Waste Management 198 (2025) 187–197

Contents lists available at ScienceDirect

Waste Management

journal homepage: www.elsevier.com/locate/wasman

Generalization abilities of foundation models in waste classification

Aloïs Babé [a,b,*], Rémi Cuingnet [b], Mihaela Scuturici [a], Serge Miguet [a]

[a] Université Lumière Lyon 2, CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Claude Bernard, Lyon 1, LIRIS, UMR 5205, Bron 69676, France
[b] Veolia Scientific & Technical Expertise Department, Maisons-Laffitte 78600, France

ARTICLE INFO

Keywords:
Waste classification
Computer vision
Foundation model
Generalization

ABSTRACT

Industrial waste classification systems based on computer vision require strong generalization abilities across location and time period in order to be deployed. This study investigates the potential of foundation models, known for their adaptability to a wide range of tasks and promising generalization capabilities, to serve as the basis for such systems. To evaluate the generalization performance of foundation models we use five waste classification datasets spanning various domains, train the models on one dataset and test them on all others. Additionally, we explore various training procedures to optimize foundation model adaptation for this specific domain. Our findings reveal that foundation models exhibit superior generalization abilities compared to standard models and that good generalization performance is correlated with the model size and the size of the model pretraining dataset. Furthermore, we demonstrate that elaborate classifier heads are not necessary for extracting discriminative features from foundation models. Both standard fine-tuning and Parameter-Efficient Fine-tuning (PEFT) improve generalization performance, with PEFT being particularly effective for larger models. Simple data augmentation techniques were found to be ineffective. Overall, application of foundation models to industrial waste classification holds very promising results.

1. Introduction

The imperative to recycle and efficiently manage waste is driven by the need to conserve resources, save energy, and reduce greenhouse gas emissions (EPA, 2013). In the United States alone, recycling of municipal solid waste contributed to a reduction of approximately 193 million metric tons of carbon dioxide equivalent in 2018 (EPA, 2013). Despite these efforts, significant amounts of recoverable materials continue to be lost, exemplified by the $11 billion worth of packaging materials landfilled in 2010 (MacKerron, 2012). Effective recycling necessitates precise sorting of waste, a process typically conducted in Materials Recovery Facilities (MRFs). Optimizing these sorting processes requires knowing the waste stream composition at various points within the MRF. Traditional methods of stream inspection, which involve either manual sampling and sorting of waste material or qualitative visual inspections, are time-consuming and fail to provide real-time data (Cuingnet et al., 2022). The development of an automatic stream composition analysis system, leveraging computer vision technologies, could enhance the efficiency and effectiveness of waste sorting by providing real-time accurate data on waste stream composition. Such a system should be capable of being deployed across diverse geographical locations and maintain its functionality over extended periods. This necessitates advanced generalization capabilities, enabling the system to adapt to changes in waste composition, in packaging materials, and in the image acquisition systems. Such robustness is crucial for ensuring the system's long-term usability and effectiveness in a variety of operational environments.

Foundation models are characterized by their vast number of parameters and are pre–trained on extensive datasets, often using self-supervision, enabling them to be adapted to a broad spectrum of tasks with minimal task-specific tuning (Bommasani et al., 2022). Initially developed in the field of natural language processing (NLP) (Brown et al., 2020), these models have shown remarkable capabilities (Touvron et al., 2023; OpenAI, 2023; Anthropic, 2024; Team et al., 2023). A notable example includes the GPT (OpenAI, 2023; Brown et al., 2020) models used by OpenAI in ChatGPT. Similarly, foundation models have begun to make significant inroads into the field of computer vision, although they are still several orders of magnitude smaller in scale compared to their NLP counterparts. While language model LLama 3 has 70 billion parameters and was trained on 15 trillion tokens (Touvron et al., 2023), Google Research has made strides with a 22 billion parameters vision model (Dehghani et al., 2023), and the largest publicly

* Corresponding author.
E-mail address: alois.babe@veolia.com (A. Babé).

