

Learning Using Privileged Information for Litter Detection

Matthias Bartolo
Dept. of Artificial Intelligence
University of Malta
Msida, Malta
matthias.bartolo@um.edu.mt

Konstantinos Makantasis
Dept. of Artificial Intelligence
University of Malta
Msida, Malta
konstantinos.makantasis@um.edu.mt

Dylan Seychell
Dept. of Artificial Intelligence
University of Malta
Msida, Malta
dylan.seychell@um.edu.mt

Abstract—As litter pollution continues to rise globally, developing automated tools capable of detecting litter effectively remains a significant challenge. This study presents a novel approach that combines, for the first time, privileged information with deep learning object detection to improve litter detection while maintaining model efficiency. We evaluate our method across five widely used object detection models, addressing challenges such as detecting small litter and objects partially obscured by grass or stones. In addition to this, a key contribution of our work is the introduction of a novel method for generating privileged information for object detection by encoding bounding box annotations as binary masks, which are used to guide and refine the detection process. Through experiments on both within-dataset evaluation on the renowned SODA dataset and cross-dataset evaluation on the BDW and UAVVaste litter detection datasets, we demonstrate consistent performance improvements across all models. Our approach not only bolsters detection accuracy within the training sets but also generalises well to other litter detection contexts. Crucially, these improvements are achieved without increasing model complexity or adding extra layers, ensuring computational efficiency and scalability. Our results suggest that this methodology offers a practical solution for litter detection, balancing accuracy and efficiency in real-world applications.

Index Terms—Litter Detection, Learning Using Privileged Information, Computer Vision, Knowledge Distillation, Object Detection

I. INTRODUCTION

Litter pollution remains a stagnant issue, with ramifications that extend beyond environmental deterioration to encompass broader socio-economic instability. With global waste output projected to rise from 2.1 to 2.6 billion tonnes annually by 2030 [1], the limitations of current management systems are becoming increasingly apparent. In response to this global challenge, recent research [2], [3] has begun to explore the application of Artificial Intelligence (AI), particularly computer vision techniques, as a means of automating the detection of litter in various environments. Similarly, Unmanned Aerial Vehicle (UAV) technology has received growing attention for its potential to assist in detecting litter across wide or inaccessible areas [4], [5].

However, despite recent progress, there still remains a clear need to improve the accuracy and efficiency of these technologies. Achieving optimal performance in diverse and dynamic environments continues to present significant challenges, es-

pecially in balancing detection accuracy with inference speed. In practical applications, litter frequently includes transparent materials or items that are either very small or partially concealed by natural elements such as grass or stones. These conditions necessitate more complex architectural frameworks and a more rigorous approach to model training, such as incorporating knowledge distillation techniques to improve generalisation while maintaining computational efficiency. It is within this context that this paper proposes the following:

- 1) A novel methodology that integrates privileged information and deep learning object detection models to improve litter detection, without increasing the number of model parameters or affecting inference time.
- 2) A performance evaluation of this methodology across five widely-used object detectors.
- 3) A detailed examination of the proposed methodology using the SODA dataset [5], alongside cross-validation on the BDW [6] and UAVVaste [4] litter detection datasets from aerial imagery.

II. RELATED WORK

Computer vision has gained attention in addressing environmental issues, especially waste detection. Litter detection stands out due to its relevance to sustainability and public hygiene, prompting the development of datasets and automated detection methods.

