Investigating the Role of Learning Using Privileged Information in Object Detection

Matthias Bartolo

Supervisor: **Dr Dylan Seychell**
Co-Supervisor: **Dr Konstantinos Makantasis**

L-Università ta' Malta
Faculty of Information & Communication Technology

Department of Artificial Intelligence

Preface

Image Attribution

Unless otherwise specified, all images not cited in this presentation will be considered the author's own work.

Abbreviations

Deep Learning (DL)
Learning Using Privileged Information (LUPI)
Unmanned Aerial Vehicle (UAV)

Overview

- Introduction
- Background & Literature Review
- Methodology
- Evaluation
- Conclusion

Investigating the Role of Learning Using Privileged Information in Object Detection

Matthias Bartolo

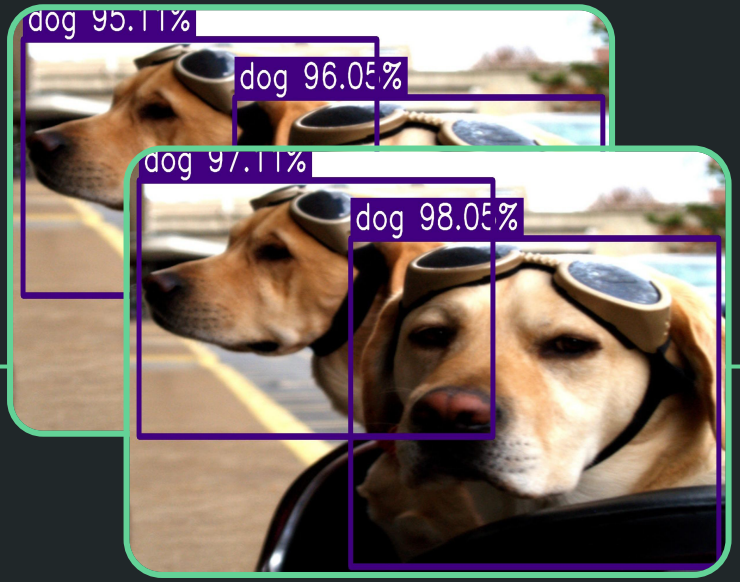
Supervisor: Dr. Dylan Seychell
Co-Supervisor: Dr. Konstantinos Makantasis

September 2025

Submitted in partial fulfilment of the requirements for the degree of Master of Science in ICT.

L-Università ta' Malta
Faculty of Information & Communication Technology

Introduction



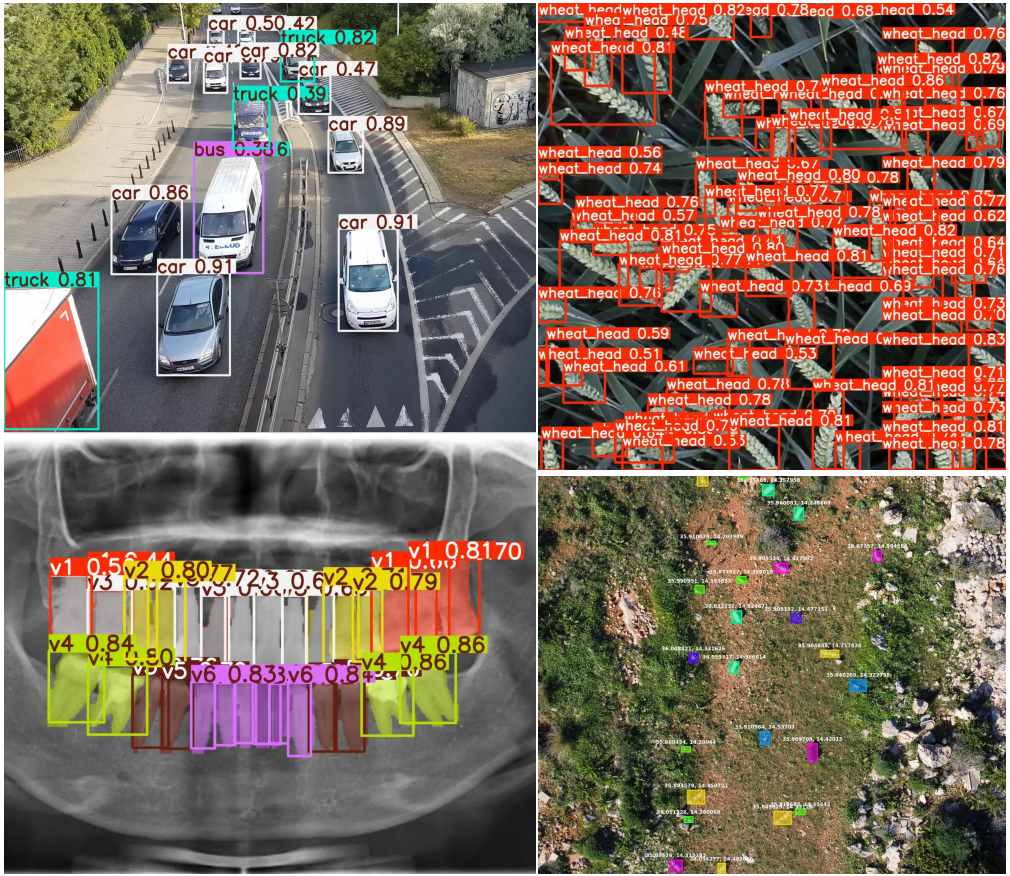
Object detection is considered a foundational problem within computer vision. . .

Object Detection

Having applications ranging from:

1. Medical analysis
2. Autonomous systems
3. Environmental monitoring
4. Security and surveillance
5. Robotics and automation

Amongst many others. . .



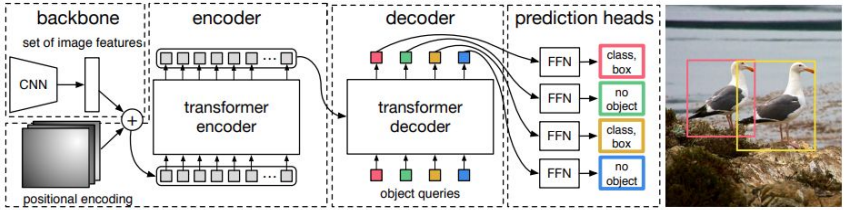
Object Detection is not a Trivial Problem

Incurring challenges such as:

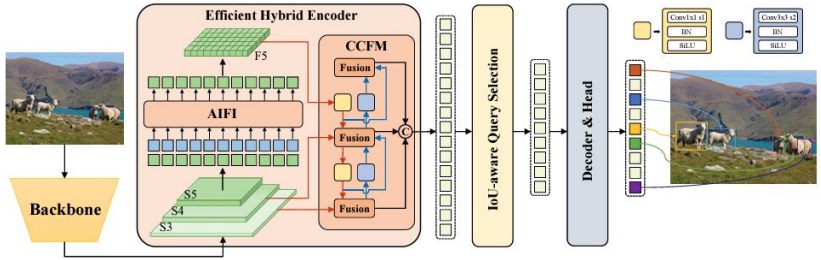
1. Complex backgrounds and interferences
2. Scale variability
3. Object occlusion
4. Class imbalance and dataset bias
5. Small object detection

Achieving High Detection Accuracy

1) Computationally Intensive Architectures



DETR Architecture (Source: [1])



RT-DETR Architecture (Source: [2])

2) Extensive Labelling Demands

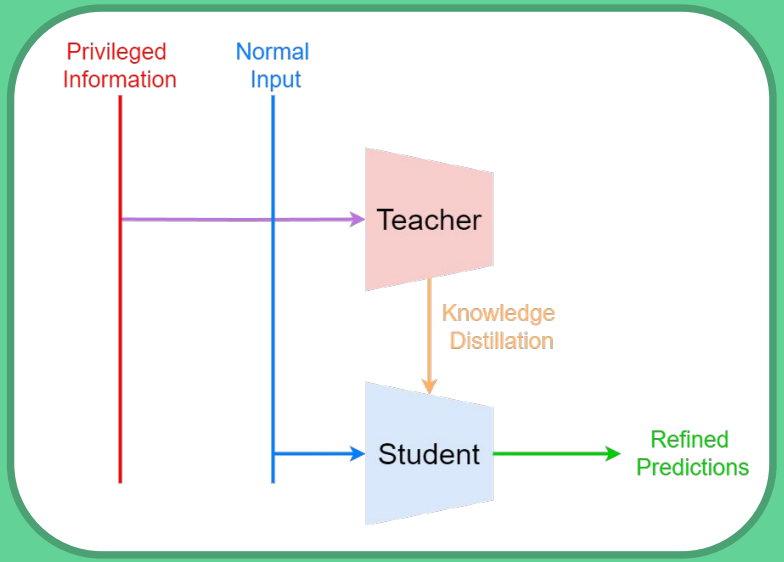
One annotated image ≈ 0.25 hrs,
≈ €5, varies by task. . .

