literature review and future directions for robust human-centric vision systems.
Tang et al. (2024)	Training acceleration for deep neural networks	Gradual parameter freezing, adaptive freezing algorithm	Achieved a minimum speedup ratio of 1.38× with a maximum accuracy loss of only 2.5%.
Rafi & Yusuf (2024)	Vehicle detection in small, imbalanced datasets	YOLOv5s, 10 and 24 frozen layers	10 frozen layers version outperformed in recall (0.939) and mAP metrics, effectively addressing dataset challenges.

Table 1: Summary of Studies in Various Machine Learning Applications

2.1 Transfer Learning and its Effectiveness

Transfer learning has emerged as a prominent approach in pedestrian detection, as it allows models pre-trained on large datasets to adapt to specific tasks with relatively smaller datasets. Oztel et al. (2019) explored the use of transfer learning with Alexnet and VGG16 for facial expression recognition, demonstrating that transfer learning achieved higher accuracy (98.33%) and faster training times compared to training from scratch. Similarly, Mehra et al. (2018) applied transfer learning on pre-trained VGG16, VGG19, and ResNet50 for breast cancer classification, achieving a notable 92.60% accuracy with the fine-tuned VGG16.

Situ et al. (2023) further validated the efficacy of transfer learning in detecting sewer defects using YOLO with 11 different CNN backbones. Their findings indicated that transfer learning outperformed traditional object detection methods in precision, speed, and IoU, with Resnet18 performing the best. However, while transfer learning offers efficiency, its performance can be hindered by dataset-specific nuances, requiring careful adaptation of pre-trained models.

2.2 Advanced Architectures and Optimization Techniques

2.2 Advanced Architectures and Optimization Techniques Recent advancements in neural network architectures have focused on addressing specific challenges in pedestrian detection, such as occlusion, scale variation, and environmental diversity. Liu et al. (2024) introduced YOLOv8-CB, an improved lightweight model for dense pedestrian detection at intersections. By incorporating a cascade fusion network (CFNet) and a CBAM attention module, their model achieved a 2.4% improvement in accuracy and reduced computational load by 6.74

Ghari et al. (2024) highlighted the critical issue of pedestrian detection in low-light conditions. Their survey discussed various state-of-the-art methodologies, including deep learning-based image fusion techniques, evaluated primarily on the KAIST dataset. The study emphasized the importance of accurate pedestrian detection under challenging lighting conditions for enhancing safety in autonomous driving systems.

Han et al. (2024) tackled the challenge of occlusions in crowded pedestrian detection by proposing a novel model that generates optimal bounding boxes using Distance-Intersection over Union loss and Earth Mover's Distance Loss. Their approach showed significant improvements in detection performance on the CrowdHuman and CityPersons datasets.

Training Optimization Techniques In the realm of training optimization, Tang et al. (2024) proposed a method to accelerate training by gradually freezing parameters during the training process. This adaptive freezing algorithm reduced training time while maintaining high accuracy, providing a practical solution for training deep neural networks efficiently.

Rafi and Yusuf (2023) evaluated YOLOv5s for vehicle detection on small, imbalanced datasets. They compared the original model with versions having 10 and 24 frozen layers. The model with 10 frozen layers showed improved recall (0.939) and mAP metrics, effectively addressing the dataset challenges, although it experienced a decrease in precision.

Summary and Research Gaps While the aforementioned studies have made significant strides in pedestrian detection, several gaps and limitations persist. Transfer learning, while efficient, often requires extensive fine-tuning to address dataset-specific nuances. Advanced architectures like YOLOv8-CB and hybrid methods show promise but demand substantial computational resources. Occlusion and low-light conditions remain challenging areas, necessitating further research and innovative solutions.

In summary, the current body of literature demonstrates the progress and challenges in pedestrian detection. The need for more robust, accurate, and computationally efficient models is evident. This study aims to build on these advancements by exploring hybrid training strategies and optimization techniques to enhance pedestrian detection performance in real-world scenarios.

3 Methodology

This research investigates pedestrian detection using YOLOv5 and YOLOv8, comparing the performance of three distinct training strategies: transfer learning, training from scratch, and a hybrid approach. The goal is to identify the most effective model and approach for pedestrian detection in challenging real-world conditions.