A. Litter Detection

In recent years, a number of litter detection datasets and methods have been introduced to support research in automated litter detection. Wang et al. [6] introduced the UAV-Bottle, or BDW, dataset in 2018, which includes 25,407 UAV-captured images focused solely on the detection of bottles across diverse environments. In addition to UAV-based litter detection, Proença and Simões [3] developed the TACO dataset in 2020. Comprising 1,500 images across 60 categories, this dataset broadened the scope of litter detection tasks and continues to be widely used in related research. In the same year, Wang et al. [7] released the MJU-WASTE dataset, which provides 2,475 images dedicated to litter segmentation within a single waste category. Similarly, Kraft et al. [4] introduced the

UAVVaste dataset in 2021, focusing on UAV-based litter detection. This dataset contains 772 UAV images and addresses the challenges of detecting small objects within a single waste category. Additionally, in terms of non-UAV based litter detection, Bashkirov et al. [2] developed the ZeroWaste dataset, while Córdova et al. [8] created the PlastOPol dataset, containing 4,503 and 2,418 images, respectively, providing real-world data that further improves litter detection research. Most recently, Pisani et al. [5], [9] presented the SODA dataset in 2024, which includes 829 images captured at various UAV altitudes across six categories. Across all of these approaches, the authors utilised the curated datasets to develop effective litter detection models, employing methodologies similar to those used in object detection, which involve training prominent deep learning detection architectures. Notable detectors that were trained in the aforementioned approaches, include YOLO [10], Faster R-CNN [11], SSD [12], and RetinaNet [13], among others. In addition, pre-processing techniques such as tiling and data augmentation were also commonly employed to bolster training robustness and accuracy [4], [9]. Nevertheless, in all of these approaches, the repeated trend of improving accuracy by exploring or developing complex architectures and learning paradigms necessitates a clearer way forward [4], [5], [9].

B. Learning Using Privileged Information in Computer Vision

The Learning using Privileged Information (LUPI) paradigm, introduced by Vapnik and Vashist [14], [15], expands traditional learning tasks by incorporating supplementary data alongside the standard input/output training pairs in machine learning. This additional information is often more pertinent to the task at hand, thereby improving prediction accuracy. The concept of LUPI allows for the *transfer of knowledge* from a teacher, trained with privileged data, to a student who only has access to the input information. In the field of Computer Vision, several problems present an asymmetric distribution of information between training and test phases [16], making LUPI particularly applicable. Sharmanska et al. [16] investigate four types of privileged information for object classification: semantic properties, bounding boxes, tags, and annotator rationale. Their study shows that applying LUPI to the SVM+ algorithm improves performance. In a similar study, Wang et al. [17] address the same issue by applying similarity constraints to capture the relationship between available and privileged information. The authors use high-resolution images and image tags as privileged data, which are accessible during training but not during testing.

C. Knowledge Distillation in Computer Vision

Knowledge distillation is a pivotal technique in machine learning that allows the transfer of knowledge from a large, complex model to a smaller, more efficient one. In the context of computer vision, as discussed by [18], there are various methods for achieving this, including response-based, feature-based, and relation-based knowledge transfer. These

approaches can be applied across a wide range of vision tasks, such as image classification, object detection, and multimodal vision models [18]. Focusing on object detection, two common distillation techniques are feature imitation and logit mimicking [19]. Interestingly, the use of valuable localisation regions to selectively distil both classification and localisation knowledge for specific areas is another key aspect of this process, as proposed in [19].

In summary, existing litter detection methods rely on complex models and large datasets to boost accuracy. In this context, applying privileged information during training without altering model structure or inference speed offers a promising alternative.

III. METHODOLOGY

This section presents our methodological framework. We begin by defining the problem and articulating our conceptual approach. Subsequently, we provide the implementation details a description of the experimental protocol.

A. Problem Definition

This study proposes a novel methodology that applies learning with privileged information to the task of object detection, specifically focusing on litter detection. Although the LUPI paradigm has previously been explored within computer vision, particularly in relation to image classification, its application to object detection remains unexplored. In this regard, the object detection problem within the LUPI framework can be rigorously described as follows: consider a training set of triplets as defined in Equation (1).

$$\mathcal{D} = (x_i, x_i^*, y_i)_{i=1}^N, \quad x_i \in X, x_i^* \in X^*, y_i \in Y. \quad (1)$$

In this formulation, X represents the space of input images, X^* denotes the space of privileged information instances, and Y comprises the space of bounding boxes with their associated class labels. Given a teacher model defined as:

$$f_{teacher} : X \cup X^* \rightarrow Y, \quad (2)$$

which accurately predicts y based on both x and x^* , our objective is to develop a student model:

$$f_{student} : X \rightarrow Y, \quad (3)$$

that effectively maps X to Y by leveraging not only the intrinsic information in X , but also the knowledge encoded within $f_{teacher}$. In other words, during training, $f_{student}$ learns to map X to Y through knowledge distillation from $f_{teacher}$ and the information contained in the labeled examples from the training set \mathcal{D} .

