International Journal of Emerging Technologies and Advanced Applications

Error Discovery and Label Utility in Cross-Domain Waste Classification

Chenwu Li[1]
[1]GuangDong Polytechnic, Foshan, China
lichenwu0808@163.com

***Abstract*—When a waste classifier encounters unfamiliar images, its most revealing mistakes may be poor examples for retraining. We examined this distinction in a single round of label acquisition for five materials. Two frozen image encoders with classification heads fitted on TrashNet supplied the scores for eight selection policies at budgets of 25, 50 and 100 labels. On RealWaste landfill photographs, we developed a policy that combines classifier disagreement with representative examples from the remaining pool. At 50 labels, mean held-out macro F1 was 0.636 for this policy and 0.593 for random sampling. We then applied the same rule to litter objects cropped from TACO using supplied bounding boxes. Across five paired candidate pools, mean macro F1 was 0.571 for the mixed policy and 0.585 for random sampling. The difference was -1.45 macro F1 points (conditional 95 percent bootstrap interval, -5.42 to $+2.65$). In a separate RealWaste analysis, selecting from the largest disagreement groups found 33.8 high-confidence errors per 50 labels, compared with 25.8 for random sampling within the same shortlist. Those batches nevertheless yielded lower held-out macro F1 (0.544 versus 0.621). Disagreement helped identify failures worth inspecting, but error counts alone did not predict the benefit of retraining.**

***Index Terms*—Active learning, Domain adaptation, Waste classification, Annotation budgets, Model disagreement**

I. Introduction

A classifier trained on photographs of isolated waste items must handle clutter, changes in object scale and unfamiliar packaging when applied to landfill images. Suppose fifty of those images can be labeled. For an audit, several photographs of plastic misclassified as metal can document a recurring error. For retraining, the same photographs may provide little information about cardboard, glass or paper. We ask how often a batch that reveals errors also improves the classifier on images outside that batch.

TrashNet provides a controlled source for examining this distinction [1]. RealWaste supplies material-labeled landfill images [2]. TACO provides annotated objects in litter scenes, allowing a different target to be constructed using reference bounding boxes [3]. The three sources support distinct experimental roles. The primary claim concerns material classification after a target label budget has been spent. On TACO the location of an object is supplied to the classifier. No detector is evaluated. In both targets, purchasing a label means revealing an existing annotation in an offline simulation.

Earlier work already establishes that a classifier trained on TrashNet loses accuracy on landfill images and benefits from target supervision [4]. A second line of work adapts vision–language classifiers with a few labeled examples per material [5]. An experiment that grants ten labels per class presupposes a way to find those classes. In an unlabeled candidate pool, a fifty-image budget buys fifty decisions, with no guarantee about how many examples of glass or cardboard will appear. This distinction matters particularly when most available photographs represent one material.

We study three questions. First, can a second source-trained predictor expose confident mistakes that low-confidence sampling overlooks? Second, does spending some labels on those disagreements and the rest on typical target appearances improve held-out macro F1? Third, does a policy chosen on landfill images retain its advantage on a new litter-crop benchmark with the same broad material names? We also examine which errors an update corrects and which new errors it introduces. Two models can correct different sets of errors and still end with similar accuracy.

The contribution is an empirical comparison of error discovery and downstream performance under eight acquisition policies, with identical candidate pools, label budgets and classifier updates. The mixed rule is a test case assembled from established selection principles. A retrospective shortlist experiment varies the allocation of queries across predicted material pairs while holding the initial shortlist fixed. The favorable development result and the unsuccessful external comparison delimit the scope of the method.

Figure 1 traces a purchased label through the two evaluations. The audit assesses the original classifier on the queried images. Adaptation is scored on a separate set of images after updating the main head.

Copyright: © 2026 The Author(s);
CC BY-NC 4.0.
Exclusive licensee IJETAA.

ISSN:3006-2985(Print)
ISSN:3006-9297(Online)