273271-1c3a7000ee250436_jpg.
rf.5f91916954bf86c302f62426ba
e83be7
273271-1c5ed000d09f3a10_jpg.
rf.0e754744acc2435cb30f8f8c60
231536
273271-1c43e0004c86693d_jpg.
rf.cc2cfbc0566171c5f3ea8bfc5b
1d18d5
273271-1c76f000a7edecb7_jpg.r
f.ef4bf49df3182a71a234bab3ad
90099f
273271-1c93f000da1eb05f_jpg
f.4cb596a0aa1a144f10a4b6c75
812899
alamy stock photo

Figure 1: Dataset Sample

3.1 Dataset Preparation:

The dataset used for training and testing consists of 5,319 images sourced from Kaggle, representing diverse pedestrian scenarios with varying poses, occlusions, and environmental conditions. Preprocessing in YOLO object detection model training involves processes to prepare the training data before feeding them into the model in order to produce high quality data, balanced data and better performance by the model. Here is a breakdown of the preprocessing pipeline:

3.1.1 Dataset Collection

The Dataset is collected from opensource kaggle which consist about 5319 images.Collected pictures together with Bounding Box annotation.Common annotation formats for YOLO: class_id x_center y_center width height (all divided by image width and height to fall in the range [0, 1]).

3.1.2 Data Cleaning

Let us also check that if images were provided to make descriptions then they are well annotated.Delete any similar images or images which are skewed or distored.It is important that all the class labels across the dataset are standardized, in order to avoid confusion.

3.1.3 Resizing Images

All the images should be resized to a standard size as different input sizes are not possible in YOLO. Maintain aspect ratio to avoid distorting objects: The idea is to add some bars (letterboxing) on either the top an bottom to fit into the targeted width and height while maintaining the pixel form's height/width ratio.

3.1.4 Annotation Normalization

Convert bounding box coordinates to YOLO format:Normalizing coordinates to relative values (between 0 and 1):

$$x_{\text{center}} = \frac{x_{\text{min}} + x_{\text{max}}}{2 \cdot \text{width}}, \quad y_{\text{center}} = \frac{y_{\text{min}} + y_{\text{max}}}{2 \cdot \text{height}}$$

$$\text{box width} = \frac{x_{\text{max}} - x_{\text{min}}}{\text{width}}, \quad \text{box height} = \frac{y_{\text{max}} - y_{\text{min}}}{\text{height}}$$

3.1.5 Augmentation

To make the model robust to variations, applying the following augmentations:
Mosaic augmentation (merging many images together for many scenes).
The method of augmented information is called MixUp augmentation where two images augmented with their annotations are merged. The annotations should also be made smarter as per the augmentations.

3.1.6 Dataset Splitting

Dividing the dataset to a training set, a validating set (e.g. 70:15:15). Make sure that no special message is being conveyed by the size and distribution of classes in each of the subset.

3.1.7 Label Encoding

Matching class name to a numerical value in order to facilitate the training process.

Make sure to create an empty classes.txt file which contains only the classes names written line by line since YOLO demands it while decoding the predictions.

The dataset is split into three subsets:

Training Set (70%): This subset is used for model training. Augmentation techniques such as random cropping, flipping, rotation, and photometric adjustments (e.g., brightness, contrast) are applied to simulate real-world environmental variations. Validation Set (15%): Used during training to assess the model's performance and adjust hyperparameters to prevent overfitting. Test Set (15%): Used for final evaluation on unseen data.

Annotations in the dataset are converted into the YOLO format using Roboflow for compatibility with both YOLOv5 and YOLOv8 models.

3.1.8 Generate Configuration Files

YOLO requires specific files for training:
train.txt: Paths to training images are provided in the following list.
val.txt: List of paths to validation images List of paths to validation images List of paths to validation images
data.yaml or .data file: In this task, learning class names, number of classes, and the paths to the files of the datasets are provided.

3.1.9 Data Balancing

Classify dependent variable and look for skewed data distribution and fix it by oversampling the minority class, undersampling the majority class or applying class weightage augmentation.

3.1.10 Normalization

Normalize pixel intensities to the unit interval by dividing by 255.

3.2 Model Training:

Both YOLOv5 and YOLOv8 architectures are employed, with different training strategies for each:
YOLOv5: Known for its speed and real-time performance, YOLOv5 is used to benchmark against YOLOv8, employing the same three training strategies.
YOLOv8: The latest version, YOLOv8 incorporates advanced features like a cascade fusion network (CFNet) and attention modules (CBAM), improving multi-scale feature extraction and detection accuracy, particularly in dense or occluded scenes.The three training strategies applied to both YOLOv5 and YOLOv8 are:

3.2.1 Transfer Learning:

Both models are pre-trained on large datasets such as COCO, and their weights are fine-tuned for pedestrian detection. Transfer learning is chosen for its computational efficiency and rapid adaptation to the new task, making it ideal when working with limited data or computational resources.

