CS231N Project: Organic Waste Quantification in Public Trash Bins in Urban Areas Using Thermal Videos

Varun Sahay
Stanford University
varuns19@stanford.edu

Pin Li
Stanford University
pinli@stanford.edu

Seoyoung Oh
Stanford University
syoh99@stanford.edu

Abstract

Homelessness is an issue in urban areas, and providing and maintaining public services for this population has been a challenge due to two main reasons: (1) this population is mobile and does not have a permanent residence. The traditional annual point-in-time surveys conducted by the Department of Housing and Urban Development (HUD) in the U.S. are too infrequent to address this concern; (2) direct tracking of this population raises significant privacy concerns. To address these limitations, this project proposes an indirect and privacy-preserving computer vision-based approach by monitoring street-level public service infrastructure using vehicle-mounted cameras. Specifically, we focus on quantifying organic food waste in public trash bins as a proxy for assessing inefficiencies in food distribution and city sanitation. Our pipeline consists of two main models: the first detects trash bins in video frames, and the second estimates the volume of organic waste using infrared imagery. We used our bin detection model to obtain the region of interest in the thermal images where the quantification model would then predict the percentage of organic wastes. We evaluated our system based on bin detection accuracy from infrared video and mean squared error in organic content quantification. This approach provides a novel, scalable, and non-intrusive solution for enabling real-time optimization of public services in urban environments.

1. Introduction

Homelessness poses a persistent challenge in urban environments, particularly in the equitable and efficient delivery of public services. Traditional data collection methods, such as the U.S. Department of Housing and Urban Development's annual point-in-time surveys, are too infrequent to capture the mobility patterns of the unhoused population [9]. Furthermore, direct tracking methods introduce significant privacy concerns. In response, we propose an indirect, privacy-preserving computer vision approach that monitors the distribution and use of public services—specifically food distribution—by analyzing trash bins along city streets using vehicle-mounted cameras. Organic food waste in public bins serves as a meaningful proxy for understanding inefficiencies in food distribution systems. By identifying patterns of discarded food across different neighborhoods, we aim to inform more effective public service planning, optimize trash collection schedules, and support broader urban sustainability goals.

Our approach addresses two primary tasks: (1) detecting trash bins in thermal video frames, and (2) estimating the quantity of organic waste within these bins. To train our models, we collected a paired RGB and thermal video dataset in a controlled laboratory-like field setup using a GoPro and a stationary infrared camera under outdoor conditions. The dataset features two distinct types of trash bins, each filled with varying levels of organic waste. While our long-term goal is to deploy this system in public settings, the current training and evaluation are conducted on this controlled dataset. The experimental setup was designed to minimize uncontrolled variables such as lighting and background clutter, while still preserving aspects of real-world variability, including different bin geometries and organic content types. This setup enables reliable supervised learning while maintaining relevance to eventual deployment scenarios, serving as a proof of concept.

In the domain of waste detection and classification, Shen et al. [6] developed a vision-based smart bin system for sorting recyclables and organics, while Minarni et al. [7] and Sathish et al. [2] explored deep learning models for real-time garbage detection using CNN and region-based methods. These works, however, primarily focus on static or indoor setups, limiting their generalizability to dynamic urban environments. In urban scene understanding, object detection models such as YOLOv5[4, 5] have shown strong performance in recognizing elements in complex street scenes. We leverage these models for robust bin detection across video frames, with potential fine-tuning for our specific use case. Thermal imaging, commonly used in industrial inspection and energy diagnostics, has rarely been applied to public waste analysis. Our work extends its use to estimate the volume of organic content, where challenges include occlusions, mixed-material waste, and low frame rates.

In summary, we present a novel computer vision pipeline that indirectly monitors food waste and public bin usage in urban environments. Our contributions include collecting a dual-modality dataset, a bin detection pipeline, and a thermal-based organic waste quantification model evaluated using field data.

2. Problem Statement

This work focuses on automating the monitoring of organic waste levels in trash bins using thermal imagery. The problem is formulated as a two-stage pipeline comprising the following core tasks:

- **Trash Bin Detection:** The first stage involves detecting trash bins within thermal video frames captured in controlled outdoor environment. This includes localizing each bin using bounding boxes. Accurate bin detection is essential to isolate the region of interest, minimizing background noise and ensuring that subsequent quantification is applied only to relevant image regions. A YOLOv5-based object detection model is employed for this task, trained specifically on thermal images of bins.

