Trash Into Treasure: Classifying Garbage from Drone Imagery Using Image Classification Algorithms

Alyssa Fong
Stanford University
Stanford, CA
amlfong@stanford.edu

Abstract

The issue of waste management has been of great interest to the computer vision community over the last three decades, with environmental problems pushing this issue to the fore of public discourse. Previous work in this field have attempted to tackle this problem from a plethora of different avenues. This paper seeks to apply computer vision models, specifically image classification models, to images taken from a drone flying over a beach. Here, the performance of YOLO-v8, DINOv2, and MobileNetv4 are compared on the same dataset. Here, we find that the YOLO-v8 model performs the best on the provided task, achieving a 100% accuracy on the test set, followed by DINOv2 with a 98.6% accuracy, finally followed by MobileNetv4 with a 97.2% accuracy. Thus, the models perform similarly across the board. These results provide evidence that a lighter model such as MobileNet can keep up with more complex models and may be useful for groups looking for low-cost options for fieldwork, but that YOLO-v8 seems to be the best model for this task.

1. Introduction

The issue of managing solid waste production has been a thorn in the sides of communities globally as consumption, waste production, and population size has exploded within the last century with the rise of industrialization [11]. In addition to being a difficult logistical problem, increasing waste production has had major adverse environmental effects on wildlife and in hydrological contexts, where disease-carrying insect vectors like mosquitoes can breed and where chemical byproducts can effectively poison residents with ammonia and methane [18].

To address this issue, communities—especially urban ones—have turned to municipal solid waste (MSW) management. At a non-technical community level, this can be implemented in the usage of distinct recycling and trash collection. In the last two decades, advancements in the field of computer vision have allowed scientists to explore the idea of using automated systems to complete this work, as research has shown that humans' ability to sort waste into defined categories is less than adequate [7]. This problem has been explored within the computer vision community, following the shift from traditional machine learning strategies to deep learning techniques.

A review completed by Lu and Chen [8] provides a roadmap for how computer vision models have been used since 1997 to attempt to address this problem, ranging from linear classifiers to region-based convolutional neural networks (R-CNNs), with increasing success on the task. However, there are major limitations on the usefulness of the work. For one, many existing datasets only work on well-defined objects on a plain background, which is not how most waste is deposited. Thus, they limit the utility of this research to being a tool to aid human sorting at the time of waste production rather than after waste collection.

However, there are some special cases where these models can be applied. One such case is in that of environmental monitoring groups such as Save our Shores [12], who organize beach cleaning days with local volunteers annually. While these efforts may make headway into making people more cognizant of their waste production, the task itself is rather arduous. One way of streamlining this task is using drone imaging to remotely detect garbage and then sending volunteers to physically collect the trash. In MSW management, implementation of this remote sensing method has been steadily increasing [13]. With waste-specific implementations showing promise in denser waste collection areas, applying computer vision methods to drone-mediate remote sensing footage to help these groups in their collection efforts seems fruitful.

2. Related Work

As mentioned above, a large review has previously been published on the application of computer vision models to

MSW, written by Lu and Chen [8]. To expand on this, this section will delve deeper into the history of work on this topic.

2.1. Early Work

The first paper in this field is a system created for robot-mediated paper recycling, published in 1997 by Faibish et al. [5] that used geometry and texture data, provided by both stereo vision and sensors, that the authors say draws from 'automated target recognition' systems. For this paper, the authors tried both linear classification and nearest-neighbor analysis, and found that the linear classifier reported the most accurate results. While this project merged both computer vision and robotics tasks into one, it was a landmark step in setting up the field to use these methods in waste management. One of its largest weaknesses, which they acknowledge in the discussion section, was that they relied heavily on multiple kinds of sensors to get the level of detail necessary to classify the images properly. Remarkably, however, Lu and Chen find that the accuracy of their model outperformed later linear classification models through the mid-2010s. Other models created for classification tasks, such as Ramli et al. [10] writing on plastic bottle classification, focused more heavily on easily visual differences, while still using the linear classifier, this time with slightly better performance than Faibish. While these papers led those collected by Lu and Chen that used linear classifiers, more complex, deep learning models have since performed much better.

