School of Computer Science
Faculty of Science and Engineering
University of Nottingham
Malaysia Campus

UG FINAL YEAR DISSERTATION REPORT

Generalisation in Pedestrian Detection: A Comprehensive Benchmarking of CNNs and ViTs Across Diverse Datasets

Student Name: Anshana Manoharan
Student ID: 20506329
Supervised by: Dr Tissa Chandesa
Submission Date: May 04 2025

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR IN SCIENCE IN COMPUTER SCIENCE WITH ARTIFICIAL INTELLIGENCE (HONS) THE UNIVERSITY OF NOTTINGHAM.

University of Nottingham
UK | CHINA | MALAYSIA

Generalisation in Pedestrian Detection: A Comprehensive Benchmarking of CNNs and ViTs Across Diverse Datasets

Submitted in May 2025, in partial fulfillment of the conditions of the award of the degree B.Sc.

Anshana Manoharan
School of Computer Science
Faculty of Science and Engineering
University of Nottingham
Malaysia Campus

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature: 

Date: 04 May 2025

Abstract

Pedestrian detection plays a pivotal role in computer vision applications such as autonomous driving and urban surveillance, yet it remains a challenging task due to occlusion, scale variation, and dataset bias. While Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have both demonstrated strong performance in general object detection, their relative strengths in the pedestrian detection domain require further investigation. This study conducts a comparative analysis of six fine-tuned object detection models—three CNN-based (Faster R-CNN, SSD, YOLOv11) and three ViT-based (DETR, Deformable DETR, RF-DETR)—trained on the EuroCityPersons and WiderPerson datasets and evaluated on the unseen INRIA dataset to assess cross-domain generalisation. Results indicate that YOLOv11 trained on WiderPerson achieved the best overall performance, with the highest precision (82.83%), F1 score (83.67%), and fastest inference time (0.31s), making it highly suitable for real-time pedestrian detection. In contrast, Deformable DETR trained on EuroCityPersons achieved the lowest Log Average Miss Rate (LAMR) of 0.0412, suggesting robustness in avoiding missed detections, albeit with high inference latency. This work provides a detailed empirical assessment of model architecture choices for pedestrian detection and offers insights into the trade-offs between accuracy, speed, and generalisability. It contributes to ongoing research by benchmarking state-of-the-art detection models on diverse datasets and informing future architecture selection for real-world deployment.

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Tissa Chandesa, for his continuous support, encouragement, and valuable guidance throughout this project. His mentorship played a key role in shaping the direction of my work. I am also grateful to the academic staff of the School of Computer Science, University of Nottingham Malaysia, for their insightful lectures and guidance on various aspects of the project. Their input helped me approach the research more critically and confidently.

Special thanks to my friends and family for their unwavering support and motivation throughout this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pedestrian detection is a crucial task in Computer Vision (CV) with applications in autonomous driving [4, 5, 6] and video surveillance [7, 8, 9]. The ability of a machine to accurately distinguish pedestrians from other objects is essential for enhancing safety and preventing accidents. However, real-world pedestrian detection presents several challenges, including blurriness (caused by video extraction), occlusion [10], scale variations, and dataset bias [11, 12].

1.1 Background

To address these challenges, early approaches focused on utilising handcrafted features and classical machine learning techniques such as HAAR Cascades [13] and Histogram of Oriented Gradients + Support Vector Machines (HOG+SVM) [14]. The first trainable object detection system was proposed by Papageorgiou & Poggio [15], followed by methods such as HOG+SVM. These approaches, while effective at the time reaching 60–80% in recall at 1 false positive per image (FPPI), still fell short of real-world application standards [14, 16].

The adoption of deep learning techniques through the proposal of Convolutional Neural Networks (CNNs) led to significant advancements in pedestrian detection. Szarvas et al. [17] were among the first to apply CNNs to this task, demonstrating potential with high detection rates at 90%. Building on this progress, Zhang et al. [18] proposed and

evaluated a version of the Faster R-CNN architecture for pedestrian detection and attained 77.12% mean Average Precision (mAP) on the KITTI dataset, emphasising its effectiveness in detecting small-scale pedestrians. Due to the rise in solving pedestrian detection tasks and the need for further research and benchmarking, the CityPersons [19] and WiderPersons [1] datasets were introduced. These datasets provided a standardised evaluation framework in the field. Soon after, the newly proposed Mask R-CNN was adapted for pedestrian detection on the WiderPersons dataset, achieving a log average miss rate (LAMR) of 13%, which is 12% lower than other mainstream detection networks [20].

Following its introduction in CV [21], Vision Transformers (ViTs) were applied to many domains such as image classification [22, 23], image recognition [24, 25], object detection [26, 27], segmentation [28, 29], among several other use cases [30, 31, 32]. In pedestrian detection, Yuan et al. [33] demonstrated that ViTs outperform state-of-the-art (SOTA) two-stage and one-stage detectors with an LAMR of 2.6% on the Caltech test set and faster inference speed.

1.2 Aims and Objectives

This dissertation investigates the generalisability and comparative performance of CNNs and ViTs in the context of pedestrian detection. Specifically, the study involves the implementation and fine-tuning of three representative models from each architecture family across multiple pedestrian detection datasets and evaluated on unseen data. The primary objective is to assess cross-dataset performance variation to examine generalisability and determine which architectural paradigm offers superior detection performance in terms of accuracy and computational efficience, thereby offering insights into the robustness of CNNs and ViTs in real-world, domain-shifted scenarios.

This dissertation aims to answer the following research questions:

RQ1: How do the selected CNNs and ViTs compare in terms of accuracy and efficiency for pedestrian detection tasks?

RQ2: Which among the selected CNN and ViT models demonstrates the best performance?

RQ3: To what extent do CNNs and ViTs generalise across datasets when trained on one dataset and tested on another?

While transformer-based models exhibit strong performance in general object detection, their efficacy for the specific task of pedestrian detection lacks systematic evaluation. By fine-tuning and benchmarking three representative CNN and ViT models across multiple datasets, including a newly proposed ViT architecture that has yet to appear in existing literature, this dissertation offers both a comprehensive performance comparison and insights into architectural trade-offs under domain-shifted conditions.

1.3 Contributions

The significance of this study lies in its potential to inform architecture selection for real-world deployment, guide future development of robust pedestrian detection systems, and extend the existing body of knowledge through a focused review of deep model performance and generalisation capabilities.

The main contributions of this dissertation are as follows:

1. Comprehensive Cross-Dataset Benchmarking: This study evaluates six state-of-the-art object detectors—Faster R-CNN, SSD, YOLOv11, DETR, Deformable DETR, and RF-DETR—trained on the WiderPerson and EuroCityPersons datasets and tested on the INRIA dataset, offering insights into their cross-domain generalisation capabilities.

2. CNN vs. Transformer Comparison: It presents a systematic side-by-side analysis of CNN-based and Vision Transformer-based models under uniform conditions, highlighting their relative strengths in terms of accuracy, generalisation, and inference time.

3. Identification of Optimal Trade-offs: The evaluation reveals that YOLOv11 trained on WiderPerson achieves the best balance between speed and accuracy,

while Deformable DETR trained on EuroCityPersons minimises missed detections (lowest LAMR), aiding decisions in time-critical vs. accuracy-critical applications.

4. Dataset Influence Analysis: The results show that models trained on WiderPerson perform better on conventional metrics like precision and accuracy, whereas EuroCityPersons yields lower LAMR, indicating its effectiveness in reducing false positives across domain shifts.

5. Empirical Validation of Training Dynamics: Analysis of training loss across epochs confirms the learning stability and convergence patterns of different models, highlighting YOLOv11 and Faster R-CNN for efficient optimisation and RF-DETR and Deformable DETR for long-term potential despite initial volatility.

This dissertation is structured as follows: Chapter 2 presents a detailed literature review on generalisation and the evolution from traditional to CNNs and ViT approaches, and the various pedestrian detection benchmark datasets. Chapter 3 outlines the methodology, detailing the selection and implementation of three CNN-based and three ViT-based object detection models, the selected datasets and the performance metrics used for evaluation. Chapter 4 describes the processing of the datasets and experimental setup used for this study. Chapter 5 discusses and evaluates the results in terms of detection accuracy, efficiency, and generalisability across datasets. Chapter 6 concludes the study with limitations, and potential directions for future research.

Chapter 2

Literature Review

2.1 Understanding Generalisation in Pedestrian Detection

Recent literature identifies several enduring and emerging challenges in pedestrian detection, particularly in safety-critical applications such as autonomous driving and surveillance. While numerous detection models demonstrate strong performance within the datasets they are trained on, their effectiveness often deteriorates under domain shifts [34, 35, 36, 37, 38, 39]. Hence, generalisation across domains remains a significant challenge in pedestrian detection.

To quantify this phenomenon, Dollar et al. [14] provided one of the earliest cross-dataset generalisation benchmarks in pedestrian detection, exposing substantial performance gaps for low-resolution images and partially occluded pedestrians. Subsequent studies have proposed various strategies to address this, such as attention-based occlusion handling mechanisms [40], and architectural enhancements for dense scenes [41], performing better than the state-of-the-art detectors at that time with a log average miss rate (LAMR) of 6.83%, 38.16%, and 13.84% on the reasonable occlusion subset, heavy occlusion subset, and a combination of both subsets. Some approaches focus on domain-specific challenges—such as low-light conditions [42], where it was also concluded that the hybrid approaches proposed were not applicable to

around-the-clock applications.