- **Organic Waste Quantification:** The second stage involves estimating the percentage of organic waste contained within each detected bin. Given a cropped thermal image of a bin, the goal is to infer a continuous value representing the volume fraction of organic waste (e.g., 0% to 85%). This is treated as a regression problem, where a quantification model is trained on labeled thermal images with known organic content levels. Accurate estimation enables real-time assessment of waste composition, which can be valuable for automated sorting, recycling, or disposal strategies.

3. Methodology

The overall workflow consists of four main stages: data collection, data preprocessing, bin detection, and organic content quantification. The various steps of this pipeline is illustrated in Figure 1.

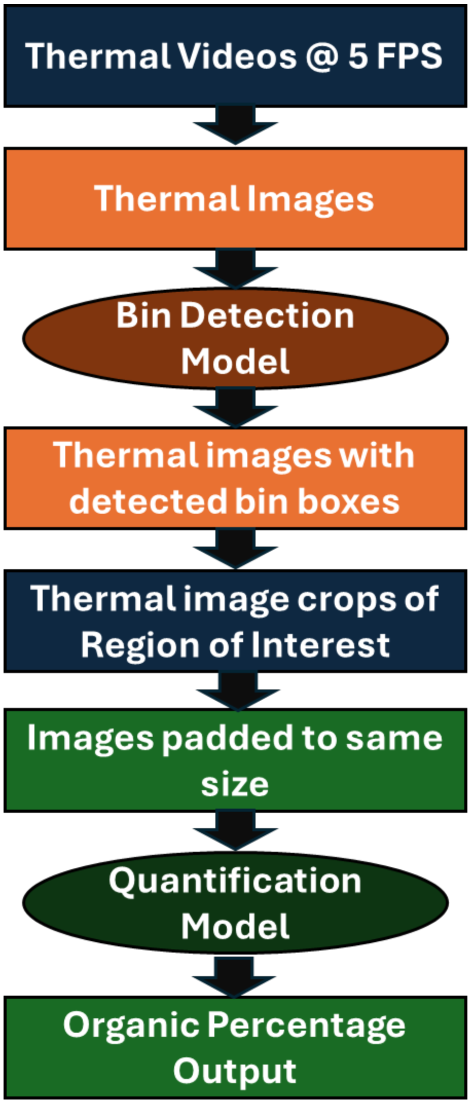
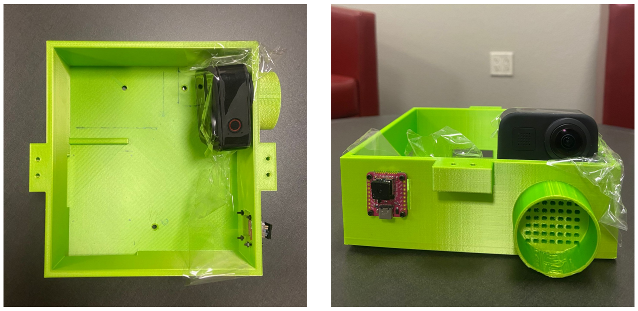

Thermal Videos @ 5 FPS
Thermal Images
Bin Detection Model
Thermal images with detected bin boxes
Thermal image crops of Region of Interest
Images padded to same size
Quantification Model
Organic Percentage Output

Figure 1: Thermal image processing and training pipeline

3.1. Data Collection

To train and evaluate our system, we collect data from outdoor environment under natural daylight with known compositions of organic waste in the bin. Figure 2 presents our team during data collection in the outdoor experimental setup. A portable and fixed-relative RGB–thermal camera pair was employed to capture synchronized video data. Figure 3 illustrates the top and front views of the camera configuration, showing the spatial arrangement of the RGB (GoPro) and thermal sensors. The GoPro was mounted on a square plastic enclosure to ensure stability, while preserving portability. Although the absolute position of the camera rig was not rigidly fixed, the relative alignment between the RGB and thermal cameras was maintained throughout all

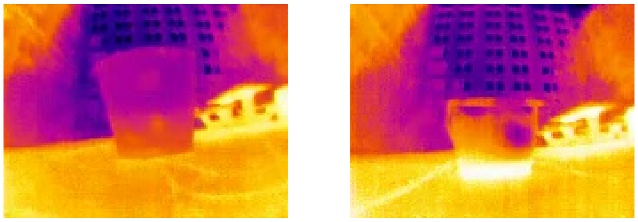
Figure 2: Our team during field data collection.