2.2. Deep Learning Models

Deep learning models have since revolutionized the field of computer vision, and in turn created more complex systems for researchers to build upon. For these models, researchers focused in two different directions: for one, fine-tuning a number of models to reveal which one included the best underlying understanding for this specific task. Another approach involved training a model without any pre-trained weights. Comparing these two approaches provided researchers in this area with a direction to focus on. One such paper, written by Bircanoğlu et al. in 2018 [3], examined how ResNet, ImageNet, DenseNet, XCeption, and MobileNet performed on these tasks, with ResNet Inception, the best-performing model without pre-trained weights, was handily outperformed by a fine-tuned model built upon DenseNet121, a model that uses Dense CNNs. On average, DenseNet and VGG performed the best out of the 12 deep learning-specific models that Lu and Chen examined. [6] [15] [14]

This research suggests that the best-performing methodology comprises of fine-tuning pre-trained model weights on specific data. While MSW analysis covers a large range of tasks, this conclusion makes sense within the observations of the larger computer vision field, where deeply trained models are able to discern patterns that specifically trained models may not. By using transfer learning from models trained on objects and living things, fine-tuned models are able to harness that enriched knowledge for better predictions.

3. Dataset and Features

The dataset used in this project comes from a set of videos taken from drone footage on beaches in Taiwan, originally recorded by National Cheng Kung University. At this time, there seem to be no publications that stem from this data, but the model that they used is available on Roboflow. Despite its publishing, there is very little public metadata about this dataset. The model, based on YOLO v8, attempted to classify objects on the beach into one of the following trash categories (with examples):

1. **Glass**: glass bottles, containers
2. **Plastic Bottle/Takeaway Cup**: Fast food drink containers, styrofoam cups.
3. **Retort pouch**: Layered metal and plastic bags, such as those that contain jerky.
4. **Take-away container**: Styrofoam or plastic containers used to carry food.
5. **Aluminum Cans**: Sealed preserve containers, etc.

Additionally, this dataset was already split into training, validation, and testing sets, which was useful in comparing the original base model to the alternatives that were explored in this project. This split was created by user Vanshika Mishra on Kaggle, which is the source from which the data for this project was downloaded [17].

Table 1, below, shows the split between classes and training, validation, and testing sets. There is a clear bias towards plastic bottles and aluminum cans in the dataset, while glass is underrepresented (less than 5% of the images).

While the specificity of these categories may seem limiting, this approach can be helpful to environmental efforts that seek to remove trash from these areas before they can be washed into the sea, where many public campaigns have shown the adverse effects that these can have on wildlife. For instance, aluminum cans and plastic bottles have been shown to suffocate ocean animals, and other sharpened pieces of waste can also harm them [9]. By automating remote sensing of certain areas of interest, a model trained on this data will be able to make the job of these groups, who are often comprised of volunteers, much more streamlined.

Table 1: The distribution of images across the five described garbage classes and along the train/validation/test divisions. Overall, there are 1,898 images in the dataset.

Garbage Class	Training	Validation	Testing	Total
Glass	81	6	6	87
Plastic Bottle (Takeaway)	567	63	20	650
Retort Pouch	339	33	16	388
Takeaway Container	129	6	4	139
Tin/Aluminum Cans	594	16	24	634

Figure 1: Examples of images in each category's training folders.

Due to the highly differential nature of these categories, it will be interesting to see how transfer learning of general models performs here, and if the models' ability to understand different shapes will be useful in this context. Examples of images from each category can be seen in figure 1. One interesting detail is that despite being taken from the same drone video, each image is a different size, an issue that will be addressed in the Methods section.

4. Methods

Drawing inspiration from the comparative nature of many previous studies, this project sought to compare three different cutting-edge deep learning models to understand which may perform best on this specialized task. The dataset used was constructed to use in an image classification model. As one of the most common computer vision tasks, there are a number of models available to perform this task. However, the added functionality of certain models to be built into drones is something that has yet to be explored with this data. Thus, the models chosen to fine-tune for testing on this dataset were all lightweight, image classification-able models: YOLO-v8, DINOv2, and MobileNetv4.