2.2 Related Surveys

Pedestrian detection has undergone significant evolution, as comprehensively captured across several landmark surveys. These works outline a progression from handcrafted features to deep learning architectures, while consistently identifying challenges in generalisation, real-time performance, and robustness under adverse conditions.

2.2.1 Evolution from Handcrafted Features to Deep Learning

Cao et al. [36] trace the trajectory from early handcrafted models (e.g., HOG, ACF) to deep CNNs, emphasising the enhanced ability of the latter to handle complex visual scenes. Similarly, Brunetti et al. [39] and Dollár et al. [38] document early pedestrian detectors and highlight their limitations in dynamic environments, particularly in video-based datasets such as Caltech [38], which reveal greater variance in occlusion and motion compared to static datasets like INRIA [3].

2.2.2 Transformer-based vs CNN-based Detectors

Recent surveys extend the discussion to include transformer-based architectures and hybrid detection paradigms. In particular, Hasan et. al. [35] evaluate a broad spectrum of model categories—spanning CNNs, transformers, multi-branch detectors, and ensemble frameworks—across diverse datasets including CrowdHuman [43], WiderPerson [1], WiderPedestrian [44], CityPersons [19], and EuroCity Persons [2]. The study finds that, while ViTs often excel in intra-dataset benchmarks, CNNs outperform them in cross-dataset generalisation, pointing to architectural biases that hinder domain robustness.

Cao et al. [36] further highlight the impact of occlusion on detector performance, noting that part-based approaches (e.g., OR-CNN [45], MGAN [46], PedHunter [47]) enhance robustness but at a notable computational cost [35],[36]. However, occlusion was

concluded to be a critical bottleneck, particularly in crowded scenes, and has become a benchmark criterion for new detection architectures [36].

Scale variance constitutes another major hurdle, particularly in detecting small or distant pedestrians. Detectors commonly exhibit a small drop in recall. Techniques such as multi-resolution inputs, feature pyramids, and context-aware mechanisms have provided partial relief [35], [36], yet detectors remain inconsistent across datasets like Caltech [38], CityPersons [19], and KITTI [48]. Galvao et al. [37] suggest that more generalisable multi-scale solutions are essential to overcome these instabilities.

Low-light and nighttime conditions introduce additional complexity, especially in autonomous driving and surveillance contexts. Ghari et. al [34], Hasan et. al. [**?**], Cao et. al. [36], Galvao et. al. [37], Dollar et. al. [38] and Brunetti et. al. [39], acknowledge the growing utility of fusion-based strategies and pixel-level enhancements. Models such as MSR [49], INSANet [50], and DaFF [51] apply halfway-fusion and modality-aware techniques to preserve discriminative cues under low illumination. Nonetheless, most hybrid solutions still underperform in real-world nighttime deployments, indicating a gap between academic innovation and deployment viability [34].

2.2.3 Real-time Constraints and Lightweight Models

Real-time inference remains elusive, especially for embedded and resource-constrained platforms. Although lightweight models like DaFF [51] demonstrate favourable trade-offs between speed and accuracy, they typically perform weaker in conditions involving occlusion or poor visibility [34]. Galvao et al. [37] and Brunetti et al. [39] argue that many proposed detectors are not ready for real-time applications, citing high computational costs and lack of standardised hardware benchmarks as significant barriers.

2.2.4 Evaluation Protocols and Dataset Biases

While detectors achieve competitive accuracy within individual datasets, their performance often deteriorates in cross-dataset scenarios due to dataset-specific

architectural tuning. For example, models trained on Caltech [38] tend to underperform on datasets like CrowdHuman [43] or CityPersons [19]. Although ViTs excel in intra-domain tasks, they lack the cross-domain robustness seen in CNN-based models [35]. To address this, Hasan et. al. [35] and Galvao et. al. [37] propose domain-agnostic representations, progressive pretraining strategies, and evaluation across diverse, densely annotated datasets.

Moreover, issues such as overfitting and inconsistent evaluation protocols remain significant challenges. Many detectors are tailored to particular datasets, distorting perceived advancements in the literature. Hasan et al. [35] advocate for the use of more realistic datasets, such as EuroCity Persons [2] and CrowdHuman [43], alongside unified standards for annotation, lighting conditions, and hardware setups to ensure reliable benchmarking and broader applicability [35], [37], [38].

2.3 Benchmark Datasets in Pedestrian Detection

The field of pedestrian detection has been supported by a variety of benchmark datasets, each offering unique challenges and advantages in terms of environmental diversity, annotation quality, and sensing modalities. The following section presents a brief overview of several widely used datasets that collectively span scenarios ranging from vehicle-mounted urban recordings to static surveillance footage, and include both visible and multispectral imaging, providing a rich foundation for developing and evaluating detection models.

2.3.1 Daytime and Urban Scene Datasets

EuroCity Persons

The EuroCity Persons dataset [44] is among the largest pedestrian detection datasets to date, with over 238,200 manually labelled person instances across 47,300 images captured from 31 cities in 12 European countries. The dataset is rich in annotations, including over 211,200 orientation labels, enabling research into fine-grained recognition

tasks. It is particularly notable for its scale and diversity, encompassing various times of day (day/night) and geographical regions, which supports generalisation studies across different conditions.

CityPersons

CityPersons [16] was built upon the Cityscapes dataset, it adds high-quality bounding box annotations for pedestrians on top of the fine pixel-level semantic segmentation data. The dataset includes 5,000 finely annotated images from 27 German and neighbouring cities. It is valued for its urban diversity, occlusion annotations, and compatibility with semantic segmentation tasks, fostering robust model training and cross-dataset generalization.

INRIA Person

INRIA Person Dataset [46] is a classic dataset that remains widely used due to its balanced representation of pedestrians in varied poses, lighting, and backgrounds. It includes 614 positive and 1,218 negative training images, and 288 positive and 453 negative test images. Despite its relatively small size, it serves as a clean benchmark for model evaluation and pretraining.

Caltech

The Caltech Pedestrian Benchmark [36], collected via a vehicle-mounted camera in Los Angeles, features approximately 43,000 images and 13,000 annotated pedestrian instances. Derived from 10 hours of driving footage, it is widely used for evaluating detection performance in North American urban settings. It includes temporal continuity and varying occlusion levels, making it suitable for studying tracking and temporal coherence.

TUD-Brussels

The TUD-Brussels [47] dataset comprises 508 image pairs with 1,326 annotated pedestrians, captured in a European city from a moving vehicle. It offers variations in pedestrian scale and viewpoint, providing a smaller yet valuable benchmark for evaluating detection algorithms in typical urban scenes.

2.3.2 Dense or Occluded Scenarios

WIDER Pedestrian

WIDER Pedestrian [43] dataset comprises 20,000 images, equally split between surveillance cameras and vehicle-mounted urban cameras. The training and validation sets together contain 66,209 annotations (46,513 pedestrians and 19,696 cyclists), with two labelled classes: walking pedestrians and cyclists. In the test set, participants are evaluated on their ability to detect all objects without category distinctions, emphasizing general object detection performance in urban environments.

WiderPerson

WiderPerson [17] dataset was designed to challenge detectors in densely populated and highly occluded scenes, it consists of 13,382 images with 399,786 annotated instances, averaging nearly 30 annotations per image. The dataset spans five annotation categories and diverse scenarios beyond just traffic, making it suitable for testing general pedestrian detection in the wild. The high density and occlusion levels present a demanding benchmark for state-of-the-art models.

2.3.3 Multispectral and Multi-Modal Datasets

KAIST

The KAIST Multispectral Pedestrian Dataset [41] offers 95,000 paired color and thermal images collected at 20 Hz from a vehicle, with over 103,000 dense annotations covering pedestrians, people, and cyclists. It supports multispectral detection research

and includes temporal correspondence akin to the Caltech dataset, facilitating studies in low-light or adverse weather conditions.

KITTI

KITTI Vision Benchmark [45] includes pedestrian detection annotations among a variety of object classes. Captured using multiple sensor modalities (RGB, stereo, LiDAR), KITTI provides a multimodal benchmark. Various subsets have been annotated for object detection, semantic segmentation, and tracking, enabling comprehensive evaluation across tasks.

2.3.4 Classical Datasets for Classification

Daimler

Daimler Pedestrian Classification dataset [50] is tailored for pedestrian classification and detection in urban driving scenes. It includes 15,560 cropped pedestrian samples and 6,744 full images without pedestrians for negative sampling in training. The test set contains over 21,790 images with 56,492 annotations, providing a comprehensive benchmark for urban pedestrian detection tasks.

2.3.5 Chapter Recap

This chapter reviews key developments and challenges in pedestrian detection, with a focus on generalisation across domains. It emphasises persistent issues such as occlusion, scale variance, and low-light conditions, which limit model robustness. The evolution from handcrafted features to deep learning and transformer-based methods is outlined, noting trade-offs between accuracy, generalisability, and computational efficiency. Benchmark datasets are summarised, emphasising their role in evaluating model performance under varied environmental and sensor conditions.