(a) Top

(b) Front

Figure 3: Different views of the GoPro and infrared camera set-up.



(a) Second type (30%)

(b) First type (85%)

Figure 4: Example thermal views of two trash bin types at different organic waste levels.

recordings. As shown in the front view, the thermal camera was affixed to the front face of the enclosure, with its sensor exposed as a small chip-like element. The thermal feed was monitored and recorded in real-time via a laptop connection. For each trash bin and fill level, we recorded video sequences by manually moving the camera setup in a circular trajectory around the bin, enabling multi-view capture from various angles.

Trash bins were loaded with controlled proportions of organic material at 0%, 10%, 30%, 50%, 70%, and 85% by volume of the bins. Two distinct types of bins were used in the experiment: (1) a smaller bin with a lid and handle, whose appearance varies significantly with viewing angle due to its asymmetric geometry, and (2) a larger, symmetrical bin without a lid. The organic content consisted of representative household food waste, including banana peels, tangerines, tomatoes, and meat products. Figure 4 illustrates how both the thermal appearance and visible geometry of the bins vary as a function of fill level, material type, and bin design.

3.2. Data Preprocessing

We collected a paired RGB and thermal (IR) video dataset, annotated with six levels of organic waste filling: 0%, 10%, 30%, 50%, 70%, and 85%. Although the RGB and thermal video streams were synchronized temporally, slight spatial misalignments were observed due to the physical separation of the two sensors. Each video stream was decomposed into individual frames at 5 FPS using FFmpeg.

While our initial plan involved performing bin detection on the RGB frames and transferring the corresponding bin locations to the thermal domain, accurate pixel-level registration between the two modalities proved challenging due to the absence of a shared lens and the spatial offset between sensors. In the interest of time, and to focus on demonstrating the core proof of concept, we opted to proceed using only the thermal frames for organic waste quantification. To isolate the region of interest in the thermal image, we trained a bin detection model on the thermal frames (described in detail in a later section). This model produced bounding boxes around detected bins for all the thermal images. We then cropped the thermal images using these bounding boxes and used the resulting regions for subsequent quantification of organic content.

After cropping the thermal images to extract the region of interest corresponding to each bin, we applied horizontal flipping as a form of data augmentation. This transformation preserves the semantic content of the images while improving the generalizability of the quantification model. Following augmentation, we obtained a total of 4,972 labeled images for training. Since the cropped regions varied in size, we standardized the input dimensions by identifying the maximum width and height across all cropped images and applying zero-padding to match this size. All images were padded to a uniform shape of 84×79, resulting in an input tensor of $(4972, 84, 79, 3)$.

3.3. Trash Bin Detection

We focus on the task of localizing trash bins in thermal video data collected under outdoor conditions. Using LabelImg [8], we manually annotated bounding boxes for the *trash_bin* class in a controlled environment. The annotations were initially exported in XML format with absolute pixel coordinates. To adapt these for YOLOv5, we developed a custom Python script to convert the bounding boxes into the required format by computing the center coordinates, width, and height, and normalizing them with respect to the image dimensions.

At this stage, annotations were binary—only the presence and location of trash bins were considered, without regard to the waste fill level. To minimize potential bias from bin type or viewing angle, we uniformly sampled 15 frames for each trash level (0%, 10%, 30%, 50%, 70%, and 85%), resulting in a total of 105 labeled images. This dataset was split into 80% training and 20% validation sets, yielding 84 training and 21 validation images.

We fine-tuned all layers of a YOLOv5 model without freezing the pretrained weights. This decision was motivated by the domain gap between the COCO dataset, on which YOLOv5 was originally trained, and our thermal imagery. Unlike RGB images, thermal data lacks color information and exhibits distinct noise characteristics, necessitating end-to-end adaptation of the model. Since our task involves only a single object class, full fine-tuning was expected to outperform partial transfer learning.

