Computer Vision Project 2

Mikołaj Nowacki
75231

Kacper Kuźnik
75267

Abstract

Detecting waste and litter in images is an important computer vision problem with significant real-world impact. Automated waste detection can support applications such as environmental monitoring, smart city maintenance, large-scale ecological assessment, and automated litter mapping.

1. Introduction

In this project, we apply modern object detection models to the task of waste detection. We begin by evaluating pre-trained detectors on a subset of the TACO dataset to assess their zero-shot performance. Next, we fine-tune a state-of-the-art RT-DETR model on the TACO subset to improve detection accuracy and enable classification into waste-specific categories. Finally, we qualitatively assess how well the fine-tuned model generalizes to more challenging aerial imagery captured by drones, highlighting the challenges of small object size, cluttered backgrounds, and variable lighting conditions.

The report covers dataset preparation, model evaluation, fine-tuning, performance metrics, and failure analysis.

We were happy to discover that the UAV Waste Dataset was prepared at our home university, Poznan University of Technology.

2. Task 1: Object Detection Setup Implementation

This section evaluates YOLOv11, DETR, and RT-DETR on the first 1000 samples of the PASCAL VOC 2007 validation set. The associated code is available here.

2.1. Quantitative Analysis: Stability and Sensitivity

YOLOv11 exhibits high sensitivity to confidence thresholds, indicative of poor calibration. While it achieves a respectable mAP@50 of 0.7178 at a low threshold (0.1), performance collapses to 0.1279 at a threshold of 0.9. This suggests YOLO correctly identifies objects but assigns them low confidence scores (0.2–0.5), leading to False Negatives during standard filtering.

In contrast, DETR demonstrates exceptional stability, maintaining an mAP of $\approx$0.79 across the 0.1–0.5 threshold range and retaining 0.74 even at 0.9. This indicates DETR predictions are highly binary (either > 0.9 or non-existent). Furthermore, DETR achieves the highest detection density (5.77 objects/image), proving that its global attention mechanism successfully resolves small or occluded objects that are typically suppressed by NMS in YOLO and RT-DETR. RT-DETR functions as a hybrid, mirroring DETR's stability at lower thresholds but suffering performance degradation at strict settings.

2.2. Qualitative Analysis

Qualitative results align with the quantitative metrics (see Figure 1). DETR demonstrates superior scene understanding, successfully localizing small, context-dependent objects (e.g., a "clock") that other models miss. Conversely, YOLOv11 shows susceptibility to texture bias, misclassifying a plain cardboard box as a "refrigerator" because local texture cues override global spatial context.

In outdoor scenes (Figure 2), DETR detects objects with high certainty (> 0.99), whereas YOLOv11 identifies the same targets with significantly lower confidence (0.70–0.77), reinforcing the finding that YOLO is "underconfident." RT-DETR, while effective at localization, exhibits specific classification failures, such as mislabeling a sheep as a "cow," suggesting that its hybrid encoder may occasionally compromise fine-grained feature resolution.

3. Task 2: Image Waste Detection In the Wild

3.1. Dataset Preparation

For this project, we used the TACO (Trash Annotations in Context) dataset subset provided for the assignment. The subset includes pre-generated COCO-style annotations and a selection of images representing multiple waste categories. The dataset was delivered with predefined Training, Validation, and Test splits. Because the dataset is in standard COCO format, we load it using the torchvision.datasets.CocoDetection class. An image transform (transforms.ToTensor()) converts images to PyTorch tensors. To understand the dataset distribution, we counted the number of annotations per class across the three splits.