Chapter 3

Methodology

This chapter outlines the procedure taken to implement this study, particularly the use of transfer learning, model selection, dataset selection and the relevant performance metrics use for evaluation.

3.1 Transfer Learning

In deep learning, transfer learning is a widely adopted strategy that exploits knowledge gained from one task to facilitate learning in a different domain. This method is especially beneficial when training models from scratch is either computationally intensive or impractical due to limited data availability [52].

In pedestrian detection, transfer learning offers a practical alternative to training models from scratch by using pre-trained weights from large-scale and diverse datasets such as ImageNet [53], MS COCO [54], and Open Images Dataset V4 [55]. Rather than randomly initialising all parameters, this approach retains the early convolutional layers of the pre-trained model, which capture generic low-level features like edges, textures, and gradients—features that are still relevant for identifying pedestrians in various environments. The later, more task-specific layers are then fine-tuned on pedestrian-focused datasets to adapt the model to the nuances of human detection in real-world scenes.

Additionally, re-training models independently would introduce considerable

computational overhead, particularly regarding hyperparameter optimisation. Hence, this study adopts this approach to train the selected models and facilitate domain adaptation to the pedestrian detection task.

3.2 Proposed Implementation Framework

To evaluate performance, three CNN and three ViT state-of-the-art object detection models are selected to represent a broad spectrum of deep learning architectures and benchmark achievements in recent object detection literature. While a comprehensive comparison was constrained by storage and computational limitations, the chosen models reflect the prevailing standards in the field.

This study evaluates only the pre-trained models sourced from official implementations hosted on widely adopted and reputable platforms such as PyTorch, HuggingFace, Ultralytics, and Roboflow. This approach is taken with the assumption that the original authors know best on how to optimise their models through domain-specific training strategies.

It is crucial to note that all selected models were originally pre-trained on the MS COCO dataset [54], ensuring a consistent and equal base for comparison. These pre-trained detectors are further fine-tuned on two selected datasets and then evaluated on an unseen dataset. This enables a systematic investigation into architectural performance and, importantly, *cross-dataset generalisation*—the extent to which models trained on one dataset can perform effectively on another. This aspect is particularly critical in real-world applications, where object detectors are rarely deployed in environments identical to their training data [56]. Robust generalisation indicates a model's ability to capture meaningful, transferable features rather than overfitting to the specific traits of the training dataset.

By assessing performance on out-of-distribution data, this dissertation provides insights into each model's resilience to domain shift, a key consideration in practical deployment scenarios such as autonomous driving, surveillance, and robotics as well as test the

performance of CNNs against ViTs.

3.3 Model Selection

The following section delves deep into the popular CNN and ViT object detectors in pedestrian detection and justification for their selection for evaluation in this study. It is important to note that although the current literature contains various state-of-the-art models for pedestrian detection, the selection choice for this study is limited to models that are pre-trained on the MS COCO dataset [54].

3.3.1 Faster Region-based Convolutional Neural Network (Faster R-CNN)

The R-CNN architecture, introduced by Girshick et al. [57], marked a major milestone in object detection by combining selective region proposals with deep CNN feature extraction. Using selective search to generate 2000 region proposals per image, R-CNN extracts features via a CNN and classifies them using SVMs, leading to a 30% performance boost over previous PASCAL VOC baselines, reaching 53.7% mAP. Despite its high accuracy, R-CNN is computationally inefficient, as each region requires a separate CNN pass and selective search is slow.

Although not originally developed for pedestrian detection, R-CNN has been adapted for related applications. Lee et al. [58] applied R-CNN in a multi-sensor system reduce region proposals and improve efficiency for autonomous environments. Park et al. [59] proposed a pedestrian knowledge bank, using refined CLIP [60] features to enhance detection in cluttered scenes, integrating this into R-CNN-based models such as Cascade R-CNN [61] and Sparse R-CNN [62].

Fast R-CNN [63] improves upon R-CNN [57] by introducing end-to-end training with a multi-task loss and RoI pooling, eliminating the need for feature caching and enabling faster inference. Despite achieving higher accuracy (66.9% mAP), it still depends on external region proposals, limiting real-time performance compared to Faster

R-CNN [64].

Faster R-CNN [64] is a two-stage object detection framework that integrates a Region Proposal Network (RPN) with a Fast R-CNN detector into a unified, fully convolutional network. The RPN generates high-quality region proposals by predicting objectness scores and bounding boxes at each spatial location, while the Fast R-CNN module classifies these proposals and refines their coordinates. By sharing convolutional features between both modules, Faster R-CNN [64] significantly reduces computational redundancy and achieves near cost-free region proposal generation. This design allows the model to operate at approximately 5 frames per second on a GPU using the VGG-16 backbone, while maintaining state-of-the-art accuracy on benchmarks such as PASCAL VOC 2007/2012 and MS COCO. Its effectiveness was further validated by its adoption in the winning entries of multiple tracks in the ILSVRC and COCO 2015 competitions. Faster R-CNN [64] has demonstrated substantial effectiveness in pedestrian detection, owing largely to its integrated Region Proposal Network (RPN), which enhances both accuracy and computational efficiency. Several studies have explored its applicability in diverse scenarios, confirming its superiority over traditional methods.

Adoption in Pedestrian Detection

Zhang et al. [18] applied Faster R-CNN to pedestrian detection by coupling a CNN-based feature extractor with an RPN refined through K-means clustering to identify candidate regions. Their approach achieved a high detection accuracy of 92.7% on the INRIA dataset, outperforming conventional algorithms and highlighting Faster R-CNN's capacity for precise localisation and classification in pedestrian-heavy environments.

Zhao et al. [65] advanced this line of work by designing a pedestrian-specific detection system using a fine-tuned VGGNet within the Faster R-CNN framework. Their system accommodates input images of arbitrary dimensions and outputs both bounding box coordinates and confidence scores for pedestrian instances. The modular architecture also allows for future integration of alternative convolutional backbones and

handcrafted features to further improve performance.

Byeon et al. [66] conducted a comparative study between Faster R-CNN and Aggregate Channel Features (ACF) using pedestrian videos from YouTube that simulated real-world driving environments. Their findings showed that Faster R-CNN achieved a precision rate 56.73% higher than ACF when manually labeled data was used. Furthermore, Faster R-CNN demonstrated a precision level approximately seven times greater than the second-best method in the study, reaffirming its robustness in real-time pedestrian detection tasks. Although ACF yielded a higher recall under specific configurations, its precision lagged significantly behind, underscoring Faster R-CNN's balance between accuracy and reliability.

Collectively, these studies affirm that Faster R-CNN remains a dominant architecture in pedestrian detection, especially in scenarios requiring accurate region proposals and high precision under varied imaging conditions, justifying it as a choice for this study.

3.3.2 Single Shot MultiBox Detector (SSD)

The SSD model [67] represents an advancement in object detection, offering a balance between detection accuracy and computational efficiency. Introduced as a single-stage detection framework, SSD eliminates the dependence on region proposal mechanisms traditionally used in two-stage detectors such as Faster R-CNN [64]. By integrating object localisation and classification within a unified convolutional neural network, SSD enables end-to-end training and inference, thereby simplifying the detection pipeline and reducing computational overhead.

SSD discretises the output space of potential bounding boxes into a set of pre-defined default boxes with varying aspect ratios and scales at each spatial location of multiple feature maps. During inference, the model predicts both the confidence scores for each object class and the offsets to adjust these default boxes to better match the ground truth object shapes. This formulation allows SSD to efficiently handle objects of diverse shapes and sizes within a single forward pass.

A critical innovation of SSD lies in its use of multiple feature maps at different

resolutions for prediction. By exploiting hierarchical features extracted from various layers of the backbone network, SSD is capable of detecting small objects using high-resolution features and larger objects using coarser, deeper features. This multi-scale prediction strategy enhances the detector's robustness across a wide range of object sizes.

Empirical evaluations demonstrate that SSD achieves competitive accuracy compared to two-stage detectors while operating at significantly higher speeds. For instance, with a 300×300 input resolution, SSD attains a mean average precision (mAP) of 74.3% on the PASCAL VOC 2007 test set, achieving real-time inference at 59 frames per second (FPS) on an Nvidia Titan X GPU. With a larger input resolution of 512×512, SSD further improves its detection performance to 76.9% mAP. These results present the effectiveness of SSD in delivering both high-speed and relatively high-accuracy detection, particularly in scenarios where low latency is critical. Many studies have extended the SSD framework to address the unique challenges of pedestrian detection, particularly the accurate localisation of small-scale and densely occluded targets.

Adoption in Pedestrian Detection

Cheng et al. [68] enhance the original SSD architecture by incorporating additional features from earlier convolutional blocks and introducing dense connections to better capture fine-grained visual information. Their approach also includes a revised matching strategy for global prior boxes and the introduction of an extra loss term tailored for dense target scenarios, resulting in improved performance on the Caltech and VOC datasets while preserving real-time inference speed.

Yan et al. [69] similarly focus on improving SSD's performance on small-scale pedestrian detection by proposing a novel two-level feature fusion mechanism and an adaptive loss function that modifies the Smooth L1 loss to improve robustness. Their method achieves state-of-the-art results on the Caltech dataset, particularly under the "Far" setting, and demonstrates a favorable speed-accuracy trade-off on the CityPersons dataset.