B. Our Approach

In this study, both $f_{teacher}$ and $f_{student}$ are implemented as neural networks, each comprising L layers. These models can be formally expressed as:

$$f_{teacher} = f_1^{(t)} \circ f_2^{(t)} \circ \dots \circ f_l^{(t)} \circ \dots \circ f_L^{(t)}, \quad (4)$$

$$f_{student} = f_1^{(s)} \circ f_2^{(s)} \circ \dots \circ f_l^{(s)} \circ \dots \circ f_L^{(s)}, \quad (5)$$

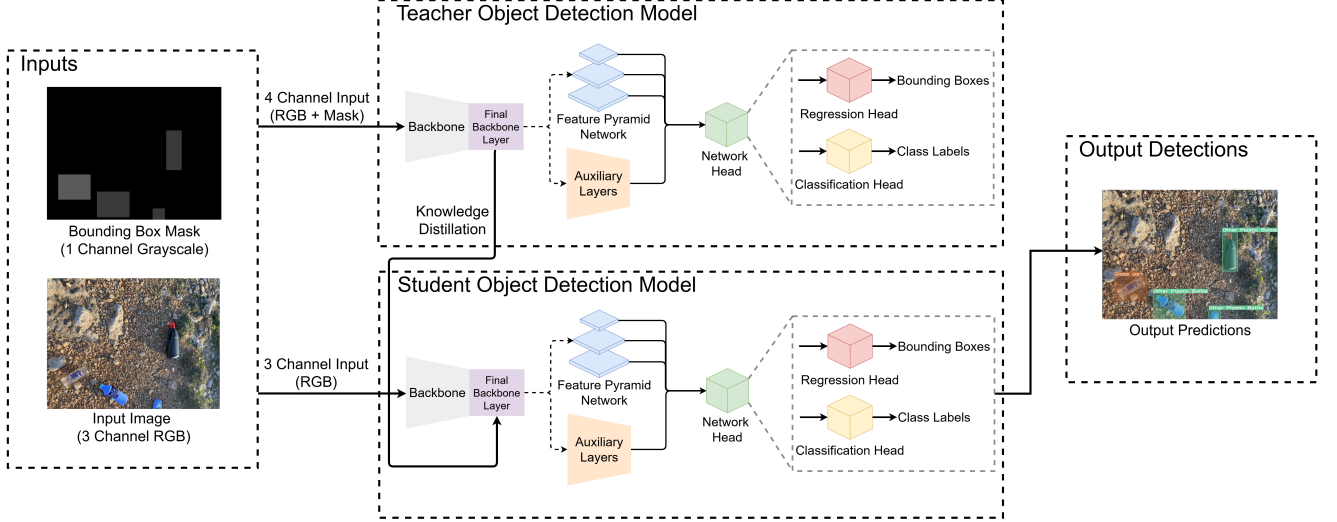


Fig. 1. Architecture of the object detection models, illustrating the use of the LUPI paradigm with RGB and bounding box mask inputs, the teacher and student networks, the final backbone layer for knowledge distillation, and the output predictions.

