Waste Classification and Management Using Computer Vision

Sue Deng
ICME
suedeng@stanford.edu

Annie Fan
ICME
anniefan@stanford.edu

Jason Sun
MBA
jasonsun@stanford.edu

Abstract

Waste misclassification contributes significantly to landfill overuse and methane emissions, yet existing recycling systems lack real-time, scalable tools for accurate sorting detection and feedback to waste generators. In this work, we develop and benchmark deep learning models for image-based waste classification using the TACO dataset. We compare custom CNNs, ResNet34, and Vision Transformers (ViTs), evaluating classification accuracy, inference performance, and deployment feasibility. Our results reveal that architectural improvements significantly outweigh data augmentation benefits, with ViTs achieving the highest F1 score through superior spatial relationship modeling, followed by ResNet34 and baseline CNNs. Surprisingly, class imbalance was not the primary limiting factor—inherent visual distinguishability varies across waste categories regardless of sample size. All models exceed real-time processing requirements, and their sub-millisecond inference times enable cost-effective edge deployment. This work demonstrates that computer vision can provide economically viable solutions for waste management infrastructure, from centralized sorting facilities to distributed smart bins, supporting improved recycling rates and reduced landfill emissions.

1. Introduction

Improper waste sorting significantly contributes to landfill expansion and greenhouse gas emissions. In the United States, landfills are the third-largest source of anthropogenic methane, a greenhouse gas over 25 times more potent than CO_2 over a 100-year period [15]. Each year, over 146 million tons of municipal solid waste are landfilled [14], much of which could be recycled or composted if properly sorted. However, contamination in recycling streams—caused by incorrectly sorted waste—frequently renders entire batches unrecoverable. U.S. curbside recycling programs report average contamination rates between 17% and 25% [8], with some studies estimating that only 21% of recyclable materials from households are successfully recovered [9].

Unlike utilities such as electricity or water, waste management lacks a direct, linear feedback mechanism for consumers. Individuals rarely see the downstream impact of their sorting decisions, making behavior change difficult. As a result, systems that automate or assist in accurate waste classification at the point of disposal can play a key role in improving diversion rates and reducing landfill dependence.

This project addresses that opportunity by developing a computer vision model capable of classifying waste items from images to waste categories (e.g. Can, Paper, etc.). We investigate the use of Vision Transformers (ViTs)—an underexplored architecture in this domain—and compare their performance to established CNN baselines like ResNet, DenseNet, and MobileNet. Our work aims to evaluate both classification accuracy and deployment feasibility, with the broader goal of contributing to more sustainable and data-driven waste management systems.

Despite the growing application of convolutional neural networks (CNNs) to waste classification, most current approaches rely on traditional architectures trained on relatively small, imbalanced datasets. These models perform well in ideal conditions but often struggle with real-world waste streams that include visual clutter, deformation, and contamination. At the same time, Vision Transformers (ViTs) have demonstrated strong performance in large-scale image recognition tasks due to their ability to model long-range spatial relationships—but remain underexplored in the context of waste sorting.

This project aims to fill that gap by evaluating the effectiveness of ViTs for multi-class waste classification. We benchmark their performance against standard CNNs in terms of classification accuracy, robustness to class imbalance, and deployment efficiency (e.g., inference latency and model size). Specifically, we investigate whether ViTs can generalize better to rare classes and messy real-world data, and whether their higher computational cost is justified by gains in accuracy that could meaningfully improve recycling recovery rates in real-world deployments in high and low compute-availability contexts.

2. Related Work

Recent work has applied deep learning, especially convolutional neural networks (CNNs), to image-based waste classification. Models such as DenseNet [3], ResNet [2], and MobileNet have been fine-tuned on public datasets like *TrashNet* [13], *TACO* [10], and *WaRP* [7]. These datasets range from clean, single-object scenes (TrashNet, 6 classes) to complex, cluttered real-world litter scenes (TACO, 60+ classes). For simple classification tasks, accuracies often exceed 95%, but performance drops on fine-grained or noisy datasets—e.g., Sayem et al. [12] reported 83.1% accuracy across 28 waste categories using a dual-stream CNN trained on WaRP.