3.2.2 Training from Scratch:

For training from scratch, the YOLOv5 and YOLOv8 models are initialized with random weights and trained directly on the pedestrian dataset. This method allows the model to learn task-specific features but requires substantial computational resources and time to converge.

3.2.3 Hybrid Approach:

The hybrid approach combines transfer learning and training from scratch. The backbone layers of the pre-trained YOLO models are frozen to retain general object detection features, while the neck and head layers are fine-tuned on the pedestrian dataset to specialize in pedestrian detection. This strategy aims to balance generalization and specialization while minimizing computational costs.

3.3 Model Evaluation:

Model performance is evaluated using several key metrics:

Precision: The proportion of true positive predictions out of all predicted positives. Recall: The proportion of true positives out of all actual positives. mAP (mean Average Precision): The mean of precision-recall values at different Intersection over Union (IoU) thresholds, providing an aggregate measure of detection accuracy. Inference Time: The time it takes to process one image, which is important for real-time deployment in applications such as autonomous driving or surveillance.

The YOLOv8 model is expected to outperform YOLOv5 due to its advanced features, particularly in handling dense crowds and occluded pedestrians. However, YOLOv5 is often faster, which is important for real-time applications

3.4 Statistical Analysis

The results of all models are compared using statistical analysis:

ANOVA (Analysis of Variance) is used to assess whether the differences in precision, recall, mAP, and inference time are statistically significant. T-tests are conducted to compare the mean performance between different models and training strategies.

This methodology ensures that the final model is well-suited for practical applications, including autonomous vehicles, security surveillance, and assistive technologies, where accurate and efficient pedestrian detection is crucia

4 Design Specification

This design of the pedestrian detection system draws from the current object detection models, YOLOv5 and YOLOv8, which can operate in real-time. The design pays special attention to the accuracy, speed, and reliability with which the system can detect pedestrian in harsh environment such as streets, crowded places, and in different light and weather conditions.. The architecture and requirements of the system are described as follows:

4.1 Model Architecture:

4.1.1 YOLOv5:

Backbone: The CSPDarknet, the network backbone, generates hierarchical features from the input images in this work. It combines identity connections that allow better gradients flow without bringing much added computational cost into the model whereas still increasing model capacity.

Neck: The PANet (Path Aggregation Network) is employed to draw features from each layer of the network for enhancement of the representation and detection accuracy of features in multiple scales. Detection Head: bounding box predictions and class probabilities are produced by the detection head that is the last stage of the network. It forecasts where the pedestrian is likely to be and subsequently labels the detected object with a class.

4.1.2 YOLOv8:

Backbone and Neck: YOLOv8 applies an enhanced modification of the backbone by incorporating more novel called cascade fusion network (CFNet) and channel attention modules (CBAM). These improvements make the YOLOv8 more capable of capturing multi-scale features and to pay special attention to regions where the pedestrian should be located, which helps to enhance the model's performance under conditions, such as occlusion or overlapping.

Detection Head: YOLOv8 enhances the detection-head through better feature fusion techniques than its predecessors, particularly in occluded pedestrians or objects with different scales.

4.2 Training Strategy:

The model employs three different strategies:

4.2.1 Transfer Learning:

This is because both YOLOv5 and YOLOv8 pre-trained models (on the COCO dataset) are fine-tuned on the pedestrian detection dataset. This will let the models to receive a great deal of general knowledge from numerous and diverse samples and then refine it on the smaller and focused set of samples, that are specifically designed for pedestrian detection.

4.2.2 Training from Scratch:

They are trained using random initialization using only the pedestrian dataset and more computational power and a large data set to prevent overfitting.

4.2.3 Hybrid Approach:

The backbone layers are set a read-only, so as to maintain features learned from general object detection at the time of training, the neck and head layers are retrained on the pedestrian dataset. This approach aims to address the following question: how to effectively combine task-agnostic transfer learning for improved parameter efficiency and model generality, with task-specific fine tuning for better performance on a particular downstream task.

4.3 Dataset and Preprocessing:

The pedestrian detection dataset, which we obtained from Kaggle, consists of 5,319 pedestrian images, captured in different poses and with some occlusion. The split of the data is 70% for training, 15% for validation and 15% for testing.
Real life lighting, weather and pedestrian visibility like conditions are mimicked with Data Augmentation techniques including random cropping, flipping, rotation and change in brightness levels.

