University of Nottingham
UK | CHINA | MALAYSIA

UG FINAL YEAR DISSERTATION

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

Good morning, I will be presenting my individual dissertation titled "generalisation in pedestrian detection: a comprehensive benchmarking of convolutional neural networks and vision transformers across diverse datasets".

Problem Statement

https://pythonawesome.com/an-open-source-framework-built-on-top-of-tensorflow-and-keras/
https://tenor.com/view/running-fast-pedestrian-passing-by-spiderman-gif-12678036
https://www.scoopwhoop.com/humour/really-lucky-people-on-earth/

Pedestrian detection is a critical task in computer vision with applications in autonomous driving and surveillance systems. Despite its importance for safety and accident prevention, achieving reliable detection in real-world scenarios is difficult due to challenges like motion blur from video data, frequent occlusions, variations in pedestrian scale, and bias across different datasets.

So let's look at the problem statement, pedestrian detection is a key task in computer vision, especially in areas like autonomous driving and video surveillance where safety is a major concern. For these systems to work effectively, they need to detect pedestrians accurately in a wide range of real-world conditions. But this is quite challenging. We often have to deal with motion blur from video frames, pedestrians being partially blocked by objects, differences in scale depending on distance, and even inconsistencies between datasets. These factors make it difficult to develop models that are both accurate and generalise well across different scenarios.

Keywords

Generalisation: The ability of a model to perform well on unseen data, not just the data it was trained on.

Convolutional Neural Networks (CNNs): Deep learning models that use convolutional layers to extract spatial features from images for tasks like detection or classification.

Vision Transformers (ViTs): Transformer-based models that process images as sequences of patches, capturing global context without convolution.

Transfer Learning: A technique where a model trained on one task is reused or fine-tuned for a different but related task to improve learning efficiency.

Before we get to the aims and objectives, let me clarify some key terms that are significant to this study.

Generalisation refers to a model's ability to perform well on data it hasn't seen during training. It's a core challenge in pedestrian detection, especially under domain shifts.

Convolutional Neural Networks, or **CNNs**, are deep learning models that use convolutional layers to extract spatial features from images. They've long been dominant in visual tasks like classification and object detection.

Vision Transformers, or **ViTs**, take a different approach. Instead of convolutions, they treat images as sequences of patches, allowing the model to capture global context more effectively. This architecture has shown promise, especially in intra-dataset scenarios.

And finally, **Transfer Learning** is a method where a pre-trained model is fine-tuned for a different but related task. It helps improve training efficiency and performance.

Aims and Objectives

This dissertation investigates the generalisability and performance of CNNs and ViTs in pedestrian detection. It involves:

- Fine-tuning and evaluating three models from each architecture on multiple pedestrian detection datasets.
- Assessing cross-dataset performance to examine generalisability.
- Comparing the accuracy and efficiency of CNNs and ViTs in real-world, domain-shifted scenarios.

For the aims and objectives, this dissertation focuses on comparing the generalisability and performance of CNNs and ViTs in pedestrian detection. The main goal is to fine-tune 3 representative models from each architecture across multiple datasets, evaluate and assess their performance on unseen data. By doing this, the dissertation aims to understand how well these models generalise across different domains.

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?

These are the key research questions guiding this dissertation.
1.How do CNNs and ViTs compare in terms of both accuracy and efficiency for pedestrian detection tasks?
2.Which specific model offers the best overall performance?
3.How well do CNNs and ViTs generalise across datasets when trained on one and tested on another?"