Model performance was evaluated using validation precision and recall (Equations 4 and 5). In particular, achieving perfect recall was prioritized to ensure that no trash bins were missed. We also report the Mean Average Precision at IoU thresholds of 0.5 (mAP@0.5) and over a range of thresholds from 0.5 to 0.95 (mAP@0.5:0.95), which collectively provide a robust measure of detection performance. Furthermore, predicted bounding boxes were visually compared against manual annotations to assess qualitative accuracy.

The trained model was then deployed on a separate test set comprising 2,417 thermal frames not used during training. Detected bounding boxes were used to crop bin regions from these images, forming the input for downstream organic waste quantification.

3.4. Organic Waste Quantification

We experimented with three different quantification models, each trained on a dataset of carefully labeled thermal images of trash bins collected in a controlled laboratory setting. These models are designed to estimate the organic waste content within each bin, serving as a composition quantification step following bin detection. The bins are first localized in thermal images using a YOLOv5-based detection model, after which the corresponding cropped regions are passed to the quantification model for organic content estimation. A total of 4,972 examples were divided into training, validation, and test sets using an 80:10:10 split.

As a baseline, we employed a standard convolutional neural network (CNN) for estimating the organic content in trash bins using thermal imagery. Building on this, we fine-tuned a ResNet-18 model pretrained on ImageNet to improve feature extraction capabilities. Finally, we explored a transformer-based architecture by fine-tuning ViT-B/16 for the same task. These models were trained as regressors to predict the proportion of organic waste present within each detected bin. The performance of each model was evaluated using mean squared error (MSE) loss mentioned in equation-1.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(Q_i^{\text{pred}} - Q_i^{\text{true}} \right)^2 \tag{1}$$

Where:

- Q_i^{pred} = Predicted quantity of organic waste in image i,
- Q_i^{true} = Ground truth quantity of organic waste in image i,
- N = Total number of evaluated images.

3.4.1 Baseline: Vanilla CNN Model

For the baseline we have implemented a baseline CNN model that consists of five convolutional blocks followed by fully connected layers. Each convolutional block includes a convolutional layer with a 3×3 kernel, batch normalization, ReLU activation, and max pooling to reduce spatial dimensions. The model was trained using the Adam optimizer.

The input and output shapes for each block are shown in Table 1.

The final feature maps are flattened and passed through three fully connected layers, ending with a sigmoid activation to produce the final output.

This architecture effectively extracts features while reducing spatial size, enabling efficient learning for the target task.

Block	Input Shape	Output Shape
Block 1	(3, 84, 79)	(16, 42, 39)
Block 2	(16, 42, 39)	(32, 21, 19)
Block 3	(32, 21, 19)	(64, 10, 9)
Block 4	(64, 10, 9)	(128, 5, 4)
Block 5	(128, 5, 4)	(256, 2, 2)

Table 1: Input and output tensor shapes at each convolutional block.

3.4.2 Finetuned ResNet-18 Model Approach

Another choice of model is adopting the architecture of the ResNet-18 [3] without using the pretrained weights and the classification layer is changed to a linear layer and sigmoid function (Equations 2 and 3). The loss function used is mean squared error 1.

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{x} + b) \tag{2}$$

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{3}$$

- $\mathbf{x}$: feature vector obtained from the model without the classification layer.
- $\mathbf{w}$: learnable weight vector.
- $b \in \mathbb{R}$: Learnable scalar bias.
- $\sigma(\cdot)$: Sigmoid function.
- $\hat{y}$: Final predicted percentage.

3.4.3 Finetuned ViT-B/16 Model Approach

The third choice of model is adopting the architecture of ViT-B/16 proposed by Dosovitskiy et al. [1] without using the pretrained weights and the classification layer is changed to a linear layer and sigmoid function (Equations 2 and 3). The loss function used is mean squared error 1.

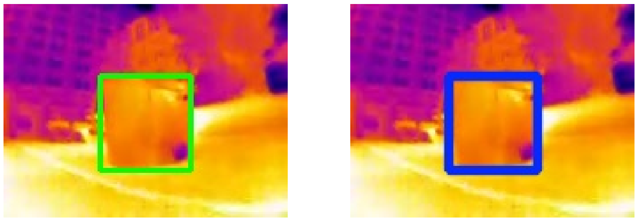
4. Results and Discussion

4.1. Trash Bin Detection Model

The model, referred to as *trash_bin_detector*, achieved a validation precision of 99.7% and a recall of 100% (Equations 4 and 5). The Perfect recall reflects the intentional design of the dataset, which includes only images with trash bins to directly evaluate detection performance. The model successfully detected all annotated instances. Table 2 summarizes performance in terms of mAP@0.5 and mAP@0.5:0.95. To qualitatively assess detection consistency, we compared predicted bounding boxes in the validation set with manual annotations. As shown in Figure 5, the model-generated bounding boxes are well aligned with ground truth labels across multiple viewpoints and fill levels.