Received: Aug. 31, 2026
Revised: Oct. 24, 2026
Accepted: Oct. 25, 2026

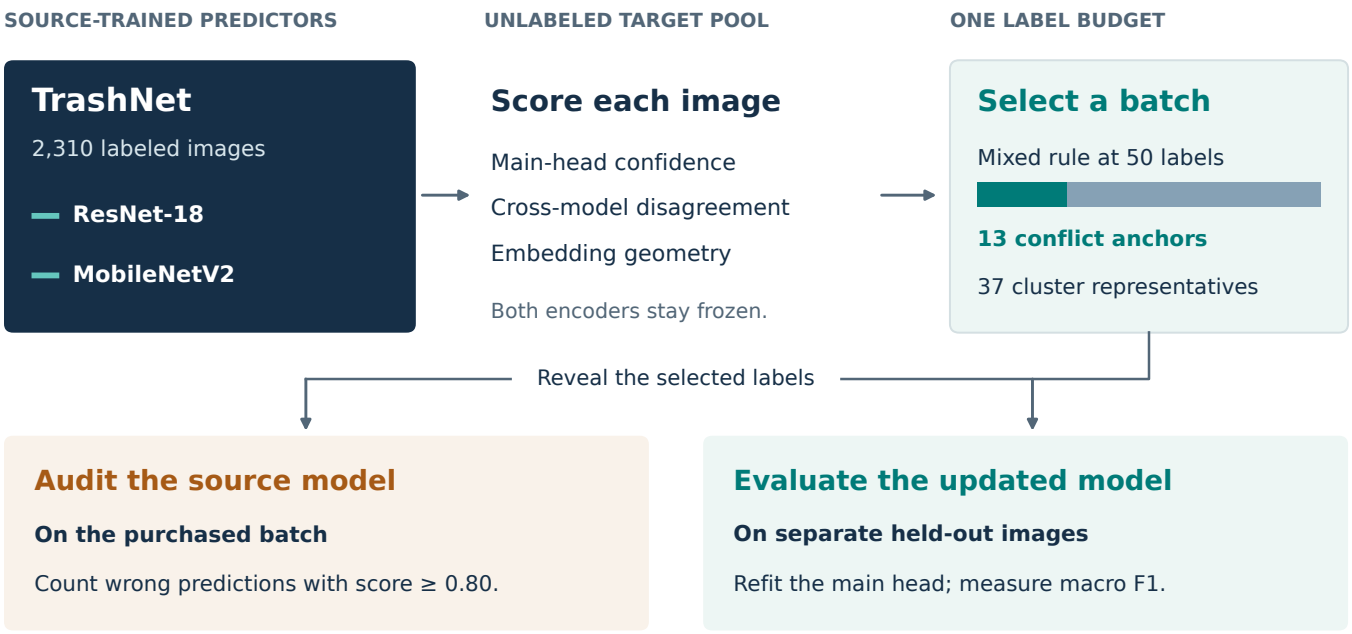

Fig. 1: From source training to target evaluation. Both heads are fitted on the five TrashNet materials. The example shows the frozen mixed policy at its primary budget. RealWaste supplies the development target; TACO supplies a separate external pool and test set. The two parts of the mixed batch are selected before their material labels are revealed.

II. Related Work

A. *Waste classification across acquisition settings*

Within-dataset experiments have compared deep models for material recognition on TrashNet [6]. For a transfer study, their accuracy answers a different question from the fraction of initially wrong target predictions that a small labeling budget can correct. A model may exploit lighting, object scale, or packaging forms that are well represented in the original split. Changing any of these can alter the ranking of candidate images even when the categories retain the same names.

Background replacement has already been used to train waste classifiers and evaluate them on real litter photographs [7]. The RealWaste study compared controlled and landfill imagery and documented material confusions [2]. Echchidmi and Bouayad reported an accuracy of 41.04 percent when transferring their compact model to a six-category RealWaste task, increasing to 79.67 percent after target fine-tuning [4]. We use five classes, a different split and frozen encoders. Their estimates establish the importance of adaptation; differences from our raw scores cannot isolate the effect of a selection policy.

Funk et al. evaluated few-shot cache-based adaptation, including on RealWaste [5]. Their question concerns how to use a support set once examples of each class are available. Here the learner chooses which labels to request before observing class membership. Work on reinforcement learning for waste classification has also considered iterative acquisition from an initially labeled subset [8]. An industrial detection benchmark compared sampling strategies across paired seeds and found that conclusions were sensitive to whether the budget counted images or bounding boxes [9]. We count one material label per image or supplied crop. Equal image counts need not represent equal human effort if one material takes longer to identify.

B. *Acquisition under distribution shift*

Least-confidence sampling orders images by the largest probability assigned to any class. A committee instead compares the predictions of several classifiers. In controlled image-classification experiments, Munjal et al. found that uncertainty, diversity and committee methods did not consistently outperform random sampling [10]. Our two classifiers use different encoders but share the same source labels. We therefore evaluate disagreement as a way to find errors, then measure what happens when those labels enter training.

Farthest-first sampling spreads a batch across the embedding space, whereas clustering allocates more representatives to dense regions. For object detection, Auh et al. adjusted the balance between predictive uncertainty and class diversity when selecting images [11]. In classification and segmentation experiments, Mittal et al. found that the relative performance of acquisition rules depended on the annotation budget and the redundancy of the candidate pool [12]. Active adversarial adaptation uses the relationship between source and target distributions [13]. CLUE clusters target embeddings with weights based on predictive uncertainty [14]. We implement an entropy-weighted centroid comparator under the same frozen representations and classifier update as our other policies. Published full training pipelines have different encoders, objectives and update schedules.

TABLE I: Image counts. RealWaste and TACO pool and test columns are after image-level duplicate screening. Source counts in the first column precede screening.

Material	TrashNet source	RealWaste pool	RealWaste test	TACO crops pool	TACO crops test
Cardboard	403	344	115	21	12
Glass	501	308	103	17	9
Metal	410	589	196	42	22
Paper	594	374	124	19	11
Plastic	482	689	230	148	79
Total	2390	2304	768	247	133

Domain generalization evaluates performance in environments unavailable during training [15]. Here the acquisition rule is fixed before external evaluation, but the classifier receives a limited number of labels from the new target. Tomani et al. found that predictions could remain overconfident after in-domain calibration when the input distribution changed [16]. We use a fixed score threshold to describe the source model's behavior. We do not interpret its value as a validated probability of correct target classification.