Related Work

Theme	Key Points	Source(s)
CNNs vs. ViTs Generalisation	ViTs perform better in intra-dataset benchmarks, while CNNs show stronger cross-dataset generalisation due to better domain robustness.	Hasan et al., Galvao et al.
Dataset Bias & Overfitting	Many detectors are tuned for specific datasets, leading to poor generalisation. ViTs are particularly sensitive to domain shift.	Hasan et al., Galvao et al., Dollar et al.
Occlusion Handling	Occlusion remains a major challenge. Part-based CNN models improve robustness but add computational cost.	Cao et al., Hasan et al.
Scale Variance	Small or distant pedestrians are harder to detect. Multi-scale strategies exist but generalisability across datasets is still limited.	Galvao et al., Cao et al., Hasan et al.
Low-Light & Night Conditions	Fusion-based CNN models (e.g., MSR, INSANet) help in low-light but still underperform in real-world scenarios; ViTs not widely benchmarked here.	Ghari et al., Hasan et al., Brunetti et al.
Real-Time Deployment	Lightweight CNNs offer speed but degrade in complex scenes. High-performing models often lack efficiency for real-time or embedded use.	Galvao et al., Brunetti et al.
Evaluation Standards	Current benchmarking practices are inconsistent. Unified evaluation protocols and more diverse datasets are needed for meaningful generalisability claims.	Hasan et al., Galvao et al., Dollar et al.

Table 1: Summary of key themes and insights from literature on pedestrian detection generalisation

This slide summarises key insights from recent surveys that examine generalisability in CNN and ViT-based pedestrian detectors.
Starting with **generalisation**, studies highlight that **ViTs excel in intra-dataset settings**, where training and testing are on the same domain. but, **CNNs demonstrate stronger cross-dataset performance**, showing better domain robustness unlike ViTs.
On the theme of **dataset bias and overfitting**, multiple works point out that **many models are over-tuned to specific datasets**. **ViTs are especially vulnerable** to domain shifts in this regard.
For **occlusion handling**, that **some CNN models improve robustness in crowded or occluded scenes**, though this comes at the cost of increased computational complexity.
Scale variance—particularly detecting **small or distant pedestrians**—is another major challenge. While multi-scale features help, generalisability across datasets remains limited.
In **low-light and nighttime conditions**, fusion-based CNNs show some promise, but still **underperform in real-world scenarios**. Interestingly, **ViTs have not yet been extensively benchmarked** under these lighting conditions.
When it comes to **real-time deployment**, **lightweight CNNs offer speed**, but

often falter in complex environments. High-performing models tend to lack the efficiency required for embedded systems or real-time use.
Lastly, **evaluation standards** are a recurring concern. Several authors stress the need for **unified benchmarks and more diverse datasets** to make generalisation claims more meaningful and reliable.

Proposed Approach: Transfer Learning

Use knowledge from pre-trained models to improve learning on new tasks. This is useful because training from scratch is impractical due to limited data or high computational costs.

- Use pre-trained weights from large datasets (e.g., ImageNet, MS COCO, Open Images V4).
- Early layers retain generic features (edges, textures); later layers are fine-tuned for pedestrian-specific patterns.

Benefits

- Avoids high computational costs of training from scratch.
- Enables efficient domain adaptation and optimised performance.
- Reduces the need for extensive hyperparameter tuning.

Let's look at the proposed approach which is transfer learning.
It is a common and effective strategy in deep learning. Instead of training from scratch, which can be both resource-intensive and data-hungry, we start with models pre-trained on large, diverse datasets like ImageNet, COCO.
The early layers of these pre-trained models capture general visual features like edges and textures, which are still useful for identifying pedestrians. We then fine-tune the later layers to adapt the model specifically to the nuances of detecting people in real-world environments.
This not only saves computational time but also reduces the complexity of retraining models and tuning hyperparameters.