https://doi.org/10.1016/j.wasman.2025.02.032
Received 2 October 2024; Received in revised form 13 February 2025; Accepted 14 February 2025
Available online 6 March 2025

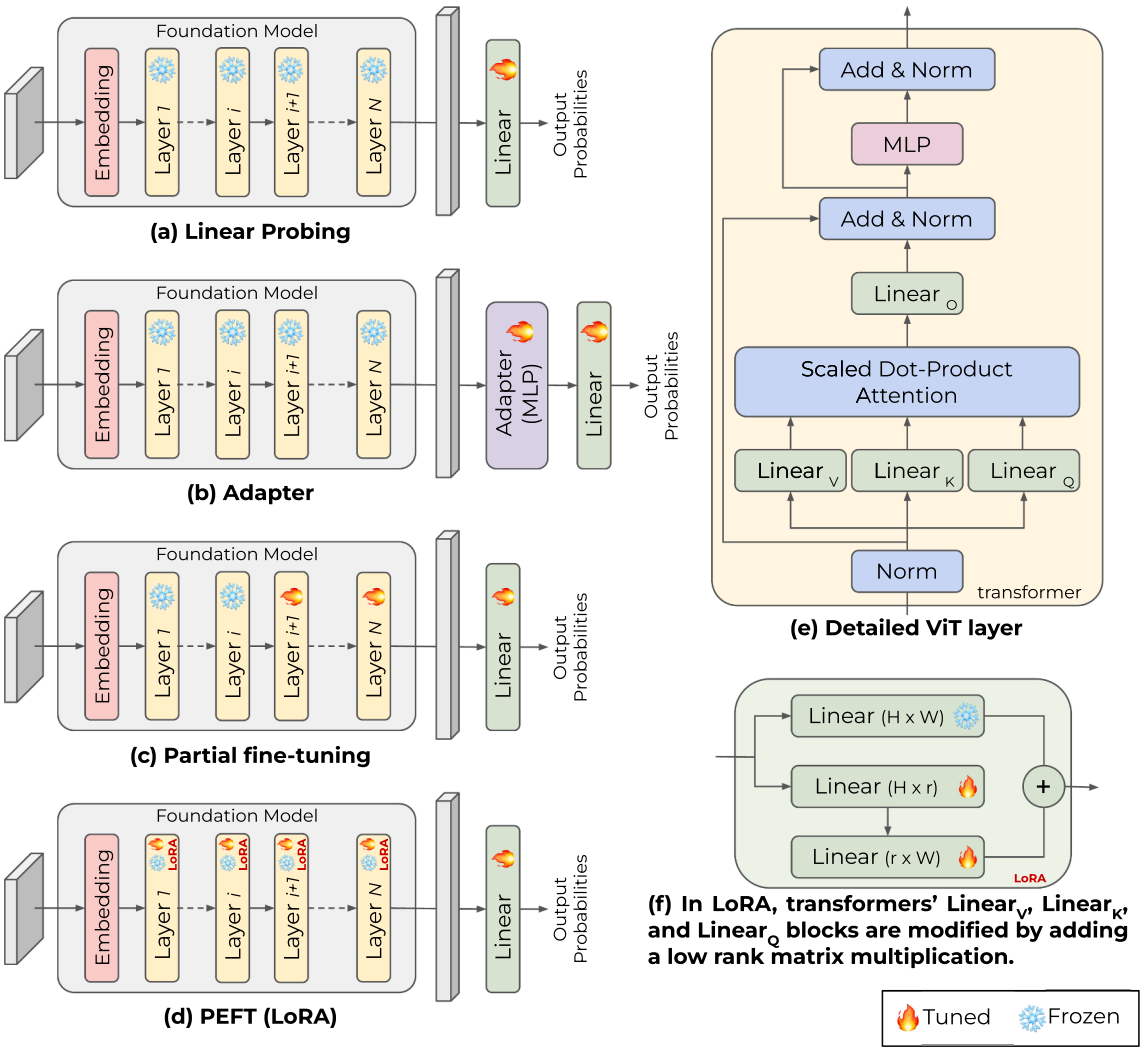

Fig. 1. Four approaches used in this study to adapt pre-trained foundation models to a new domain. (a) Linear probing: training a single fully connected layer on top of frozen model features. (b) Adapter modules: learning a more complex network over frozen features. (c) Partial fine-tuning: unfreezing and training some of the model's last layers. (d) Parameter-Efficient Fine-Tuning (PEFT): enabling small parameter adjustments across all model layers, here Low-Rank Adaptation (LoRA, Hu et al., 2022) (f) was used to adapt Vision Transformer (ViT) (e) based foundation models.