III. Data, Task and Separation of Evidence

A. Five aligned materials

The label set is cardboard, glass, metal, paper and plastic, in that order throughout. The material names refer to dataset annotations. The protocol assumes beforehand that every candidate belongs to one of the five eligible categories. It estimates the value of material labels conditional on this screening and, for TACO, on a supplied object box. The source archive contains 2,527 JPEG files. Its remaining 137 images are assigned to a residual trash category. The scope of that category can change with the collection protocol. We exclude it and the target-only categories to focus on five named materials. Table I gives the image counts. All five remain eligible at every budget. This is a closed-set classification study. Operational categories such as hazardous material, contaminated packaging, unknown items and composite products would require additional definitions.

Image-level screening gives priority to source images, then processes target images in sorted order. An image is dropped if the SHA-256 digest of its decoded pixels has already appeared, or if its 64-bit horizontal difference hash differs from a retained hash by at most four bits. Difference hashes use luminance resized to nine by eight pixels. Five disjoint bit bands retrieve candidate matches before exact Hamming-distance testing. A separate cross-target check found no matching pixel digests or difference hashes within four bits between the 382 available TACO crops and 3,092 eligible RealWaste images. The screening removes 80 of 2,390 source images and 20 of 3,092 eligible RealWaste images. This greedy approximation can merge distinct low-detail objects or miss a repeated physical item photographed from a new angle. A real object identifier would provide stronger separation.

B. Independent construction of a litter-crop target

TACO scenes can contain multiple labeled objects of different materials. Assigning the entire image a single label would confound localization with classification. We first map an explicit subset of fine categories to the five materials. Examples include corrugated carton, glass bottles, drink cans, normal paper and clear plastic bottles. Drink cartons and ambiguous wrappers are excluded. The reproducibility files record the exact category identifiers and retained scenes. These rules define an evaluable subset of the public collection, not a new annotation of material composition.

For each scene, we keep at most one eligible object. Its smaller bounding-box dimension must be at least 48 pixels after scaling the longer image dimension to 640 pixels. Among eligible objects we choose the largest box by area, with smaller annotation ID breaking ties. We request the official thumbnail and transform its bounding-box coordinates using the actual thumbnail width and height. This produces a rectangular crop that retains local context inside the box. It also favors larger objects and systematically misses small litter. Use of reference boxes assumes access to locations at evaluation time.

Of 652 eligible scenes, 382 supplied usable thumbnails; the other 270 lacked a thumbnail URL in the annotations snapshot. Their full-resolution images were not substituted. The results therefore apply to the available-thumbnail subset. Two additional crops were excluded by source-first duplicate screening. Of the 380 retained scenes, 247 enter the candidate pool and 133 enter the stratified test partition. Plastic accounts for 79 of the test images and glass for nine. We carry those denominators into the analysis of uncertainty.

C. Order of analysis

The first RealWaste pilot compared four selection policies. The study was not publicly preregistered. Its results and error patterns motivated the mixed policy in this paper. The 768-image RealWaste test partition had already been examined during that pilot. We retain it as a development benchmark and explicitly treat the expanded comparisons there as exploratory. The mixing fraction is chosen on an inner split of the RealWaste development pool. Following the distinction between model development and independent evaluation discussed by Kapoor and Narayanan [17], we reserve the external comparison for the separately assembled TACO subset.

The TACO category mapping, crop geometry and sample availability were inspected while constructing the benchmark. The external macro F1 comparison at fifty labels and the model settings were recorded before computing external model scores. External target labels were then revealed only for selected candidate images during each fit. The external test annotations were used for stratification and final scoring. This transfers a selection rule to a new target and permits a bounded amount of supervision there. The distinction from zero-label transfer is important: the source-only row is the zero-target-label comparison.

IV. Model and Label Acquisition

A. Source scores and fixed adaptation objective

The image encoders are ResNet-18 and MobileNetV2 [18], [19], initialized from official ImageNet-trained weights. Each RGB image is resized to 224 by 224 pixels and normalized with the standard ImageNet channel statistics. The final classification layer is removed. The resulting pooled embedding is normalized to unit Euclidean norm and multiplied by ten before logistic regression. Rectangular inputs are resized directly, so aspect ratio is not preserved. Using the same transformation throughout keeps this distortion fixed across policies.

Let $f(x)$ be the main representation and $g(x)$ the auxiliary one. Multinomial logistic heads fitted on all retained source examples yield class distributions $p(x)$ and $q(x)$. ResNet-18 supplies p. Both encoders and both acquisition heads are fixed while selecting a batch. Only the main head is refit after target labels are revealed. We fit the heads with scikit-learn 1.7.2, using regularized logistic regression with $C = 1$ and at most 500 optimizer iterations. No image augmentation or end-to-end fine-tuning takes place.

Let n_s denote the number of retained source examples and B the label budget. For the adapted model, source observations receive weight $1/2$ each, and each of the B target observations receives weight $n_s/(2B)$. The total source and target weights are equal at every budget. In the weighted data loss, the two domains therefore contribute

$$L_{\text{data}} = \tfrac{1}{2} \sum_{i=1}^{n_s} \ell_i + \tfrac{n_s}{2B} \sum_{j=1}^{B} \ell_j. \tag{1}$$

Here ℓ is multinomial cross-entropy. The estimator adds its usual weight penalty. In particular, a heavily biased selection can strongly influence the update. The primary comparison fixes these weights. A retrospective sensitivity check varies the target share on RealWaste only, holding the purchased labels fixed.