Complementing these efforts, Sun et al.[70] introduce E-SSD, which augments SSD with a deconvolutional feature-fusion module and an attention mechanism that combines spatial and channel attention to enhance feature selectivity. Evaluated on the UCAR pedestrian dataset as well as standard benchmarks like Caltech and COCO Persons, E-SSD achieves superior performance on small pedestrian targets while maintaining detection efficiency.

Collectively, these works highlight SSD's adaptability and continued relevance in pedestrian detection, especially when augmented with context-aware features, attention mechanisms, and improved loss formulations for small-object and occlusion-heavy scenarios.

This shows that SSD has had a profound impact on the object detection landscape. Its open-source implementation has further facilitated widespread adoption and adaptation in both academic research and industrial deployment, which is apart from its adoption in pedestrian detection, another factor towards choosing it as a model for evaluation in this study.

3.3.3 You Only Look Once version 11 (YOLOv11)

YOLOv11 [71] represents a significant advancement in the field of real-time object detection, building on the efficiency and unified design principles of its predecessors in the YOLO series. It was designed to deliver high performance across a range of computer vision tasks, including not only object detection but also instance segmentation, pose estimation, and oriented object detection. This multi-task capability broadens its applicability across diverse domains, from autonomous systems and surveillance to industrial automation.

One of YOLOv11's key strengths is its scalability, offering a suite of model sizes from lightweight nano versions to more powerful extra-large variants [71]. This design allows practitioners to balance speed and accuracy based on deployment constraints, making it suitable for both resource-constrained edge devices and high-performance computing platforms. Particularly, the nano variant achieves improved inference speed without a

significant increase in model size, making it ideal for real-time applications.
Another notable aspect of YOLOv11 is its emphasis on efficiency without compromising detection accuracy. Comparative benchmarks demonstrate that YOLOv11 consistently outperforms its predecessors in mean Average Precision (mAP) while maintaining or improving upon real-time processing capabilities. Its performance gains are especially evident in scenarios involving small or occluded objects, which remain challenging for many detectors.

Adoption in Pedestrian Detection

Despite being a recent addition to the object detection landscape, YOLOv11[71] has begun to gain traction in the pedestrian detection domain, with a few emerging studies demonstrating its potential in complex real-world scenarios. Sui et al. [72] introduce FA-YOLO, an enhanced version of YOLOv11 tailored for pedestrian detection, incorporating a Feature Enhancement Module (FEM) and an Adaptive Sparse Self-Attention (ASSA) mechanism to strengthen feature representation and suppress redundant information. Their model achieves state-of-the-art results on the WiderPerson and RTTS datasets, surpassing baseline YOLOv11 by 3.5% and 3.0% in precision, respectively.
Similarly, Li et al.[73] apply YOLOv11 to the challenging task of miner detection in environments with low lighting and partial occlusion, integrating an Efficient Channel Attention (ECA) module and new loss formulations to enhance robustness and accuracy. Their approach demonstrates superior performance over existing detection models, accentuating YOLOv11's adaptability to harsh visual conditions.
YOLOv11 [71] has emerged as a versatile and efficient object detection framework, with strong potential across various computer vision tasks. Although its adoption in pedestrian detection is still limited due to its recent introduction, the discussed studies showcase its effectiveness in real-time performance, spatial attention, and robustness to challenging conditions. These promising capabilities motivated its selection for evaluation in this study, aiming to contribute to the limited yet growing body of

research on YOLOv11 in pedestrian detection.

3.3.4 DEtection TRansformer (DETR)

DETR [74] represents a significant departure from traditional object detection frameworks by reformulating object detection as a direct set prediction task. Unlike conventional detectors, which rely heavily on complex, hand-crafted components such as anchor boxes, region proposals, and non-maximum suppression, DETR introduces a streamlined end-to-end pipeline based on a transformer encoder-decoder architecture. It uses a fixed set of learned object queries and a global set-based bipartite matching loss, enabling the prediction of unique object instances without the need for post-processing heuristics.

By capturing global image context and object relations, DETR simplifies the detection pipeline while maintaining competitive accuracy with established detectors such as Faster R-CNN [64]. Notably, DETR achieves strong performance on the COCO dataset and demonstrates robust generalisation to related tasks like panoptic segmentation, offering a unified framework for multiple vision challenges. Its ability to directly model object relationships in parallel through attention mechanisms has laid the groundwork for a new class of detection models, especially those seeking to integrate detection with other vision tasks in a cohesive architecture.

Adoption in Pedestrian Detection

Despite its original design for general object detection, DETR and its variants have recently been adapted to pedestrian detection, particularly in crowded and complex environments. Wu et al. [75] propose an improved RT-DETR model that integrates HiLo attention and a nonlinear feature fusion module, alongside a novel InnerMPDIoU loss function, demonstrating enhanced accuracy and robustness on the CityPersons dataset, with a 4.2% gain in mAP50 over the baseline.

Lin et al. [76] identify the challenges DETR faces in dense pedestrian scenes, such as inefficient bipartite matching and performance drop compared to traditional methods

like Faster R-CNN. To address these, they introduce the PED detector with a tailored decoder and a visible-part-focused mechanism, showing competitive performance on CityPersons and CrowdHuman.

Gao et al. [77] further tackle DETR's limitations in crowded settings by incorporating a rank-based contrastive learning strategy that enforces better discrimination between similar pedestrians and background features, achieving state-of-the-art results with 38.9% miss rate on the CrowdHuman dataset.

Collectively, these works highlight DETR's growing presence in the pedestrian detection domain, particularly when adapted with architectural or training enhancements to meet the unique challenges of crowd density, occlusion, and feature ambiguity. Owing to its relevance and promising results demonstrated in the aforementioned studies, DETR was selected as one of the models for evaluation.

3.3.5 Deformable DETR

Deformable DETR [78] represents a significant refinement over the original DETR [74], addressing key limitations that hindered DETR's broader applicability, particularly its slow convergence and difficulty in capturing fine-grained spatial information. While DETR reformulated object detection as a direct set prediction task, eliminating traditional components like anchor generation and non-maximum suppression, it struggled with computational inefficiencies and poor performance on small objects due to the dense and global nature of its attention mechanism.

To overcome these constraints, Deformable DETR introduces a more efficient attention mechanism that restricts computation to a sparse set of sampling points around each query reference, rather than the entire image feature map. This design dramatically accelerates training—requiring approximately ten times fewer epochs than DETR—while simultaneously improving accuracy, particularly in the detection of small and densely packed objects. This improvement is crucial for real-world object detection scenarios where objects often appear in cluttered or low-resolution contexts.

Adoption in Pedestrian Detection

Recent literature reflects a growing interest in adapting Deformable DETR for pedestrian detection, particularly in complex and densely populated scenes. These studies emphasise the model's capacity to handle occlusion, scale variance, and feature sparsity more effectively than traditional detectors.

Han et al. introduce the Improved Deformable-DETR (IDPD) [79], which enhances feature representation and training efficiency through dynamic convolutional necks and a hybrid decoding loss that includes one-to-many auxiliary supervision and contrastive de-noising. Their method achieves state-of-the-art results on the CrowdHuman dataset, surpassing both the Deformable DETR [78] baseline and conventional CNN-based models.

Deng et al. [80] further optimise Deformable DETR for efficiency by incorporating EfficientNet and a cross-fertilisation module that preserves occluded and small-scale pedestrian features, yielding performance gains with significantly reduced computational cost.

Similarly, Yuan et al. [81] propose a multi-scale deformable transformer encoder that effectively captures fine-grained features across multiple levels, out-performing two-stage detectors on challenging benchmarks like CityPersons and Caltech.

Collectively, Deformable DETR's is increasing relevance in pedestrian detection, where its adaptive attention mechanisms and flexibility support robust performance in real-world, high-density scenarios.

3.3.6 RoboFlow DETR (RF-DETR)

RF-DETR [82] represents a recent advancement in transformer-based object detection, specifically designed to balance real-time performance with competitive accuracy. It introduces a lightweight architecture that maintains the core advantages of transformer models while being optimised for deployment on resource-constrained edge devices. It is notable for being the first real-time transformer-based model to surpass 60 Average Precision (AP) on the Microsoft COCO benchmark, placing it among the

top-performing models in this category.

Beyond standard benchmarks, RF-DETR demonstrates strong generalisation capabilities, achieving state-of-the-art results on the RF100-VL benchmark, which evaluates model adaptability across diverse, real-world domains. This highlights its suitability for practical applications where domain shifts and deployment constraints are common. The model delivers competitive inference speed relative to existing real-time detectors, positioning it as a robust choice for scenarios requiring a balance of accuracy, adaptability, and efficiency.

Because RF-DETR is a newly proposed model, it has not yet been adopted in any of the current literature, which was a key reason as to why this model was chosen for evaluation in this study.

3.4 Dataset Selection

Given the characteristics of the aforementioned benchmark datasets in Chapter 2.3, this study utilises the EuroCity Persons dataset [2] and WiderPerson dataset [1] for training object detection models, while the INRIA dataset [3] is used only for evaluation as the "unseen" dataset. The WiderPerson dataset [1] is selected for its high annotation density (see Figure 3.1) and substantial occlusion challenges, making it particularly suitable for training models in densely populated urban environments.