4.1. YOLO-v8

As mentioned above, the original creators of the dataset intended for the images to be used in a You-Only-Look-Once (YOLO) model, specifically YOLO-v8. This approach to object detection was revolutionary in the quest to create models that were both fast and reliable. The original YOLO model contains 24 convolutional layers, trained on ImageNet classification, that are then fed into 2 fully-connected layers, that predict bounding boxes and labels from a single pass of the image, which classifies it as a single-stage object detector, similar to SSD or RetinaNet.

Later models built on this backbone, with large bounds being made in YOLO-v3, with the replacement of the original architecture with Darknet-53, which is a complex convolutional neural network (CNN) that comprises of 53 convolutional layers. This deep architecture allowed YOLO-v3 to perform better on object detection tasks with small objects, while still maintaining its speed.

The release of YOLO-v5 and YOLO-v8 by the company Ultralytics have made YOLO even more easily accessible to the general public, and remain as popular models for object detection tasks. Importantly, the introduction of YOLO-v8 allowed for the model to support a wider berth of tasks, including the image classification that this project harnesses. In addition, YOLO-v8 has also integrated transformer-based attention into its architecture, improving even more upon the leaps that YOLO-v3 made over the original model.

4.2. DINOv2

DINOv2 is a state-of-the-art object detection model created by Meta engineers that relies heavily on self-distillation, which can be explained more handily as a student-teacher model. This concept, known popularly as "Be Your Own Teacher," was introduced by Zhang et al. [19]. It relies on knowledge distillation, which is the idea of training a model to match the outputs of another, larger model, which in practice provides similar outcomes much more cheaply [2]. In the case of DINOv2, an identically-constructed teacher and student vision transformer model are fed variations of the same image, and the

student learns from the teacher to match its output via backpropagation. The student's performance is then used to update the weights of the teacher's model via a moving average, making it more stable than the student's. Importantly, this is a self-supervised model, which means that it will perform well without pre-labeled training data. While that feature is less applicable in this more traditional case, this is one of the major selling points for using DINOv2.

In this project, DINOv2 is used as a backbone and combined with a simple linear classifier head that takes in the features that DINOv2 discovers to robustly predict the category of waste shown in the image. I define a range of augmentations that can be performed to make the self-distillation process more robust, including resizing, flipping, rotations, perspective shifts, and color changes.

4.3. MobileNetv4

MobileNet is known as a state-of-the-art architecture that is used for relatively resource-light computational settings such as mobile devices. The family of MobileNet architectures are built on the concept of depthwise separable convolutions, which decomposes a convolution into two steps:

1. A depthwise grouped convolution that performs a convolution for all M channels in an input.

2. A point-wise convolution that applies N filters in a 1x1 convolution.

This dramatically reduces the computational load of running a model without sacrificing its efficacy, allowing complex models to be applied in a wider set of contexts. Mobile devices such as cell phones are one such context, which in this case may be a volunteer organizer's cell phone that is being fed imagery from a drone.

MobileNetv4's updates are very similar to those of YOLO-v8 in that it introduces transformer layers into the model architecture. It also mirrors YOLO-v8 in that this new version updates the single model to perform a wider range of tasks on its own, including classification which is used in this project. Further optimizations for mobile hardware and more condensed architecture are also included in this new version.

In all three cases, the smallest available model was used as a way to reflect the computational capabilities in practice for these models.

5. Results

In running these models, I aimed to use similar batch sizes, epoch counts, learning rates, and optimizers across all three. As this project focused on building off of existing models, I used one of the most common optimizers, AdamW. For this optimizer, it is common practice to set the learning rate to 0.001, as platforms such as PyTorch and Keras set their default learning rate to that value. Additionally, I used batch sizes of 32, which may not have worked very well given the disparities with the smaller waste categories (such as glass). Still, this batch size settled as a happy medium between the cons of both small and large batch sizes. In terms of epoch count, I allowed the original YOLO-v8 model to run for 100 epochs or until it recognized no improvement in performance, which ended at 27 epochs. Thus, for both DINOv2 and MobileNetv4 the epoch count was also set at this number to understand their comparative performances.

	validation		test	
Backbone	Acc	mAP	Acc	mAP
YOLO-v8	100%	1.000	100%	1.000
DINOv2	97.6%	0.969	98.6%	0.960
MobileNetv4	93.1%	0.997	97.2%	1.000

Table 2: Accuracy & Mean Average Precision (mAP) 50 values for the three models.