B. Disagreement, ranking and representation

For candidate i, define $c_i = \max_k p_{ik}$ and $d_i = 1$ when the predicted classes of the two source heads differ. For a deterministic priority order we use

$$s_i = 2d_i + c_i. \tag{2}$$

All disagreements precede agreements, with larger main-model scores first within each set. The auxiliary model can expose confident mistakes, but it can also dissent when the main prediction is correct. Thus disagreement can be evaluated by counting purchased labels that reveal an actual error. An alternative comparator chooses the smallest c_i directly.

Table II summarizes the eight rules. For a finite embedding set, farthest-first selection starts at a seeded random point. Subsequent choices minimize the maximum cosine similarity to the already chosen items. This tends to cover extremes. The centroid rule instead fits B clusters with one seeded k-means++ initialization and at most 100 iterations. From each cluster it buys the actual image nearest the center. In the entropy variant, the clustering weight of image i is $-\sum_k p_{ik} \log p_{ik}$. The nearest representative is still an observed image. Weighting changes the center locations and potentially the partition.

TABLE II: Eight acquisition policies. All receive the same candidate pool, label count and adaptation rule.

Policy	Rule
Random	Uniform permutation, first B images.
Least confident	Lowest B main-head maxima.
Farthest first	Cosine farthest-first across the pool.
Conflict rank	First B by the disagreement priority.
Conflict spread	Farthest-first within the top $2B$, truncated to pool size.
Centroids	Nearest member of each unweighted cluster.
Entropy centroids	Nearest member of each entropy-weighted cluster.
Mixed	Conflict anchors followed by representatives of the remaining pool.

C. A mixed policy and its optimization

The mixed policy spends $k = \lfloor \alpha B + 0.5 \rfloor$ labels on disagreement anchors. We take the first $2k$ candidates in the priority ordering and apply farthest-first sampling to select k anchors. We remove these images, cluster the remaining pool into $B - k$ groups and buy one representative per group. No target label, including the anchor labels, is read before the entire batch is chosen. This is a single-batch policy. Its second stage can select an image visually close to an anchor because it clusters the remaining distribution without conditioning its centers on anchor locations.

The initial pilot suggested that spending an entire batch on conflicts might overrepresent a restricted set of appearances. We therefore tested fractions of 0.25, 0.50 and 0.75. We held out 33 percent of the 2,304-image RealWaste development pool with stratification, and simulated acquisition from the remainder. Each fraction was evaluated at all three budgets and the first three fixed seeds. Averaging the nine macro F1 values per fraction gave 0.6191, 0.6090 and 0.6131, respectively. We fixed the fraction at 0.25 for both reporting domains. This selects 6, 13 and 25 conflict anchors at budgets of 25, 50 and 100. The development objective was the mean across budgets; there was no per-budget tuning or external retuning.

V. Evaluation

A. Paired candidate pools and outcome measures

The five availability seeds are 17, 29, 43, 71 and 101. For each seed, we permute the domain's development pool and expose the first 80 percent to all acquisition methods. This yields 1,843 RealWaste candidates or 197 TACO candidates per seed. Within a seed the policies see the same unlabeled images in the same order. We reveal the reference label only after selection. Each budget is a new single-batch simulation from the original source model. A 50-label fit therefore need not contain the 25-label set. The expanded comparison comprises 240 fits across two domains. The original 60 RealWaste pilot runs are represented within that comparison and are not counted as additional independent trials.

For a material k let TP_k, FP_k and FN_k denote the three corresponding counts in the test confusion matrix. Our primary outcome weights the five materials equally [20]:

$$F_{\text{macro}} = \frac{1}{5}\sum_{k=1}^{5}\frac{2TP_k}{2TP_k + FP_k + FN_k}. \tag{3}$$

Accuracy weights individual test images equally. The two scores can rank policies differently on TACO because plastic is common. We compute each score separately per run and then average over paired seeds. Confidence is descriptive: a high-confidence error has $c_i \geq 0.80$ and an incorrect main-model prediction. We report the total error count as well, since least-confidence sampling and high-confidence error counting naturally emphasize opposite ends of the ranking.

To track what changes after buying labels, we count held-out images that the source model missed and the adapted model corrects. We separately count previously correct images that become wrong. Their difference divided by test size is exactly the change in accuracy. It does not determine the change in macro F1 because the distribution of errors among classes matters. To characterize batch coverage we measure mean cosine similarity from each remaining candidate to its nearest selected image. All these diagnostics describe the behavior of the frozen representation and the selected batches.

B. Primary contrast and uncertainty

The external primary contrast subtracts random sampling from the mixed policy at fifty labels. The point estimate averages five paired macro F1 differences. For an interval conditional on the observed models, seeds and class counts, we resample held-out scenes with replacement within each true material, using the same sampled indices for both policies and all five seeds. The 2.5th and 97.5th percentiles of 2,000 resampled mean differences form the reported interval. The resampling seed is 3407. Fixing the class totals avoids empty rare classes in bootstrap replicates. It also conditions away uncertainty about class prevalence.

The five availability seeds reuse the same test set. Their dispersion describes sampling and initialization variability within this experiment. Treating the five values as independent deployment sites would grossly overstate the replication. Comparisons at other budgets, selection-rule ablations, per-class findings and the exploratory RealWaste interval are descriptive. The study does not select a favorable test budget after scoring.