Complementing this, the EuroCity Persons dataset [2] provides geographic and temporal diversity, enhancing the robustness of the training process. Due to storage limitations, only the validation subset of EuroCity Persons [2] was downloaded and subsequently split into training and validation sets using an 80:20 ratio. Despite this constraint, the two selected datasets offer complementary characteristics; the downloaded subset of EuroCity Persons[2] contains comparatively sparse annotations (see Figure 3.2) to WiderPerson[1], contributing to diversity in training conditions.

Figure 3.1: Annotated samples from the WiderPerson dataset [1]

Figure 3.2: Annotated samples from the EuroCity Persons dataset [2]

For the performance evaluation, the INRIA dataset [3] was chosen due to its moderate annotation density and varied background scenes (see Figure 3.3), serving as a reliable benchmark for assessing the generalisation ability of the trained models to unseen data. This combination of datasets facilitates a balanced training strategy and supports a comprehensive evaluation of cross-dataset generalisation performance.

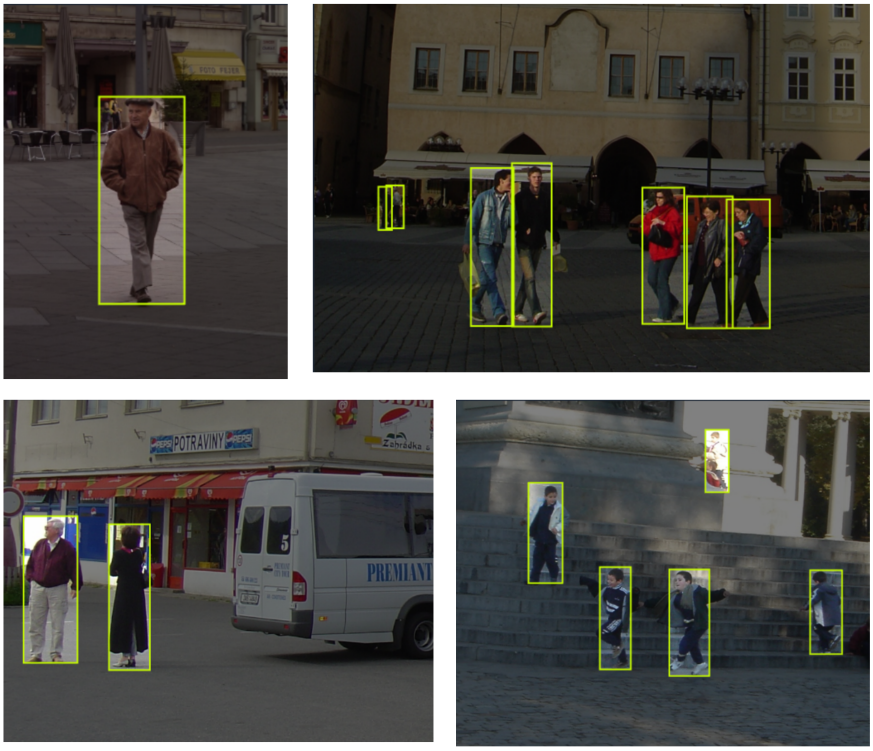

Figure 3.3: Annotated samples from the INRIA dataset [3]

3.5 Performance Metrics

Object detection algorithms are critically assessed using various evaluation metrics. Common performance evaluation and object detection metrics include the following (Refer Figure 3.4):

		Ground Truth	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 3.4: Confusion matrix illustrating True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

1. Accuracy is defined as the proportion of the predicted labelled samples over the

total samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

2. Precision is the fraction of correctly predicted positive samples that were actually predicted positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

3. Recall is the proportion of the positive samples that were predicted correctly as a percentage of the actual positive samples:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

4. F1 Score is the harmonic average of accuracy and recall:

$$\text{F1 Score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.4}$$

5. Log-Average Miss Rate (LAMR) is used to evaluate how consistently a detection algorithm performs across different confidence thresholds. To compute it, the miss rate (the ratio of FN and TP + FN) is first calculated at multiple false positive per image (FPPI) rates. Then, the logarithm of each miss rate is taken, and these values are averaged, typically using a geometric mean, to produce a single representative score:

$$\text{LAMR} = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(\text{MR}_i)\right) \tag{3.5}$$

where:

- MR_i is the miss rate at the i-th FPPI reference point.
- N is the number of reference points (commonly 9 points spaced logarithmically between 0.01 and 1).

This logarithmic averaging approach ensures that the metric captures performance variations more sensitively, especially in challenging detection scenarios.

6. Intersection over Union (IoU) measures the positional accuracy, in particular the overlap rate between the predicted bounding box and the ground truth bounding box annotated in the original image:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|} \tag{3.6}$$

where, A denotes the predicted box region and B denotes the truth box region. The threshold for IoU is commonly set to assess the accuracy of the detected bounding box. A value of 0.5 or higher for IoU indicates a successful detection, while an IoU of 1 represents the ideal prediction. (Refer Figure 3.5).

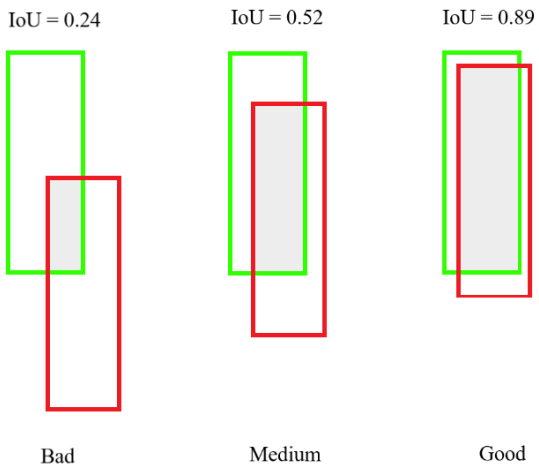

Figure 3.5: Examples of IoU thresholds on predicted and ground truth boxes.

7. Inference Time refers to the amount of time taken by the model to produce a prediction on a singular test data. It is typically measured in seconds or milliseconds, depending on the model's complexity and the hardware used. It is a critical performance metric, especially in real-time applications such as autonomous driving or surveillance systems, where quick decision-making is essential. For this evaluation, inference time is measured by calculating the average time taken to process a batch of images after finetuning the model.

These performance metrics play a crucial role in assessing object detection algorithms, enabling the comparison of different models and evaluating their generalisation capabilities. Accuracy measures the overall correctness of predictions, while precision and recall assess the model's ability to detect objects accurately and completely. IoU measures how precisely the predicted bounding boxes align with the ground truth. Additionally, LAMR provides insight into the model's robustness by averaging miss rates over a range of false positives per image, with a logarithmic emphasis that highlights weaknesses in challenging detection scenarios. Together, these metrics deliver a multifaceted assessment of an object detection model, helping identify architectural limitations and guiding future improvement strategies.

3.5.1 Chapter Recap

This chapter discusses methodology adopted for this study, detailing the implementation process. It covers the application of transfer learning as a foundational approach and the rationale behind model and dataset selection. Additionally, it outlines the evaluation framework, including the performance metrics employed to assess model effectiveness and generalisation across datasets.

Chapter 4

Implementation

This chapter contains the dataset pre-processing details and experimental setup necessary for the performance benchmarking in this study.

4.1 Dataset Processing

The WiderPerson dataset [1], being of a manageable size, was fully downloaded, and its official training and validation subsets were utilised to train all selected models. In contrast, due to storage constraints, only the daytime validation subset of the EuroCity Persons dataset [2] was acquired. This subset was subsequently partitioned into training and validation sets using an 80:20 split. For both datasets, annotation files were converted into COCO format to ensure compatibility with the training pipeline. Additionally, a corresponding YAML configuration file was created to support training with YOLO-based models. Annotations corresponding to ignore regions were excluded from the datasets to prevent bias during training and evaluation.
As this study aims to evaluate and compare the overall performance of the selected models under standard conditions, data augmentation techniques were deliberately omitted. This decision ensures that model comparisons reflect architectural capabilities rather than gains from augmentation strategies.

4.2 Experimental Setup

All experiments in this study were conducted using the Google Colab environment, which provided a convenient platform for dataset preprocessing, data restructuring, model fine-tuning, and evaluation. The Tesla T4 GPU was used, with a shared 6MB L2 cache and 16GB of high-bandwidth memory.

4.2.1 Training Setup

Each model was trained on both the *WiderPerson* and *EuroCity Persons* datasets for a total of 5 epochs. As mentioned in Chapter 3, hyperparameter tuning would have introduced high computation overhead, so training hyperparameters, including batch size, learning rate, momentum, and weight decay, were selected in a way accommodate the computational demands of each architecture. Table 4.1 summarises the specific settings for each model-dataset pair.

Table 4.1: Training hyperparameters for each model on the two datasets.