available vision foundation model, DINOv2 (Oquab et al., 2024), features 1 billion parameters and was trained on 100 billion tokens. Nevertheless these foundation models have demonstrated state-of-the-art generalization capacities, even competing with supervised models (Oquab et al., 2024; Radford et al., 2021; Kirillov et al., 2023).

Therefore, foundation models and their generalization abilities seem like good candidates for a computer vision based waste classification system. To date, those abilities have predominantly been tested on standard datasets such as ImageNet, CIFAR, and Pascal VOC 2007 (Oquab et al., 2024; Radford et al., 2021; Vogt-Lowell et al., 2023; Wang et al., 2024), which, while useful, do not necessarily reflect the complexities of real-world domain-specific applications. This research aims to contribute to bridge this gap by evaluating the generalization ability of foundation models on the task of waste classification.

To achieve this, we have tested a range of foundation models of varying sizes on several waste classification datasets. The generalization capabilities are assessed by training each model on a single dataset and then evaluating its performance across all others. Furthermore, to determine the optimal training procedures for a foundation model-based classifier, we will explore the impact of various adaptation techniques presented in Fig. 1. These include the type of head classifier used, the extent of fine-tuning with both standard methods and Parameter-Efficient Fine-Tuning (PEFT) techniques (Hu et al., 2022), and the role of data augmentation.

2. Related work

2.1. Foundation models producing all purpose features

Foundation models producing all purpose visual features build upon self-supervised learning techniques developed for models such as MoCo (He et al., 2020), SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), SwAV (Caron et al., 2020), DINO (Caron et al., 2021), MAE (He et al., 2022) or iBOT (Zhou et al., 2022). These methods laid the groundwork for learning robust representations that could be effectively transferred to various downstream tasks, but those early models were all trained on ImageNet (Russakovsky et al., 2015), which contains only 1.3 million images. The CLIP (Radford et al., 2021) model marked a milestone by being the first vision model to be trained on a very large dataset, consisting of 400 million image-text pairs, with a high number of parameters (300 millions), and can thus be considered as a foundation model (Bommasani et al., 2022). CLIP employs two encoders—one for images and another for text—and trains them using contrastive learning. During training the model is fed with batches of (image, text) pairs and is

trained to maximize the similarity of the embeddings from real (image, text) pairs while minimizing the similarity across incorrect pairings. A similar approach is adopted by ImageBind (Girdhar et al., 2023), which extends the modality to include audio, heat maps, depth, and IMU data alongside images and text. Another foundation model, DINOv2 (Oquab et al., 2024), focuses solely on images. Its training process involves feeding two different random transformations of an input image to a teacher and a student network with the same architecture. The student's parameters are learned via a combination of DINO (Caron et al., 2021) and iBOT (Zhou et al., 2022) losses and the teacher parameters are updated with an exponential moving average of the student's parameters. A centering and sharpening of the momentum teacher outputs is used to avoid model collapse; a regularizer to spread features and a short high-resolution training phase were also added. The training dataset for DINOv2, LVD-142 M, was constructed from an initial pool of 1.2 billion images sourced from the web, followed by deduplication and retrieval of images that closely align with those in 15 curated datasets (like ImageNet). It is also worth mentioning the emergence of diffusion models, which have demonstrated remarkable capabilities in the domain of image generation (Saharia et al., 2022; Rombach et al., 2022; Ramesh et al., 2022). Among these, the Latent Diffusion Mode (LDM) (Rombach et al., 2022) stands out by incorporating an autoencoder, which functions as a type of feature extractor. This autoencoder is used to facilitate the generation process by working within a compressed latent space. Recent research (Pnvr et al., 2023) has successfully explored using the LDM's autoencoder as a feature extractor.