Our Proposed Solution:

"Can **Learning Using Privileged Information (LUPI)** be effectively integrated into object detection models, and if so, how feasible and generalisable is such an approach?"

A brief definition of LUPI:
Information not available during testing, used only in training

Inspired from the Japanese proverb:
"Better than a thousand days of diligent study is one day with a great teacher"



Aims and Objectives

Objective 1 (O1) Develop a methodology for integrating LUPI into object detection for litter detection

Objective 2 (O2) Evaluate its adaptability and performance across renowned detection architectures

Objective 3 (O3) Test the approach on recognised litter datasets, within and across dataset evaluation

Objective 4 (O4) Analyse accuracy vs. computational cost, and assess wider applicability

Publications

The key milestones of this study, including research on object detection, were published in internationally peer-reviewed conferences and journals:

[1] M. Bugeja, M. Bartolo, M. Montebello, and D. Seychell, “MRTMD: A Multi-Resolution Dataset for Evaluating Object Detection in Traffic Monitoring Systems,” *IEEE Access*, vol. 13, pp. 134460–134483, 2025. doi: 10.1109/ACCESS.2025.3585986.

[2] M. Bartolo, K. Makantasis, and D. Seychell, “Learning Using Privileged Information for Litter Detection,” in *Proc. 2025 13th European Workshop on Visual Information Processing (EUVIP)*, 2025.

IEEE Access®
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

EUVIP

IEEE®

Background & Literature Review

Overview of Relevant Background

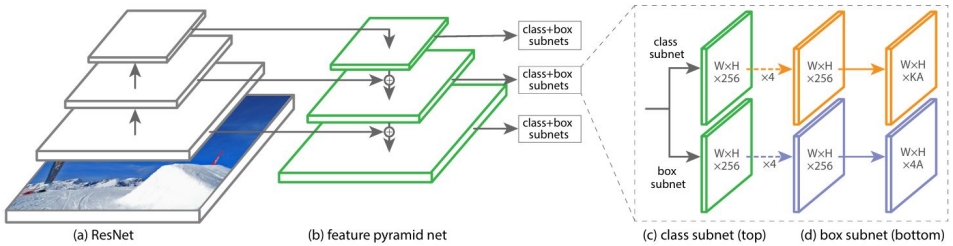
Background & Literature Review will consist of:

1) Deep learning object detection approaches
2) UAV-based litter detection methods
3) LUPI in computer vision
4) Evaluation metrics

Deep Learning Object Detectors

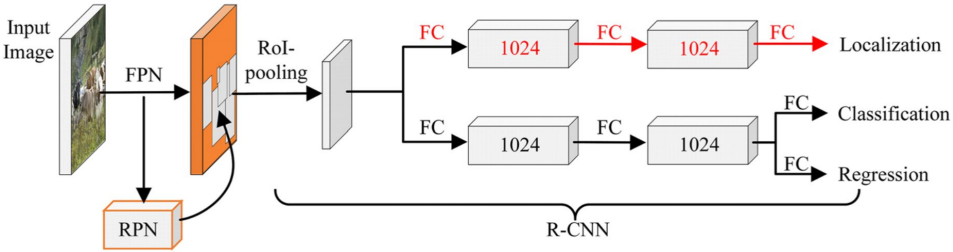
- **One-Stage Detectors**
  - Overfeat, YOLO, YOLO-World, SSD, RetinaNet, SSDLite, FCOS
- **Two-Stage Detectors**
  - R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, R-FCN, EfficientDet
- **Transformer-Based Detectors**
  - DETR, DINO, Grounding DINO, RT-DETR, PaliGemma, Florence-2
- **Other DL Approaches**
  - Reinforcement Learning, DCN, CenterNet, SAHI

e.g.



RetinaNet Architecture (Source: [4])

e.g.



Faster R-CNN Architecture (Source: [5])

(Image Source: [3])

Why UAV-based Litter Detection?

- **Application of detectors:** evaluating deep learning methods in real-world environments
- **Small, varied objects:** presenting challenges for standard object detection
- **Cluttered, complex scenes:** extending models beyond conventional datasets
- **Practical deployment:** lightweight models suitable for UAVs and real-time use

(Source: [1])

UAV-based Litter Detection Methods

Name	Year	No of. Images	UAV	AGL Altitudes	Dataset Details	No of. Categories	Available
BDW Dataset [17]	2018	25,407	Yes	10m–30m	Detection	1 (Litter)	Yes
UM Geo. Survey [112]	2018	472	Yes	30m	Data Collection	5 (Litter)	No
SuperDock [114]	2019	100	Yes	5m–10m	Detection	1 (Litter)	No
Styrofoam Monitoring [115]	2019	N/S[1]	Yes	15m	Detection, Segmentation	1 (Litter)	No
Small Litter Detection [110]	2019	744	Yes	5m–10m	Detection	1 (Litter)	No
TACO Dataset [7]	2020	1,500	No	N/A[2]	Detection, Segmentation	60 (Litter) [28 Super]	Yes
MJU-Waste Dataset [118]	2020	2,475	No	N/A[2]	Segmentation	1 (Litter)	Yes
UAVVaste Dataset [19]	2021	772	Yes	low-altitude	Detection, Geolocation	1 (Litter)	Yes
ZeroWaste Dataset [8]	2022	10,715	No	N/A[2]	Detection, Segmentation	4 (Litter)	Yes
PlasOPol Dataset [18]	2022	2,418	No	N/A[2]	Detection	1 (Litter)	Yes
HAIDA Dataset [124]	2022	1,319	Yes	1m–10m	Detection, Geolocation	2 (Litter)	Yes
Bangladeshi Dataset [125]	2023	4,418	No	N/A[2]	Detection	10 (Litter)	No
Beach Litter Dataset [126]	2023	4,126	Yes	10m–60m	Detection, Geolocation	67 (Litter) [7 Super]	No
TrashNet [127]	2024	2,524	No	N/A[2]	Detection	6 (Litter)	Yes
SODA Dataset [13]	2024	829	Yes	1m, 5m–30m	Detection, Segmentation	6 (Litter) [4 Super]	Yes

Annotations Unavailable

Summary of datasets and approaches for litter detection, including UAV and non-UAV imagery

LUPI Approaches in Computer Vision

1) **LUPI in Object Localisation**
   - Feyereisl et al. *"Object localization based on structural svm using privileged information"* **(2014)**
   - Sun et al. *"A new method for structured learning with privileged information"* **(2018)**
2) **LUPI in Computer Vision**
   - Sharmanska et al. *"Learning to rank using privileged information"* **(2013)**
   - Sharmanska et al. *"Learning to transfer privileged information"* **(2014)**
   - Wang et al. *"Learning with privileged information for multi-label classification"* **(2018)**
3) **Knowledge Distillation in Computer Vision**
   - Zheng et al. *"Localization distillation for object detection"* **(2022)**
   - Habib et al. *"A comprehensive review of knowledge distillation in computer vision"* **(2024)**

Evaluation Metrics

1) Intersection over Union (IoU)
2) Precision
3) Recall
4) F1 Score
5) Average Precision (AP)
6) Mean Average Precision (mAP)
7) Mean Average Recall (mAR)
8) Confusion Matrix

$$\text{IoU}(b, g) = \frac{\text{area}(b \cap g)}{\text{area}(b \cup g)}.$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{IoU}(b, g) > \text{threshold}}{\text{IoU}(b, g) > \text{threshold} + \text{FP}}.$$

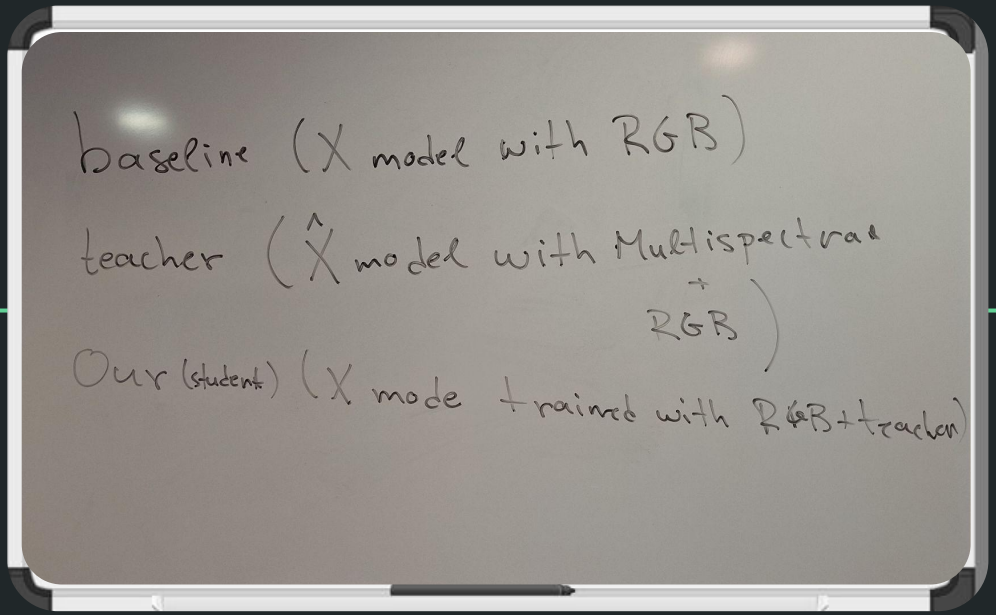
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{IoU}(b, g) > \text{threshold}}{\text{IoU}(b, g) > \text{threshold} + \text{FN}}.$$

$$\text{Average Precision (AP)} = \sum_{k=0}^{n} (R_k - R_{k-1}) \cdot P_k.$$

$$\text{Mean Average Precision (mAP)} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i.$$