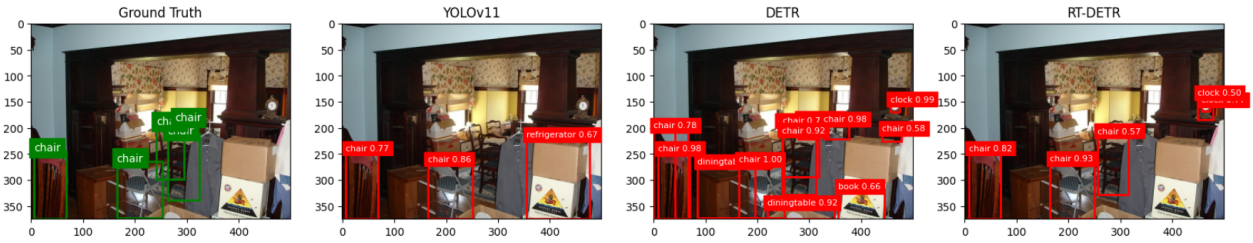


Figure 1. Indoor environment comparison.

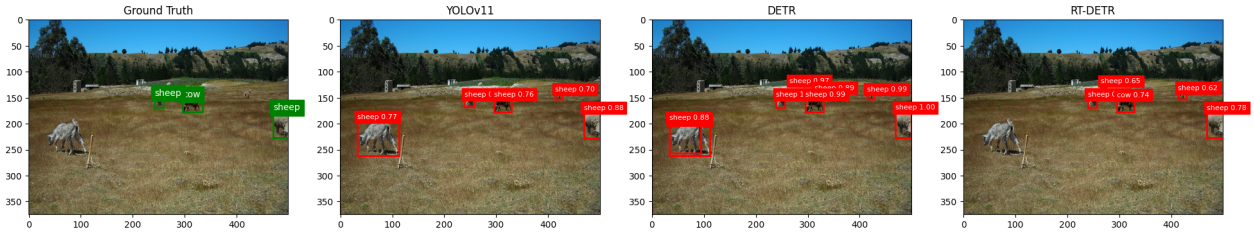


Figure 2. Outdoor environment comparison.

A helper function iterates over the COCO annotations and aggregates counts using collections.Counter. The counts from Train, Validation, and Test are merged into a single dataframe. A summary of the resulting statistics is shown in jupyter notebook. To gain an understanding of the dataset's visual complexity, we visualized example images from each class using bounding box overlays. The visualization script selects a few sample images per class and draws their annotations using matplotlib.

Visual inspection

This analysis reveals several key challenges encountered during object detection:

- **Cluttered scenes:** Many images contain multiple waste items or unrelated background objects, increasing the difficulty of accurate detection.
- **High background variability:** Images are captured in diverse environments such as streets, parks, and beaches, leading to substantial variations in texture, color, and overall appearance.
- **Object size variation:** Some items (e.g., cigarettes and bottle caps) are extremely small relative to the image resolution, making them harder for the detector to localize reliably.
- **Occlusion:** Waste items are frequently overlapping or partially hidden behind other debris, which reduces model visibility of object boundaries.
- **Lighting conditions:** Images display strong variations in illumination, including shadows, high contrast, overexposure, and low-light scenarios, all of which negatively affect detection performance.

3.2. Zero-shot Object Detection

In this subtask, we evaluate the zero-shot generalization ability of three pretrained object detection models - YOLO, DETR, and RT-DETR on the TACO waste detection dataset. None of these models were fine-tuned on TACO. They were used relying only on knowledge learned from the COCO 2017 dataset.

Model Inference

For each model, we perform inference on every image in the test split of the TACO subset. Since the COCO 2017 categories used by YOLO, DETR, and RT-DETR do not overlap with the waste classes present in TACO, we adopt a **class-agnostic evaluation protocol**. Detection quality is therefore assessed purely based on bounding box localization, independent of any semantic class predictions.

Let objs_i denote the set of ground-truth waste objects in image i. For each model, the predicted bounding boxes are sorted by confidence score, and the top-k detections are selected, where

$$k = |\mathrm{objs}_i|.$$

This ensures that an image with three ground-truth waste objects produces three predicted detections for evaluation, regardless of the predicted classes or model architecture. Each model is evaluated in inference mode, and computations are performed on GPU.