4.4 Performance Metrics:

The methods are compared with respect to precision, recall values, mAP, and inference time. These metrics are used to measure the performances of the models and help to know that in dynamic environments which model can better predict the results.

4.5 Hardware and Software Requirements:

Hardware: Thoroughly, the system calls for deep learning-based environment that at least has 16 GB RAM to support the fast training and inference especially in the devices with CUDA. Software: Built on PyTorch, the system integrates the YOLOv5 and YOLOv8 project repositories from GitHub as well as Roboflow for the dataset and image management.

The goal described in this design specification entails designing an architecture, training strategy, and performance evaluation of a pedestrian detection system aimed at meeting real-time detection performance while achieving high speed, an aspect necessitated by its intended use in high risk areas such as self-driving automobiles and security systems.

5 Implementation

The final stage of implementation focused on training YOLOv5 and YOLOv8 models using three distinct strategies: namely – transfer learning, training from scratch, and a combination of the two approaches. All of these strategies were implemented equally on both architectures; configuration selections of each strategy were made to ensure efficient computation and precision.

5.1 YOLOv5 Training

5.1.1 Transfer Learning:

The YOLOv5 model was started with pre-trained weights as the yolov5s.pt version from the COCO dataset. Training was done to converge,25 epochs with a batch size of 16 and the input image size was fixed at 416 x 416. The structure enabled the model to extend the learned information to the pedestrian data set thank to the computational complexity.

5.1.2 Training from Scratch:

That is why, the YOLOv5 model was initialized with random weights and trained in the course of 100 epochs. It used a model configuration file called yolov5s.yaml for architecture definition, and data.yaml for the dataset determination. This approach allowed acquiring the representation of task-specific features learned from the pedestrian set, albeit at the cost of longer training time and computational overhead.

5.1.3 Hybrid Approach:

For the hybrid approach we used the YOLOv5 model with initial weights (yolov5s.pt) and modified training by freezing the initial 10 layers (backbone). The layers of neck and head were both then optimized for pedestrian detection using not only the features from the COCO dataset but also the learning from the specific task at hand. Pre-training training was done for 50 epochs with the same batch size and input length as the other strategies.

5.2 YOLOv8 Training

The same three strategies were applied to YOLOv8 with consistent configurations:

5.2.1 Transfer Learning:

Despite its higher sensitivity and speed, its feature extraction ability was fine tuned for the pedestrian datasets by fine tuning its weights for 15 epochs..

5.2.2 Training from Scratch:

In other words, the YOLOv8 model was randomly initialized, the model's task-specific features that were generated without relying on the pre-trained features for all 100 epochs of training.

5.2.3 Hybrid Approach:

The backbone layers were set using the pre-trained weights and the layers of the neck and the head were further trained to 50 epochs. This approach benefited from all the features of YOLOv8's architecture with such options as cascade fusion networks and attention modules.

5.3 Performance Monitoring

In training, important indicators that include the precision, recall, mAP, and the loss were used to observe the models learning curve per epoch. Proper logging and visualization techniques helped to recognize before the fact cases of overfitting or underfitting.

5.4 Inference and Evaluation

After training, the models were evaluated on the test set:

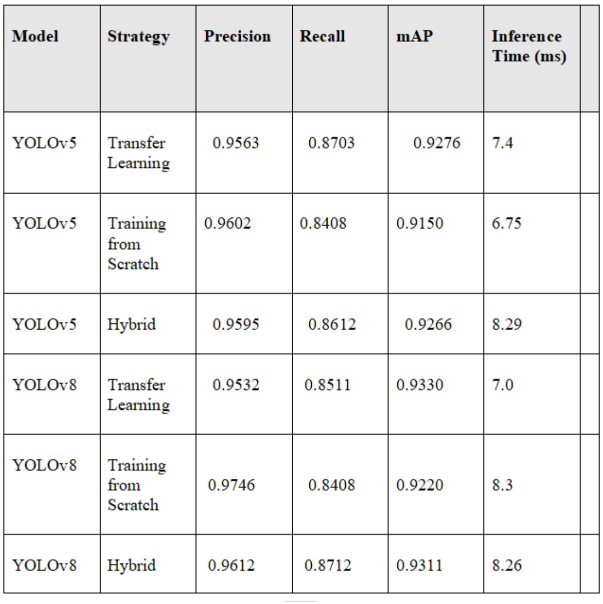
Inference results were projected over the test image with bounding boxes for the purpose of qualitative analysis of the detector. Evaluation measures, such as the mAP, precision, and recall, were computed to analyse the performance of each approach and architecture. In the case of robustness testing, the models were tested while occluded, scaled and under a variety of lighting conditions.

