

Municipal Waste Classification with DINOv2 Embeddings: RGB and Thermal Pipeline Performance

Abstract

We present two single-modal classification pipelines for automated conveyor belt waste sorting into four material classes (glass, metal, paper, plastic) using a frozen DINOv2 ViT-L/14 backbone. Each pipeline trains only a lightweight attention pooling layer and MLP classification head ($\sim 528K$ parameters, 0.17% of the backbone), operating on pre-cached $[CLS]$ token features. Evaluated via 5-fold stratified cross-validation on 550 labeled tracklets from 19 experiment videos, the RGB pipeline achieves **95.1%** macro F1 (27 errors) while the thermal pipeline achieves **90.6%** macro F1 (48 errors). Glass is near-perfectly classified in both modalities ($F1 \geq 0.97$), while paper remains the most challenging class, particularly for thermal ($F1 = 0.804$). The thermal modality, despite lower standalone accuracy, captures complementary material properties that benefit downstream fusion.

1. Introduction

This report summarizes the performance of two single-modal classification pipelines for conveyor belt waste sorting into four material classes: **glass**, **metal**, **paper**, and **plastic**. Both pipelines use a frozen DINOv2 ViT-L/14 backbone ($\sim 304M$ parameters) as a feature extractor, with only a lightweight attention pooling layer and MLP classification head trained ($\sim 528K$ trainable parameters each, 0.17% of backbone). All results are from **5-fold stratified cross-validation** on 550 labeled tracklets from 19 experiment videos.

2. RGB Pipeline

2.1. Architecture

The RGB pipeline (Fig. 1) processes video through a frozen Detectron2 Mask R-CNN detector and OC-SORT multi-object tracker. Per-tracklet masked crops (gray fill 128, resized to 518×518) are passed through a frozen DINOv2 ViT-L/14 to extract $[CLS]$ token features. A trainable attention pooling layer aggregates 8 uniformly sampled frame features into a single tracklet representation, classified by an

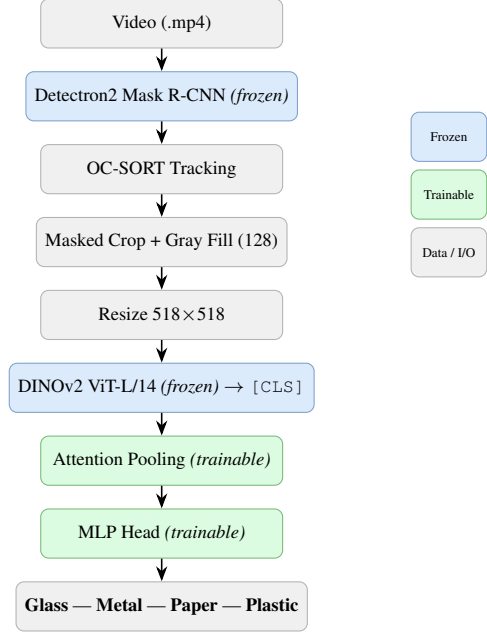


Figure 1. RGB classification pipeline. Detection and DINOv2 are frozen; only the attention pool and MLP head are trained ($\sim 528K$ parameters).

Table 1. RGB pipeline: 5-fold stratified CV (550 tracklets).

Metric	Value
Mean Accuracy	0.9509 ± 0.0093
Mean Macro F1	0.9507 ± 0.0110
Pooled Accuracy	0.9509
Total Errors	27 / 550

MLP head.

2.2. Results

Tables 1 and 2 summarize the RGB results. The pipeline achieves 95.1% macro F1 with only 27 errors across 550 tracklets. Glass is near-perfect ($F1=0.985$), while paper and plastic show the most room for improvement.

Table 2. RGB per-class metrics (pooled across all 5 folds).

Class	Prec.	Rec.	F1	N
Glass	0.978	0.992	0.985	132
Metal	0.921	0.977	0.948	131
Paper	0.957	0.908	0.932	98
Plastic	0.951	0.926	0.938	189

		Predicted Label			
		Glass	Metal	Paper	Plastic
True Label	Glass	131	1	0	0
	Metal	0	128	0	3
	Paper	0	3	89	6
	Plastic	3	7	4	175

Figure 2. RGB confusion matrix (pooled, 5 folds). 27 errors; glass near-perfect (131/132). Most confusion: plastic↔metal (10) and paper↔plastic (10).

2.3. Confusion Matrix

The confusion matrix (Fig. 2) reveals that glass is nearly perfectly classified (131/132). Most errors involve plastic confused with metal (7+3=10 errors) and paper confused with plastic (6+4=10 errors), reflecting visual similarity between these materials.