Metric	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Score	99.7	100.0	99.5	86.4

Table 2: Validation performance of Trash Bin Detection Model (*trash_bin_detector*) in %



(a) Manual annotation

(b) Model prediction

Figure 5: Comparison of manual and predicted bounding boxes for a thermal frame.

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{4}$$

$$\text{Recall} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{5}$$

Where:

- N_{TP}: Number of correctly detected trash bins. The predicted bounding box overlaps with a ground truth bin with Intersection-over-Union (IoU) ≥ 0.5.
- N_{FP}: Number of incorrect detections (e.g., spurious detections in background regions, or IoU < 0.5 with any actual bin).
- N_{FN}: Number of missed detections (e.g., The model failed to detect a visible bin in the frame).

From the separate test set consisting of 2,522 thermal frames not used in training, the trained model successfully detected trash bins in 2,486 frames with average confidence level of 81.95%. All 36 failures corresponded to bins with a 70% fill level and were excluded from subsequent analysis. Bounding boxes from successful detections were matched with their corresponding thermal images and used to generate cropped bin images for downstream analysis.

4.2. Organic Waste Quantification Model

Table 3 summarizes the training and test losses for the three different quantification models. All models demonstrate strong performance with minimal differences between training and test loss, indicating effective generalization and no signs of overfitting. However, notable differ-

Model	Training MSE Loss	Test MSE Loss
Baseline CNN	1.66e-5	3.46e-5
Finetuned ResNet-18	2.80e-5	9.00e-6
Finetuned ViT-B/16	1.30e-5	9.80e-5

Table 3: Comparison of training and test MSE loss across different quantification models.

ences in test performance reveal important insights about each model's characteristics.

The finetuned ViT-B/16 model, despite achieving the lowest training loss, exhibited a substantially higher test loss compared to the CNN-based models. This behavior is likely due to the Vision Transformer's reliance on self-attention mechanisms and its relatively weak inductive biases such as locality and translation equivariance, which are inherent in convolutional architectures. As a result, ViTs generally require very large datasets to generalize well. Given the limited size of our labeled dataset, the ViT model was more prone to overfitting, leading to poorer generalization on unseen data.

Conversely, the finetuned ResNet-18 outperformed the baseline CNN model, achieving lower test loss despite similar training loss levels. This improvement can be attributed to ResNet's deeper architecture and use of residual connections, which enhance gradient flow and enable learning of more complex, robust features. These characteristics help ResNet-18 better capture discriminative patterns relevant to organic waste quantification, resulting in superior generalization compared to the simpler baseline CNN.

Overall, these results suggest that while transformer-based architectures hold promise, convolutional models—especially those with advanced designs like ResNet—are currently better suited for our relatively small dataset size and task.

To visualize model performance, we plotted the predicted percentages against the corresponding ground truth organic waste levels (0%, 10%, 30%, 50%, 70%, and 85%) in Figure 6 for the three models. A box plot overlay at each true percentage level captures the distribution of predicted values. This visualization serves two main purposes: (1) it allows qualitative assessment of prediction accuracy by showing how closely the predicted values cluster around the ground truth, and (2) it reveals any biases or variance in predictions across different levels of organic content. Ideally, tight, symmetric box plots centered near the ground truth indicate consistent and unbiased model performance.

From the visualization, the baseline CNN and finetuned ResNet-18 models show tightly clustered predictions with low variance across most fill levels, particularly at intermediate and high percentages. This consistency reflects their robust ability to estimate organic content accurately. The ResNet-18 predictions tend to be slightly closer to the true values, corroborating its superior quantitative performance shown in Table 3.