Methodology

“Everything begins with an idea.”
– Earl Nightingale



Overview of Problem Definition & Proposed Approach

Problem

- Object detection decomposed into ***localisation*** and ***classification***
- **Outputs:** *bounding boxes* coupled with *categorical labels*
- **UAV litter detection poses challenges:** small object sizes, cluttered backgrounds, high variability
- **Objective:** improve detection performance through LUPI

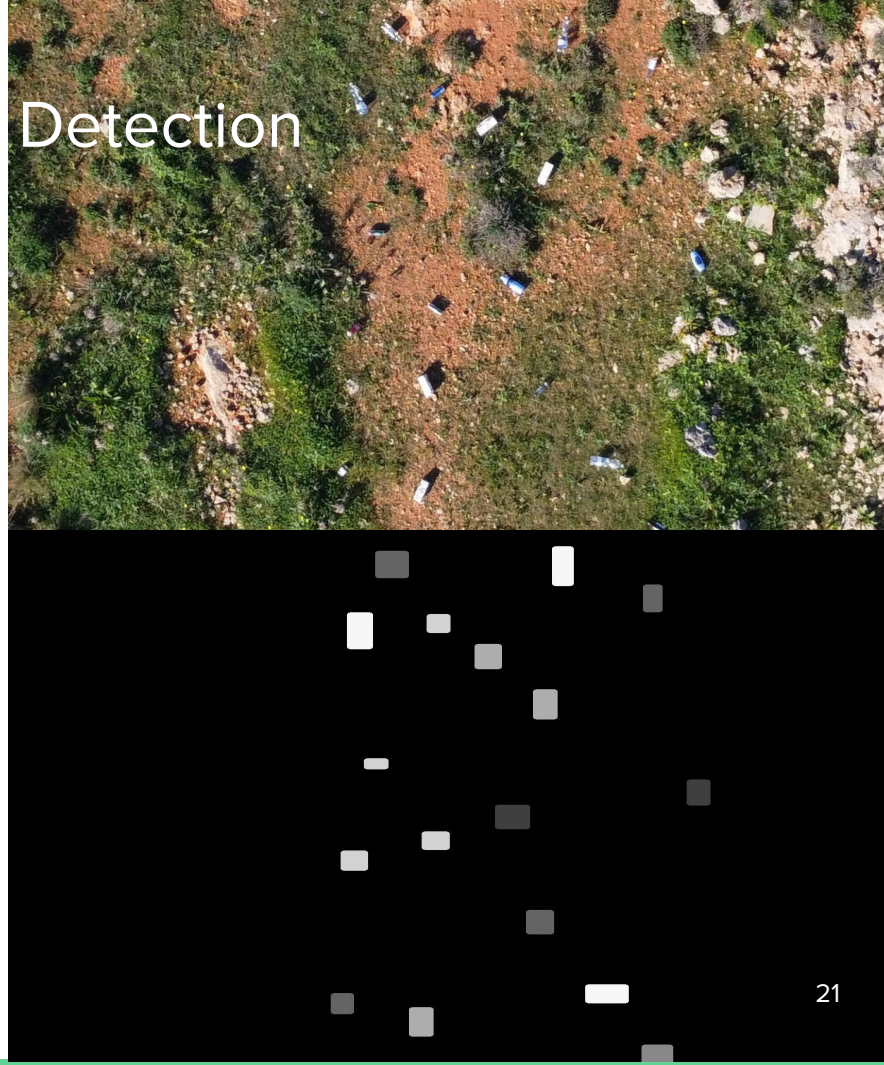
Approach

- **Teacher–student** framework implemented as ***deep neural networks***
  - **Teacher:** trained with standard input plus privileged information (training only)
  - **Student:** trained with standard input; used at inference
- **Knowledge distillation** applied at feature-based layers
- Student network guided by teacher's ***latent representations***

$$L_S = (1-\alpha)\cdot L(f_{\text{student}}(x,y)) + \alpha \cdot D(f_i^{(t)}, f_i^{(s)}),$$

Privileged Information for Object Detection

- Explored *depth*, *saliency*, and *masks* as privileged inputs
- Literature mainly uses segmentation masks
- Adopted **box masks** for:
  - Handling overlapping objects more effectively
  - Suitability when segmentation labels are unavailable
  - Lower annotation error
- Inspired by the **Spotlight** principle of spatial attention
- Implemented as grayscale images with lighter boxes for object classes



Chosen Deep Learning Architectures

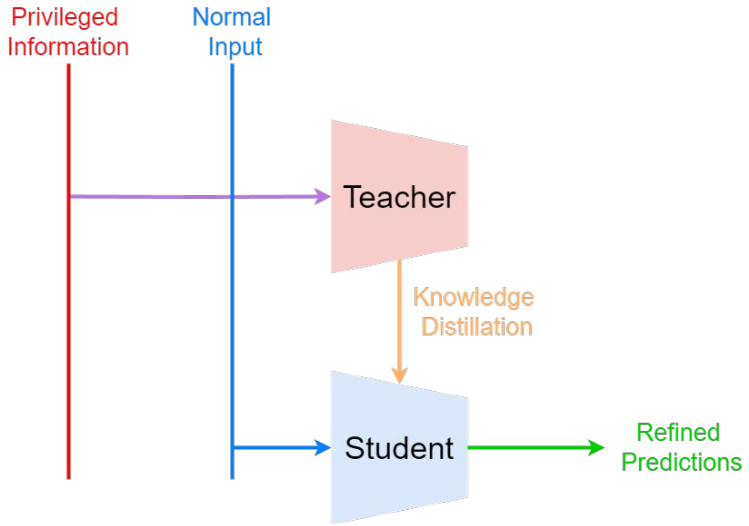
For the purpose of this study, **five** detectors encompassing *one-* and *two-stage* models were chosen

The chosen architectures, all part of the ***torchvision*** library, consist of:

1) **Faster R-CNN** (Region-based Convolutional Neural Network)
2) **SSD** (Single Shot MultiBox Detector)
3) **RetinaNet**
4) **SSDLite**
5) **FCOS** (Fully Convolutional One-Stage Object Detection)

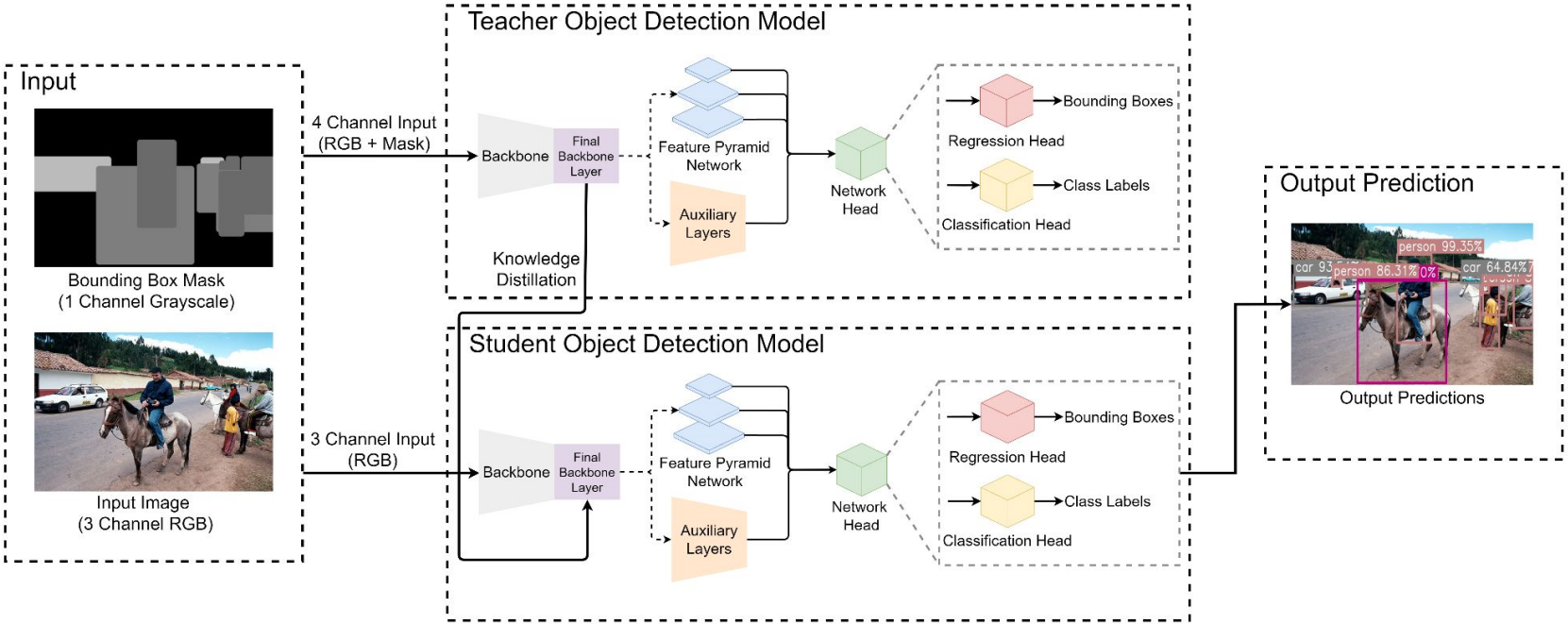
LUPI for Object Detection (1) Teacher and Student