3. Thermal Pipeline

3.1. Architecture

The thermal pipeline (Fig. 3) reuses the same tracklet detection and tracking from the RGB pipeline but extracts features from thermal frames instead. RGB-space masks are warped to thermal coordinates via a per-experiment homography matrix, and single-channel grayscale thermal images are replicated to 3 channels for DINOv2 compatibility.

3.2. Results

Tables 3 and 4 show that the thermal pipeline achieves 90.6% macro F1 with 48 errors. Glass remains strong (F1=0.970), but paper drops significantly to F1=0.804, reflecting similar thermal signatures between paper and plastic materials.

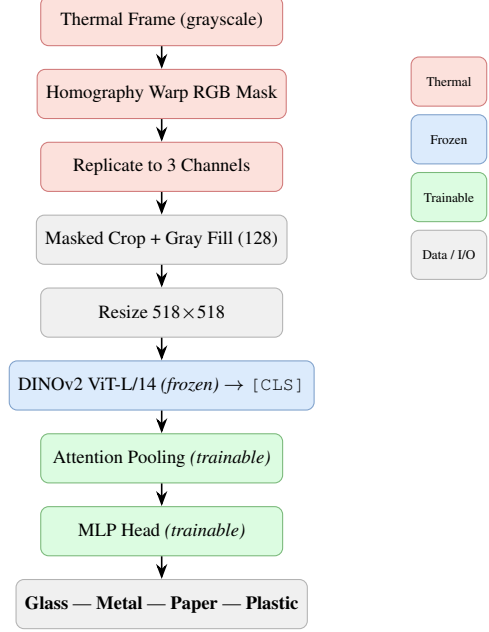


Figure 3. Thermal classification pipeline. Red blocks indicate thermal-specific steps: homography-based mask warping and grayscale-to-3-channel replication.

Table 3. Thermal pipeline: 5-fold stratified CV (550 tracklets).

Metric	Value
Mean Accuracy	0.9127 ± 0.0384
Mean Macro F1	0.9056 ± 0.0405
Pooled Accuracy	0.9127
Total Errors	48 / 550

Table 4. Thermal per-class metrics (pooled across all 5 folds).

Class	Prec.	Rec.	F1	N
Glass	0.956	0.985	0.970	132
Metal	0.938	0.924	0.931	131
Paper	0.792	0.816	0.804	98
Plastic	0.929	0.905	0.917	189

3.3. Confusion Matrix

The thermal confusion matrix (Fig. 4) shows that paper↔plastic confusion accounts for 27 of 48 total errors (11 paper→plastic, 16 plastic→paper), reflecting inherently similar thermal signatures for these materials.

4. Comparative Summary

Figure 5 compares per-class F1 scores across both modalities. Key findings:

- **RGB outperforms thermal overall** (95.1% vs. 90.6%)

		Predicted Label			
		Glass	Metal	Paper	Plastic
True Label	Glass	130	1	0	1
	Metal	4	121	5	1
	Paper	2	5	80	11
	Plastic	0	2	16	171

Figure 4. Thermal confusion matrix (pooled, 5 folds). 48 errors; paper \leftrightarrow plastic confusion dominates (27 errors).

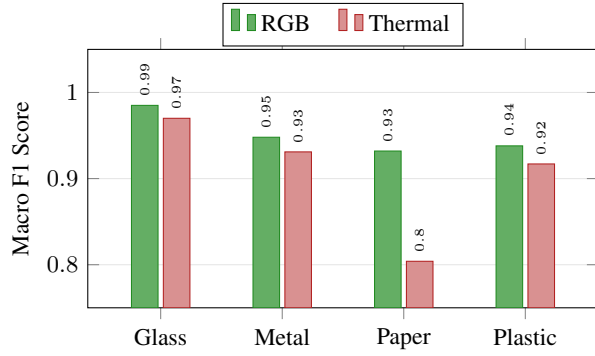


Figure 5. Per-class F1 comparison: RGB vs. Thermal. RGB outperforms on all classes; largest gap on paper (0.932 vs. 0.804).

macro F1), with both modalities achieving strong results using the same frozen DINOv2 backbone and identical trainable architecture.

- **Glass is near-perfect** in both pipelines ($F1 \geq 0.97$), indicating highly distinctive visual and thermal signatures.
- **Paper is the hardest class** for both modalities, but especially for thermal ($F1 = 0.804$ vs. 0.932), reflecting inherently similar thermal signatures between paper and plastic.
- **Thermal shows higher variance** across folds (± 0.04 vs. ± 0.01), suggesting that thermal features are more sensitive to the specific train/test partition.
- Despite lower standalone accuracy, the **thermal modality captures complementary** material properties (thermal conductivity, emissivity) that benefit a downstream late fusion approach.