2.2. Generalization performances of foundation models

These foundation models exhibit generalization capabilities that are now state-of-the art. The generalization capabilities of CLIP and DINOv2 have been rigorously tested from their inception, albeit primarily on ImageNet-like datasets such as Imv2, Im-A, Im-R, ImC, and Sketch (Radford et al., 2021; Oquab et al., 2024) and never for a domain-specific application. Under these conditions, Zero-shot CLIP has demonstrated a significantly higher robustness to distribution shift compared to standard ImageNet models (Radford et al., 2021). DINOv2, on the other hand, has reported state-of-the-art weakly supervised generalization performances on some of these datasets (Oquab et al., 2024). Several studies have highlighted the superior performance of foundation models in generalization. For instance, it has been claimed that high train-test similarity alone cannot account for CLIP's performance (Mayilvahanan et al., 2024), implying that the model achieves a certain level of robustness. Similarly, it has been demonstrated the superior performance of few-shot CLIP over a few-shot vision-only model in limited data environments that contain realistic distribution shifts (Vogt-Lowell et al., 2023) and evaluations of CLIP under natural distribution shift have shown robust performance (Wang et al., 2024). Furthermore, CLIP models have shown competitive performance in out-of-distribution (OOD) detection across commonly used benchmarks such as iNaturalist and ImageNet-O (Tu et al., 2024), generally exhibiting better factor-level robustness than other models.

However, some studies have challenged these generalization abilities. For example, the generalization of foundation models across geography has been tested, revealing significant disparities in performance linked to the geographic origin of data (Richards et al., 2024). For Imagebind, the authors tested the generalization capabilities of their model on audio and depth modalities and found that it outperforms its competitors (Girdhar et al., 2023).

2.3. Foundation models for real-world domain-specific applications

The adaptation of foundation models to real-world domain-specific applications has been a growing research subject. For instance, Chen et al. (2023) have adapted foundation models to plant phenotyping, providing a comprehensive discussion on the optimal use-cases of different foundation models and tuning methods. In another study, Baharoon et al. (2023) conducted a thorough evaluation of DINOv2 across various scenarios for multiple radiology modalities. The study demonstrated DINOv2's ability to either outperform or compete with supervised methods trained end-to-end.

Furthermore, Zhang et al. (2024) benchmarked state-of-the-art vision models, including ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), DINOv2 (Oquab et al., 2024), and ConvNeXt (Liu et al., 2022), for microscopic material analysis. The study found that DINOv2 exhibited the most stable performance, thereby demonstrating its capacity to generalize to rarely seen domains. Lastly, Huix et al. (2024) applied foundation models to medical image classification, where DINOv2 was the only foundation model that consistently outperformed the standard practice of learning a standard model pretrained on ImageNet. All those studies adapt foundation models to their specific task, but do not test the generalization abilities of adapted models, to the best of our knowledge such a study has never been carried out.

2.4. Computer vision for waste classification

The literature on computer vision applications in waste management spans three primary research domains: recyclable material identification, trash pollution detection, and solid waste classification (Wu et al., 2023). Within these domains, the tasks predominantly focus on classification, object detection, and image segmentation (Wu et al., 2023). Convolutional Neural Networks (CNNs) have been the dominant approach for tackling these tasks, with various architectures (Aral et al., 2018; Bircanoğlu et al., 2018) such as ResNet, MobileNet, DenseNet (Mao et al., 2021) and EfficientNet (Malik et al., 2022) being extensively utilized. Recent studies have also explored the potential of hybrid models (Ahmad et al., 2020), vision transformers (Alrayes et al., 2023), and advanced techniques like feature fusion and attention mechanisms to enhance model performance (Deng et al., 2021; Zhang et al., 2021). In the specific areas of detection and segmentation, object detectors such as YOLO (Demetriou et al., 2023; Singh et al., 2024), Faster R-CNN (Karbasi et al., 2018; Cuingnet et al., 2022), and Mask R-CNN (Koskinopoulou et al., 2021) have been prominently featured in recent research.