Model Selection

CNN Models	**Faster R-CNN** High accuracy and robustness (92.7% on INRIA - Zhang et al.). Modular and adaptable to different backbones (Zhao et al.). 56.73% higher precision than ACF (Byeon et al.). **Selected for:** Precision-critical performance and versatility.	**SSD** Enhanced for dense targets and real-time (Cheng et al.). Feature fusion improved small-scale detection (Yan et al.). Attention-boosted SSD for efficiency (Sun et al.). **Selected for:** Real-time speed and scalable design.	**YOLOv11** Feature-enhanced and robust in low light (Sui et al., Li et al.). Real-time and adaptable to complex environments. **Selected for:** Promising early performance and speed.
ViT Models	**DETR** Improved for crowded scenes (Wu et al.). Custom decoders and fusion for dense detection (Lin et al.). Contrastive learning improved crowd handling (Gao et al.). **Selected for:** Attention-based architecture in complex scenes.	**Deformable DETR** Achieved SOTA on CrowdHuman (Han et al.). Fine-grained features and dynamic necks (Deng et al., Yuan et al.). **Selected for:** Occlusion handling and strong high-density performance.	**RF-DETR** First real-time transformer to exceed 60 AP on COCO. Strong cross-domain generalisation (RF100-VL). **Selected for:** Emerging potential in real-time ViT detection.

The rationale for selecting the models are as follows:
Faster R-CNN was chosen for its high accuracy and robustness. Proved to be ideal for precision-critical detection in diverse environments.
SSD (Single Shot MultiBox Detector): was selected for its real-time performance and adaptability, with improvements for dense and small-scale pedestrian detection.
YOLOv11 was chosen for its real-time capabilities, and robustness in occlusion-heavy and low-light scenarios.
DETR (short for detection transformer) was selected for its innovative attention-based architecture and success in dense, crowded scenes.
Deformable DETR was chosen for its flexibility in handling occlusion, scale variance, and dense crowds. It obtained state-of-the-art performance in high-density pedestrian benchmarks.
RF-DETR was selected as it is a newly proposed variant of the DETR model, which proved to perform better and reached the highest mean average precision on the MS COCO benchmark. It was also chosen because it is not yet explored in the current literature.

Dataset Selection

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

First, **WiderPerson** is selected as a primary training dataset due to its **dense annotations and high occlusion levels**, as seen in the sample here. These characteristics make it especially suitable for training models aimed at detecting pedestrians in **crowded urban environments**.

Dataset Selection

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

Next for training, the **EuroCity Persons** dataset complements WiderPerson by introducing **geographic and temporal diversity** as you can see here. It covers various European cities, seasons, and lighting conditions. This dataset helps expose models to **less densely populated scenes.**

Dataset Selection

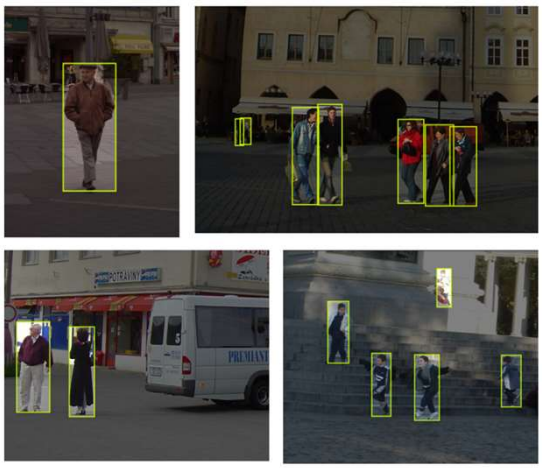


Figure 3.3: Annotated samples from the INRIA dataset [3]

INRIA dataset was selected as the test set. It is recognised for its **moderate annotation density** and **diverse background scenes**, making it a reliable benchmark for assessing generalization.

Performance Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.2)$$

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

		Ground Truth: Positive	Ground Truth: Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
Prediction	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).

Here are some of the widely adopted **evaluation metrics in object detection used in this study:**

Accuracy measures the overall proportion of correct predictions out of all samples.

Precision indicates the percentage of predicted positive detections that are actually correct.

Recall shows how many of the actual positives the model successfully identified.

To balance precision and recall, the **F1 Score** is used—it's the harmonic mean of the two.

Performance Metrics

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

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).

Inference Time: for this evaluation, inference time is measured by calculating the average time taken to process a batch of images after finetuning the model.

Another important metric is the **Log-Average Miss Rate**, or **LAMR**, which captures the model's performance across various false positive per image (FPPI) thresholds. It calculates the miss rate—(1-recall)—at multiple confidence levels, applies a logarithmic scale, and averages the results. Lower the LAMR, less missed predictions, better the performance.
Inference Time is also measured to assess how quickly the model produces predictions after fine-tuning.