Here, “ \circ ” represents the function composition operation, while $f_i^{(t)}$ and $f_i^{(s)}$ denote the i -th layer of the teacher and student models, respectively. We establish the constraint that the l -th layer of both the teacher and student networks contain an identical number of hidden neurons. Consequently, knowledge can be distilled from the teacher to the student model by minimizing the dissimilarity:

$$D(f_l^{(t)}, f_l^{(s)}). \quad (6)$$

This minimization is performed for each triplet (x_i, x_i^*, y_i) in the training set \mathcal{D} . Specifically, given a triplet (x_i, x_i^*, y_i) , we require that the latent representation at the l -th layer of the student closely approximates the corresponding latent representation at the l -th layer of the teacher. Since the teacher utilizes both x_i and x_i^* , we hypothesize that its l -th layer latent representation contains more informative features than the representation generated by the student, which relies solely on x_i . In our methodology, we incorporate this requirement into the training process of the student model by modifying the loss function as follows:

$$L_s = (1 - \alpha) \cdot L(f_{student}(x, y)) + \alpha \cdot D(f_i^{(t)}, f_i^{(s)}). \quad (7)$$

In this equation, $L(f_{student}(x, y))$ represents the standard object detection loss, and α determines the relative influence of the teacher on the student’s learning process. It is important to emphasize that during the training phase, the student model leverages knowledge derived from $\{x_i^*\}_{i=1}^N$ by emulating the teacher’s latent representations, whereas during the testing phase, it relies exclusively on $x \in X$ to generate predictions.

C. Implementation

Given the object detection problem within the LUPI paradigm, the methodology for applying it to litter detection

is as follows: Each object detection model uses both a teacher and a student network with identical layers, differing only at the input. The teacher receives a four-channel input-three-channel RGB plus a privileged information channel-while the student gets only the standard RGB input.

Selecting the privileged channel is challenging, especially for encoding bounding box information. Inspired by the Attention Spotlight principle in the human visual cortex [20], a grayscale mask is generated for all bounding boxes, with each object class represented by a distinct shade. Preliminary tests showed this approach yielded the best results and was adopted as the privileged channel. Other forms, like saliency and depth prediction [21], did not show significant improvements.

Knowledge distillation from teacher to student occurs at the final backbone layer, where a feature representation vector is generated and Cosine Distance [22] is used. This vector is incorporated into the student’s loss function, as defined in (7).

The methodology was evaluated on five well-known object detection architectures-Faster R-CNN [11], RetinaNet [13], FCOS [23], SSD [12], and SSDLite [24]-across individual and multiple datasets. The approach, adaptable to any detection model, is shown in Figure 1.

D. Experimental Setup

To evaluate the methodology both within and across datasets, the publicly available SODA, BDW, and UAVVaste datasets were used. SODA was selected for training due to its varied-altitude images, offering practical, real-world data.

For preprocessing, SODA’s 829 images were tiled using a 3x3 grid (unlike the 5x5 in [9]) based on hyper-parameter tuning, then resized to 1280x1280 pixels for high-resolution input. Privileged bounding box masks were generated on the tiled RGB images as grayscale masks. Min-Max normalization was applied to both RGB and mask images, standardizing pixel

values to $[0, 1]$. BDW and UAVVaste datasets were also resized to 1280x1280 pixels, but not tiled, as this was not part of their preprocessing. These datasets were used only for cross-dataset evaluation.

No data augmentation techniques were applied, as they were beyond the study's scope. All detectors were trained with the Adam optimizer at a constant 0.0001 learning rate and no weight decay, based on preliminary tests showing Adam's fast convergence and consistency. Early stopping with a patience of 8 was used to prevent overfitting, and all models were trained for 100 epochs. For post-processing, Non-Maximum Suppression (NMS) with an IoU threshold of 0.5 was applied to reduce background predictions.