Efforts to enable real-world deployment have focused on lightweight models like MobileNetV2, which achieved approximately 90.7% accuracy and 0.6s inference latency on Raspberry Pi hardware [4]. Detection architectures such as YOLOv8 have also been adapted for conveyor belt sorting in MRFs, with inference times under 20ms per frame [7]. Hybrid segmentation-classification pipelines, such as SAM + MobileNetV2, offer modularity and performed robustly on mixed waste streams with accuracy ranging from 86–97% [6]. ResNet variants (e.g., ResNet-50) have also served as strong backbones in prior studies, often achieving over 94% accuracy on small-to-medium datasets like TrashNet [2].

Despite the rapid progress of CNN-based methods, relatively few studies have explored the use of Vision Transformers (ViTs) for waste classification. Recent work by Sayed et al. [11] and Kumar et al. [5] shows that ViTs can offer competitive accuracy—particularly when pretrained on large-scale data—but often require more training samples and computation than traditional CNNs. For example, ViT-base achieved around 88% accuracy on a 28-class waste dataset, slightly outperforming baseline CNNs. However, these models remain underused due to their computational demands and the limited size of most trash datasets. Our motivation in this work is to evaluate the viability of ViTs in waste classification by leveraging modern data augmentation and transfer learning to compensate for small dataset size, with the goal of testing their generalization on contaminated, cluttered, and real-world trash images.

Table 1: Reported Accuracy and Inference Speed by Model Type

Model	Dataset	Accuracy (%)	Time	Notes
ResNet-50	TrashNet	93–96	∼120ms	Strong baseline
DenseNet-121	TrashNet	95–99	∼100ms	High accuracy, clean data
MobileNetV2	Huawei Trash	90.7	0.6s	Edge deployment
YOLOv8n	WaRP	85–90 mAP	∼15ms	Real-time detection
SAM + MobileNetV2	Mixed	86–97	0.3–1s	Segmentation + classification

3. Methods

3.1. Baseline: CNN

We utilized a Convolutional Neural Network (CNN) as the baseline model to establish performance benchmarks. The baseline CNN model consists of three convolutional blocks, each containing a convolutional layer, batch normalization, ReLU activation, and max pooling, followed by a fully connected classification layer.

3.2. ResNet-50 Architecture

ResNet-50 (Figure 1) is a 50-layer deep convolutional neural network that uses residual blocks with skip connections to enable efficient training of very deep models. Each block employs a bottleneck structure with three convolutional layers. These bottleneck blocks are stacked in four main stages, following an initial convolution and pooling layer. In the end, it uses a softmax activation to predict probabilities across 27 superclasses.

Mathematically, a residual block can be described as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

where $\mathbf{x}$ is the input to the block, $\mathcal{F}$ is the residual function (a stack of three convolutional layers with weights $\{W_i\}$), and $\mathbf{y}$ is the output. The bottleneck structure is:

$$\mathcal{F}(\mathbf{x}) = W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot \mathbf{x}))$$

where W_1, W_2, and W_3 are the weights of the 1×1, 3×3, and 1×1 convolutions, and σ denotes the ReLU activation.

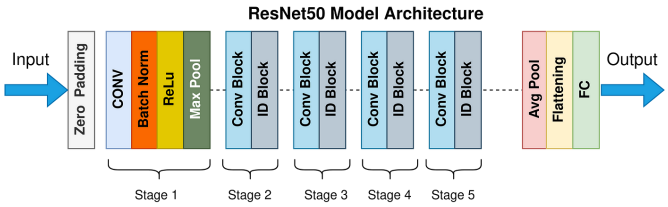


Figure 1: ResNet-50 architecture

3.3. Vision Transformer (ViT)

Vision Transformer (ViT) (Figure 2) is a transformer-based architecture that processes images as sequences of 16×16 pixel patches. Each patch is linearly embedded and combined with positional encodings before passing through transformer encoder layers containing multi-head self-attention mechanisms. The final [CLS] token embedding is fed through an MLP head to produce predictions, using softmax activation to classify images into 27 waste superclasses.

Formally, given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, it is split into N patches, each flattened and linearly projected:

$$\mathbf{z}_0 = [\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}$$

where $\mathbf{x}_p^i$ is the i-th patch, $\mathbf{E}$ is the patch embedding matrix, and $\mathbf{E}_{pos}$ is the positional encoding.

Each transformer encoder layer applies:

$$\mathbf{z}'_l = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}$$

$$\mathbf{z}_l = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l$$

where MSA is multi-head self-attention and LN is layer normalization.

