

Municipal Waste Classification with DINOv2 Embeddings

RGB and Thermal Pipeline Performance

February 18, 2026

1 Overview

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

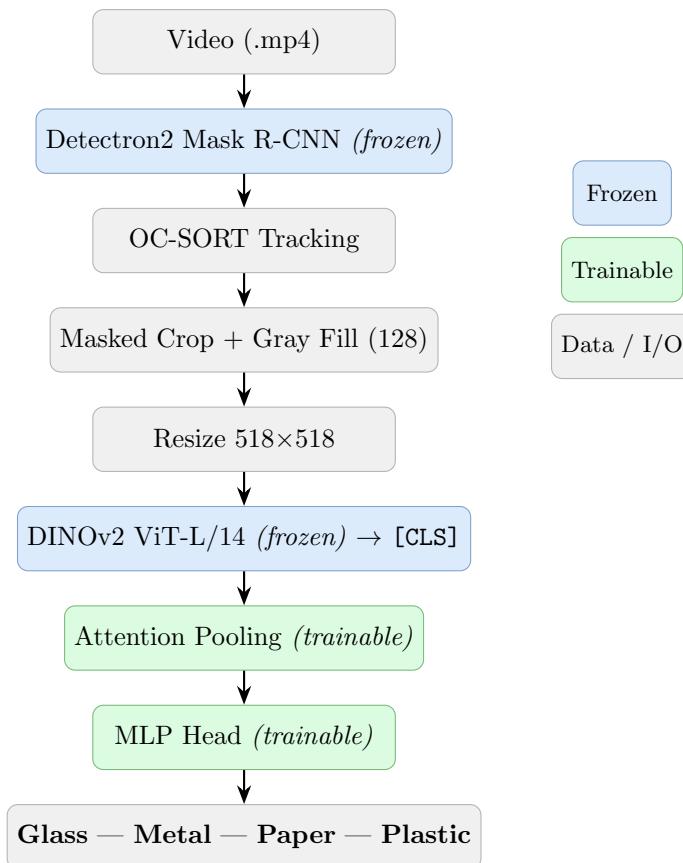


Figure 1: RGB classification pipeline. Detection and DINOv2 are frozen; only the attention pool and MLP head are trained ($\sim 528K$ parameters).

2.2 Overall Results

Table 1: RGB pipeline: 5-fold stratified CV results (550 tracklets, 19 videos).

Metric	Value
Mean Accuracy	0.9509 ± 0.0093
Mean Macro F1	0.9507 ± 0.0110
Pooled Accuracy	0.9509
Pooled Macro F1	0.9508
Total Errors	27 / 550

2.3 Per-Class Metrics

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

Class	Precision	Recall	F1	Count
Glass	0.9776	0.9924	0.9850	132
Metal	0.9209	0.9771	0.9481	131
Paper	0.9570	0.9082	0.9319	98
Plastic	0.9511	0.9259	0.9383	189

2.4 Confusion Matrix

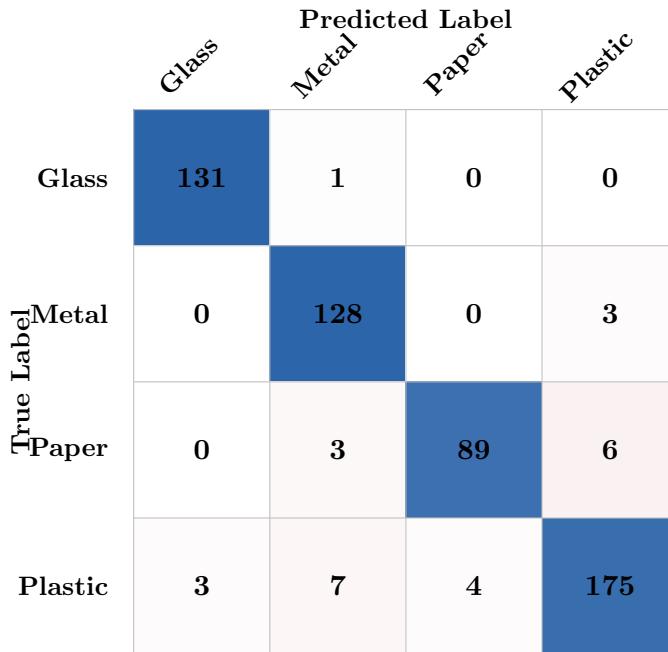


Figure 2: RGB confusion matrix (pooled across 5 folds). Only 27 misclassifications out of 550. Glass is near-perfect (131/132). Most confusion: plastic↔metal (10) and paper↔plastic (10).