IV. RESULTS

A. Evaluation Metrics

To evaluate the proposed methodology as outlined in Subsection III-D, standard object detection metrics were adopted within the experimental framework. These included the COCO Detection metrics [25], which follow the benchmark Mean Average Precision (mAP) at IoU thresholds of 0.5 and 0.75, as well as the averaged metric across a range from 0.5 to 0.95. To complement the COCO metrics, three additional evaluation metrics were employed to facilitate a more thorough and nuanced assessment of the model's performance. Mean Precision evaluated how well the model identified correct detections whilst omitting false positives. Recall measured how completely the model detected all relevant objects. Finally, the F1 Score, calculated as the harmonic mean of Precision and Recall, served to evaluate each model's performance by balancing its accuracy with its ability to detect all relevant objects.

B. Performance Evaluation on the SODA Dataset

Three experiments were conducted using the SODA dataset. The first involved training and evaluating the selected detectors on a 3 by 3 tiled version of the dataset, as detailed in Subsection III-D, specifically for multi-label small litter detection. The second followed the same setup but assessed binary detection instead. The third focused on training and evaluating the detectors on a subset of images captured at an altitude of one meter, treating it as a binary litter detection task without tiling.

Although FCOS and RetinaNet have demonstrated superior results on the COCO benchmark—partly due to their more recent development and improved architectures—this pattern was also observed in the results of the first experiment for multi-label small litter detection. As shown in Figures 2, and 3, the results showed that FCOS emerged as one of the best performing detectors. Interestingly, Faster R-CNN outperformed RetinaNet in this specific task. Meanwhile, SSD and SSDLite yielded the lowest performance, yet all models still significantly benefited from the application of a teacher model, leading to notable improvements in detection accuracy.

An analysis was also carried out to investigate the impact of the teacher model on the performance of the student model,

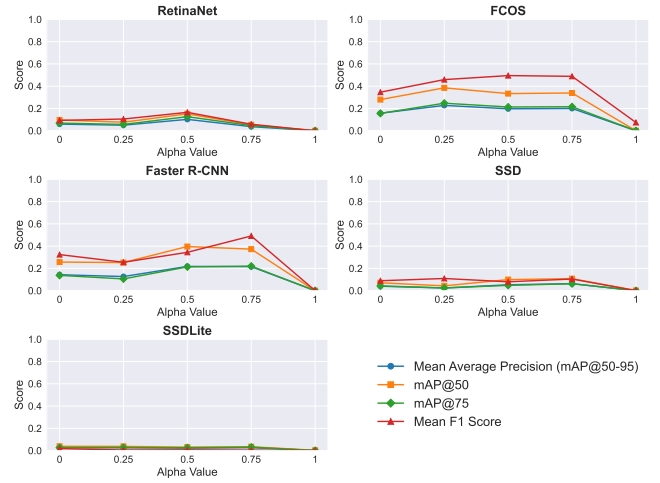


Fig. 2. The Effect of the α Parameter on Student Model Performance (SODA Dataset - Tiled Multi-label Detection).

based on the influence parameter α as defined in (7). For each of the selected models, student versions were trained using five different values for α : 0, 0.25, 0.5, 0.75, 1, as was done in [22]. As shown in Figure 2, the α parameter had a noticeable effect on overall performance. On average, values between 0.25 and 0.5 resulted in higher mAP, while a value of 0.75 tended to yield better F1 Scores. It is also important to note that applying full teacher influence ($\alpha = 1$) frequently led to worse performance compared to omitting the teacher model altogether.

It is also worth highlighting that, when comparing the teacher models, the privileged information channel provided by the bounding box mask proved to be informative. This input enabled most models to more effectively learn the underlying target concept, as demonstrated in Table I. Interestingly, Faster R-CNN proved to be the most effective teacher model overall, demonstrating the greatest ability to grasp the true target concept, particularly in terms of small litter detection. However, FCOS and RetinaNet produced comparable results, suggesting that their architectures were also well suited to guiding student models. In contrast, SSD and SSDLite yielded weaker results as teacher models, which can be attributed in part to their simpler architecture. Nevertheless, these models still performed better than the baselines.