1) **Teacher Network**
   - **Inputs:** RGB images + privileged data (e.g., bounding box masks)
   - Input layer modified to ***4*** channels
   - Pre-trained backbone with **Kaiming Normal initialisation** for input layer
2) **Student Network**
   - **Inputs:** RGB images only
   - Same architecture as teacher
   - **Loss:** detection + distillation
3) **Knowledge Distillation**
   - Align features at **final backbone layer**
   - Compare latent features via **cosine distance**
   - Distillation loss weighted by ***α***



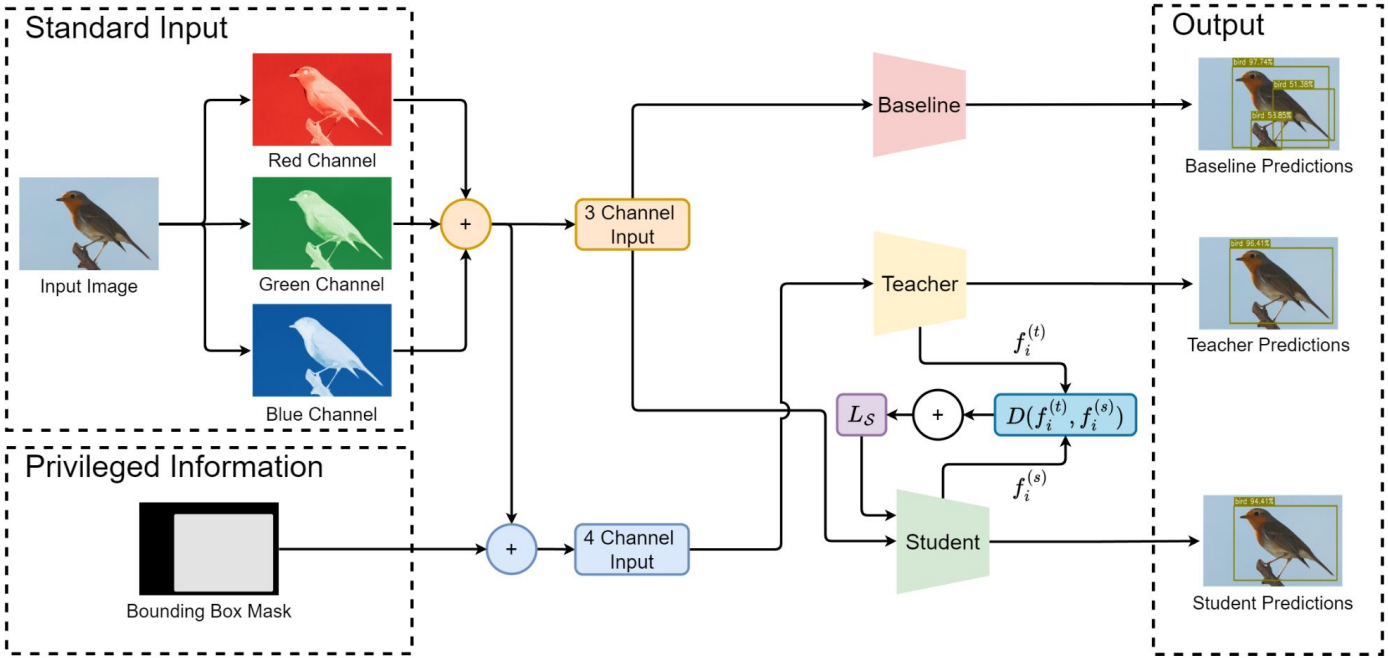
Teacher guides student using privileged information

LUPI for Object Detection (2) General Architecture



General architecture illustrating the teacher processing RGB and privileged information while the student receives RGB input only, with knowledge distilled from the final backbone layer

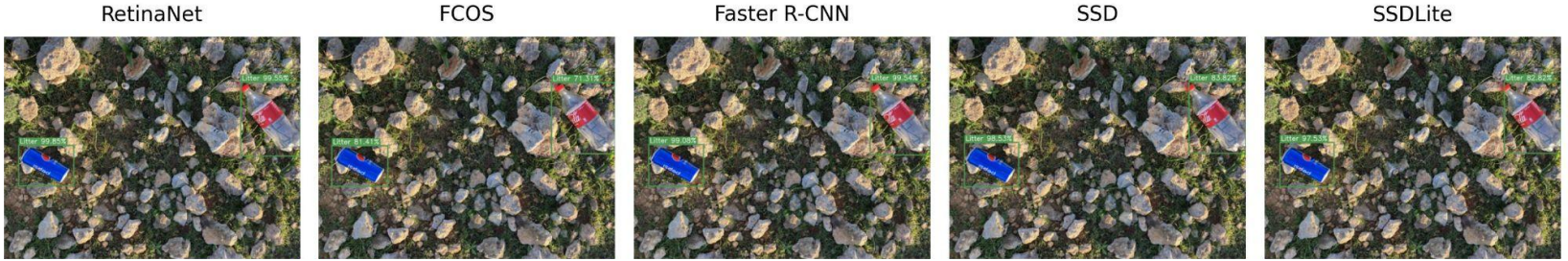
LUPI for Object Detection (3) Detailed Architecture



Detailed architecture showing the teacher with RGB and privileged inputs, the student with RGB input only guided by feature-level knowledge distillation, and a baseline RGB-only model for comparison

Training Parameters, Pre-processing & Post-processing

1) **Training Parameters**
   - **100** epochs, **Adam** optimiser, **LR = *1×10^{-3}***
   - **Early stopping** (8 epochs) and **model checkpointing**
   - **Pre-trained COCO** weights, detection heads adapted for dataset
2) **Pre-processing Steps**
   - **Min-max normalisation** ([*0,1*])
   - **Resize** to 800×800
   - **Channel-wise standardisation** (zero-mean, unit-variance)
3) **Post-processing Steps**
   - **Non-Maximum Suppression** (IoU 0.5)
   - Ensures ***fair comparison across models***



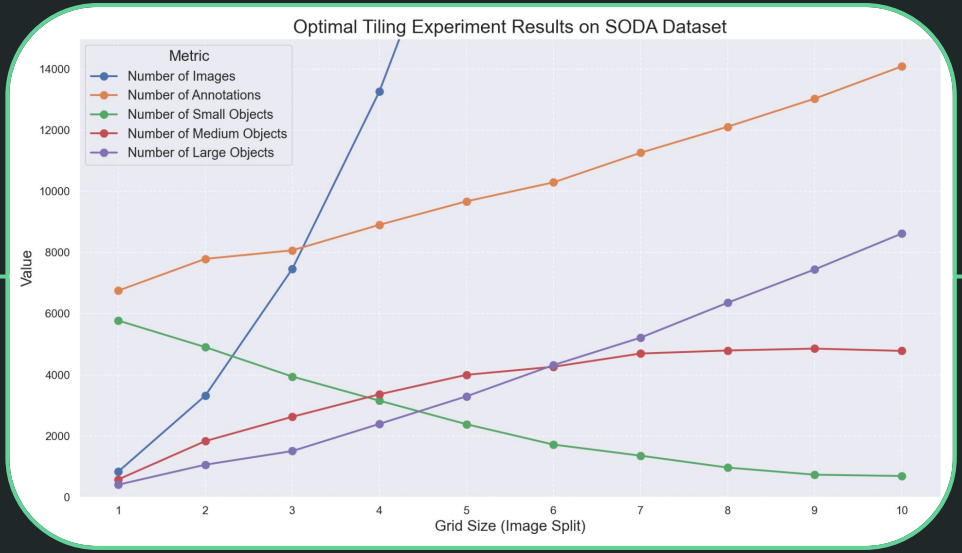
Datasets

1) SODA: Small Objects at Different Altitudes
2) BDW: Bottle Detection in the Wild
3) UAVVaste
4) Pascal Visual Object Classes 2012