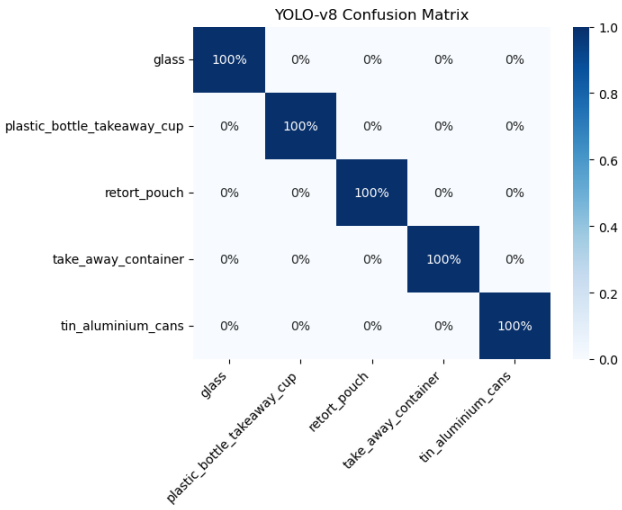

Figure 2: YOLO-v8 Confusion Matrix

Overall, all three models performed extremely well on the data, with overall test and validation set accuracy and mean average precision (mAP)-50 reported for all three models in table 2, which are common metrics for performance on classification tasks. Here, we can see that all models had an accuracy greater than 97% on the testing set with YOLO performing the best at 100% accuracy.

The normalized confusion matrices for each model show a similar story of high performance, as seen in figures 2, 3, 4, where all but three categories had perfect performance:

- DINOv2 glass classified as a takeaway container
- MobileNetv4 plastic bottle/takeaway cup classified as