Model	**Dataset**	**Batch Size**	**Learning Rate**	**Momentum**	**Weight Decay**
Faster R-CNN	WiderPerson	2	5e-3	9e-1	5e-4
Faster R-CNN	EuroCity Persons	2	5e-3	9e-1	5e-4
SSD300	WiderPerson	4	1e-3	9e-1	5e-4
SSD300	EuroCity Persons	4	1e-3	9e-1	5e-4
YOLOv11	WiderPerson	1	1e-2	9.37e-1	5e-4
YOLOv11	EuroCity Persons	1	1e-2	9.37e-1	5e-4
DETR	WiderPerson	2	5e-5	N/A	Default
DETR	EuroCity Persons	2	5e-5	N/A	Default
Deformable DETR	WiderPerson	1	5e-5	N/A	Default
Deformable DETR	EuroCity Persons	1	5e-5	N/A	Default
RF-DETR	WiderPerson	2	1e-4	Default	Default
RF-DETR	EuroCity Persons	2	1e-4	Default	Default

For DETR and Deformable DETR, which employ the AdamW optimiser, momentum is not applicable. The default values for the optimiser's parameters (e.g., weight decay, beta coefficients) were used. In the case of RF-DETR, since explicit optimiser settings were not documented in the literature (with this model being a relatively new proposed architecture), default values were adopted as provided by the official implementation.

4.2.2 Evaluation

Evaluation will be conducted on the final fine-tuned models on the INRIA dataset using a subset of 20 different images of different scenarios, including occlusion, scale difference, crowd density and sparsity of pedestrians.

4.3 Chapter Recap

This chapter detailed the implementation process of the study, including dataset preparation and experimental setup. The WiderPerson and a subset of the EuroCity Persons datasets were preprocessed and converted into COCO format for compatibility, with annotations for ignore regions excluded. No data augmentation was applied to maintain a consistent evaluation of model architectures. All models were trained in Google Colab using a Tesla T4 GPU, with training conducted over five epochs per dataset. Hyperparameters were selected based on each model's computational needs, and standard settings were used for optimisers where appropriate or undocumented.

Chapter 5

Results and Evaluation

5.1 Analysis and Discussion

This section critically evaluates the performance of six state-of-the-art object detection models—Faster R-CNN, SSD, YOLOv11, DETR, Deformable DETR, and RF-DETR—each trained independently on the WiderPerson and EuroCityPersons datasets and tested on the INRIA pedestrian detection dataset. The results are assessed through multiple metrics including Precision (Equation 3.2), Recall (Equation 3.3), F1 Score (Equation 3.4), Accuracy (Equation 3.1), IoU (Equation 3.6), LAMR (Equation 3.5), and Inference Time. This section also analyses the training loss progression of each model over the 5 epochs (see Figure 5.3).

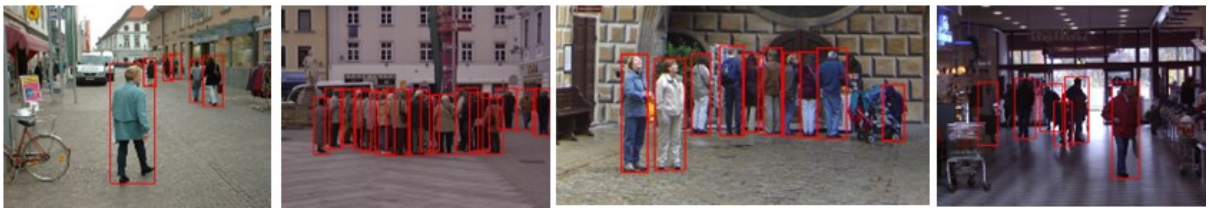

Figure 5.1: Ground truth bounding boxes from a subset of the INRIA dataset used for evaluation.

Table 5.1: Performance metrics for all six trained architectures

Model	**Trained On**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**	**Mean IoU (%)**
Faster R-CNN	WiderPerson	56.69	91.75	70.08	53.94	75.78
	EuroCityPersons	33.09	91.75	48.63	32.13	78.05
RF-DETR	WiderPerson	65.67	90.72	76.19	61.54	80.69
	EuroCityPersons	66.18	92.78	77.25	62.94	**80.82**
YOLOv11	WiderPerson	**82.83**	84.54	**83.67**	**71.93**	76.74
	EuroCityPersons	68.07	83.51	75.00	60.00	78.55
DETR	WiderPerson	21.74	92.78	35.23	21.38	76.53
	EuroCityPersons	19.24	93.81	31.93	19.00	72.61
Deformable DETR	WiderPerson	58.28	90.72	70.97	55.00	76.99
	EuroCityPersons	27.68	**95.88**	42.96	27.35	76.58
SSD	WiderPerson	0.18	7.22	0.34	0.17	55.53
	EuroCityPersons	0.00	0.00	0.00	0.00	0.00

5.1.1 General Observations

Among all models (see Table 5.1), *YOLOv11 trained on WiderPerson* delivered the best overall detection performance on the INRIA subset (see Figure 5.1), with the highest Precision (82.83%), F1 Score (83.67%), and Accuracy (71.93%). Its Recall (84.54%) was also competitive, and it had the lowest inference time (0.31 seconds per image), making it the most practical model for real-time pedestrian detection scenarios. However, its LAMR (see Table 5.2) was relatively higher (0.1546 and 0.1649), meaning more missed detections at varying false positive rates compared to transformer-based models. In terms of training loss, all models exhibited a consistent reduction in training loss across the five epochs, indicating that gradient descent optimisation was functioning correctly. However, the rate and stability of convergence varied significantly between models and datasets. Figure 5.4 and Figure 5.5 illustrate the predicted bounding boxes of 4 images of the evaluated INRIA subset.

Table 5.2: LAMR comparison of the trained models on the INRIA dataset.

Model	**WiderPerson**	**EuroCity Persons**
Faster R-CNN	0.0825	0.0825
RF-DETR	0.0928	0.0722
YOLOv11	0.1546	0.1649
DETR	0.0722	0.0619
Deformable DETR	0.0928	**0.0412**
SSD	0.9278	1.0000

Table 5.3: Inference time comparison of the trained models on the INRIA dataset.

Model	**WiderPerson**	**EuroCity Persons**
Faster R-CNN	9.52	9.16
RF-DETR	2.93	2.56
YOLOv11	**0.31**	0.38
DETR	5.61	5.06
Deformable DETR	22.32	22.77
SSD	1.40	1.44

5.1.2 Vision Transformer Architectures

The transformer-based models—DETR, Deformable DETR, and RF-DETR—demonstrated diverse strengths (see Table 5.3 for Inference Times):

1. RF-DETR yielded a strong balance of performance and efficiency. When trained on EuroCityPersons, it produced the highest F1 Score (77.25%) among transformer models and a relatively low LAMR (0.0722). Its inference time (2.56s) was also moderate compared to other transformers, showing its potential as an efficient and accurate alternative.

2. Deformable DETR trained on EuroCityPersons yielded the *lowest LAMR* (0.0412) across all models, highlighting its superior detection consistency across different false positive rates. However, it lagged in Precision (27.68%), Accuracy (27.35%), and had the highest inference time (22.77s), suggesting it may not be suitable for time-sensitive applications.

3. DETR had the lowest Precision and Accuracy overall, despite maintaining a low LAMR (0.0619 with EuroCityPersons). This shows that while it avoids many false positives, it also fails to confidently detect true positives, reflected in its poor F1 Scores (35.23% and 31.93%).

5.1.3 Convolution Neural Network Architectures

1. Faster R-CNN showed high Recall (91.75%) regardless of training dataset, but Precision dropped significantly when trained on EuroCityPersons (33.09%), resulting in poor Accuracy (32.13%). The model's high LAMR (0.0825) and inference time (9s) reinforce its status as a reliable but slower and less generalisable detector.

2. SSD failed to generalise entirely. It was observed that it delivered negligible values across all metrics when trained on either dataset. It generated predictions for bounding boxes that were too small (see Figure 5.4 (c) and (d)) for representation in the metrics leading to frequent cases of 0s in the performance metrics and an LAMR of 1.0. This is expected as the SSD model was tailored for small object detections (see Chapter 3). Despite its low inference time (1.4s), its LAMR (0.9278 and 1.0) confirms its ineffectiveness in cross-dataset pedestrian detection.

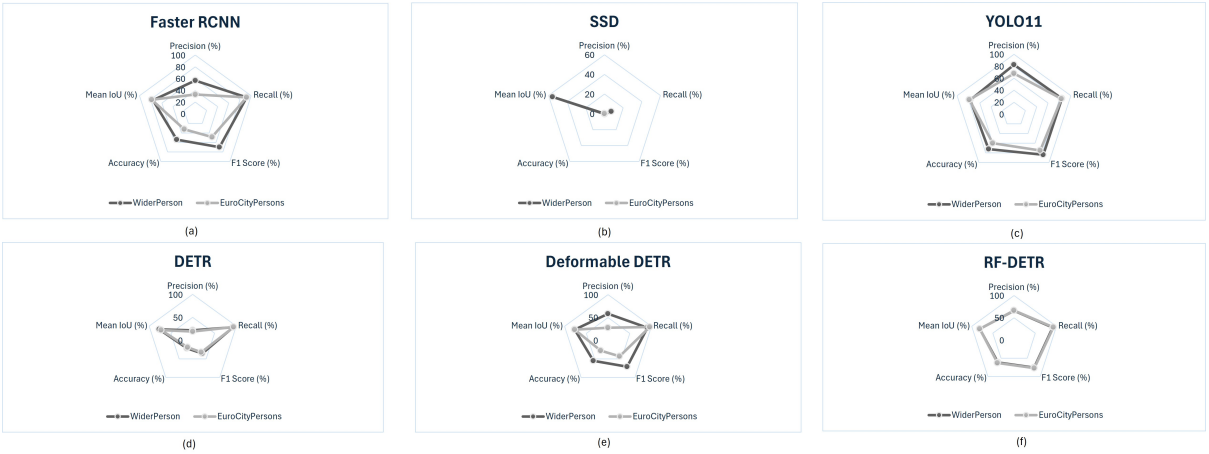

Figure 5.2: Comparison of the evaluation metrics from Table 5.1 for all finetuned models.