TABLE I
COMPARISON OF TEACHER MODELS ACROSS KEY DETECTION METRICS ON SODA DATASET (TILED MULTI-LABEL DETECTION)

Model	mAP@50-95	mAP@50	mAP@75	Precision	Recall	F1 Score
RetinaNet	0.88	0.92	0.91	0.76	0.97	0.85
FCOS	0.91	0.95	0.94	0.91	0.97	0.94
Faster R-CNN	0.95	0.99	0.98	0.96	0.99	0.97
SSD	0.36	0.49	0.45	0.59	0.76	0.63
SSDLite	0.11	0.13	0.13	0.00	0.37	0.01

Across all three experiments conducted on the SODA dataset, as illustrated in Figures 3, 4, and 5, there is a clear and consistent improvement when applying the proposed methodology to litter detection. This applies both to the localisation

and classification components that define the detection task. In the first experiment (Figure 3), it was shown that applying the proposed methodology to address the problem of small litter detection, together with the use of tiling, led to a significant improvement in both mAP and F1 Score when comparing the performance of the student models to their respective baselines.

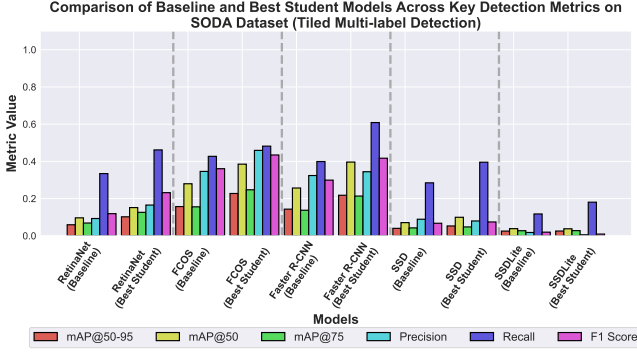


Fig. 3. Comparison of Baseline and Best Student Models Across Key Detection Metrics on SODA Dataset (Tiled Multi-label Detection).

Similarly, in the second experiment (Figure 4), which focused on binary small litter detection, the methodology again demonstrated improved results compared to the baselines. While all models benefited from the approach, the improvements were more pronounced when comparing baseline models with their student counterparts. Models such as Faster R-CNN, FCOS, and RetinaNet exhibited notable improvements, whereas SSD and SSDLite achieved smaller improvements.

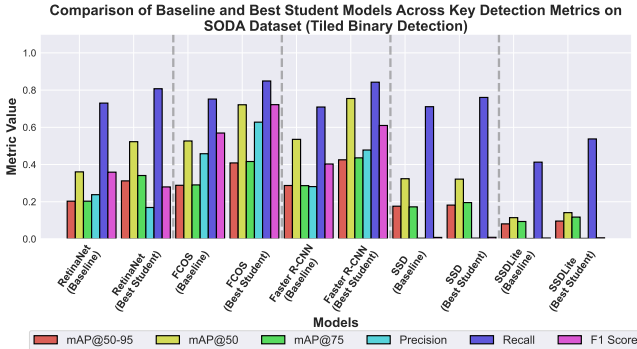


Fig. 4. Comparison of Baseline and Best Student Models Across Key Detection Metrics on SODA Dataset (Tiled Binary Detection).

The third experiment aimed to assess whether the proposed methodology would still yield an improvement when applied to the task of close-range litter detection. At an altitude of one meter, the litter appears relatively large, effectively framing the task as a standard object detection problem. The results, as shown in Figure 5, indicate a clear improvement, which in most cases is more pronounced than in the previous experiments. This suggests that the methodology remains proficient even when object scale is no longer a limiting factor.