Performance Metrics

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

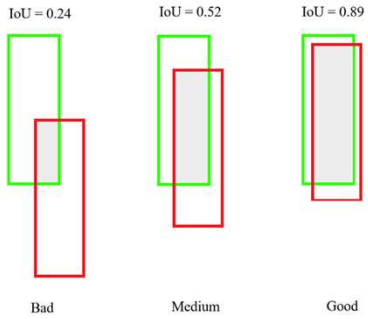


Figure 3.5: Examples of IoU thresholds on predicted and ground truth boxes.

Intersection over Union, or **IoU**, evaluates how well the predicted bounding boxes align with ground truth annotations. A value of 0.5 or above typically indicates a successful detection.

Experimental Setup

Dataset Preparation

- WiderPerson dataset fully used (training and validation); EuroCityPersons daytime validation subset used with an 80:20 split.
- Annotations converted to COCO format; ignore regions removed.
- YOLO-compatible YAML files created.

Training Protocol

- Models trained for 5 epochs on each dataset.
- No data augmentation applied to ensure fair model comparison.
- Hyperparameters tuned per model to fit Colab GPU constraints (see Table 4.1).

Evaluation

Final models tested on 20 diverse INRIA images, covering occlusion, scale, and density variations.

Environment

Google Colab with Tesla T4 GPU (16GB memory, 6MB L2 cache).

In terms of training, WiderPerson was fully used, and only daytime validation subset of EuroCityPersons was available due to storage limits. This subset was split 80:20 for training and validation.
All annotations were converted to COCO format and compatible config files were created for YOLO-based models.
For controlled comparison, no data augmentation was used—so the differences in performance reflect model architecture alone.
Each model was fine-tuned for five epochs.
For cross-dataset evaluation, each model was tested on a curated subset of 20 INRIA images capturing diverse pedestrian scenarios such as occlusion and crowding.
Finally, training and evaluation were conducted entirely in Google Colab, using a Tesla T4 GPU with 16GB of memory.

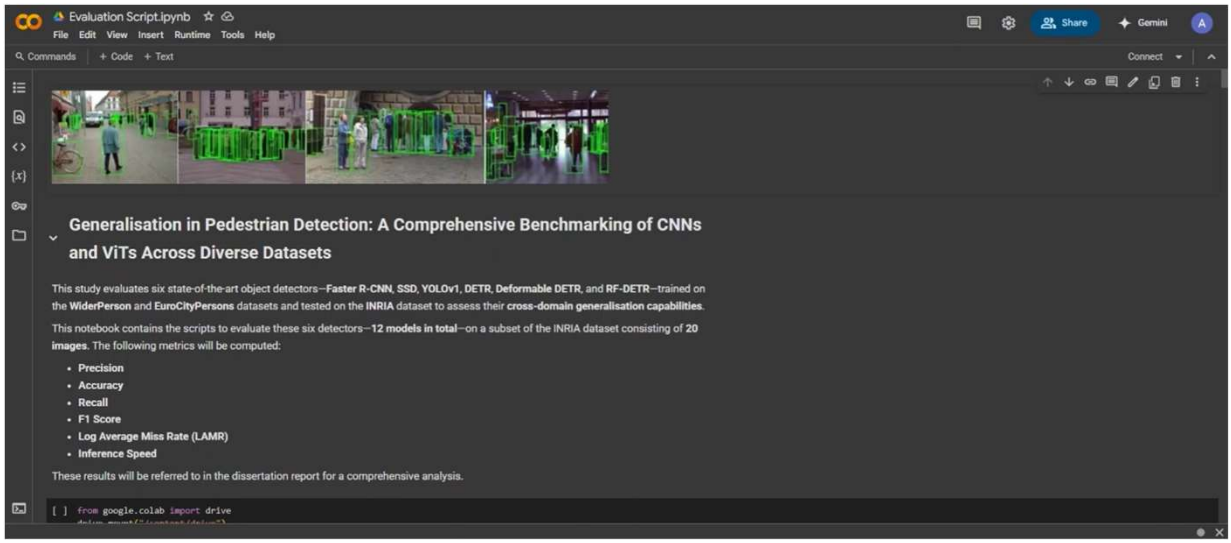
Experimental Setup

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

Here is the training-set up, the hyperparameters chosen here are optimised for running in the Google Colab to keep up with the RAM and session limit.