VI. Results

A. Confident errors are easy to find on RealWaste

Without target supervision the primary source model obtains accuracy 0.5469 and macro F1 0.4844 on the 768 RealWaste test images. In the development pool, 566 of 1,043 mistakes have scores of at least 0.80. Among the 1,511 high-confidence predictions, more than one in three is wrong. For 113 images the source head predicts metal although the reference material is plastic. Another 72 glass items are confidently called plastic. These directions of confusion suggest that a policy which queries only the least confident images can leave important error modes unaudited.

TABLE III: RealWaste. Held-out macro F1 (mean and sample standard deviation across five paired candidate pools). Columns give label budgets.

Policy	25	50	100
Random	0.548 ± 0.037	0.593 ± 0.021	0.641 ± 0.027
Least confident	0.530 ± 0.019	0.582 ± 0.009	0.613 ± 0.005
Farthest first	0.542 ± 0.023	0.580 ± 0.012	0.623 ± 0.016
Conflict rank	0.584 ± 0.007	0.605 ± 0.011	0.664 ± 0.011
Conflict spread	0.566 ± 0.018	0.604 ± 0.013	0.625 ± 0.008
Centroids	0.591 ± 0.021	0.621 ± 0.016	0.673 ± 0.014
Entropy centroids	0.567 ± 0.021	0.592 ± 0.031	0.657 ± 0.017
Mixed	0.586 ± 0.024	0.636 ± 0.017	0.661 ± 0.019

A retrospective neighborhood audit quantifies how these mistakes relate in the main embedding. For each high-confidence error, we find the five nearest other development images by cosine distance. On average 44.2 percent of these neighbors are also high-confidence errors. The mean drops to 29.8 percent when error indicators are permuted 200 times within each material class. The permutation preserves each material's number of errors. It supports the interpretation that errors concentrate in particular appearances within the feature space. It does not establish that a particular cue, such as transparency, caused those errors.

At fifty labels, conflict spread uncovers 26.0 high-confidence errors on average, compared with 11.2 for random acquisition. Least-confidence selection finds 34.2 errors of any score and no high-confidence errors. Its failure to reach the high-confidence stratum follows from its ranking rule when sufficient low-score examples exist. The more informative comparison is downstream: exposing more total source errors does not give least-confidence sampling a corresponding performance advantage.

B. A mixed batch improves the development result

The mixed policy obtains macro F1 of 0.6362 at fifty labels, compared with 0.5926 for random, 0.6042 for conflict spread and 0.6214 for unweighted centroids. All five paired differences against random are positive at this budget, ranging from 0.0109 to 0.0754. The conditional image-bootstrap interval for the mean difference is [0.0257, 0.0621]. As the choice of policy was motivated by this domain, the interval describes a development result.

The effect has a concrete interpretation at the level of individual predictions. At fifty labels the mixed policy corrects a mean of 119.2 previously wrong test predictions and introduces 35.4 new errors. Random selection corrects 96.0 and introduces 34.4. The mean net difference is 22.2 correctly classified images out of 768, consistent with the increase from 0.6271 to 0.6560 accuracy. The corresponding class recalls reveal a redistribution: glass recall rises from 0.305 under random sampling to 0.466 with mixed selection, whereas plastic recall falls from 0.629 to 0.597.

Table III reports the mean and between-pool standard deviation at each budget. All five fits at a given budget use the same test images. Unweighted centroids give the highest observed

TABLE IV: TACO crops. Held-out macro F1 (mean and sample standard deviation across five paired candidate pools). Columns give label budgets.

Policy	25	50	100
Random	0.546 ± 0.032	0.585 ± 0.025	0.565 ± 0.048
Least confident	0.543 ± 0.015	0.558 ± 0.015	0.587 ± 0.020
Farthest first	0.550 ± 0.016	0.581 ± 0.035	0.625 ± 0.026
Conflict rank	0.564 ± 0.036	0.582 ± 0.018	0.599 ± 0.022
Conflict spread	0.531 ± 0.038	0.559 ± 0.025	0.596 ± 0.030
Centroids	0.555 ± 0.023	0.547 ± 0.020	0.568 ± 0.031
Entropy centroids	0.545 ± 0.027	0.566 ± 0.022	0.566 ± 0.033
Mixed	0.557 ± 0.041	0.571 ± 0.052	0.579 ± 0.047

mean macro F1 at both 25 and 100 labels on RealWaste. At 100 labels, conflict spread discovers 61.2 high-confidence errors and attains 0.6252 macro F1; random selection discovers 24.6 such errors and attains 0.6410. Strong enrichment is therefore compatible with worse generalization. Expanding the number of purchased labels also changes their weights in the fitted classifier, so marginal gains need not be linear in the count of new corrections.

To test whether the development advantage depended on assigning the target half the weight, we repeated the fifty-label fits at target shares of 0.10, 0.25 and 0.50, keeping total weight, selections and test images fixed. Mean macro F1 for random versus mixed was 0.572 versus 0.602, 0.582 versus 0.620, and 0.593 versus 0.636, respectively. The sign persisted across these choices on RealWaste. This retrospective comparison does not test transfer of the weight setting to a third site.

C. External result on litter crops

The source-only model achieves accuracy 0.4812 and macro F1 0.4526 on the 133 held-out TACO crops. At fifty labels random sampling increases macro F1 to 0.5853. The frozen mixed rule reaches 0.5708. Its mean paired advantage over random is -0.0145, with a conditional 95 percent interval of $[-0.0542, 0.0265]$. The mixed rule wins two of the five paired runs. The five differences, in the stated seed order and on a 0–100 macro F1 scale, are -0.57, $+1.77$, -8.09, $+2.11$ and -2.49. The mean therefore conceals substantial pool sensitivity. This comparison does not support a transferable advantage for the chosen mixing fraction. The interval includes both a modest benefit and a more consequential loss.