Sample images from the BDW dataset (Source: [6])

Sample images from the Pascal VOC 2012 dataset (Source: [7])

Evaluation


Optimal Tiling Experiment Results on SODA Dataset
Metric
Number of Images
Number of Annotations
Number of Small Objects
Number of Medium Objects
Number of Large Objects
Value
0
2000
4000
6000
8000
10000
12000
14000
1
2
3
4
5
6
7
8
9
10
Grid Size (Image Split)

Evaluation Strategy

Experiment	Dataset/s	Metric/s	Objective/s	Purpose
Optimal Tiling	SODA	Small Object Ratio per Image	preliminary for **O1**, **O2**, **O3**	Validate tiling preprocessing; no reason presented in literature
Within-Dataset Evaluation	Subsets of SODA	mAP, Precision, Recall, F1 Score, mAR, Confusion Matrix, Training Time, Model Size, FPS, GFLOPS	**O1**, **O2**, **O3**, partially **O4**	Test methodology on UAV-based litter detection
Cross-Dataset Evaluation	BDW, UAVVaste	mAP, Precision, Recall, F1 Score	**O3**	Assess generalisation to other UAV litter detection datasets
Object Size Performance Analysis	Subsets of SODA, BDW, UAVVaste	mAP (small, medium, large)	**O2**, **O3**	Evaluate small-object detection improvements
Pascal VOC 2012 Evaluation	Pascal VOC 2012	mAP, Precision, Recall, F1 Score, mAR, Confusion Matrix, Training Time, Model Size, FPS, GFLOPS	**O4**	Test generalisability on a larger dataset with a broader range of object categories
Ablation Study on Alpha (α)	Subsets of SODA, Pascal VOC 2012	mAP, F1 Score	**O2**, partially **O4**	Assess impact of critical alpha (α) parameter
Visual Comparison	All datasets	Qualitative comparison	–	Visually inspect improvements beyond metrics
Visual Interpretability	Primarily SODA, but also evaluated on other datasets with similar results	Grad-CAM variants	–	Explore model decision processes and attention

Summary of evaluation strategy including experiments, datasets, key metrics, objectives, and purpose of each experiment

Optimal Tiling Experiments

Purpose

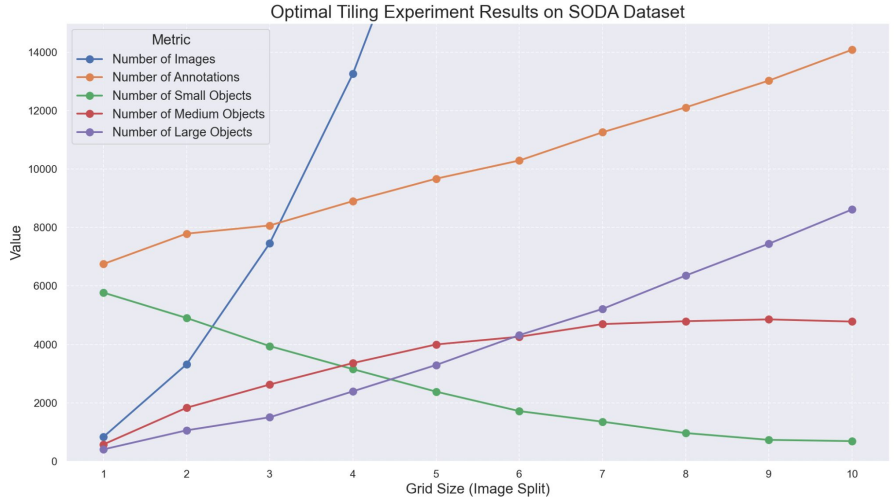
- Find best tiling configuration for the SODA dataset
- Balance visibility vs. cost
- No justification in literature for prior *5×5* choice

Method

- Tested grid sizes 1–10
- COCO size categories
- Elbow method to find optimal

Key Takeaways

- Finer grids → better visibility, higher cost
- *3×3* grid selected as optimal configuration



Results covering all altitudes

Within-Dataset Evaluation

Purpose

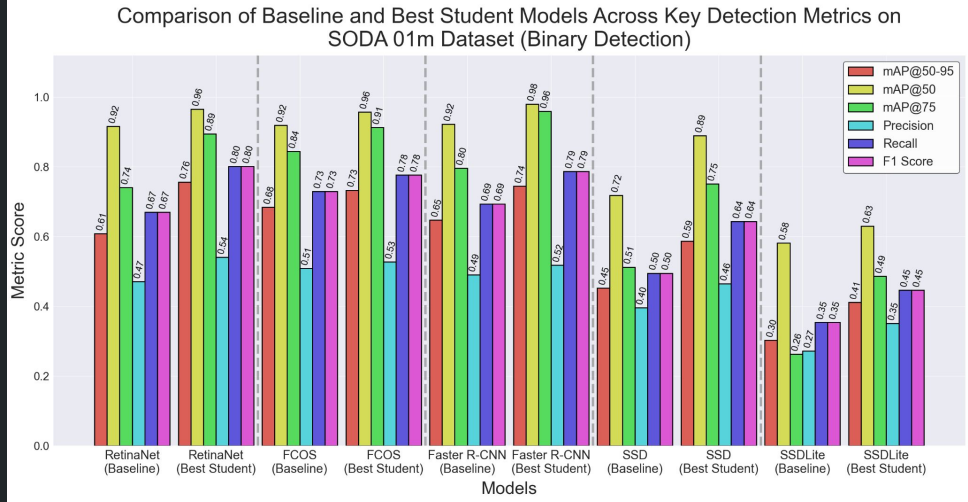
- Test LUPI framework for UAV litter detection
- Handle small objects & varied altitudes

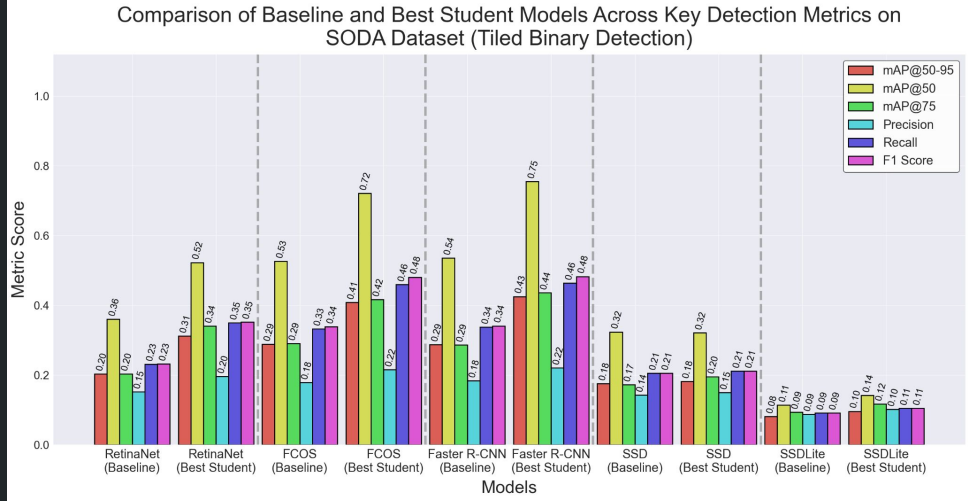
Method

- **3** experiments:
  - Binary at 1 metre (no tiling)
  - Binary across altitudes (*3×3* tiling)
  - Multi-label across altitudes (*3×3* tiling)
- **5** architectures × teacher + students
- ***α*** values *{0, .25, .5, .75, 1}*

Key Takeaways

- Student models usually greater than baseline
- **Best models:** RetinaNet (close-range), Faster R-CNN (multi-alt), FCOS (multi-label)
- Performance boosts come with ⬆ training time, but no extra cost at inference