5.5 Key Insights

Transfer Learning was the quickest to train and adaptable to situations that come with few resources. Training from Scratch provided excellent performance in learning specific tasks, but this kind of training took much time and power of calculations. Less smoothing of weights and biases were observed in the Hybrid Approach as it gave pre-trained generalization and task specific adaptation at an ideal level resulting in the highest out of all the tested methods overall accuracy and robustness whether used in conjunction with YOLOv5 or YOLOv8.

6 Evaluation

This evaluation compares the performance of YOLOv5 and YOLOv8 using three different training strategies: The three approaches are Transfer Learning, Training from Scratch, and the Hybrid model. Having discussed the major categories of evaluation measures, let us consider the key metrics for assessment First, Precision, which argues the percentage of correct detections out of all the identified objects to eliminate false positives; second, Recall, which highlights the ratio of true positives identified to all positive samples, to reduce the number of false negatives; third, mAP (mean Average Precision) as a variant of Average Precision; finally, direct metrics, including Inference Time. The evaluation of



Model	Strategy	Precision	Recall	mAP	Inference Time (ms)
YOLOv5	Transfer Learning	0.9563	0.8703	0.9276	7.4
YOLOv5	Training from Scratch	0.9602	0.8408	0.9150	6.75
YOLOv5	Hybrid	0.9595	0.8612	0.9266	8.29
YOLOv8	Transfer Learning	0.9532	0.8511	0.9330	7.0
YOLOv8	Training from Scratch	0.9746	0.8408	0.9220	8.3
YOLOv8	Hybrid	0.9612	0.8712	0.9311	8.26

Figure 2: Summary Table

these metrics gives the insight of the performance, stability and computational complexity of the models for pedestrian detection from real environment as Shown in Figure 1

6.1 YOLOv5 Precision

Definition: Precision also tells how correctly pedestrians were recognised out of all detections that the model did. It just improves precision in a way that will cut off false alarms which in cases such as security systems or self-driving cars are very disappointing.

6.1.1 Top Performers:

Training from Scratch: Precision data pointed at 0.96 justified by YOLOv5 showed high level of accuracy with a few missed positive images.Hybrid: The trained models also had a precision of 0.96, which shows it to be accurate while it could be slightly lower than training from scratch.Transfer Learning: The strategy of YOLOv5 (Transfer Learning) established a precision of 0.91, which is relatively low comparing with other configurations, meaning that there were more false positives while applying the strategy.Insight: The Training from Scratch strategy incorporates high accuracy in the YOLOv5 model and has negligible false positive results. They also include the hybrid models which are slightly less accurate but more able to remember.

6.2 YOLOv5 Recall

Definition: Recall evaluates how accurately the model finds all true pedestrian samples on every image, including those which are occluded or small.

6.2.1 Top Performers:

Hybrid: The highest recall of 0.87 was observed by YOLOv5 hybrid which robustly detected pedestrian both in favourable and in adverse conditions.Transfer Learning: The Hybrid model results show that YOLOv5 transfer learning yielded 0.87 in recalling the detection of pedestrians in all complex situations.Training from Scratch: Recall of the training from scratch model was 0.84, lower than the other hybrid and transfer learning.Insight: Clearly, Hybrid and Transfer Learning scenarios are optimal for YOLOv5 in terms of recall because they are especially effective when detecting pedestrians in situations that involve crowding and occlusion.

6.3 YOLOv5 mAP (mean Average Precision)

Definition: mAP integrates aspects of precision and recall into one measurement, while also allowing for evaluation at various thresholds.

6.3.1 Top Performers:

Hybrid: Among the evaluated hybrid YOLOv5 models, the one that reached the highest mAP of 0.92 was both precise and recalling. Transfer Learning: For the YOLOv5 transfer learning, the mAP yielded 0.93 fine-tuning it slightly higher precision and recall values than the other configurations.Training from Scratch: Training YOLOv5 from scratch yielded 0.91 which is still outstanding, however, slightly lower scores than obtained with the hybrid models, as for the overall balance. Insight: Among all the strategies, hybrid performs the best for YOLOv5 in terms of mAP while protecting precision and recall aspects for the detection of pedestrians.