Table IV gives the complete budget comparison. The external ordering also depends on the endpoint. At fifty labels, farthest-first selection has higher mean accuracy than random (0.7263 versus 0.7128), with slightly lower macro F1 (0.5807 versus 0.5853). At one hundred labels it has the highest observed mean macro F1, 0.6254. The small target pool changes what these budgets mean: one hundred queries exhaust more than half of the 197 available candidates, whereas on RealWaste they consume approximately five percent. A shortlist of up to twice the budget can therefore encompass the entire available external pool. This explicitly limits the separation between unrestricted diversity and shortlist diversity at the largest budget.

TABLE V: Matched-shortlist comparison on RealWaste. Groups are distinct ordered prediction pairs. High is the mean number of confident source errors found.

B	Selection	Groups	High	Macro F1
25	Concentrated	2.8	15.0	0.5325
	Interleaved	10.8	8.6	0.5578
	Shortlist random	9.0	11.2	0.5684
50	Concentrated	2.0	33.8	0.5444
	Interleaved	13.8	19.4	0.5910
	Shortlist random	11.8	25.8	0.6208
100	Concentrated	2.4	76.2	0.5925
	Interleaved	16.0	51.2	0.6447
	Shortlist random	14.0	63.0	0.6489

Figure 2 shows how the paired fifty-label comparison changes between targets. At the primary budget, the mixed policy discovers 21.0 confident errors against 18.0 for random. It corrects 38.8 previously wrong test predictions and introduces 9.6 new errors, compared with 39.2 corrections and 8.4 new errors for random. All five materials appear in the selected batch under both policies in every run at this budget. The external result cannot be attributed simply to one method missing an entire material in its 50 queries.

D. A matched-shortlist test of concentration

To investigate why error counts can mislead, we conducted a retrospective RealWaste control. We held the candidate pool, $2B$-image disagreement shortlist, classifier and update weights fixed. Within the shortlist we grouped candidates by the ordered pair of the main and auxiliary predicted material. These are suspected confusion groups available before labeling; a true confusion group would require the unknown reference category. Within each group, candidates retained their original priority order.

A concentrated policy fills the batch from the largest groups first, resolving equal group sizes by the leading confidence. An interleaved policy takes the first example from each group before returning for another. A third control selects a uniform random subset of the same shortlist. It differs from uniform selection over the entire pool. Table V summarizes the control. Each policy was evaluated with all five seeds and all three budgets, adding 45 retrospective fits. The experiment uses the previously examined RealWaste population. We do not use it to change any TACO result.

Figure 3 shows the two outcomes for the same selected batches. The concentrated rule finds the most high-confidence errors at every budget and has the lowest downstream macro F1 of the three. At fifty labels it spends the entire budget across an average of just two ordered groups. Interleaving reaches almost fourteen groups and improves macro F1 by 0.0466. Shortlist random performs better still, despite representing fewer groups than interleaving. Group coverage alone therefore does not explain utility. The choice to privilege the highest-score members within each group and the balance of labels ultimately supplied to the update are additional plausible explanations. The experiment perturbs a sampling rule, not an isolated causal property of each image.

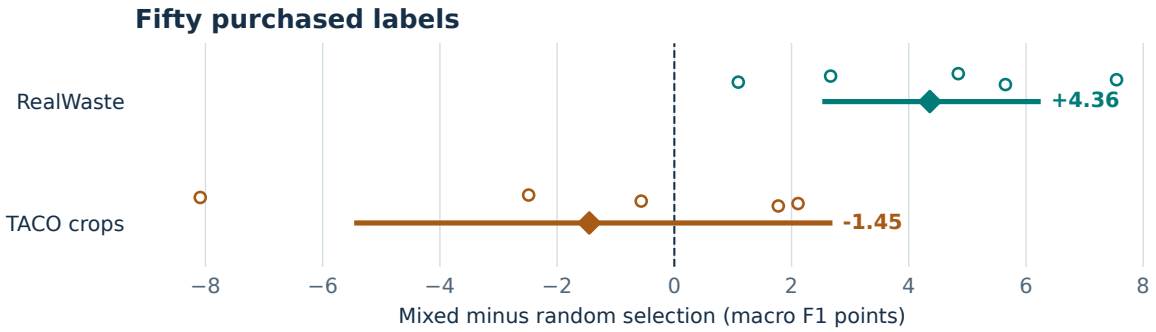

Fig. 2: Effect of mixed acquisition relative to random sampling at fifty labels. Open circles show five paired candidate pools, diamonds show their mean, and bars give conditional 95 percent image-bootstrap intervals. Positive values favor mixed acquisition. The RealWaste comparison is exploratory.