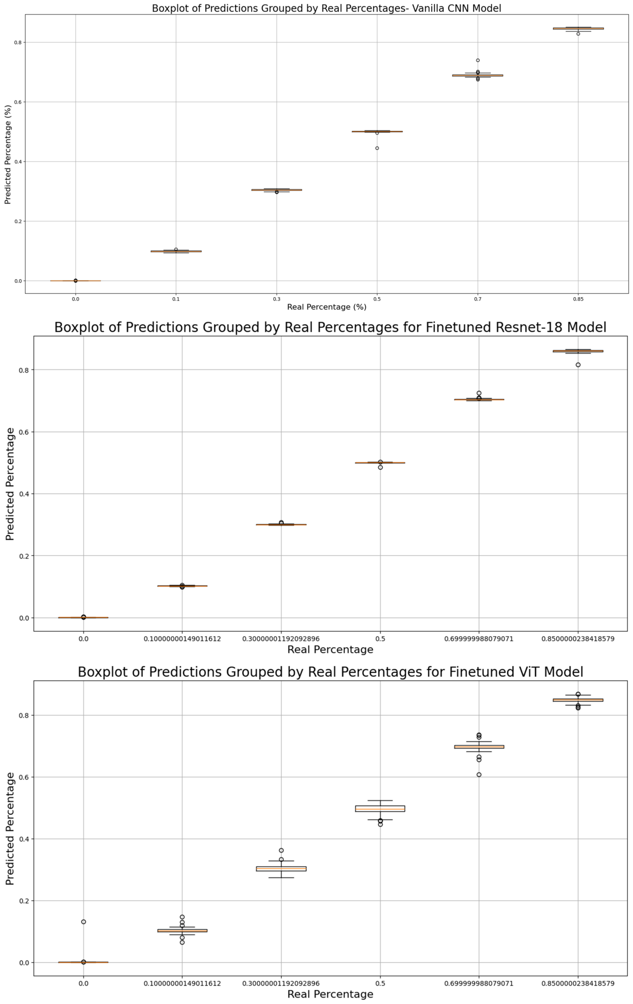


Figure 6: Box plots of ground truth versus predicted organic waste percentages for the three quantification models.

In contrast, the ViT-B/16 model exhibits greater spread and occasional deviations from the ground truth, consistent with its higher test loss and reduced generalization. This suggests that, despite good training performance, the ViT struggles to maintain stable predictions on unseen data, likely due to dataset size limitations and weaker inductive biases. Overall, the box plots reinforce the quantitative results, confirming that convolutional architectures—especially ResNet-18—provide more reliable and accurate organic waste quantification on our thermal image dataset.

Figure 7 illustrates the prediction results of the best-performing quantification model, the finetuned ResNet-18, on the test set. The plot shows that the predicted percent-

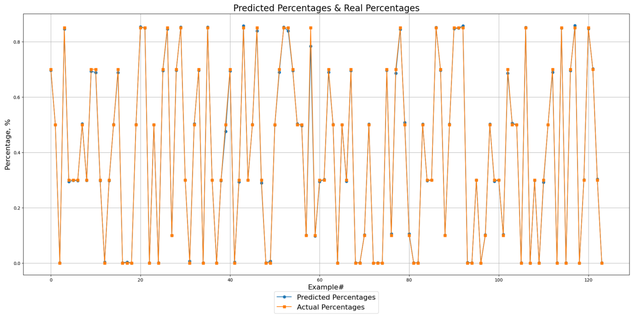


Figure 7: Results for the Best Performing Quantification Model-Finetuned ResNet-18 Model On the Test Set

ages closely match the actual ground truth values for all test examples (a subset of the total set is plotted), demonstrating the model's high accuracy in estimating organic waste content from thermal images. This visual evidence corroborates the quantitative metrics presented earlier and underscores the model's robustness and generalizability for practical applications in organic waste quantification.

5. Conclusion

This project points a novel direction for indirectly assessing urban public service usage by quantifying organic waste without collecting any personally identifiable information. By detecting trash bins and estimating the level of organic waste, we aimed to offer insights for tracking homelessness and managing the urban environment. We developed a two-stage computer vision pipeline using thermal video data: the first stage performs object detection to localize trash bins, and the second stage estimates organic waste levels through regression-based analysis of cropped thermal images.

We used a YOLOv5-based model trained on the manually labeled thermal imagery leveraging LabelImg to perform trash bin detection. The detected bounding boxes were then used to extract regions of interest for the subsequent quantification stage. For the quantification stage, we experimented with multiple deep learning algorithms including a baseline CNN, a fine-tuned ResNet-18, and a fine-tuned Vision Transformer (ViT-B/16). This allowed us to evaluate trade-offs in accuracy while choosing the optimal model. Among these, the CNN and ResNet-18 achieved particularly low prediction error and demonstrated strong alignment with ground truth across a wide range of fill levels.