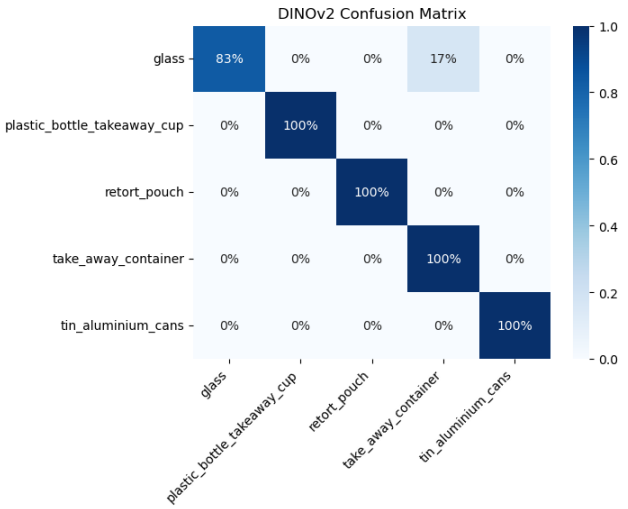

Figure 3: DINOv2 Confusion Matrix

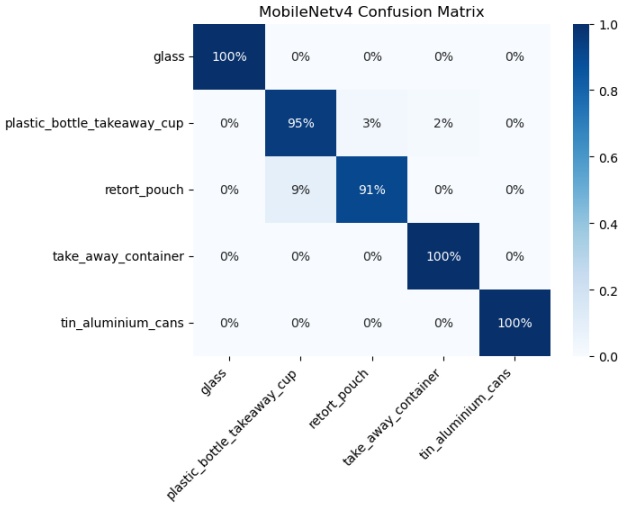

Figure 4: MobileNetv4 Confusion Matrix

retort pouch and takeaway container

- MobileNetv4 retort pouch classified as plastic bottle/takeaway cup

One oddity of these results is the perfect performance of the YOLO model. While this would explain why the dataset authors decided to use the YOLO framework, I would be interested in using more images from the same beach, preferably at different times of day or seasons, to see whether this is a coincidence with this specific set of images, or if the YOLO framework is by far the most useful model for this task. Closer examination of the augmented images as seen in Figure 5 shows that some of the transformations seem to mimic these different conditions, and the model still seems to work as well as it does.

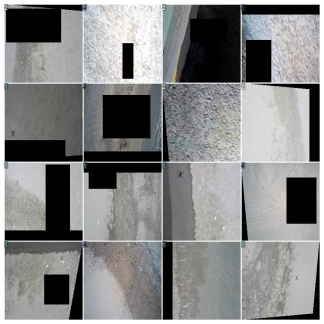

Figure 5: A sample from a batch from YOLO-v8 training

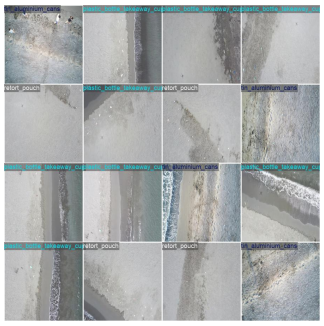

Figure 6: Testing images with their respective *true* labels.

Figures 6 and 7 showcase the success with which YOLO labeled parts of a testing batch. While these are all classified properly, it is very difficult to discern some of the waste. This introduces the question about a limitation of the dataset to only the five categories, as it's possible that real, unannotated drone footage may not capture any waste in a frame. In that novel case, it's certain that any of these models trained on this dataset may not be applicable in a real-life context.

These results showcase that novel classification models are making deep headways into at least one facet of the complex link between MSW and computer vision and pave a way for environmental efforts to coalesce with computer vision researchers to create low-cost solutions for these problems. Their promise, however, raises a wealth of new questions that may be able to be solved using these models and others, related or novel.

6. Further Work

In this paper, multiple new classification models were tested against one another to attempt to understand which of these three major new models would perform best on this specific classification task. However, all three models are part of different families of model architecture, which opens up questions about what other models might be available to apply to this dataset.

One such model is PAWS, which was released alongside

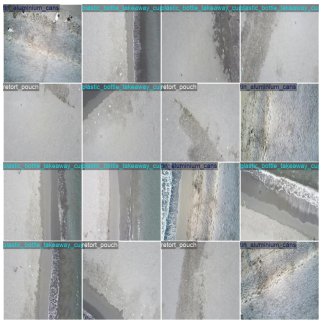

Figure 7: Testing images with their respective *predicted* labels by YOLO-v8.

DINO by Meta's AI division. This model differs slightly from DINO in that it is not fully self-supervised, but rather uses a small amount of labelled input data to learn how to label unlabeled data [4]. While this was an avenue I was unable to fully explore, an expansion of this dataset would be a great way to test the efficacy of this model in more complex situations.

Another interesting avenue to consider is the application of YOLOv9 through 11, which have been released over the past year, far past the publication of this dataset. At the moment, since YOLOv9 and YOLOv10 are currently only used for object detection tasks according to the Ultralytics site, they was not a candidate for this project [16]. However, a further project attempting to build a classification head over the YOLOv9 backbone would be another interesting extension, similarly to what was built here over DINO. It would also be interesting to see how YOLOv11, which was very recently released, performs against YOLOv8, as it is their first major update to include a classification head.