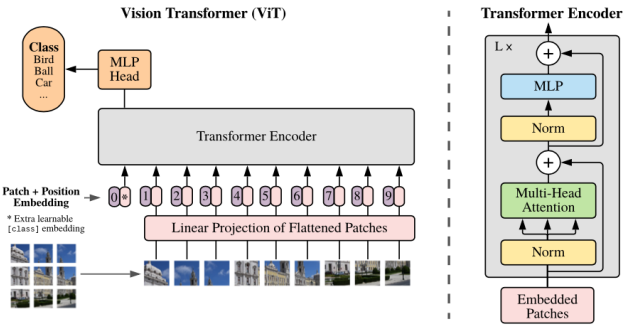


Figure 2: ViT architecture

3.4. Evaluation Metrics

Common applications for waste classification algorithms are in waste material recovery facilities (MRFs) to run live for robotic sorting operations. As a result, sorting accuracy, the inference latency, inference FLOPs, and confusion matrix will be important to monitor. Due to dataset and population class imbalance, monitoring precision-recall (PR) curves will also be necessary to tune the model for desired outcome.

Accuracy:

$$\text{Accuracy}_{\text{super}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i^{\text{super}} = y_i^{\text{super}})$$

Where:

- N is the total number of annotations
- $\hat{y}_i^{\text{super}}$ is the predicted supercategory for annotation i
- y_i^{super} is the ground-truth supercategory for annotation i

Precision and Recall:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix:
The confusion matrix $C \in \mathbb{N}^{27 \times 27}$ is defined as:

$$C_{ij} = \sum_{n=1}^{N} \mathbb{I}(y_n = i \wedge \hat{y}_n = j)$$

where C_{ij} counts the number of samples with true class i and predicted class j.

FLOPs: For reference, the number of floating point operations (FLOPs) for a convolutional layer is:

$$\text{FLOPs}_{\text{conv2d}} = 2 \cdot H_{out} \cdot W_{out} \cdot C_{in} \cdot C_{out} \cdot K^2$$

and for multi-head self-attention:

$$\text{FLOPs}_{\text{MSA}} = 2 \cdot N^2 \cdot d \cdot h$$

where h is the number of attention heads.

3.5. Loss Function

We employed Binary Cross-Entropy with Logits Loss as our primary loss function due to the multi-class nature of our classification task. The loss function is mathematically defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}[y_{i,c}\log(\sigma(x_{i,c})) + (1-y_{i,c})\log(1-\sigma(x_{i,c}))]$$

where N is the batch size, C is the number of classes, $y_{i,c}$ is the true label, $x_{i,c}$ is the raw logit output, and σ represents the sigmoid function. This formulation effectively handles class imbalance and provides stable gradients during backpropagation, making it particularly suitable for our diverse waste category dataset.

4. Dataset and Features

We used the Trash Annotations in Context (TACO) dataset[1], an open-source dataset designed for waste detection and classification. TACO contains 1,500 images with 4,784 annotations across 60 categories organized into 28 super-categories, which serve as our classification labels. The dataset features images from various environments with an average of 3.19 annotations per image, making it particularly suitable for real-world waste classification applications. An example of the image data is as follows (Figure 3). The dataset was randomly split into training (70%), validation (20%), and test (10%) sets.

Figure 3: Image Example

4.1. Dataset Statistics & Characteristics

Our exploratory data analysis revealed a significant class imbalance across super-categories (Figure 4), with "plastic bag & wrapper" representing the most frequent category (850 annotations) while "Battery" had only 2 annotations. This imbalance necessitated careful data processing and augmentation to ensure effective model training.

The dataset also exhibits substantial variation in image dimensions (Figure 5), with widths ranging from approximately 1000 to 6000 pixels and heights from 500 to 5000 pixels. This diversity in image resolution requires standardized resizing for consistent input to our CNN model.