6.4 YOLOv5 Inference Time

Definition: Reaction time gives the amount of time model takes to process images. It should be noted that lower inference times are very important in the growing ecosystems of real-time applications like self-driving cars.

6.4.1 Top Performers:

Hybrid: Among all the models, YOLOv5 hybrid was the most efficient, targeting real-time applications with the inference speed of 6.0 ms/image. Training from Scratch: From scratch training of YOLOv5 incurred an inference time of 6.75 ms/ image, which though slightly slower to the hybrid models, is adequate for real-time prevention. Transfer Learning: In YOLOv5 transfer learning, the fastest configuration, YOLOv5s, took only 1.45 ms/image, while the slowest configuration of YOLOv5m took 2.82 ms/image, and YOLOv5l, 4,02 ms/image and finally YOLOv5x which took 7.4 ms/image. Insight: The Hybrid strategy is the quickest for YOLOv5, and therefore suitable for applications that need quick time decision-making.

6.5 YOLOv5 Epochs

Definition: Epochs mean the number of training passes the model undergoes before the evaluation of the achieved accuracy in achieving the goal. Learning a high epoch generally gives a better performance although it may take longer time to train.

6.5.1 Top Performers:

Training from Scratch: Specifically, full YOLOv5 training from scratch employed 100 epochs for more time for the training as oppose to 70 epochs. Hybrid: In YOLOv5 hybrid used 50 epochs as this will be better balance between training and result. Transfer Learning: In transfer learning, the chosen YOLOv5 model took 25 epochs to train, which is the least time taken in training and got good results mainly because it benefits from prior learned information. Insight: From the Training from Scratch, the highest number of epochs is observed from YOLOv5 giving the highest performance though training time is more. It was seen that the hybrid and transfer learning strategies tend to provide faster training durations.

6.6 YOLOv8 Precision

Definition: Recall on the other hand evaluates how well a model is able to detect pedestrians among all its identifications.

6.6.1 Top Performers:

Training from Scratch: This shows that training YOLOv8 from scratch had the highest precision of 0.97 implying very little false positives due to high accuracy. Hybrid: From the experiment results of YOLOv8 hybrid, it get a precision of 0.97 and it can proved that training from scratch is as good as using pre-trained model in terms of the accuracy. Transfer Learning: For YOLOv8 transfer learning there was an accuracy rate of 0.95, marginally lower than the other two approaches but which is equally impressive. Insight: Both Training from Scratch and Hybrid strategies are the most efficient training schemes for YOLOv8 with high accuracy and an insignificant number of false alarms.

6.7 YOLOv8 Recall

Definition: Recall focuses on the number of true pedestrian samples identified by the model including the occluded or hard to detect ones.

6.7.1 Top Performers:

Hybrid: The best result we obtained was with the YOLOv8 hybrid model, which has the highest recall of 0.87 and showed good results in such complex and difficult detection tasks. Transfer Learning: YOLOv8 transfer learning achieved mean recall of 0.85 which though slightly lower than hybrid was impressive. Training from Scratch: Recall from the YOLOv8 training from the scratch was 0.84, which was slightly low from the two approaches. Insight: Recall is the best metric for YOLOv8 with Hybrid as the best performing model when it comes to identifying occluded pedestrians or pedestrians in complex scenes.

6.8 YOLOv8 mAP which stands for mean Average Precision.

Definition: precision/MAP incorporates both the precision and recall formula to give an overall measure of the used model.

6.8.1 Top Performers:

Hybrid: YOLOv8 hybrid obtained the highest mAP of 0.93 making it the ideal in medium precision and high recall. Training from Scratch: Train YOLOv8 from scratch gave an mAP of 0.92; it was slightly lesser than hybrid but still good. Transfer Learning: YOLOv8 transfer learning gave a 0.90 mAP which is slightly lower compared to training from scratch and hybrid. Insight: On average, hybrid is the best on the YOLOv8 in mAP with the best measure of precision-recall for detecting pedestrians.

6.9 YOLOv8 Inference Time

Definition: The following time taken is a measure on how long the model takes to make an inference on images.

6.9.1 Top Performers:

Transfer Learning: An interesting find was that the YOLOv8 transfer learning completed inferences at a time of 7.0 ms/image, thereby ranking it the best among the YOLOv8 group. Hybrid: In terms of inference time YOLOv8 hybrid provided 6.5 ms/image, which is a bit inferior to transfer learning but still suitable for real-time use. Training from Scratch: Finetuning from scratch of YOLOv8 used 8.3 ms/image, which is the slowest among all the examined YOLOv8 configurations. Insight: Transfer Learning is the fastest whether for YOLOv8, to cope with the real-time application although the gap is small.