Demonstration



Results: Training Convergence Trends

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

Here are the **training losses** over five epochs.
YOLOv11 consistently reduced loss, showing efficient convergence on both datasets.
Faster R-CNN achieved the **lowest final loss** on EuroCityPersons at 0.4505.
Deformable DETR also converged significantly, particularly on EuroCityPersons, dropping from 9.14 to 1.06.
RF-DETR demonstrated steady but slower improvement on both datasets.
In contrast, **SSD** faced instability and began with the **highest initial loss**, only moderately improving.

Results: Detection Performance Metrics

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

YOLOv11, trained on the WiderPerson dataset, delivered the best overall performance. It achieved the highest **Precision, F1 Score**, and **Accuracy.** Its **Recall** was also competitive at 84.54%.
RF-DETR, trained on EuroCityPersons, stood out with the highest F1 Score for transformer-based models.
Faster R-CNN achieved the highest **Recall** overall, but its **Precision** and **Accuracy** dropped significantly when trained on EuroCityPersons, meaning poor generalisation.
In contrast, **DETR** and **Deformable DETR** showed low **Precision** and **Accuracy.**
Finally, **SSD** failed to generalize completely, with extremely poor metric values due to its bounding boxes being too small to register in evaluations. This underperformance was consistent across both datasets.

Results: Efficiency and Miss Rate Analysis

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

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

Deformable DETR recorded the **lowest LAMR**, making it ideal for applications where minimal missed detections are critical. However, it had the **highest inference time**—22.77 seconds per image—and low precision, limiting its practical use.

RF-DETR offered a strong balance. With a low LAMR of 0.0722 and an inference time of 2.56 seconds, it emerged as an efficient and accurate transformer-based model.

YOLOv11, on the other hand, had the **fastest inference time**, just 0.31 seconds per image. Though its LAMR was slightly higher at around 0.16, it was still acceptable for real-time use cases.

Faster R-CNN had a high LAMR of 0.0825 and slow inference at 9 seconds, reflecting its limitations in time-sensitive settings.

SSD, while relatively fast at 1.4 seconds, had extremely high LAMR values—close to or equal to 1—making it unsuitable for cross-dataset detection tasks.

Results: Detection Performance Metrics

And here are the illustrated performance metrics from the table of results. Interestingly, models trained on **WiderPerson** outperformed those trained on EuroCityPersons in **Precision** and **Accuracy**, likely due to the dataset's higher pedestrian density and annotation quality. However, for **LAMR**, models trained on EuroCityPersons—especially RF-DETR and Deformable DETR—performed better, suggesting robustness against false positives.

Figure 1: Illustrative results from CNN detectors(WiderPerson, EuroCityPersons): Faster RCNN, SSD, YOLOv11.

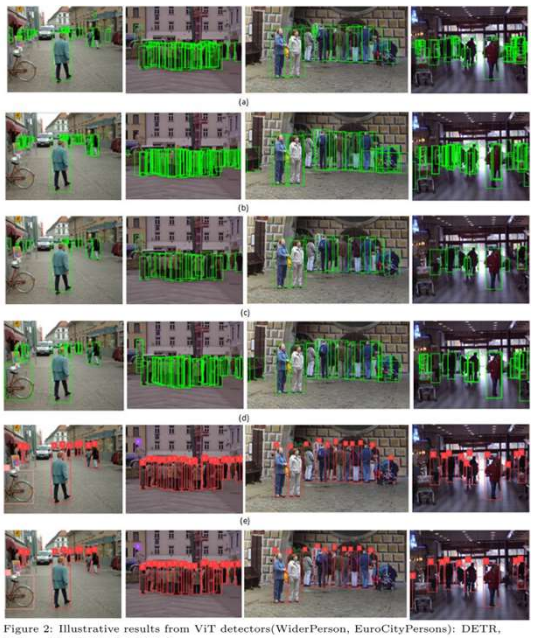


Figure 2: Illustrative results from ViT detectors(WiderPerson, EuroCityPersons): DETR, Deformable DETR, RF-DETR.