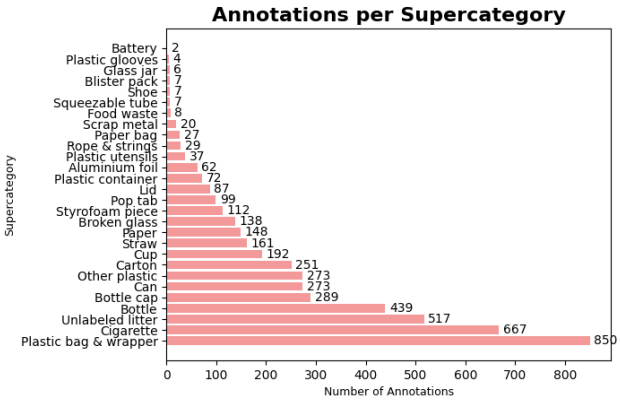


Figure 4: Annotations per Supercategory

4.2. Data Preprocessing & Data Augmentation

- Data Cleaning: We removed annotations labeled as "Unlabeled litter" to improve classification precision. This resulted in the removal of 517 annotations affecting 269 images (10.81% of all annotations), leaving us with a cleaner dataset of 4,267 annotations.

- Dimension Standardization: Given the significant variation in image dimensions, we implemented consistent resizing (224 x 224) to standardize inputs for CNN, ensuring optimal feature extraction across all images.

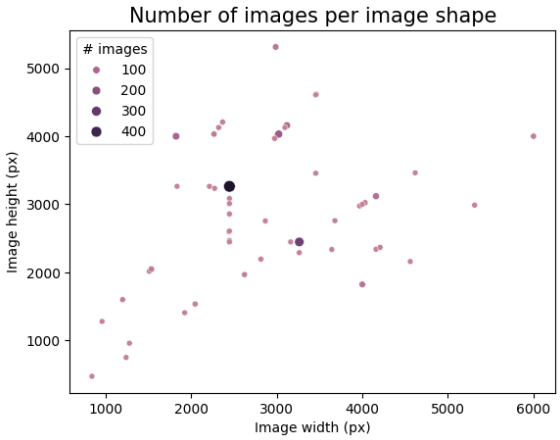


Figure 5: Number of Images per Image Shape

- Data Augmentation: To address class imbalance, we applied data augmentation techniques to supercategories with fewer than 100 annotations. For each image in these underrepresented classes, we generated 10 augmented variants using a sequential pipeline, which combines multiple transformations applied in random order, including variability in noise, blur, orientation, brightness, contrast, and rotation.

5. Experiments/Results/Discussion

5.1. Baseline Model + Non-Augmented Data

The baseline CNN model was trained for 10 epochs using a batch size of 32 and the Adam optimizer with learning rate 1×10^{-4}. The data used for these experiments is the original, non-augmented dataset containing 1,500 images (prior to any data augmentation). The results of this baseline model are presented in Table 2.

Performance on the small subset suggests signs of overfitting, indicating that the baseline architecture has sufficient capacity and could benefit from training on a larger dataset. However, when scaled to the full dataset, the model's training performance declined, and generalization deteriorated significantly.

Several factors may contribute to this poor performance. First, as revealed in our exploratory data analysis (EDA), class imbalance is a serious issue. To address this, we plan to retrain the baseline model on the augmented dataset and incorporate stronger regularization techniques to mitigate overfitting. Second, the classification task itself is inherently challenging: trash objects are often small and can blend into complex backgrounds. To address these challenges, we will apply transfer learning using ResNet and

Vision Transformers to leverage pre-trained representations and improve feature extraction for such difficult cases.

Table 2: Baseline CNN Model Performance

Metric	Validation (%)	Test (%)
Precision	1.69	0.51
Recall	0.31	0.07
F1 Score	0.52	0.13

5.2. Baseline Model + Augmented Data

To address the poor performance observed with the non-augmented dataset, we applied data augmentation techniques to expand our training data. The augmented dataset contained 9,090 additional images generated through various transformations, bringing the total combined training dataset to 10,140 images. The baseline CNN model architecture remained unchanged, maintaining the same hyperparameters: 10 epochs, batch size of 32, and learning rate of 1×10^{-4} with the Adam optimizer.

The training results demonstrated substantial improvement over the non-augmented baseline. Training precision progressed from 0% in the first epoch to 50.1% by the final epoch, while validation precision increased from 0% to 40.8%. The final test precision achieved 24.7%, with recall of 9.4% and F1 score of 12.3%, representing significant improvements from the non-augmented baseline which achieved only 0.5% precision, 0.07% recall, and 0.1% F1 score.