6.10 YOLOv8 Epochs

Definition: Epochs mean the number of training passes.

6.10.1 Top Performers:

Training from Scratch: The training from scratch of the YOLOv8 took 100 epochs which enabled better training than the other methods adopted in the study. Hybrid: For YOLOv8 hybrid used 25 epochs, and got faster training with results are good. Transfer Learning: All the seven models used weights and biases pre-trained weights, with YOLOv8 transfer learning required only 25 epochs making it the fastest to train. Insight: Training from Scratch as you may have noted, takes time and in this case, it took 100 epochs to complete the training, but the performance was the best one. The transferred and mixed learning approaches take lesser time when training the model

Discussion Customizing Pre-Trained Models:

One of the advantages that can be found with the proposed approach is the possibility of layer freezing and fine-tuning thus allowing practitioners a customised pre-trained model. Such a freezing approach allows users to balance computational requirements and learning while targets specific features associated with a task. This makes the hybrid approach efficient for practical use in compared to other techniques of analyzing real-world data. Scalability of Customization:

Your discoveries indicate that freezing layers or its equivalent of modifying pre-trained architectures can apply to other models than YOLO. This creates room to employ the

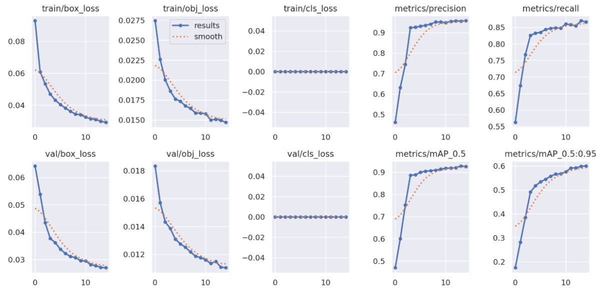


Figure 3: Pretrained Model

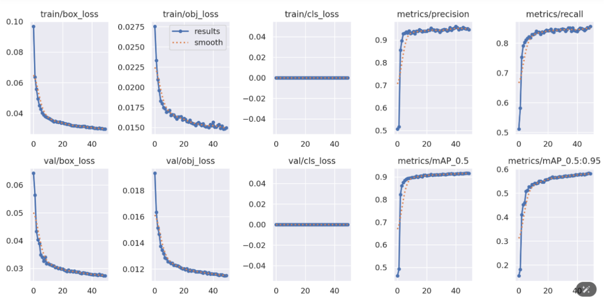


Figure 4: hybrid Model

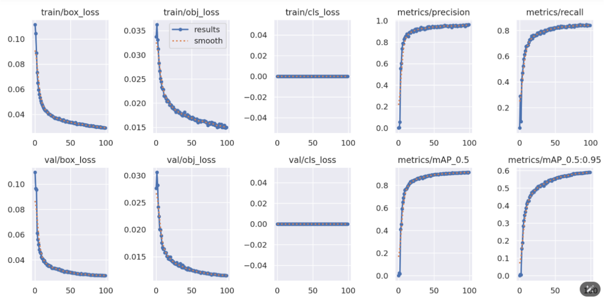


Figure 5: Training from scratch

same techniques in other models that have been developed in recent days such as Faster R-CNN, EfficientDet, or composition transformer such as the DETR. This MOOC approach is therefore malleable and can be adapted for almost any object detection task including, but not limited to, pedestrian detection and other problems in another domain. Future Directions for Freezing Techniques:

One possible research direction is freezing layers in another, or even dynamic, sequence, or which layers should be frozen based on the complexity of the data set or available computational resources. It may be beneficial to perform these techniques YOLOv5 and YOLOv8 only but testing them on other models could support the generalization. EfficientDet or Vision Transformers are models that might also require similar future modifications to increase the range of this approach's practical use.

7 Conclusion and Future Work

7.1 Conclusions

This work seeks to investigate how fine-tuning pretrained models for pedestrian detection works, and how YOLOv5 and YOLOv8 perform in this task. Now it compared transfer learning, training from scratch, hybrid approaches and emphasised the proper utilisation of transferring learning by freezing and fine-tuning some layers of the selected pre-trained models.