Despite these advancements, the field faces significant challenges

Table 1
Models used in this study, with their number of parameters, the dimension of their output features, and the average inference time for one image on a 16 GB NVIDIA Tesla P100.

Model	Number of parameters	Features dimension	Inference time (ms)
DINO ViT-S/16	21,665,664	384	8.07
DINO ViT-S/8	21,670,272	384	22.69
DINO ViT-B/16	85,798,656	768	15.16
DINO ViT-B/8	85,807,872	768	59.81
DINOv2 ViT-S/14	22,056,576	384	10.02
DINOv2 ViT-B/14	86,580,480	768	19.00
DINOv2 ViT-L/14	304,368,640	1024	53.21
DINOv2 ViT-g/14	1,136,480,768	1536	203.91
CLIP-RN50	38,316,896	1024	8.20
CLIP-RN101	56,259,936	512	8.91
CLIP-ViT-B/16	86,192,640	512	8.52
CLIP-ViT-B/32	87,849,216	512	8.30
CLIP-RN50x4	87,137,080	640	11.10
CLIP-RN50x16	167,328,912	768	19.06
CLIP-ViT-L/14	303,966,208	768	16.20
CLIP-RN50x64	420,380,352	1024	43.95
ImageBind	629,680,640	1024	67.59
Stable Diffusion	34,163,592	3136	27.11
ConvNeXt-large	197,767,336	1536	27.23
ResNet-152	60,192,808	2048	13.44
ViT L/14	304,326,632	1280	209.83

(a) Research hall (Hall)
4600 objects, 24 classes

(b) Aluminum stream (Alu)
6323 objects, 5 classes

(c) News & Pams stream (N&P)
7134 objects, 7 classes

(d) Zero Waste (ZW)
18000 objects, 6 classes

(e) TrashNet (TN)
1889 objects, 4 classes

Fig. 2. Datasets used in this study, for each dataset an example instance is provided, along with the number of individual pieces of waste and the number of waste classes in the dataset.

related to the suitability and specificity of datasets. Current datasets often lack the granularity (e.g., specific types of plastics like PET, HDPE, PP, PS) needed for municipal solid waste sorting in Material Recovery Facilities (MRFs), with many featuring overly broad or irrelevant class labels (Wu et al., 2023). Notably, there are only two public datasets designed for municipal waste sorting on conveyor belts—the ZeroWaste (Bashkirova et al., 2022) and WaRP (Yudin et al., 2024) datasets. The need for a robust system has been demonstrated by (Cuingnet et al., 2022): a drop in performance was noted when the model was tested on the same training waste stream but six months later. Furthermore, to the best of our knowledge, foundation models have not yet been explored within the framework of computer vision systems applied to waste management.

3. Material and methods

3.1. Foundation models

In this study we compared four families of foundation models DINOv2, CLIP, ImageBind and Stable Diffusion.

DINOv2 (Oquab et al., 2024) was pre-trained on LVD-142 M, a 142 million images dataset. In this work we will use the original version ViT-g/14, pre-trained from scratch, and smaller version obtained with distillation, keeping roughly the same pre-training procedure but using the bigger model as the teacher network (Oquab et al., 2024). For comparison we will also use different versions of the original DINO (Caron et al., 2021) model that were all pre-trained from scratch on ImageNet (Russakovsky et al., 2015).

CLIP (Radford et al., 2021) was pre-trained on a private dataset of 400 million (image, text) pairs collected from publicly available sources on the Internet. The different model versions consist of five ResNet models and three ViT models. For the ResNet models, RN50x4, RN50x16, and RN50x64 are three models scaled using the EfficientNet methodology (Tan and Le, 2019), utilizing approximately 4, 16, and 64 times the compute of a ResNet-50, respectively.