Despite this notable improvement, the overall performance remains suboptimal for practical deployment. Detailed analysis of the confusion matrix (Figure 8) reveals that class imbalance is not the primary factor limiting performance, contrary to initial hypotheses. Specifically, several high-support classes demonstrate surprisingly poor recall rates: Plastic bag & wrapper (49 samples) achieves only 16.3% recall, Bottle (40 samples) reaches 17.5% recall, and Bottle cap (27 samples) attains 11.1% recall. Conversely, some low-support classes perform remarkably well: Pop tab (10 samples) achieves 20% recall and Lid (10 samples) reaches 30% recall.

This performance pattern contradicts what would be expected if class imbalance were the dominant issue. If insufficient training data were the primary limitation, high-support classes should outperform low-support classes consistently. The observed results suggest that certain waste categories possess inherently more distinguishable visual features regardless of sample size, while others present fundamental recognition challenges even with adequate training data.

The mixed performance across support levels—such as Paper (11 samples) achieving 0% recall despite low support,

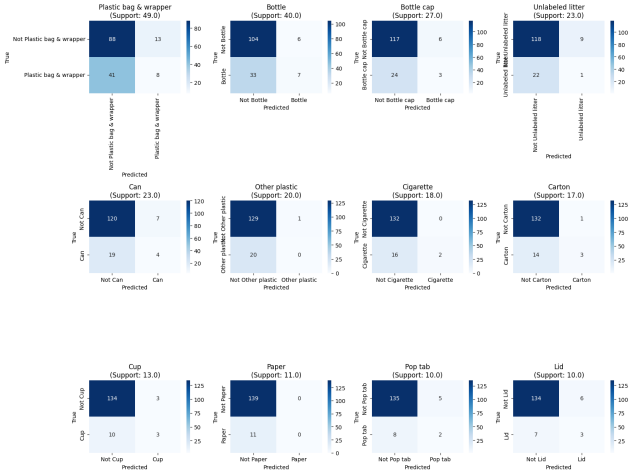


Figure 6: CNN Baseline Confusion Matrix

and Can (23 samples) reaching 17.4% recall with medium support—further confirms that model architecture limitations, rather than data quantity alone, are constraining performance. The baseline CNN's limited representational capacity appears insufficient to capture the complex visual patterns necessary for robust waste classification.

These findings indicate that architectural improvements are essential for advancing performance. More sophisticated models with enhanced feature extraction capabilities, such as Vision Transformers (ViT) with their attention mechanisms and pre-trained representations, may better handle the nuanced visual distinctions required for accurate waste categorization across diverse environmental conditions and object orientations.

Table 3: Baseline CNN Model Performance

Dataset	Precision (%)	Recall (%)	F1 (%)
Non-Augmented	0.51	0.07	0.13
Augmented	24.71	9.41	12.29

5.3. ResNet + Augmented Data

Next, we utilized ResNet for the architectural improvements. We chose the pretrained model because of its representations learned from large-scale image datasets. The ResNet model consists of a pretrained ResNet backbone and a fully connected classification layer. Since our CNN baseline exhibited clear overfitting issues, we incorporated a learning rate scheduler and early stopping to regulate training dynamics. To further resolve overfitting and reduce computational consumption, we used ResNet34 instead of ResNet50. Each experiment is trained with a starting learning rate 5e-4 and trained for 20 epochs.

Table 4 displays the results of our ResNet models. The

ResNet implementation demonstrated substantial improvements over the baseline, with ResNet34 achieving 63.4% precision, 45.0% recall, and 49.8% F1 score. The train and validation loss curves in Figure 8 shows that ResNet shows improved ability in learning compared to the CNN model. However, the curves also indicate persistent overfitting behavior, with validation loss plateauing around epoch 3 while training loss continued its downward trajectory, despite utilizing regularization techniques. The moderate recall values suggest that while ResNet achieved reasonable precision, it struggled to identify a substantial portion of true positive instances across waste categories.

The ResNet confusion matrix reveals significant improvements over the CNN model. High-support classes that struggled with the CNN model demonstrated moderate gains: Plastic bag & wrapper improved from 22% to 41% recall, Bottle increased from 13% to 48% recall, and Bottle cap rose from 19% to 33% recall. Additionally, ResNet maintained strong performance in previously well-performing low-support categories, with Pop tab and Lid achieving perfect 100% recall. This pattern indicates that ResNet's enhanced feature extraction capabilities successfully addressed many of the representational limitations observed in the baseline CNN, though certain categories with inherently challenging visual characteristics continue to present classification difficulties regardless of architectural sophistication.