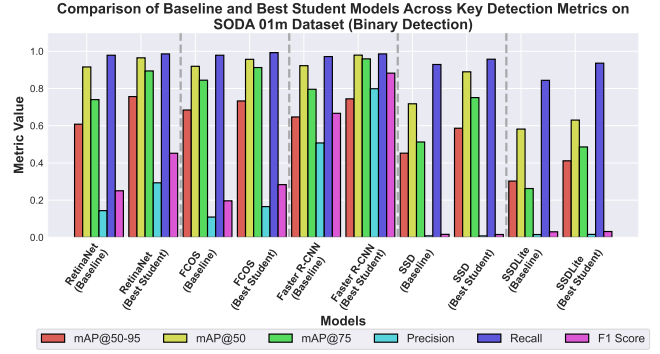


Fig. 5. Comparison of Baseline and Best Student Models Across Key Detection Metrics on SODA 01m Dataset (Binary Detection).

C. Cross-Dataset Performance Evaluation

In addition to evaluating the trained models on the dataset used during training, two further experiments were carried out to assess how well the models would perform on external litter detection datasets. Specifically, the binary litter detection models were tested on the BDW and UAVVaste datasets, both of which also frame the problem as binary litter detection. Due to the characteristics of the BDW dataset, where bottle litter appears at a larger scale, the models trained on the SODA dataset at one meter altitude were used for inference. Conversely, the binary SODA tiled models were applied to the UAVVaste dataset, given its focus on small-scale litter.

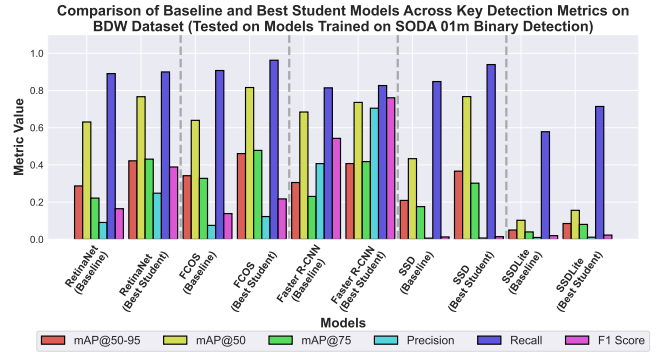


Fig. 6. Comparison of Baseline and Best Student Models Across Key Detection Metrics on BDW Dataset (Tested on Models Trained on SODA 01m Binary Detection).

In both cases (Figures 6, and 7), the student models continued to outperform their corresponding baselines, demonstrating that the benefits of the proposed methodology extend beyond the original training data. While SSD and SSDLite followed a similar trend to previous experiments, showing only marginal gains, the overall advantage of adopting the proposed approach remains evident.

The results of these experiments demonstrate substantial improvements across five object detection models applied to litter detection, with consistent advancements observed throughout. Notably, no architectural changes were made between the baseline and student models, nor was there any

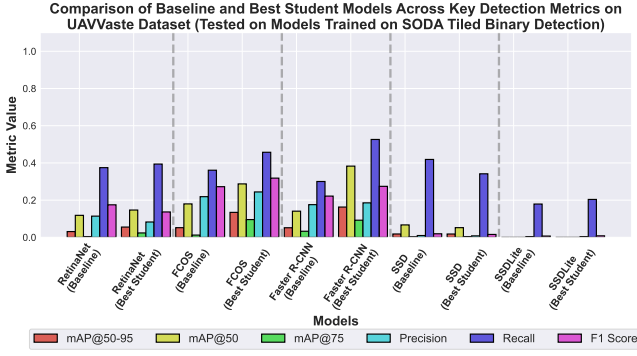


Fig. 7. Comparison of Baseline and Best Student Models Across Key Detection Metrics on UAVVaste Dataset (Tested on Models Trained on SODA Tiled Binary Detection).

increase in parameters or layers. Nevertheless, performance improved, albeit with slightly longer training times due to the added cost of generating privileged information and training a teacher model. As each result reflects a single experimental run, statistical analysis was not applicable.