ImageBind (Girdhar et al., 2023) learned a joint embedding across six diverse modalities: images, text, audio, depth, thermal, and IMU data. Initially, the model employs aligned text and image encoders pre-trained on the LAION-2B dataset, a subset of LAION-5B (Schuhmann et al., 2022) containing only English image-text pairs. To integrate additional modalities, separate encoders for audio, depth, thermal, and IMU data were pre-trained while the initial image and text encoders remained frozen.

Stable Diffusion (Rombach et al., 2022) is a latent diffusion model trained on a subset of LAION-5B (Schuhmann et al., 2022) and filtered for unsafe content. Here, only the encoder to the latent diffusion space will be used.

Traditional models ResNet, ViT and ConvNeXt were also used as a baseline. ResNet (He et al., 2016) introduced residual connections, enabling the training of networks with an unprecedented number of layers, reaching up to 1000. ViT (Dosovitskiy et al., 2021) was developed as an alternative to convolutional neural networks (CNNs), while ConvNeXt (Liu et al., 2022) is a convolutional model inspired by transformers. These models have demonstrated state-of-the-art performance on image recognition benchmarks at the time of their publication. All three models are pre-trained on ImageNet. We reported in Table 1 all the models that were studied along with their number of parameters and the dimension of their output features space.

3.2. Datasets

This study utilized five datasets, presented in Fig. 2, encompassing a

Table 2
Accuracies of various frozen foundation models using K-NN and linear classifiers across multiple datasets. The classifiers were added on top of the foundation models and trained exclusively on the Research Hall dataset, with the foundation models' parameters kept frozen. Evaluation was conducted on the test sets of all five datasets.

	K-NN					Linear				
Model	**Hall**	**Alu**	**N&P**	**ZW**	**TN**	**Hall**	**Alu**	**N&P**	**ZW**	**TN**
DINOv2 ViT-g14	**70.9**	78.0	48.5	20.5	**88.3**	**81.7**	80.2	52.5	33.5	87.2
ImageBind	68.9	**78.8**	49.8	21.8	83.0	80.4	**82.0**	54.7	**55.3**	**91.0**
CLIP ViT-L14	68.3	70.6	50.5	14.7	85.1	80.0	77.5	**56.2**	25.5	88.8
ResNet 152	64.1	71.0	**51.5**	15.0	68.6	72.8	76.3	53.6	37.7	70.7
ConvNeXt large	61.3	61.1	48.7	17.0	68.6	77.8	77.1	51.2	31.6	74.5
ViT L/16	60.0	61.6	48.0	**28.4**	62.8	77.2	78.2	49.9	22.0	75.5
Stable Diffusion	19.1	25.5	34.8	11.4	20.2	24.3	27.8	29.2	19.6	34.0

diverse range of waste types, locations, and time periods. Three datasets are provided by us, capturing real-world waste streams in material recovery facilities (MRFs). Two of them includes respectively images from a News and Pams stream in France (N&P) and an aluminum (Alu) stream in England, captured by cameras and lighting systems placed over conveyor belts. The third dataset (Hall) features isolated waste items from MRFs individually placed on a conveyor belt in a research hall, resulting in clean backgrounds. Additionally, two publicly available datasets were leveraged: ZeroWaste (ZW) (Bashkirova et al., 2022), containing images of waste on conveyor belts in a paper stream in Massachusetts, and TrashNet (TN) (Yang and Thung, 2016), a dataset of manually collected recycled objects around Stanford with isolated waste against a white background. We used a modified version of the TrashNet dataset, removing 2 waste classes than are not found in MRFs (*trash* and *glass*) in order to be able to perform the label conversion we describe in the next paragraph. As we focus on classification, for the Hall, Alu, N&P and ZeroWaste datasets, we consider that a data point is defined as a bounding box of an object (piece of trash). These datasets introduce domain gaps due to variations in waste types, locations, and image capture methods. The detailed distribution of classes for each dataset can be found in the supplementary material.