Exploring Model Configurations

To obtain the highest possible zero-shot performance, several model configurations and hyperparameters were evaluated. Specifically, we varied confidence score thresholds,

which influence the ranking of detections.

The following thresholds were chosen for evaluation:

YOLO: $\{0.10, 0.25, 0.50\}$ DETR/RT-DETR: $\{0.30, 0.50, 0.70\}$

Since only the top-k detections are ultimately used, the threshold primarily influences confidence ordering, rather than the total number of predictions.

Performance Metrics

Model predictions are evaluated using the `torchmetrics` implementation of mean Average Precision (mAP). Because this subtask is class-agnostic, all predicted bounding boxes are assigned label 0, and all ground-truth objects are also assigned label 0. This reduces the evaluation to a pure geometric localization task. We report:

$$\text{mAP} \quad \text{and} \quad \text{mAP}_{50},$$

Improved Bounding Box Selection

As an additional strategy, we implemented an improved matching procedure aimed at better localization. Instead of selecting only the top-k boxes, we consider the **top-10** detections ranked by confidence. For each ground-truth bounding box, the predicted box with the highest IoU is selected:

$$\hat{b}_j = \arg \max_{b \in \text{Top-10}} \text{IoU}(b, g_j),$$

where g_j is a ground-truth box.

Zero-shot Detection Results

Table 1 reports the zero-shot performance across all models and configurations. Each row corresponds to a specific model and confidence threshold. RT-DETR achieves the highest overall performance at a confidence threshold of 0.3.

Table 1. Zero-shot mAP results on TACO test split (class-agnostic evaluation).

Model / Config	mAP	mAP50
DETR-conf0.3	0.0572	0.0863
DETR-conf0.5	0.0525	0.0777
DETR-conf0.7	0.0409	0.0593
RT-DETR-conf0.3	**0.0772**	**0.1013**
RT-DETR-conf0.5	0.0604	0.0721
RT-DETR-conf0.7	0.0400	0.0487
YOLO-conf0.1	0.0517	0.0630
YOLO-conf0.25	0.0334	0.0389
YOLO-conf0.5	0.0212	0.0234

Discussion

The results indicate that transformer-based detectors generalize better in a zero-shot setting. RT-DETR achieves the best mAP values, suggesting stronger localization ability even for objects outside its training categories. DETR provides stable performance across mid-range thresholds. YOLO, while efficient, performs worse in zero-shot scenarios, likely due to stronger class-specific biases and confidence calibration issues.

Lower confidence thresholds consistently lead to higher mAP across models, confirming that COCO-trained detectors assign low confidence to waste objects despite being able to localize them reasonably well. This effect is particularly notable for YOLO.

Overall, transformer-based architectures appear to be more robust for zero-shot waste detection, although all models struggle significantly compared to their performance on COCO.

3.3. Fine-Tuning an Object Detector

In this subtask, we fine-tune an object detection model on a subset of the TACO dataset in order to improve performance on waste detection in real-world images. Pre-trained detectors are typically optimized for generic object categories, and therefore fine-tuning enables the model to specialize in the visual characteristics of waste objects while also learning to classify them into the six TACO-derived categories.

Model Selection

We selected the RT-DETR v2 architecture, using the checkpoint `PekingU/rtdetr_v2_r50vd` available on HuggingFace.

Training Procedure

The model was fine-tuned on the TACO training split using the HuggingFace `Trainer` API. The validation split was used to select the best model checkpoint based on the mean Average Precision (mAP). We incorporated lightweight data augmentation using the Albumentations library, including horizontal flips, brightness/contrast changes, and mild color jittering to improve generalization.