3 Thermal Pipeline

3.1 Architecture

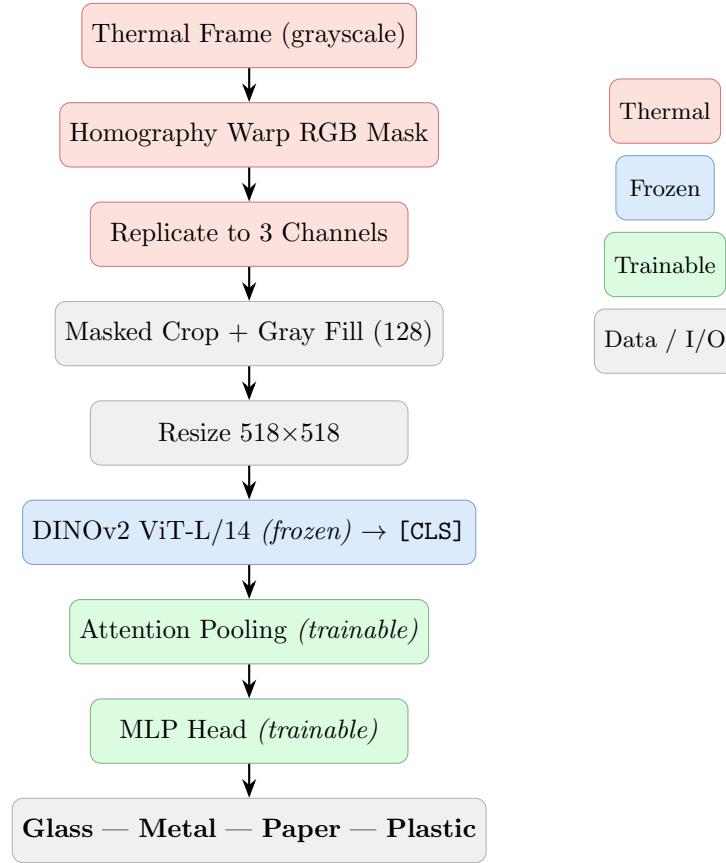


Figure 3: Thermal classification pipeline. Thermal-specific steps (red) include homography-based mask warping from RGB space and grayscale-to-3-channel replication for DINOv2 compatibility.

3.2 Overall Results

Table 3: Thermal pipeline: 5-fold stratified CV results (550 tracklets, 19 videos).

Metric	Value
Mean Accuracy	0.9127 ± 0.0384
Mean Macro F1	0.9056 ± 0.0405
Pooled Accuracy	0.9127
Pooled Macro F1	0.9055
Total Errors	48 / 550

3.3 Per-Class Metrics

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

Class	Precision	Recall	F1	Count
Glass	0.9559	0.9848	0.9701	132
Metal	0.9380	0.9237	0.9308	131
Paper	0.7921	0.8163	0.8040	98
Plastic	0.9293	0.9048	0.9169	189

3.4 Confusion Matrix

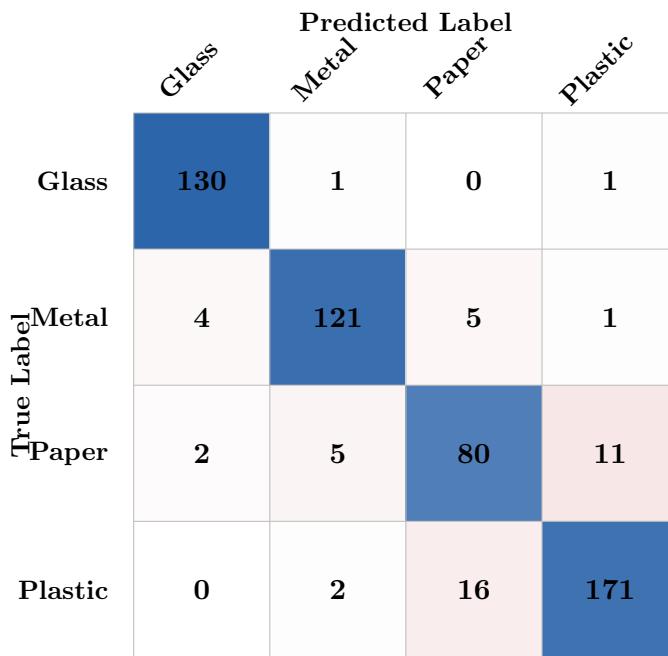


Figure 4: Thermal confusion matrix (pooled across 5 folds). 48 misclassifications. Paper↔plastic confusion dominates (27 errors), reflecting similar thermal signatures for these materials.

4 Comparative Summary

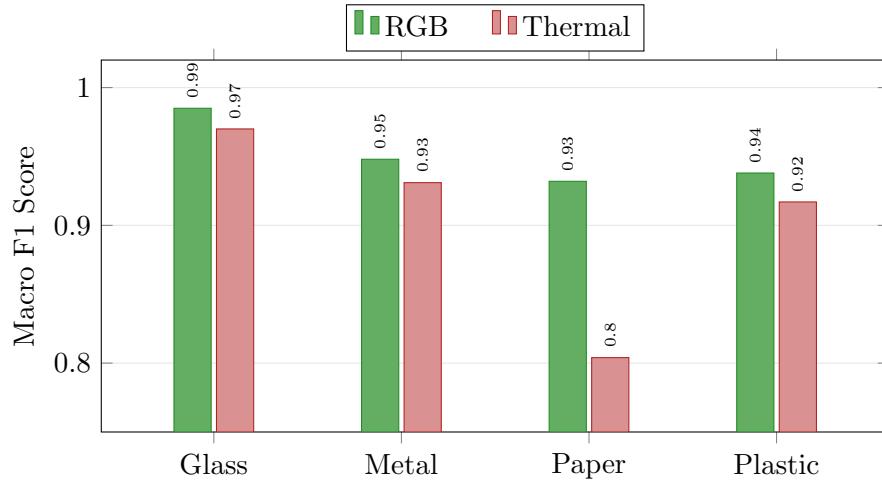


Figure 5: Per-class F1 comparison: RGB vs. Thermal. RGB outperforms thermal on all classes, with the largest gap on paper (0.932 vs. 0.804).

Key takeaways: RGB outperforms thermal overall (**95.1%** vs. **90.6%** 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 ($F1 \geq 0.97$). 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. The thermal pipeline also exhibits 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 can benefit a fusion approach.