Here are the predicted images from each model.

Contributions

This dissertation supports model selection for real-world use, aids the development of robust pedestrian detection systems, and adds to existing research by evaluating deep models' performance and generalisability across datasets.

1. **Cross-Dataset Benchmarking**
   - Evaluated 3 CNNs and 3 ViT models, i.e. Faster R-CNN, SSD, YOLOv11, DETR, Deformable DETR, RF-DETR
   - Trained on WiderPerson & EuroCityPersons, tested on INRIA
2. **CNN vs. ViT Comparison**

   Uniform evaluation of CNNs vs. ViTs in terms of accuracy, generalisation, and speed

There are five key contributions of this dissertation:
First, it provides a comprehensive benchmarking of six state-of-the-art detectors—trained on 2 different datasets, and evaluated on an unseen dataset to test cross-domain generalisation.
Second, it offers a fair, side-by-side comparison of CNN-based and transformer-based models, assessing accuracy, inference time, and generalisation under consistent conditions.

Contributions

3. **Optimal Trade-Offs**
   - YOLOv11 (WiderPerson): Best speed-accuracy balance
   - Deformable DETR (EuroCityPersons): Lowest LAMR (0.0412), fewer missed detections
4. **Dataset Impact**
   - WiderPerson: Higher precision & accuracy
   - EuroCityPersons: Lower LAMR, better for reducing false positives
5. **Training Dynamics**
   - YOLOv11 & Faster R-CNN: Stable, efficient training
   - RF-DETR & Deformable DETR: Volatile start, stronger long-term potential

Third, it identifies practical trade-offs. YOLOv11 trained on WiderPerson showed the best speed-accuracy balance, while Deformable DETR trained on EuroCityPersons had the lowest Log Average Miss Rate of 0.0412, making it better at avoiding missed detections.
Fourth, the results reveal dataset influence—WiderPerson led to higher accuracy and precision, while EuroCityPersons helped reduce false positives, as shown by lower LAMR.
And finally, the training behaviour of the models was analysed. YOLOv11 and Faster R-CNN converged efficiently with stable loss curves, while RF-DETR and Deformable DETR showed initial volatility but promising in long-term performance.

Conclusion

- Provided a systematic benchmark of CNN vs. ViT object detectors for pedestrian detection under domain shift.
- Identified key trade-offs between speed, accuracy, and generalisability across real-world urban datasets.
- Offered practical guidance for model selection based on deployment needs (e.g., real-time use vs. precision-critical applications).
- Emphasised the impact of dataset characteristics on generalisation performance, supporting better dataset curation decisions.
- Established a baseline for cross-domain pedestrian detection, contributing to safer AI in autonomous systems.

To conclude, this dissertation delivered a systematic benchmarking of CNN and Vision Transformer-based object detectors, specifically evaluating how well they perform under domain shifts which is a crucial concern in real-world pedestrian detection scenarios.
The findings offer actionable guidance for model selection: for instance, real-time applications might benefit more from models like YOLOv11, whereas precision-critical systems could use transformer-based models like Deformable DETR or RF-DETR.
Additionally, the study reinforced how dataset characteristics — such as density, occlusion, and geographic variation — directly affect a model's generalization. This guides future dataset curation strategies.
Ultimately, this work contributes a valuable performance baseline for cross-domain pedestrian detection, with the long-term aim of supporting safer and more reliable AI systems in safety-critical domains.

Q&A Session

Thank You!