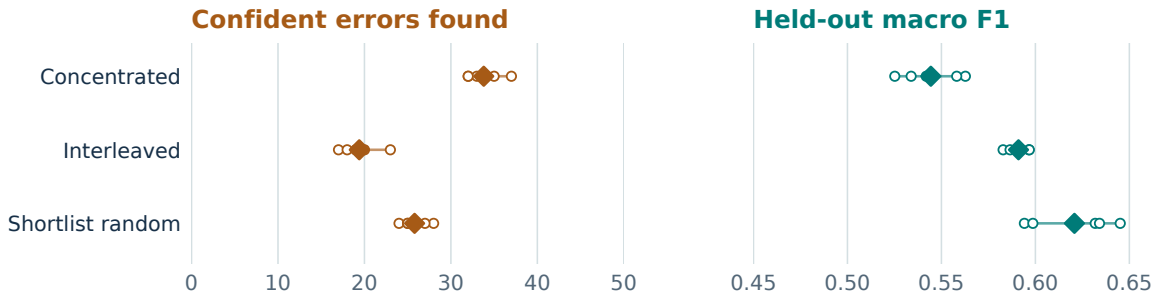

Fig. 3: Two outcomes for the same fifty labels in the RealWaste shortlist control. Each circle is one candidate pool, the diamond is the five-pool mean, and each line spans the observed range. Concentrating purchases finds the most confident mistakes while yielding the lowest mean held-out macro F1.

VII. Discussion

A. *Which labels would an auditor and a trainer buy?*

An auditor investigating false material assignments might reasonably choose conflict rank. On the external pool it identifies 33.4 confident errors in fifty queries, nearly twice the yield of uniform sampling. This count answers how many existing failures a reviewer would uncover. A practitioner choosing labels for a future classifier needs the held-out gain, including any newly introduced errors. In the same experiment conflict rank finishes at 0.5819 macro F1, close to and below random sampling. The acquisition rule prioritizes high-confidence disagreements, and the audit counts high-confidence errors. That shared confidence criterion partly explains enrichment. Held-out macro F1 evaluates the update on separate images.

One possible explanation concerns how similar training examples affect the update. Several images with nearly the same feature vector and previously incorrect label contribute gradients in closely related directions. Repeated labels may reinforce a genuinely important correction. But when the batch receives half of the total training weight, repetitions can also move a boundary at the expense of neighboring classes. This is a plausible interpretation of the matched-shortlist results. The present experiment does not estimate the marginal counterfactual effect of each purchased label. A direct test would compare adding one specific image to many otherwise identical batches.

Representative selection introduces a different trade-off. At fifty labels on RealWaste, unweighted centroids achieve a mean remaining-pool similarity of 0.811, compared with 0.788 for random and 0.760 for farthest first. A typical candidate is closer to an unweighted representative. Yet on TACO the centroid policy has the largest value of this coverage measure among the eight methods at fifty labels and falls below random in macro F1. An embedding can separate texture, viewpoint or scene composition more strongly than the material boundary of interest. A batch may represent that embedding accurately while missing visually subtle distinctions useful to the classifier.

B. What travels between the targets?

The closed-set task, class names, model weights, acquisition functions and update rule travel between the experiments. Image composition and class proportions change. RealWaste photographs represent discarded material in a landfill. The external images are selected object crops, and their bounding boxes suppress some of the localization challenge. A classifier trained on white-background objects may respond differently to a tight crop than to an entire cluttered photograph. The external benchmark adds independent evidence about the rule while changing several aspects of the input distribution simultaneously.

Absolute budgets are appropriate for a team that can label fifty objects. They induce very different sampling fractions here. This is particularly consequential for conflict spread, whose candidate set expands with the budget. Two complementary future comparisons would hold the budget fraction fixed and hold the externally available class counts fixed. The latter would be a retrospective diagnostic: class-balanced purchasing before observing any labels would require information unavailable in a normal unlabeled pool.

The small external classes also constrain conclusions. A single changed glass prediction alters glass recall by one ninth, approximately eleven percentage points in one run. Averaging five model fits can make a table look smoother without increasing the number of independent glass objects. Table I reports each class denominator. The reproducibility files give the paired scores and per-class recalls for all runs.

C. Practical implications and limits

A practical adaptation study should log three quantities separately: what the acquired images looked like to the original model, which truths the acquired labels revealed, and what changed on held-out images after updating. The first can guide acquisition, the second can audit a deployed baseline offline, and the third evaluates the value of annotation for generalization. Using a single quantity as a surrogate for all three can favor a policy for the wrong objective.

The classes encode the original dataset labels. Mixed-material items, transparent bottles and damaged packaging may be difficult to identify from a single RGB view. We did not obtain independent composition measurements or adjudicate disputed images. Similarly, the observation that a classifier initially labeled plastic as metal cannot identify the visual feature responsible for the mistake. Controlled imaging or independently verified material labels would be necessary to attribute an error to appearance ambiguity.

Our comparison holds the backbone and target-update weights fixed to separate acquisition effects. Fine-tuning, class-sensitive losses, foundation-model features or a support-set adapter could alter the relative value of unusual labels. A committee with common ImageNet pretraining can agree on a shared error. Budgeted calls to human annotators may also involve differing time costs and repeated inspection. Purchasing reference labels offline has none of the ambiguity-resolution costs of a human labeling study.

Finally, one landfill collection and one available-thumbnail litter subset provide limited environmental replication. Photographs of the same product design may persist after image-level screening. Geographic, camera and collection effects are not independently identified. Model scores are uncalibrated in the target domains. Neither the image-level accuracy nor the number of high-confidence mistakes estimates conveyor sorting throughput, recycling purity or field deployment performance.