Cross-Dataset Evaluation

Purpose

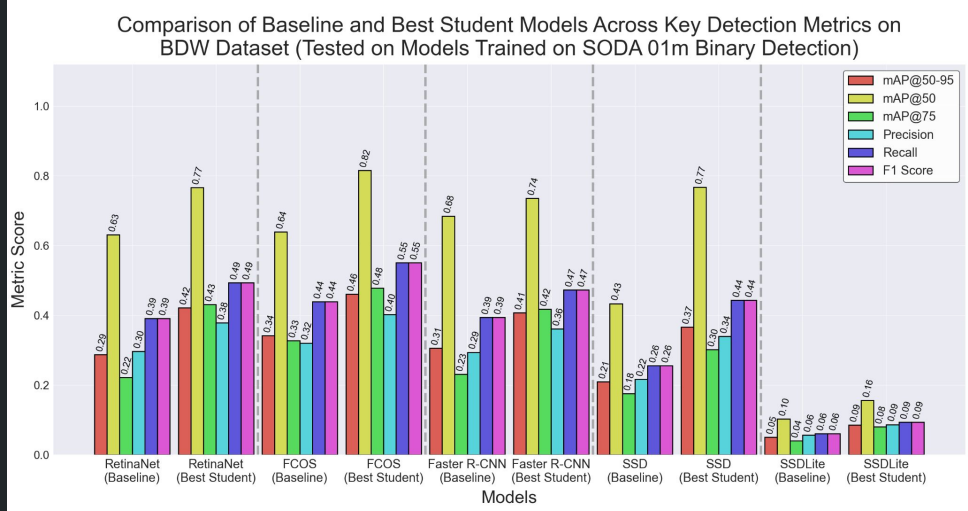
- Assess generalisation of SODA-trained models
- Compare LUPI-trained students against baselines

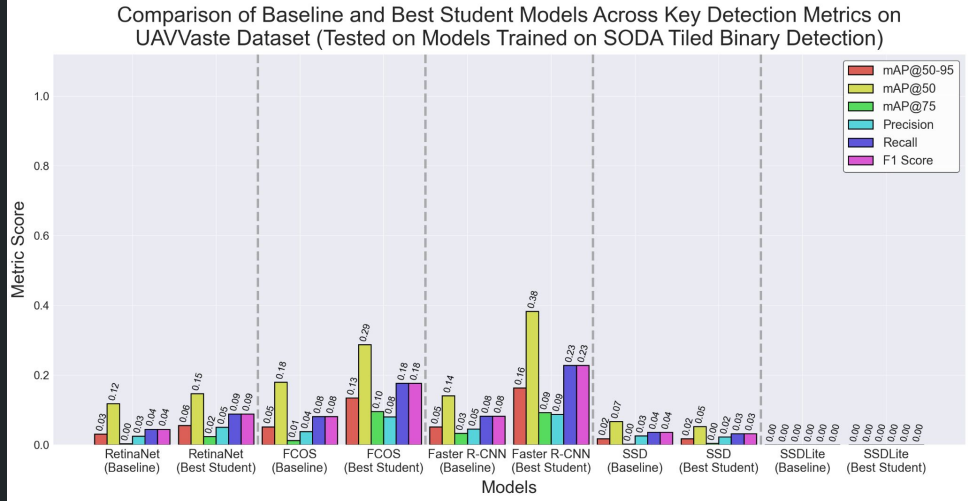
Method

- **BDW:** models from SODA at 1 metre
- **UAVVaste:** models from SODA *3×3* tiling
- Binary litter detection across both datasets

Key Takeaways

- LUPI-trained students consistently outperform baselines
- Strongest results from FCOS, Faster R-CNN, and RetinaNet
- SSD/SSDLite perform poorly, limited transfer
- LUPI improves adaptability without increasing model size





Performance by Object Size

Purpose

- Evaluate the impact of LUPI on small, medium, and large objects
- Focus on detecting tiny litter, which is especially challenging

Method

- **COCO** size-based mAP/mAR metrics
- Within- and cross-dataset evaluations (SODA, BDW, UAVVaste)
- Binary & multi-label tasks with tiling

Key Takeaways

- Students generally outperform baselines, especially on medium and large objects
- Small-object improvements are modest and vary by architecture
- ResNet-FPN backbones deliver the strongest performance; SSD/SSDLite are the weakest

Model	Type	mAP@50–95	mAP^{Small}	mAP^{Medium}	mAP^{Large}	mAR^{Small}	mAR^{Medium}	mAR^{Large}
RetinaNet	Baseline	0.20	0.08	0.26	0.41	0.10	0.29	0.47
	Best Student	**0.31**	**0.11**	**0.43**	**0.59**	**0.14**	**0.50**	**0.65**
FCOS	Baseline	0.29	0.15	0.37	0.49	0.19	0.44	0.54
	Best Student	**0.41**	**0.27**	**0.50**	**0.62**	**0.34**	**0.57**	**0.67**
Faster R-CNN	Baseline	0.29	0.13	0.36	0.55	0.16	0.44	0.62
	Best Student	**0.43**	**0.28**	**0.48**	**0.68**	**0.33**	**0.55**	**0.74**
SSD	Baseline	0.18	**0.02**	0.25	0.43	**0.03**	0.30	0.49
	Best Student	**0.18**	0.01	**0.25**	**0.49**	0.01	**0.30**	**0.56**
SSDLite	Baseline	0.08	0.00	0.01	0.39	0.00	0.00	0.47
	Best Student	**0.10**	0.00	**0.01**	**0.47**	0.00	**0.01**	**0.54**

Table 4.9 Comparison of the baseline and best-performing student models across COCO detection metrics on the 3×3 SODA dataset across all altitudes for binary litter detection, with results reported separately for each object size category.

Pascal VOC 2012 Evaluation

Purpose

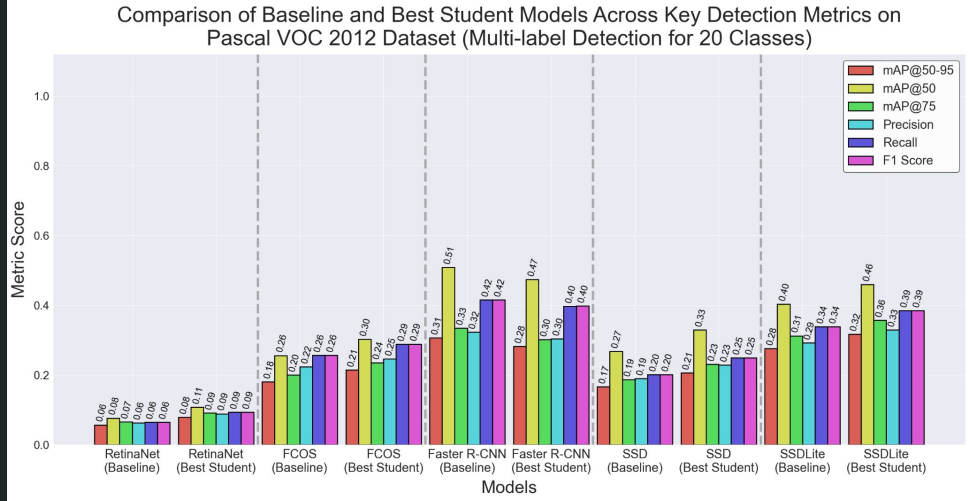
- Test generalisability of LUPI across 20-class dataset
- Examine student vs baseline performance on complex multi-purpose scenes

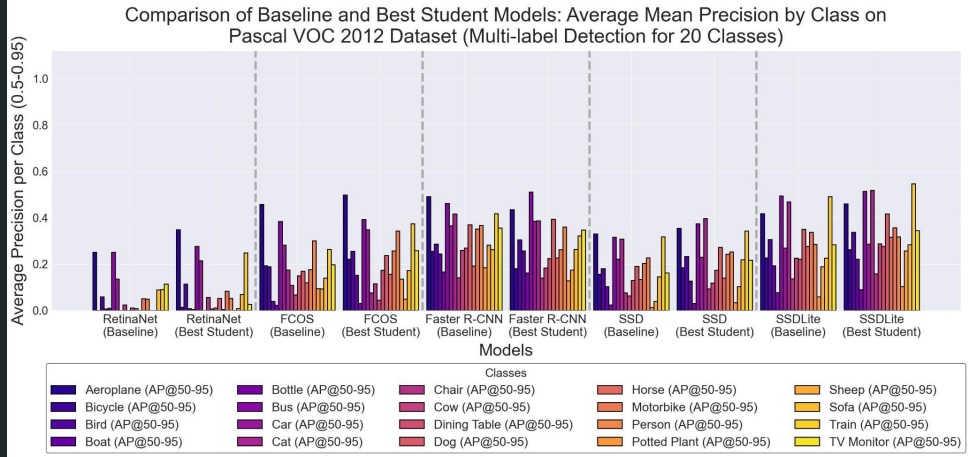
Method

- Multi-label detection on **Pascal VOC 2012**

Key Takeaways

- Students generally match or slightly outperform baselines
- SSD/SSDLite showed largest gains; Faster R-CNN sometimes worse
- High recall, moderate precision; small improvements meaningful in 20-class context
- Training time ↑ for students; model size & inference speed largely unchanged





Ablation Study on Alpha (α)