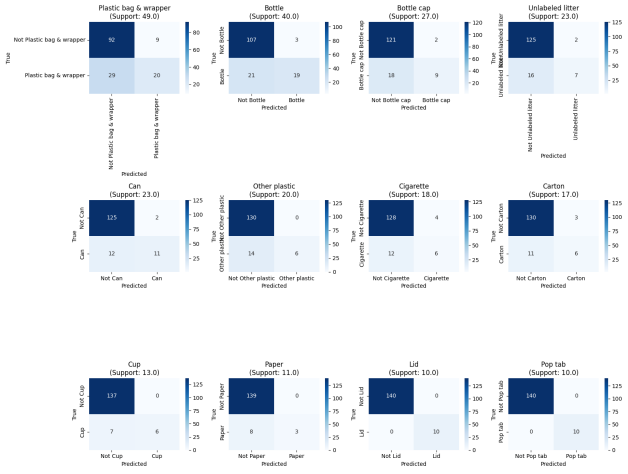


Figure 7: ResNet Confusion Matrix

While the ViT model outperforms ResNet in overall evaluation metrics, the confusion matrices in Figure 8 reveal a more nuanced picture. ViT shows marked improvements in certain categories such as "Bottle" (recall: 47.5% vs. 40.0%) and "Bottle cap" (precision: 81.8% vs. 77.8%), as well as "Unlabeled litter" (recall: 30.4% vs. 26.1%), where it demonstrates better true positive rates and fewer misclassifications. However, for other categories like "Plastic bag & wrapper" (recall: 44.4% vs. 53.1%), "Cigarette" (precision: 55.6% vs. 83.3%), and "Other plastic" (precision: 30.0% vs. 75.0%), the improvements are less consistent, with ViT occasionally exhibiting higher false positive rates or a drop in class-wise precision and recall compared to ResNet. This mixed performance suggests that while ViT captures global patterns better at the aggregate level, it may still struggle with specific fine-grained classes that are visually similar or underrepresented in the training data.

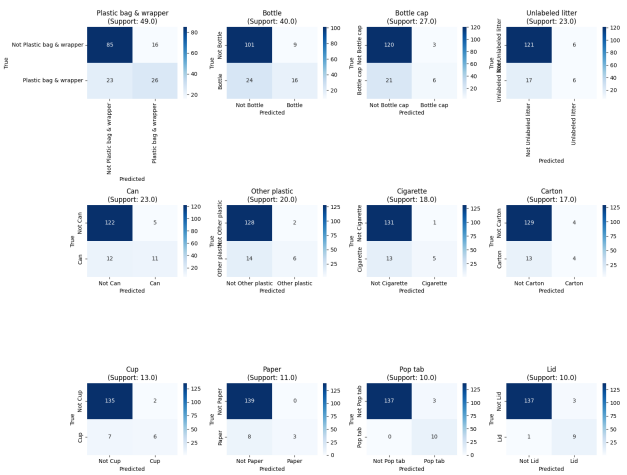


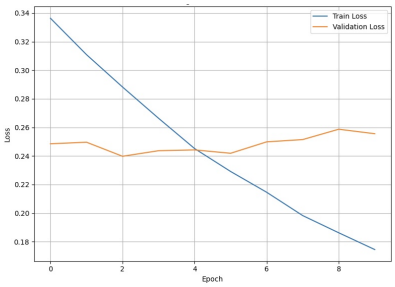
Figure 8: ViT Confusion Matrix

5.4. ViT + Augmented Data

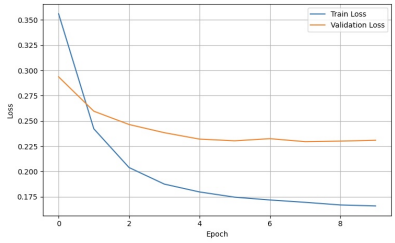
Similar to ResNet, we utilized the backbone of the pretrained ViT with a fully connected layer for classification. The ViT model was trained using identical experimental settings to ResNet, with a learning rate of 5e-4, 20 training epochs, and the same learning rate scheduler and early stopping mechanisms. According to Table 4, ViT achieved the highest overall performance among all tested architectures, reaching 69.1% precision, 59.7% recall, and 60.7% F1 score, representing notable improvements over both CNN and ResNet. These results demonstrate that ViT's transformer-based architecture with self-attention mechanisms can more effectively capture complex spatial relationships and discriminative features within waste images compared to traditional convolutional approaches. However, the training and validation loss curves in Figure 8(c) reveal that ViT also exhibited overfitting behavior.