VIII. Conclusion

At a fixed labeling budget, a batch rich in identifiable source-model mistakes can be a poor batch for retraining. Concentrating fifty RealWaste queries in a few predicted confusion groups exposed 33.8 confident errors on average and produced macro F1 of 0.5444. A uniform sample from the same shortlist exposed fewer such errors and yielded 0.6208. A mixed conflict-and-representation policy improved the fifty-label RealWaste result over unrestricted random sampling, but the pre-specified external comparison on TACO favored random by 0.0145 macro F1 and was inconclusive within its conditional interval. The most defensible use of disagreement in this study is to find cases for inspection. Selecting labels for adaptation requires its own held-out evaluation under the intended budget and target distribution.

Data, Code and Research Transparency

TrashNet, RealWaste and TACO are available from their respective producers [1]–[3]. The accompanying reproducibility files contain data-source links, the category map, retained image identifiers, selected identifiers, run-level results, predictions and analysis scripts. Images and pretrained weights remain available at their source repositories. The overview depicts the implemented protocol. All quantitative figures and tables use outputs from the local experiments. Source documents supply background and comparisons, with attribution. No new human annotation, field collection or hardware deployment forms part of the study.

An AI coding and writing assistant (OpenAI Codex) was used in the research workflow for experimental design, code generation and execution, literature retrieval, analysis, drafting and typesetting. The local experiment outputs provide the basis for the numerical results. This anonymous review draft leaves author-specific funding, competing interests, contributions and final approval to the submitting authors.

References

[1] M. Yang and G. Thung, “Classification of trash for recyclability status,” Stanford Univ., Stanford, CA, USA, CS229 Project Rep., 2016. Accessed: Oct. 24, 2026. [Online]. Available: https://cs229.stanford.edu/proj2016/report/ThungYang-ClassificationOfTrashForRecyclabilityStatus-report.pdf

[2] S. Single, S. Iranmanesh, and R. Raad, “RealWaste: A novel real-life data set for landfill waste classification using deep learning,” *Information*, vol. 14, no. 12, 2023, Art. no. 633, doi: 10.3390/info14120633.

[3] P. F. Proença and P. Simões, “TACO: Trash annotations in context dataset,” ver. 1.0, Zenodo, Jun. 9, 2019, dataset, doi: 10.5281/zenodo.3587843.

[4] M. Echchidmi and A. Bouayad, "Compact waste image classification with multi-student CNNs and edge-oriented model selection," *Front. Artif. Intell.*, vol. 9, 2026, Art. no. 1804734, doi: 10.3389/frai.2026.1804734.

[5] J. V. Funk, P. Bäcker, L. Roming, J. Josekutty, G. Maier, and T. Längle, "Evaluation of vision–language models for waste classification: Zero- and few-shot with a training-free tip-adapter," *Cleaner Waste Syst.*, vol. 13, 2026, Art. no. 100475, doi: 10.1016/j.clwas.2026.100475.

[6] M. M. Hossen *et al.*, "A reliable and robust deep learning model for effective recyclable waste classification," *IEEE Access*, vol. 12, pp. 13809–13821, 2024, doi: 10.1109/ACCESS.2024.3354774.

[7] A. Patrizi, G. Gambosi, and F. M. Zanzotto, "Data augmentation using background replacement for automated sorting of littered waste," *J. Imaging*, vol. 7, no. 8, 2021, Art. no. 144, doi: 10.3390/jimaging7080144.

[8] K. Maleknya, R. Alizadehsani, S. Pedrammehr, and K. Shirini, "A solid waste classification using reinforcement learning boosted by data augmentation and hyperparameter optimization," *Sci. Rep.*, vol. 16, no. 1, 2026, Art. no. 18132, doi: 10.1038/s41598-026-46347-7.

[9] T. T. Le, T. N. Chu, and M. H. Nguyen, "A controlled benchmark of active learning strategies for class-imbalanced waste detection on SortWaste with YOLO," *Int. J. Adv. Comput. Sci. Appl.*, vol. 17, no. 10, pp. 1429–1442, 2026, doi: 10.14569/IJACSA.2026.01710123.

[10] P. Munjal, N. Hayat, M. Hayat, J. Sourati, and S. Khan, "Towards robust and reproducible active learning using neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 223–232, doi: 10.1109/CVPR52688.2022.00032.

[11] J. Auh, C. Cho, and S.-T. Kim, "Adaptive active learning with dynamic uncertainty-diversity balancing for object detection," *J. Intell. Robot. Syst.*, vol. 112, no. 2, 2026, Art. no. 45, doi: 10.1007/s10846-026-02380-2.

[12] S. Mittal *et al.*, "Realistic evaluation of deep active learning for image classification and semantic segmentation," *Int. J. Comput. Vis.*, vol. 133, no. 7, pp. 4294–4316, 2025, doi: 10.1007/s11263-025-02372-z.

[13] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 728–737, doi: 10.1109/WACV45572.2020.9093390.

[14] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8485–8494, doi: 10.1109/ICCV48922.2021.00839.

[15] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023, doi: 10.1109/TPAMI.2022.3195549.

[16] C. Tomani, S. Gruber, M. E. Erdem, D. Cremers, and F. Buettner, "Post-hoc uncertainty calibration for domain drift scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10119–10127, doi: 10.1109/CVPR46437.2021.00999.

[17] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023, Art. no. 100804, doi: 10.1016/j.patter.2023.100804.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[20] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.