Overall, our models achieved high precision and recall in bin detection and demonstrated strong generalization in estimating waste content across varying fill levels. The successful integration of object detection and organic waste quantification modeling represents a meaningful step towards a scalable, real-time monitoring solution. Our results suggest the potential of thermal video and deep learning to address data gaps in public infrastructure planning while respecting individual privacy.

6. Future Works

6.1. Translation to Real-World Settings

To extend the applicability of our system, we plan to fine-tune the *trash_bin_detector* model on videos collected in real urban environments. Unlike our controlled experimental setup, public trash bins in practice contain mixed organic and inorganic waste, and exhibit greater variation in bin shapes and surrounding scenes. Fine-tuning with diverse, real-world data will allow the model to better handle domain shifts and increase its robustness under less constrained conditions.

6.2. Dataset Diversity

Our current dataset was carefully constructed under controlled conditions to ensure clarity and consistency during training. However, it includes only two bin types and limited waste configurations, which may not fully capture the diversity of public waste disposal. Expanding the dataset to include a wider range of bin shapes, mixed waste content, and environmental backgrounds will help improve the generalization ability of the model and reduce overfitting to specific visual patterns.

6.3. Multimodal Fusion and Noise Handling

While our initial attempts to project bounding boxes between domains were limited by spatial misalignment and differing resolutions, current approach show reliable detection accuracy using thermal video. We expect the future architectures could explore joint processing using fusion or cross-modal attention. Furthermore, thermal videos recorded in outdoor environments may include domain-specific noise such as heat reflections or ambient temperature interference. Incorporating denoising strategies and robust data augmentation can further enhance model stability in real-world deployments.

7. Contributions and Acknowledgments

7.1. Code

The code can be accessed through this link: CS231n Project.

7.2. Contributions

Varun Sahay: Conceptualization, collected data, obtained and organized the frames from the video inputs, cropped and augmented the obtained data for quantification models to use, set up the baseline CNN model for composition predictions, analyzed predictions results, participated in writing the report.

Pin Li: Collected data, set up the finetuned ResNet-18 model and finetuned ViT-B/16 model for composition predictions, analyzed composition predictions results, participated in writing the report.

Seoyoung Oh: Collected data, labeled the bounding boxes and set up the trash bin detection model, obtained the detected trash bins in bounding boxes, analyzed trash bin detection results, participated in writing the report.

7.3. Acknowledgments

The tool from https://github.com/tzutalin/labelImg is used during the manual labeling of trash bin detection process.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] S. P. Gundupalli, S. Hait, and A. Thakur. Multi-material classification of dry recyclables from municipal solid waste based on thermal imaging. *Waste Management*, 70:13–21, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[4] G. Jocher, A. Chaurasia, J. Borovec, NanoCode012, Laughing, tkianai, SkalskiP, A. Hogan, Z. Wang, et al. YOLOv5 by ultralytics. https://github.com/ultralytics/yolov5, 2020. Accessed: 2025-05-17.

[5] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[6] Y.-L. Shen, D.-Y. Lan, P.-J. He, Y.-P. Qi, W. Peng, F. Lü, and H. Zhang. Nondestructive optical and spectroscopic techniques combined with machine learning for identifying solid waste: A review. *Trends in Analytical Chemistry*, 186:118195, 2025.

[7] M. Shiddiq, D. S. Arief, Zulfansyah, K. Fatimah, D. Wahyudi, D. A. Mahmudah, D. K. E. Putri, I. R. Husein, and S. A. Ningsih. Plastic and organic waste identification using multispectral imaging. *Materials Today: Proceedings*, 87:338–344, 2023. 3rd International Conference on Chemical Engineering and Applied Sciences.

[8] Tzutalin. Labelimg: Label image for object detection. https://github.com/tzutalin/labelImg, 2015. Accessed: 2025-06-02.

[9] U.S. Department of Housing and Urban Development. The 2024 annual homelessness assessment report (ahar) to congress: Part 1 – point-in-time estimates of homelessness, 2024. Office of Community Planning and Development.