Purpose

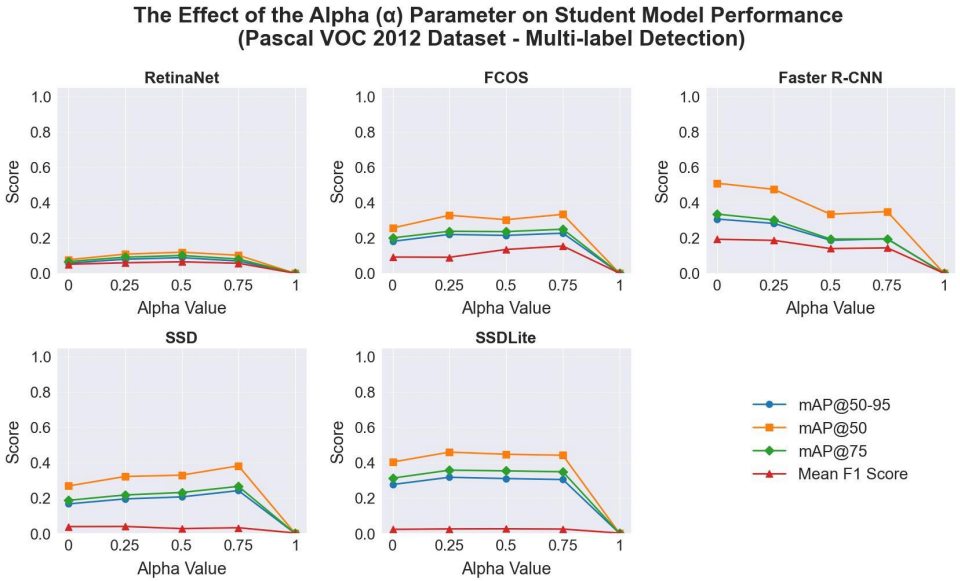
- Evaluate effect of α on student performance
- Balance reliance between teacher guidance and ground-truth annotations

Method

- Tested $\alpha = \{0, .25, .5, .75, 1\}$
- Experiments on SODA (binary & multi-label) and Pascal VOC 2012

Key Takeaways

- Best results generally for α=0.25–0.5
- α=1 (teacher only) drastically reduces performance
- Some models/tasks benefit from α=0.75
- Avoid over-reliance on the teacher, or student learning suffers



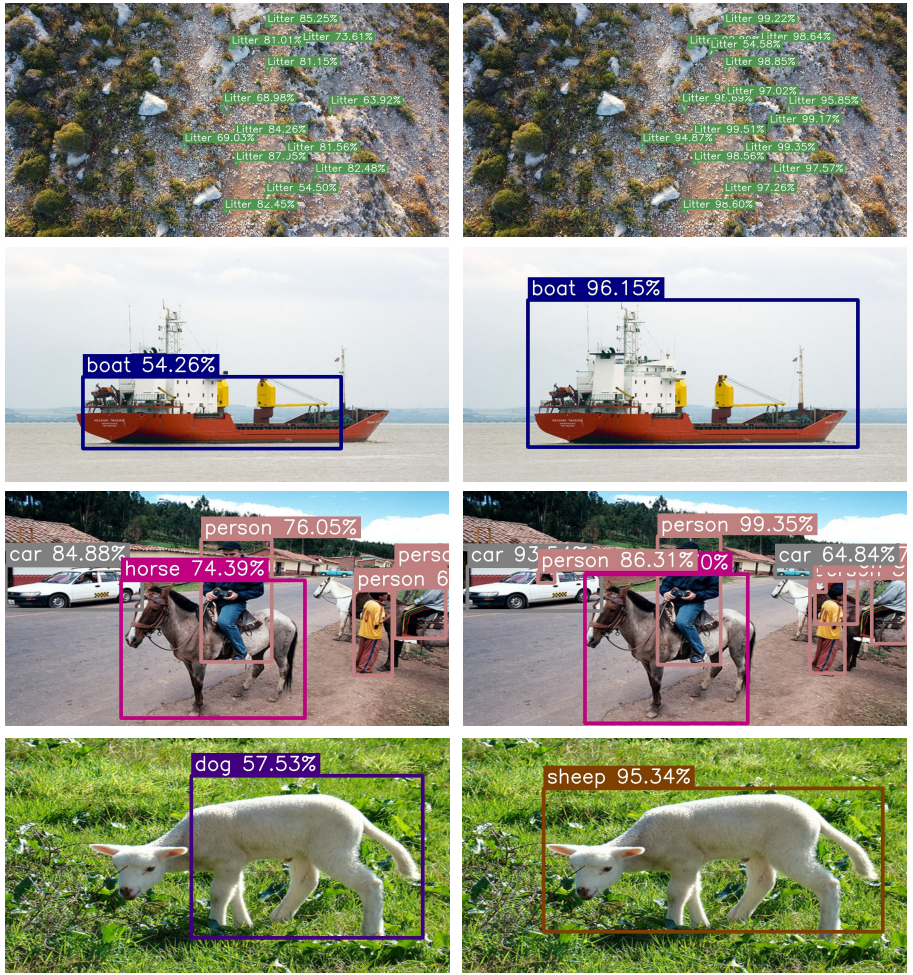
Visual Comparison

Student models detect more objects with higher confidence

Improvements seen in both litter and general datasets

Bounding boxes and **labels** are more accurate

Visual results **align** with quantitative metrics



Baseline

Best Student

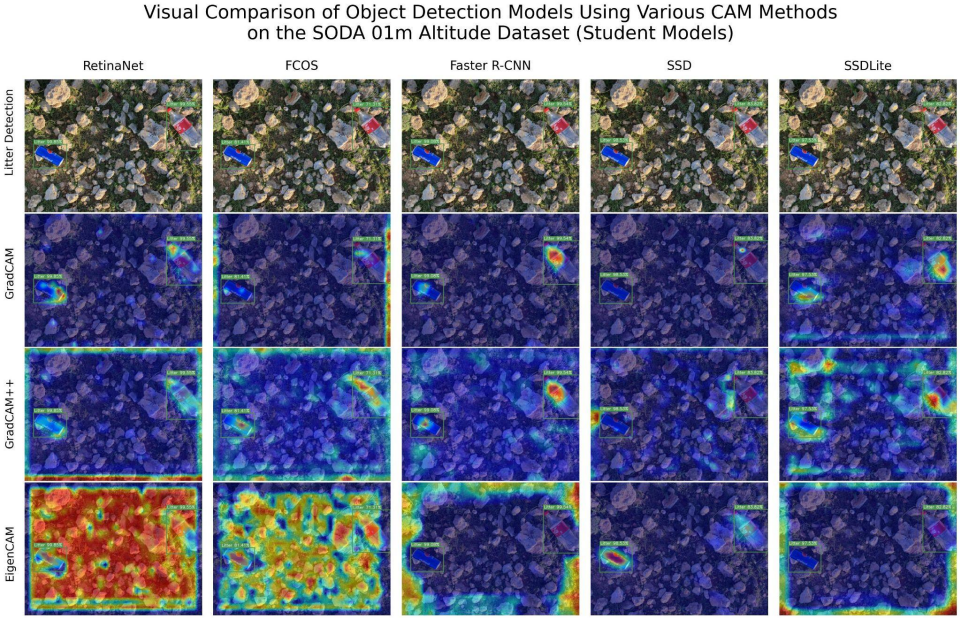
Visual Interpretability

Student models correct baseline mistakes and detect missed objects

Grad-CAM and Grad-CAM++ show **reduced background distraction** in student models

Eigen-CAM indicates broader, more **balanced attention** across objects

Student models **focus on relevant areas** and **disperse attention** from irrelevant regions, improving predictions



Discussion

Privileged information enhances detection performance but has **inherent limitations**:

i) ***Overlapping objects*** *of the same class*
ii) *Larger objects* ***occluding*** *smaller ones*
iii) ***Restricted colour*** *differentiation in datasets with numerous categories*