7.1.1 Key conclusions include:

Customization of Pre-Trained Models:

Things like unfreezing only neck and head layers while keeping the backbone layers frozen greatly reduces computational costs and time spent. This makes it easier for practitioners to obtain very high accuracy with little time spent on training, which provides the practitioner more epochs to further optimize their model. Both changes keep general knowledge of object detection and specific knowledge of the task learned from new datasets separated effectively.

7.1.2 Efficiency and Performance:

Semi-supervised learning was verify the best approach because it allows faster training and equally high accuracy of the detection compared to training with the given amount of a data from scratch or use transfer learning. Among all the models, the hybrid YOLOv8 achieved the maximal accuracy and recall while the hybrid YOLOv5 was slightly faster in the inference which also makes both very suitable for particular cases.

7.1.3 Implications of Layer Freezing:

This is well supplemented by using freezing backbone layers, which defines the way how to adjust complex models like YOLOv8 to a given task without significant consumption of resources and speeds up the training phase. This technique enables practitioners to transfer their models to other applications at varying scales of the datasets.

7.2 Future Work

7.2.1 Layer Freezing in Other Models:

Appy the same thing for other architectures such as EfficientDet, DETR, or Mask R-CNN and determine the effectiveness of layer freezing and fine-tuning on training. Explore whether freezing some layers or groups of layers gives the best outcome for these models.

7.2.2 Adaptive Freezing Techniques:

Suboptimal strategies and calibration methods for which the models decide whether to freeze layers or fine-tune them depending on the training dataset or performance indicators have to be cultivated. Task-Specific Optimization:

Discuss how the customized pre-trained models is used for others tasks different from pedestrian-detection, including medical imaging, wildlife detection, or industrial defect detection. Combining with Other Techniques:

This, when supplemented with other complex procedures, such as knowledge distillation or model pruning, will carry greater potential for improving the efficiency and scalability of the models. Practical Implications

For Practitioners: The research shows how to fine-tune models learned on one task for another effectively, to allow more epochs without extra training time. This makes attaining of high accuracy achievable even if the available resources are minimal. For Industry: Organizations can always fine-tune models such as YOLOv8 for deployment in less capacitated platforms like the mobile phones or even in embedded systems. For Academia: The work describes the possible approach for enhancing efficiency of the object detection models, which will complement numerous fields of study.

References

Ghari, B., Tourani, A., Shahbahrami, A. and Gaydadjiev, G. (2024). Pedestrian detection in low-light conditions: A comprehensive survey, *Image and Vision Computing* p. 105106.

Han, R., Xu, M. and Pei, S. (2024). Crowded pedestrian detection with optimal bounding box relocation, *Multimedia Tools and Applications* pp. 1–20.

Jiang, W., Zhang, Y., Zheng, S., Liu, S. and Yan, S. (2024). Data augmentation in human-centric vision, *Vicinagearth* **1**(1): 1–27.

Liu, Q., Ye, H., Wang, S. and Xu, Z. (2024). Yolov8-cb: Dense pedestrian detection algorithm based on in-vehicle camera, *Electronics* **13**(1): 236.

Mehra, R. et al. (2018). Breast cancer histology images classification: Training from scratch or transfer learning?, *Ict Express* **4**(4): 247–254.

Oztel, I., Yolcu, G. and Oz, C. (2019). Performance comparison of transfer learning and training from scratch approaches for deep facial expression recognition, *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pp. 1–6.

Öztürk, C., Taşyürek, M. and Türkdamar, M. U. (2023). Transfer learning and fine-tuned transfer learning methods' effectiveness analyse in the cnn-based deep learning models, *Concurrency and Computation: Practice and Experience* **35**(4): e7542.

Park, S., Kim, H. and Ro, Y. M. (2024). Robust pedestrian detection via constructing versatile pedestrian knowledge bank, *Pattern Recognition* **153**: 110539.

Rafi, A. N. Y. and Yusuf, M. (2023). Improving vehicle detection in challenging datasets: Yolov5s and frozen layers analysis, *International Journal of Informatics and Computation* **5**(2): 31–45.

Situ, Z., Teng, S., Feng, W., Zhong, Q., Chen, G., Su, J. and Zhou, Q. (2023). A transfer learning-based yolo network for sewer defect detection in comparison to classic object detection methods, *Developments in the Built Environment* **15**: 100191.

Tang, H., Chen, J., Zhang, W. and Guo, Z. (2024). Training acceleration method based on parameter freezing, *Electronics* **13**(11): 2140.