Table 4: ResNet and ViT Performance on Test Data

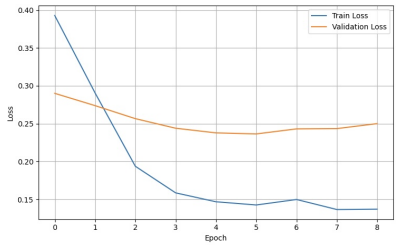
Model	Precision (%)	Recall (%)	F1 Score (%)
ResNet34	63.4	45.0	49.8
ViT	69.1	59.7	60.7



(a) CNN



(b) Resnet



(c) ViT

Figure 9: Train vs. Validation Loss Curves

5.5. Model Size and Inference Characteristics

An important aspect of these models in production is that they need to run locally at acceptable framerates for decision making like robotic sorting. A model developed by Ghazi et al. demonstrated a real-time inference speed of 33.61 frames per second, indicating its viability for deployment in industrial waste sorting environments where high-throughput performance is required [1]. Here, we examined the model parameter size, inference latency/framerate and memory requirement of each of the three models tested on a Google T4 GPU.

Characteristic	ViT-B/32	ResNet34	Custom CNN
Total Params	87.7M	21.4M	51.5M
Trainable Params	30.96M (35.3%)	21.4M (100%)	51.5M (100%)
Model Size (MB)	334.4	81.7	196.4
Batch32 Inference Time (ms/img)	2.60	1.34	0.642
Throughput (FPS)	384.5	748.3	1558.2
GPU Memory (MB)	492.8	110.4	644.6

Table 5: Model Size Comparisons

The Custom CNN demonstrates exceptional computational efficiency with the highest throughput at 1558.2 FPS - more than 4x faster than ViT-B/32 and over 2x faster than ResNet34. This is particularly impressive given its inference time of only 0.642 ms/image. ResNet34 strikes an excellent balance, being the most parameter-efficient model at just 21.4M parameters while maintaining competitive performance with 748.3 FPS throughput. Notably, even the ViT-B/32, despite having the highest parameter count at 87.7M and slower inference at 2.60 ms/image, still achieves 384.5 FPS - well above the 30-60 FPS threshold typically required for real-time industrial sorting applications. This suggests that the superior classification performance of ViT (60.7% F1 score) may justify its computational overhead in centralized MRF deployments where accuracy directly impacts recovery rates and contamination levels.

While all models comfortably exceed real-time requirements on a T4 GPU ($500-800 used), there's significant potential for cost-effective edge deployment through model optimization. The sub-3ms inference times across all architectures suggest that with quantization and pruning techniques, these models could run on much cheaper hardware like NVIDIA Jetson Nano ($99) or Google Coral ($60), representing a 10x cost reduction while still maintaining real-time performance. For residential applications requiring ultra-low power consumption, the Custom CNN's efficiency makes it particularly attractive - with INT8 quantization potentially reducing its size to under 50MB while maintaining acceptable accuracy on sub-$20 microcontrollers. Even the ViT model, through knowledge distillation or model compression, could potentially be deployed on mid-range edge devices ($100-200) for scenarios where its superior accuracy justifies slightly higher hardware costs. This scalability across the hardware spectrum enables a tiered deployment strategy: Large ViT models in centralized facilities where T4-class or more performant GPUs are already amortized across high-volume sorting, and lightweight versions of ViT or ResNets in thousands of battery-powered smart bins providing real-time contamination feedback to residents at minimal per-unit cost.

6. Conclusion/Future Work

This study demonstrates the potential and challenges of deep learning for waste classification using real-world images from the TACO dataset. Our comprehensive evaluation of CNNs, ResNet34, and Vision Transformers (ViTs) reveals important insights about model architecture, data requirements, and deployment feasibility. While data augmentation improved CNN baseline performance from 5.0% to 18.8% test accuracy, this architecture proved fundamentally limited in capturing the complex visual patterns required for waste classification. Contrary to initial hypotheses, class imbalance was not the primary limiting factor—several high-support classes performed poorly while

some low-support classes achieved strong recall, suggesting that inherent visual distinguishability varies significantly across waste categories.