Student models consistently **outperform baseline** models across tasks

Improvements in mAP range from *0.02* to *0.15* depending on task complexity

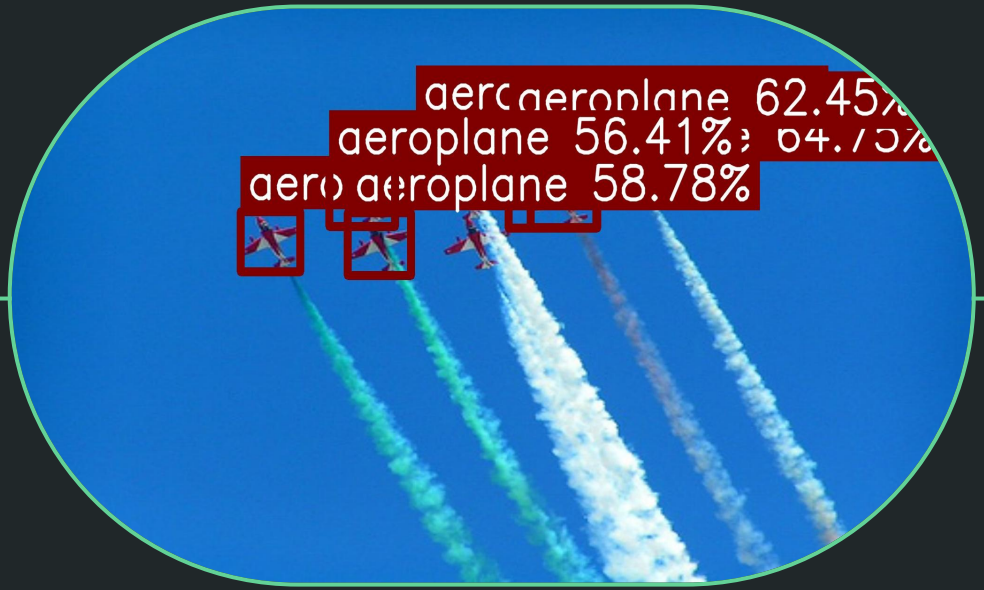
Evaluated across **five** architectures: Faster R-CNN, SSD, SSDLite, RetinaNet, FCOS

Increased training time, with **no impact on model size**, parameter count, or inference speed

Visual and feature-based analyses indicate improved object coverage, higher confidence scores, and more focused attention

Conclusion

“The end of one journey is simply the beginning of another.” **– A. A. Milne**



Revisited Objectives

All objectives fulfilled across the conducted experiments

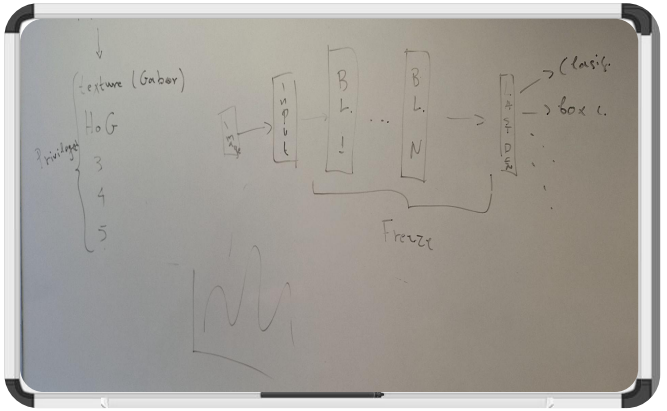
The **research question** on whether LUPI enhances object detection has been **addressed with a high degree of confidence**, showing consistent improvements

The **proposed approach** was applied to **five** renowned object detection architectures

A **total of *120*** models were trained and evaluated, including baseline, teacher, and student variants

Student models **improved accuracy** while preserving model size and inference speed

Ablation and visual/feature-based analyses **confirm robustness and interpretability**



Main Contributions

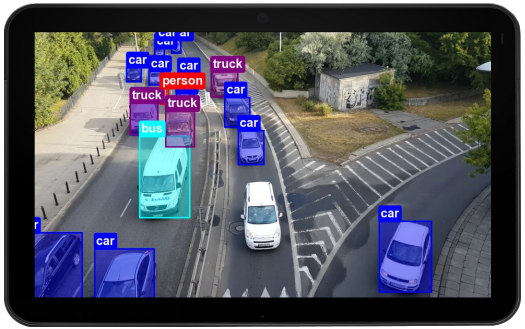
1) Introduction of LUPI to object detection
2) Improved litter detection and localisation
3) Model-agnostic performance improvement
4) Generalisation across litter detection datasets
5) Generalisation across object detection datasets

Practical Applications & Impact

Potential applications of this work include, *but are not limited* to:

1) *UAV-based litter detection and geolocation systems*
2) *Traffic monitoring systems*
3) *Video surveillance monitoring*
4) *Underwater exploration systems*

Amongst many others. . .



DJI Air 2S (Source: [8])

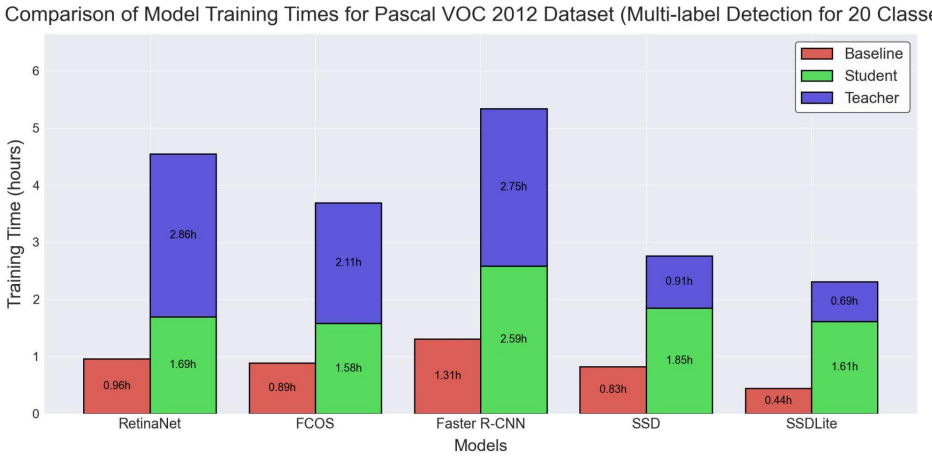
Limitations & Future Work

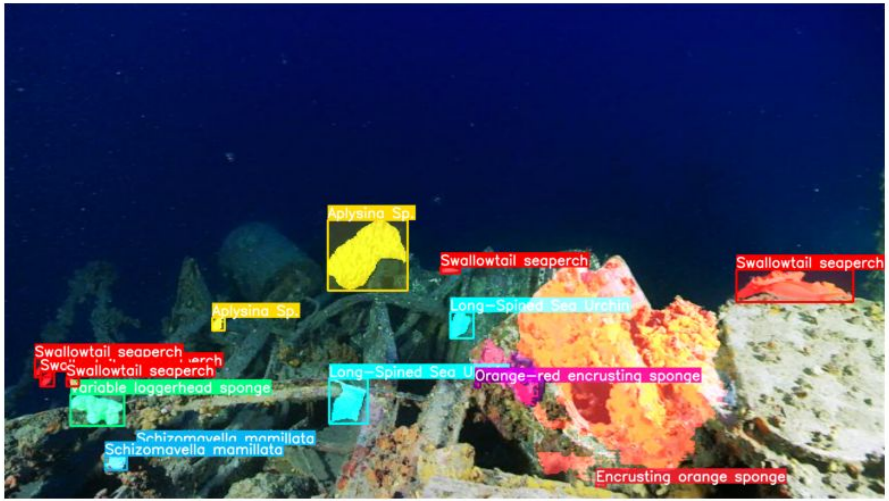
Limitations

- Additional **training overhead** due to teacher–student setup
- **Privileged information** may not always be available or easy to generate
- The design of privileged inputs is **task-dependent** and requires careful construction

Future Work

- Evaluate on more **advanced architectures** (YOLOv12, DETR, RT-DETR)
- Explore improved **encoding strategies** for privileged information
- Expand **knowledge distillation** techniques to classification and regression
- **Fine-tune** architecture-specific hyperparameters
- Extend the approach to **segmentation** tasks





Questions and Comments?

They are welcome. . .

References (1)

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” CoRR, vol. abs/2005.12872, 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[2] Y. Zhao et al., “DETRs Beat YOLOs on Real-time Object Detection,” arXiv preprint arXiv:2304.08069, 2023. [Online]. Available: https://arxiv.org/abs/2304.08069

[3] D. Pisani, D. Seychell, C. J. Debono and M. Schembri, “SODA: A Dataset for Small Object Detection in UAV Captured Imagery,” *IEEE International Conference on Image Processing (ICIP),* Abu Dhabi, United Arab Emirates, pp. 151-157, 2024.

References (2)

[4] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” CoRR, vol. abs/1708.02002, 2017. [Online]. Available: http://arxiv.org/abs/1708.02002

[5] X. Wang, G. Wei, S. Chen, and J. Liu, “An efficient weakly semi-supervised method for object automated annotation,” *Multimedia Tools and Applications*, vol. 83, pp. 1–24, Jun. 2023. DOI: 10.1007/s11042-023-15305-0.

[6] J. Wang, W. Guo, T. Pan, H. Yu, L. Duan, and W. Yang, “Bottle Detection in the Wild using Low-altitude Unmanned Aerial Vehicles,” in *2018 21st International Conference on Information Fusion (FUSION)*, 2018, pp. 439–444. DOI: 10.23919/ICIF.2018.8455565.

References (3)

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) *Results*, http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[8] DJI, “DJI Air 2S,” DJI. [Online]. Available: https://www.dji.com/mt/support/product/air-2s. [Accessed: 30-Jan-2025].

Thank you!

matthias.bartolo.21@um.edu.mt
Matthias Bartolo