Figure 5.3: Comparison of the training loss across 5 epochs for all selected models on the selected datasets.

5.1.4 WiderPerson vs. EuroCityPersons

In terms of training loss on the WiderPerson dataset (refer Figure 5.3), YOLOv11 showed efficient convergence, reducing loss from 1.8983 to 1.5863, while Faster R-CNN steadily improved from 1.0194 to 0.8478, aided by its pre-trained backbone. SSD faced challenges with a high initial loss of 15.84, dropping to 12.88. DETR improved moderately from 2.3325 to 1.4724, and Deformable DETR converged rapidly from 1.0774 to 0.8379. RF-DETR showed gradual improvement from 7.6899 to 5.4317. On EuroCityPersons, YOLOv11 showed a slightly slower reduction (2.0959 to 1.7890) due to urban complexity. Faster R-CNN achieved the *lowest loss* (0.4505), while SSD

struggled with instability, dropping from 20.71 to 7.62. DETR showed slower improvement (3.229 to 2.1488), and Deformable DETR improved significantly from 9.147 to 1.063, despite initial instability. RF-DETR demonstrated steady progress from 7.0483 to 6.4343.

Models trained on WiderPerson generally outperformed those trained on EuroCityPersons in terms of Precision and Accuracy (refer Table 5.1 and the plots in Figure 5.2). This can be attributed to the higher pedestrian density and annotation consistency in the WiderPerson dataset, allowing better feature learning and generalisation to INRIA's sparser scenes. However, in terms of LAMR, models trained on EuroCityPersons (particularly RF-DETR and Deformable DETR) performed better, suggesting robustness to false positives.

5.1.5 Suitability in Real-World Applications

For *real-time applications*, YOLOv11 stands out as the top candidate due to its unmatched speed and strong accuracy metrics. For *low-miss applications* (e.g., autonomous driving), Deformable DETR offers unmatched LAMR but requires optimisation to improve speed and precision. RF-DETR presents the best *overall balance* of performance, generalisation, and efficiency, making it suitable for moderately real-time yet accurate deployment.

Figure 5.4: Predictions from (a) Faster RCNN model trained on WiderPerson dataset, (b) Faster RCNN model trained on EuroCity Persons dataset, (c) SSD trained on WiderPerson dataset, (d) SSD trained on EuroCityPersons Dataset, (e) YOLOv11 trained on WiderPerson dataset, (f) YOLOv11 trained on EuroCityPersons Dataset

Figure 5.5: Predictions from (a) DETR model trained on WiderPerson dataset, (b) DETR model trained on EuroCity Persons dataset, (c) Deformable DETR trained on WiderPerson dataset, (d) Deformable DETR trained on EuroCityPersons Dataset, (e) RF-DETR trained on WiderPerson dataset, (f) RF-DETR trained on EuroCityPersons Dataset

5.1.6 Key Takeaways

This evaluation reveals that no single model dominates across all metrics. YOLOv11 is preferable for fast and accurate inference, RF-DETR is a well-rounded transformer model, and Deformable DETR is unmatched in minimising missed detections. Meanwhile, SSD performed weakly as it failed to detect pedestrians across datasets. The training dataset also plays a critical role in generalisation, with WiderPerson providing better results in most conventional metrics, while EuroCityPersons occasionally leads in LAMR. Key takeaways from the training loss trends affirm that YOLOv11 and Faster R-CNN are highly stable and efficient across both datasets. SSD, despite improving loss-wise, did not translate learning into effective detection performance, revealing issues with generalisation. Transformer-based models, especially Deformable DETR and RF-DETR, showed promising convergence given more epochs, though initial training phases remain volatile. These findings support the evaluation metrics, reinforcing the conclusion that YOLOv11 is best suited for fast, stable learning, while Deformable DETR holds potential for high-accuracy, long-term optimisation.

Chapter 6

Conclusion and Future Work

6.0.1 Summary of Research Objectives and Contributions

This dissertation undertook a comprehensive benchmarking of CNN- and ViT-based object detection models for pedestrian detection, with a particular focus on generalisability across datasets. The study was motivated by the critical role of pedestrian detection in safety-critical applications such as autonomous driving, and it aimed to address persisting challenges posed by occlusion, scale variation, and dataset bias. By fine-tuning and evaluating six state-of-the-art object detection models—three CNN-based (Faster R-CNN, SSD, YOLOv11) and three ViT-based (DETR, Deformable DETR, RF-DETR)—across EuroCityPersons and WiderPerson datasets, and testing on the unseen INRIA dataset, this work provides insights into cross-domain performance and trade-offs between speed, accuracy, and robustness.

RQ1, which investigates how CNNs and ViTs compare in terms of accuracy and efficiency for pedestrian detection, is addressed through a uniform evaluation of all six models using standardised metrics such as precision, accuracy, inference time, and LAMR. This comparison revealed that while CNN-based models such as YOLOv11 offer higher inference efficiency, ViT-based models like Deformable DETR excel in reducing missed detections, particularly in densely crowded or occluded scenarios.

RQ2, concerning which model performs best overall, is answered by identifying that YOLOv11 trained on WiderPerson delivers the most favourable trade-off between

detection accuracy and speed, making it ideal for real-time deployment. Conversely, Deformable DETR trained on EuroCityPersons achieves the lowest LAMR, suggesting its strength in applications where minimising false negatives is critical.

RQ3, which examines the generalisation capabilities of CNNs and ViTs across datasets, is directly addressed through cross-dataset benchmarking. Models were trained on EuroCityPersons and WiderPerson, then evaluated on the INRIA dataset to simulate domain shift. The findings demonstrate that models trained on WiderPerson generally achieve higher precision and accuracy, whereas those trained on EuroCityPersons exhibit robustness, as reflected in lower LAMR values. This accentuates the importance of dataset characteristics in influencing generalisation performance.

Together, these findings contribute to a nuanced understanding of the design and deployment trade-offs between CNN and transformer-based architectures in pedestrian detection. They offer actionable guidance for selecting appropriate models depending on application requirements—whether speed, generalisation, or detection sensitivity is the priority.

6.0.2 Project Development and Implementation

The project commenced during a research internship, which provided the foundational understanding of deep learning architectures—specifically, the operational paradigms and trade-offs between Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). This foundation led to the implementation of object detection models during the interim period, with initial efforts focused on training Faster R-CNN and Mask R-CNN from scratch. However, the significant computational overhead and long training durations necessitated a strategic shift towards fine-tuning pre-trained models, which enabled more efficient experimentation within the project's time constraints.

6.0.3 Challenges and Limitations

Managing computational resources proved to be a central challenge throughout this dissertation. Although the university-provided high-performance computing (HPC)

facilities were initially considered, their storage limitations hindered the uploading of large-scale pedestrian datasets. As a result, the project shifted to Google Colab, which offered GPU acceleration and greater flexibility in managing data via multiple linked Google Drive accounts. Nonetheless, the limited compute time and storage on Colab constrained the scope of the experiments. In particular, the inability to train models at scale or to evaluate them on additional datasets—such as Caltech and CityPersons—restricted the breadth of conclusions that could be drawn regarding model adaptability.

6.0.4 Future Work and Research Directions

Future research should aim to expand the dataset variety and scale of evaluation to better understand model generalisability. Incorporating additional pedestrian benchmarks such as Caltech and CityPersons could yield more comprehensive insights into the representational strength of training datasets like WiderPerson and EuroCityPersons. Additionally, further experimentation on model robustness under varying levels of occlusion and crowd density could help delineate the comparative advantages of CNN- versus ViT-based architectures. With access to more extensive computational resources, subsequent studies can conduct deeper and broader evaluations, thereby building on this dissertation's empirical foundation to advance the design of scalable, generalisable pedestrian detection systems.

Bibliography

[1] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22(2):380–393, 2019.

[2] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1844–1861, 2019.

[3] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, pages 1582–1588, 2006.

[4] D. Du and Y. Xie. Vehicle and pedestrian detection algorithm in an autonomous driving scene based on improved yolov8. *Journal of Transportation Engineering, Part A: Systems*, 151(1):04024095, 2025.

[5] S. Hou, M. Yang, W. S. Zheng, and S. Gao. Multispectral transformer fusion via exploiting similarity and complementarity for robust pedestrian detection. *Pattern Recognition*, page 111383, 2025.

[6] K. Lu, C. Zhu, M. Liu, and X. C. Yin. Oss-ocl: Occlusion scenario simulation and occluded-edge concentrated learning for pedestrian detection. *Pattern Recognition Letters*, 2025.