Training was performed for 10 epochs with the following hyperparameters:

- Learning rate: 5×10^{-5}
- Batch size: 8
- Warmup steps: 300
- Gradient clipping: 0.1
- Precision: bfloat16
- Model selection metric: validation mAP

Evaluation on the Test Set

After training, we evaluated the best model on the TACO test split. Global and per-class metrics were computed using the

`torchmetrics.detection.MeanAveragePrecision` implementation with the `class_metrics=True` flag enabled.

Table 2 summarizes the final performance.

Metric	Value
mAP	0.2467
mAP@50	0.3300
mAP@75	0.2531
mAP (small)	0.1350
mAP (medium)	0.2695
mAP (large)	0.4695
mAP (Plastic bag & wrapper)	0.3179
mAP (Cigarette)	0.0538
mAP (Bottle)	0.3899
mAP (Bottle cap)	0.1108
mAP (Can)	0.3096
mAP (Carton)	0.2982

Table 2. RT-DETR v2 fine-tuning results on the TACO test dataset.

Failure Analysis

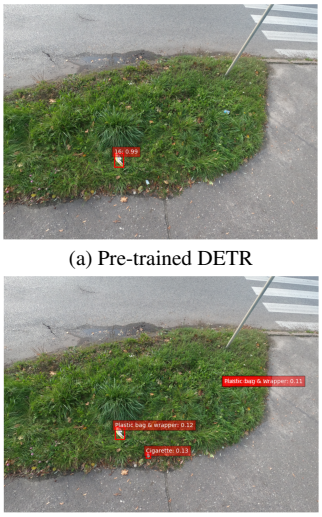
A qualitative inspection of detection results revealed several recurring failure modes:

- **Small objects** (e.g., cigarettes, bottle caps) were frequently missed due to limited pixel resolution and class imbalance.
- **Localization errors** occurred in cluttered scenes with overlapping waste items.
- **False positives** were triggered by visually similar textures or reflective surfaces.
- **Truncated or partially visible objects** were prone to low-confidence predictions or misclassification.

Overall, the fine-tuned RT-DETR model achieved very good performance on the TACO test set, substantially outperforming the zero-shot detectors evaluated in the previous subtask. Most waste objects were correctly localized and classified, demonstrating the effectiveness of fine-tuning on domain-specific data.

However, optimal performance required careful adjustment of the confidence threshold for predicted boxes. Lowering or raising the threshold had a significant impact on accuracy

- When the threshold is set **too high**, many valid detections are discarded, which leads to **missed objects** and lower recall. This is particularly problematic for small or partially occluded waste items.
- When the threshold is set **too low**, the model retains many low-confidence predictions, resulting in an increased number of **false positives** and lower precision. While recall improves, the quality of the detections decreases.



(a) Pre-trained DETR

(b) Fine-tuned RT-DETR

Figure 3. Comparison of a Pre-trained DETR and a Fine-Tuned RT-DETR model.

4. Task 3: Generalizing to Aerial/Drone Images

The transition to UAV imagery introduces a significant challenge: at TACO's ground-level, large objects contrast sharply with the drone's top-down, small-scale inputs. Consequently, confidence thresholds had to be lowered below 10% to achieve plausible recall, indicating that models must rely on weak textural cues rather than strong geometric certainty in this new domain.

Figure (3) illustrates the impact. The pre-trained DETR (Fig. 3a) fails to detect a distinct blue bottle, likely due to its reliance on canonical side-views. Conversely, the fine-tuned RT-DETR (Fig. 3b) successfully localizes the waste, suggesting it learned material properties rather than just geometry. However, this sensitivity causes False Positives in textured backgrounds, emphasizing the need for domain-specific training to address the small object detection challenge effectively.

5. Conclusion

Transformer-based detectors generalize better than YOLO for zero-shot waste detection, but performance improves significantly after fine-tuning on the TACO dataset. RT-DETR achieved accurate localization and classification, though small and occluded objects remain challenging. Extending to aerial imagery highlights the need for domain-specific training. Overall, fine-tuned transformers offer a strong approach for real-world automated waste detection.