While this project focused on examining other methods of image classification, another project that can be built off of this one is to use DINO and YOLO as they were meant to be—object detection models. While this project is a good first step in examining how individual frames might be classified, training these models on complete drone videos may provide more practical considerations for drone-based scanning and waste management practices. In these more practical cases, it may also be worth adding a sixth category that covers the case in which there is no waste captured. Beach cleanup tasks where a drone classifies everything it flies over as having waste, though possible, would negate the need for this kind of technology.

On a similar note, as mentioned in the previous section, extensions of this dataset would be incredibly helpful in determining the best model for this task. Creating a diverse dataset of beaches from around the world, including different types of sand at different times of the year and at varying distances from coastlines would make this set more robust. As drone imaging technologies continue to grow cheaper [1], this kind of international effort becomes much more feasible. In doing this, it will likely require an expansion of the categories beyond the idea mentioned prior of a "no waste" class. Waste is diverse around the world, and it would be a wonderful challenge to examine if these models can hold up as research into this topic increases in complexity.

References

[1] J. . How much does a drone cost in 2023? here's a price breakdown, 12 2022. 6

[2] Z. Allen-Zhu and Y. Li. Three mysteries in deep learning: Ensemble, knowledge distillation, and self-distillation, 01 2021. 3

[3] C. Bircanoglu, M. Atay, F. Beser, O. Genc, and M. A. Kizrak. Recyclenet: Intelligent waste sorting using deep neural networks. *2018 Innovations in Intelligent Systems and Applications (INISTA)*, 07 2018. 2

[4] P. Bojanowski, M. Rabbat, A. Joulin, N. Ballas, M. Caron, and M. Assran. Advancing the state of the art in computer vision with self-supervised transformers and 10x more efficient training, 04 2021. 6

[5] S. Faibish, H. Bacakoglu, and A. Goldenberg. An eye-hand system for automated paper recycling. 11 1997. 2

[6] J. D. Lau Hiu Hoong, J. Lux, P.-Y. Mahieux, P. Turcry, and A. Aït-Mokhtar. Determination of the composition of recycled aggregates using a deep learning-based image analysis. *Automation in Construction*, 116:103204, 08 2020. 2

[7] H. Liu, H. Shang, and J. Yu. Why waste sorting implementation remains ineffective: A theoretical analysis based on the generation process of waste sorting behavior in china. *Waste Management*, 201:114812, 06 2025. 1

[8] W. Lu and J. Chen. Computer vision for solid waste sorting: A critical review of academic research. *Waste Management*, 142:29–43, 04 2022. 1, 2

[9] N. O. M. D. Program. What is marine debris? — orr's marine debris program, 03 2024. 2

[10] S. Ramli, M. M. Mustafa, D. A. Wahab, and A. Hussain. Plastic bottle shape classification using partial erosion-based approach. pages 1–4, 05 2010. 2

[11] H. Shah, R. Yasmeen, M. Sarfraz, and L. Ivascu. The repercussions of economic growth, industrialization, foreign direct investment, and technology on municipal solid waste: Evidence from oecd economies. *Sustainability*, 15:836–836, 01 2023. 1

[12] S. O. Shores. About. 1

[13] N. Sliusar, T. Filkin, M. Huber-Humer, and M. Ritzkowski. Drone technology in municipal solid waste management and landfilling: A comprehensive review. *Waste Management*, 139:1–16, 02 2022. 1

[14] C. Srinilta and S. Kanharattanachai. Municipal solid waste segregation with cnn. *2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*, 07 2019. 2

[15] L. Sun, C. Zhao, Z. Yan, P. Liu, T. Duckett, and R. Stolkin. A novel weakly-supervised approach for rgb-d-based nuclear

waste object detection. *IEEE Sensors Journal*, 19:3487–3500, 05 2019. 2

[16] Ultralytics. Models, 11 2023. 6

[17] M. Vanshika. Drone garbage detection dataset. Kaggle, 2024. Accessed: June 05, 2025. 2

[18] S. E. Vergara and G. Tchobanoglous. Municipal solid waste and the environment: A global perspective. *Annual Review of Environment and Resources*, 37:277–309, 11 2012. 1

[19] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *International Conference on Computer Vision*, 10 2019. 3