[7] M. Bilal, A. Khan, M. U. K. Khan, and C. M. Kyung. A low-complexity pedestrian detection framework for smart video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(10):2260–2273, 2016.

[8] A. Raza, S. A. Chelloug, M. H. Alatiyyah, A. Jalal, and J. Park. Multiple pedestrian detection and tracking in night vision surveillance systems. In *CMC*, volume 75, pages 3275–3289, January 2023.

[9] U. Gawande, K. Hajari, and Y. Golhar. Pedestrian detection and tracking in video surveillance system: issues, comprehensive review, and challenges. *Recent Trends in Computational Intelligence*, pages 1–24, 2020.

[10] X. Huang, Z. Ge, Z. Jie, and O. Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10750–10759, 2020.

[11] Z. Ouardirhi, S. A. Mahmoudi, and M. Zbakh. Enhancing object detection in smart video surveillance: a survey of occlusion-handling approaches. *Electronics*, 13(3):541, 2024.

[12] C. Y. Tsai, R. Y. Wang, and Y. C. Chiu. Sw-yolox: A yolox-based real-time pedestrian detector with shift window-mixed attention mechanism. *Neurocomputing*, 606:128357, 2024.

[13] S. Paisitkriangkrai, C. Shen, and J. Zhang. Fast pedestrian detection using a cascade of boosted covariance features. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1140–1151, 2008.

[14] P. Doll'ar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311. IEEE, June 2009.

[15] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38:15–33, 2000.

[16] "O. Kaplan and E. Saykol. Comparison of support vector machines and deep learning for vehicle detection. In *RTA-CSIT*, pages 64–69, 2018.

[17] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata. Pedestrian detection with convolutional neural networks. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005*, pages 224–229. IEEE, June 2005.

[18] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 443–457. Springer International Publishing, 2016.

[19] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.

[20] C. Liu, H. Wang, and C. Liu. Double mask r-cnn for pedestrian detection in a crowd. *Mobile Information Systems*, 2022(1):4012252, 2022.

[21] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] C. F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021.

[23] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023.

[24] L. Meng, H. Li, B. C. Chen, S. Lan, Z. Wu, Y. G. Jiang, and S. N. Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[26] D. Lamichhane. Advanced detection of ai-generated images through vision transformers. *IEEE Access*, 2024.

[27] H. Chen, C. Li, G. Wang, X. Li, M. M. Rahaman, H. Sun, and M. Grzegorzek. Gashis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognition*, 130:108827, 2022.

[28] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, and Q. Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1):96–107, 2023.

[29] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[30] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz. Vision transformers in medical computer vision—a contemplative retrospection. *Engineering Applications of Artificial Intelligence*, 122:106126, 2023.

[31] H. Chen et al. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[32] L. Zhou et al. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018.

[33] J. Yuan, P. Barmpoutis, and T. Stathaki. Effectiveness of vision transformer for fast and accurate single-stage pedestrian detection. *Advances in Neural Information Processing Systems*, 35:27427–27440, 2022.

[34] B. Ghari, A. Tourani, A. Shahbahrami, and G. Gaydadjiev. Pedestrian detection in low-light conditions: A comprehensive survey. *Image and Vision Computing*, page 105106, 2024.

[35] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao. Pedestrian detection: Domain generalization, cnns, transformers and beyond. *arXiv preprint arXiv:2201.03176*, 2022.

[36] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao. From handcrafted to deep features for pedestrian detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4913–4934, 2021.

[37] L. G. Galvao, M. Abbod, T. Kalganova, V. Palade, and M. N. Huda. Pedestrian and vehicle detection in autonomous vehicle perception systems—a review. *Sensors*, 21(21):7267, 2021.

[38] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[39] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018.

[40] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4975, 2019.

[41] X. Huang, Z. Ge, Z. Jie, and O. Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10750–10759, 2020.

[42] B. Ghari, A. Tourani, A. Shahbahrami, and G. Gaydadjiev. Pedestrian detection in low-light conditions: A comprehensive survey. *Image and Vision Computing*, page 105106, 2024.

[43] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[44] C. C. Loy, D. Lin, W. Ouyang, Y. Xiong, S. Yang, Q. Huang, ..., and W. Zhou. Wider face and pedestrian challenge 2018: Methods and results. *arXiv preprint arXiv:1902.06854*, 2019.

[45] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.

[46] Q. Hoang, T. D. Nguyen, T. Le, and D. Phung. Mgan: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations (ICLR)*, 2018.

[47] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10639–10646, 2020.

[48] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[49] J. U. Kim, S. Park, and Y. M. Ro. Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1157–1165, 2022.

[50] S. Lee, T. Kim, J. Shin, N. Kim, and Y. Choi. Insanet: Intra-inter spectral attention network for effective feature fusion of multispectral pedestrian detection. *Sensors*, 24(4):1168, 2024.

[51] A. Althoupety, L. Y. Wang, W. C. Feng, and B. Rekabdar. Daff: Dual attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2997–3006, 2024.

[52] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[55] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[56] X. Zhang, H. Huang, D. Zhang, S. Zhuang, S. Han, P. Lai, and H. Liu. Cross-dataset generalization in deep learning. *arXiv preprint arXiv:2410.11207*, 2024.

[57] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[58] Seokju Lee. Hierarchical pedestrian detection using omnidirectional camera and laser sensor fusion. 2015.

[59] Sungjoon Park, Hyunwoo Kim, and Yong Man Ro. Robust pedestrian detection via constructing versatile pedestrian knowledge bank. *Pattern Recognition*, 153:110539, 2024.

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[61] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, 2018.

[62] Peng Sun, Ruijie Zhang, Yuxin Jiang, Tao Kong, Chenfeng Xu, Weijia Zhan, others, and Ping Luo. Sparse r-cnn: An end-to-end framework for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15650–15664, 2023.

[63] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[65] Xiaotong Zhao, Wei Li, Yifang Zhang, T Aaron Gulliver, Shuo Chang, and Zhiyong Feng. A faster rcnn-based pedestrian detection system. In *2016 IEEE 84th vehicular technology conference (VTC-Fall)*, pages 1–5. IEEE, 2016.

[66] Yeong-Hyeon Byeon and Keun-Chang Kwak. A performance comparison of pedestrian detection using faster rcnn and acf. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 858–863. IEEE, 2017.

[67] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander 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, pages 21–37. Springer, 2016.

[68] Yongren Cheng, Changhong Chen, and Zongliang Gan. Enhanced single shot multibox detector for pedestrian detection. In *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, pages 1–7, 2019.

[69] Chaoqi Yan, Hong Zhang, Xuliang Li, and Ding Yuan. R-ssd: Refined single shot multibox detector for pedestrian detection. *Applied Intelligence*, 52(9):10430–10447, 2022.

[70] Chang Sun, Yibo Ai, Xing Qi, Sheng Wang, and Weidong Zhang. A single-shot model for traffic-related pedestrian detection. *Pattern Analysis and Applications*, 25(4):853–865, 2022.

[71] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

[72] Hang Sui, Huiyan Han, Yuzhu Cui, Menglong Yang, and Binwei Pei. Fa-yolo: A pedestrian detection algorithm with feature enhancement and adaptive sparse self-attention. *Electronics*, 14(9):1713, 2025.

[73] Yadong Li, Hui Yan, Dan Li, and Hongdong Wang. Robust miner detection in challenging underground environments: An improved yolov11 approach. *Applied Sciences*, 14(24):11700, 2024.

[74] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[75] Tianyong Wu, Xiang Li, and Qiuxuan Dong. An improved transformer-based model for urban pedestrian detection. *International Journal of Computational Intelligence Systems*, 18(1):68, 2025.

[76] Matthieu Lin, Chuming Li, Xingyuan Bu, Ming Sun, Chen Lin, Junjie Yan, Wanli Ouyang, and Zhidong Deng. Detr for crowd pedestrian detection. *arXiv preprint arXiv:2012.06785*, 2020.

[77] Feng Gao, Jiaxu Leng, Ji Gan, and Xinbo Gao. Rc-detr: Improving detrs in crowded pedestrian detection via rank-based contrastive learning. *Neural Networks*, 182:106911, 2025.

[78] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[79] Wenjing Han, Ning He, Xin Wang, Fengxi Sun, and Shengjie Liu. Idpd: Improved deformable-detr for crowd pedestrian detection. *S@articlehan2024idpd, title=IDPD: Improved deformable-DETR for crowd pedestrian detection, author=Han, Wenjing and He, Ning and Wang, Xin and Sun, Fengxi and Liu, Shengjie, journal=Signal, Image and Video Processing, volume=18, number=3, pages=2243–2253, year=2024, publisher=Springer ignal, Image and Video Processing*, 18(3):2243–2253, 2024.

[80] Su Deng and Jinping Li. Efficient dense pedestrian detection based on transformer. In *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 7, pages 992–997. IEEE, 2024.

[81] Jing Yuan, Panagiotis Barmpoutis, and Tania Stathaki. Multi-scale deformable transformer encoder based single-stage pedestrian detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2906–2910. IEEE, 2022.

[82] Isaac Robinson, Peter Robicheaux, and Matvei Popov. Rf-detr. https://github.com/roboflow/rf-detr, 2025. SOTA Real-Time Object Detection Model.