V. CONCLUSION

This study proposed a novel methodology that integrates privileged information and knowledge distillation to improve litter detection, all without increasing model parameters or affecting inference time. The methodology was tested across five widely used object detectors, addressing different detection challenges, including small litter detection and standard object detection for objects at varying scales. A key contribution of this work is the introduction of a novel technique for encoding bounding box information, which is fed to the model as a binary mask. This approach was found to be informative, aiding the model in guiding the detection process more effectively. The results demonstrated consistent improvements when applying this methodology, both within the models trained on the SODA dataset and through cross-dataset evaluations on the BDW and UAVVaste litter detection datasets. These findings illustrate that the proposed methodology not only boosts performance on the dataset it was trained on, but also generalises well to other litter detection datasets. Importantly, these improvements were achieved without the need to increase model complexity or add new layers, making the approach both efficient and practical. As a natural extension of this work, future experiments could explore the generalisation capability of the approach on broader and more diverse benchmarks, including Pascal VOC and COCO.

REFERENCES

- [1] S. Kaza, L. C. Yao, P. Bhada-Tata, and F. V. Woerden, *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050*. Washington, DC: World Bank, 2018.
- [2] D. Bashkurova, M. Abdelfattah, Z. Zhu, J. Akl, F. Alladkani, P. Hu, V. Ablavsky, B. Calli, S. A. Bargal, and K. Saenko, "Zerowaste dataset: Towards deformable object segmentation in cluttered scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2022, pp. 21 147–21 157.
- [3] P. F. Proença and P. Simões, "Taco: Trash annotations in context for litter detection," *arXiv preprint arXiv:2003.06975*, 2020.
- [4] M. Kraft, M. Piechocki, B. Ptak, and K. Walas, "Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle," *Remote Sensing*, vol. 13, no. 5, 2021.
- [5] D. Pisani, D. Seychell, C. J. Debono, and M. Schembri, "Soda: A dataset for small object detection in uav captured imagery," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024, pp. 151–157.
- [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.
- [7] T. Wang, Y. Cai, L. Liang, and D. Ye, "A multi-level approach to waste object segmentation," *CoRR*, vol. abs/2007.04259, 2020.
- [8] M. Córdova, A. Pinto, C. C. Hellevik, S. A.-A. Alaliyat, I. A. Hameed, H. Pedrini, and R. d. S. Torres, "Litter detection with deep learning: A comparative study," *Sensors*, vol. 22, no. 2, 2022.
- [9] D. Pisani and D. Seychell, "Detecting litter from aerial imagery using the soda dataset," in *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, 2024, pp. 897–902.
- [10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016, cite arxiv:1506.02640.
- [11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.
- [12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, 2016, vol. 9905, pp. 21–37.
- [13] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE Computer Society, 2017, pp. 2999–3007.
- [14] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks : the official journal of the International Neural Network Society*, vol. 22, pp. 544–57, 07 2009.
- [15] V. N. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer," *J. Mach. Learn. Res.*, vol. 16, pp. 2023–2049, 2015.
- [16] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 825–832.
- [17] S. Wang, S. Chen, T. Chen, and X. Shi, "Learning with privileged information for multi-label classification," *Pattern Recognition*, vol. 81, pp. 60–70, 2018.
- [18] G. Habib, T. Jan Saleem, S. M. Kaleem, T. Rouf, and B. Lall, "A comprehensive review of knowledge distillation in computer vision," 2024.
- [19] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, and M.-M. Cheng, "Localization distillation for object detection," 2022.
- [20] S. A. McMains and D. C. Somers, "Multiple spotlights of attentional selection in human visual cortex," *Neuron*, vol. 42, no. 4, pp. 677–686, 2004.
- [21] M. Bartolo and D. Seychell, "Correlation of object detection performance with visual saliency and depth estimation," 2024.
- [22] K. Makantasis, K. Pinitas, A. Liapis, and G. N. Yannakakis, "From the lab to the wild: Affect modeling via privileged information," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 380–392, 2024.
- [23] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, 2019, pp. 9626–9635.
- [24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019.
- [25] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision - ECCV 2014 - 13th European Conference, September 6-12, 2014*, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693. Cham, Switzerland: Springer, 2014, pp. 740–755.