ResNet34 demonstrated substantial improvements, achieving 49.8% F1 score while maintaining excellent computational efficiency at 748.3 FPS on a T4 GPU. Vision Transformers achieved the highest performance at 60.7% F1 score, validating their superior ability to model long-range spatial relationships in cluttered waste images. Critically, all three architectures exceed real-time processing requirements by significant margins, with even ViT achieving 384.5 FPS, well above the 30-60 FPS threshold for industrial sorting. This computational headroom, combined with the potential for 10x hardware cost reduction through edge deployment on devices like NVIDIA Jetson Nano or Google Coral, makes widespread deployment economically feasible.

Future work should address the persistent overfitting observed across all architectures through advanced regularization techniques, self-supervised pretraining on unlabeled waste facility footage, and synthetic data generation for rare categories. Incorporating multispectral imaging could enable finer plastic grade discrimination, while active learning pipelines could continuously improve models using production feedback. The strong performance-to-cost ratio demonstrated here suggests that computer vision can play a transformative role in waste management infrastructure, from centralized MRFs to distributed smart bins, ultimately contributing to improved recycling rates and reduced methane emissions from landfills.

7. Contributions Acknowledgements

S.D., A.F., J.S. conceived the project and selected the dataset. J.S. designed the experimental methodology and evaluation framework. S.D., A.F. designed and implemented the data preprocessing pipeline and augmentation strategies. S.D., A.F. implemented and trained the CNN baseline, ResNet34, and Vision Transformer models. S.D., A.F. conducted the performance evaluation and generated confusion matrices. S.D., A.F. analyzed model inference characteristics and computational requirements. J.S. structured the comparative analysis between architectures. S.D., A.F., J.S. interpreted the results and identified key findings. S.D., A.F., J.S. wrote the paper.

We thank the TACO dataset creators for making their annotations publicly available; the PyTorch team for pretrained model implementations; and the CS231N teaching staff for compute resources and project guidance. This project was completed specifically for CS231N with no overlap with other courses.

References

[1] J. Ghazi, A. K. Moen, and R. Jenssen. Vision-based sorting in mixed food-inorganic waste stream. *Resources, Conservation and Recycling*, 210, 2024.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[4] S. Jin et al. Garbage detection and classification using a new deep learning-based machine vision system as a tool for sustainable waste recycling. *Waste Management*, 162:123–130, 2023.

[5] R. Kumar et al. Swin transformer meets yolo: Efficient litter detection in real-world street environments. *Waste Management*, 2023.

[6] A. I. Myronenkov et al. Versatile waste sorting in small batch and flexible manufacturing industries using deep learning techniques. *Scientific Reports*, 15:87226, 2025.

[7] I. Ogrezeanu et al. Waste recognition in a recycling plant using deep learning: evaluation of warp dataset. *Waste Management*, 2022.

[8] T. R. Partnership. 2020 state of curbside recycling report. `https://recyclingpartnership.org/2020-state-of-curbside-report`, 2020. Accessed: 2025-05-17.

[9] T. R. Partnership. Report shows only 21% of u.s. residential recyclables are captured. `https://recyclingpartnership.org/recyclables-report-2024`, 2024. Accessed: 2025-05-17.

[10] P. F. Proença and P. Simões. Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*, 2020.

[11] G. I. Sayed et al. A hybrid vision transformer for sustainable waste classification. *Environmental Science and Pollution Research*, 2024.

[12] F. R. Sayem et al. Enhancing waste sorting and recycling efficiency: robust deep learning-based approach for classification and detection. *Neural Computing and Applications*, 37:4567–4583, 2025.

[13] G. Thung and M. Yang. Trashnet dataset. `https://github.com/garythung/trashnet`, 2017.

[14] U.S. EPA. Advancing sustainable materials management: 2020 fact sheet. `https://www.epa.gov/facts-and-figures-about-materials-waste-and-recyclin` 2020. Accessed: 2025-05-17.

[15] U.S. EPA. Inventory of u.s. greenhouse gas emissions and sinks: 1990–2022. `https://www.epa.gov/ghgemissions`, 2024. Accessed: 2025-05-17.