Models were trained on the research hall dataset, which has the most granular class labels. Labels from other datasets can be viewed as unions of these detailed labels. Hence, the predicted probability of a class is obtained by summing all the probabilities of the corresponding fine-grained research hall dataset classes.

3.3. Experimental protocol

This study aims to evaluate the generalization abilities of foundation models for real-world domain-specific computer vision tasks, specifically waste classification. To rigorously test generalization capabilities, we adopted a cross-dataset evaluation approach, training models on one dataset (the research hall dataset) and testing them on all remaining datasets. This allowed us to assess the models' ability to learn classification features that are robust to distribution shifts across diverse datasets, due to various geographies, time periods, and acquisition processes.

We also aimed at founding the optimal training pipeline for a foundation model-based classifier. The standard approach to extract features from the input data using foundation models is to use them "as is", without further fine-tuning. These features are then used to train a separate classifier head for the specific task at hand (Fig. 1a). We evaluated the foundation models we presented in Section 3.1 using this direct pipeline. Additionally, we explored potential modifications to enhance the pipeline's performance. One approach involved incorporating an adapter network between the foundation model and the classifier head (as depicted in Fig. 1b), which we achieved by adding layers to the classifier head. Another method we explored was standard fine-tuning (as shown in Fig. 1c), and we also examined Parameter

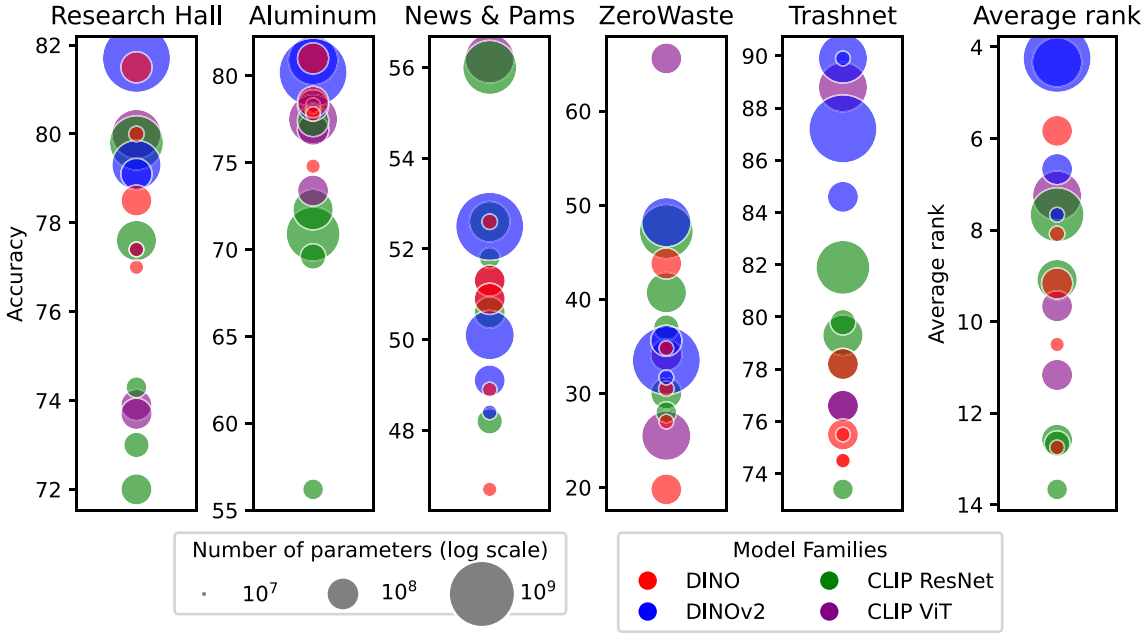

Fig. 3. Linear probing accuracy of pre-trained DINO and CLIP models of various sizes. Models were linearly probed on the Research Hall dataset and evaluated on test sets of five datasets. The last column shows the average rank of each model over all datasets. Exact numerical values are presented in Supplementary Table 6.

Table 3
Accuracy of various frozen foundation models with an added MLP classifiers with 1,2 and 3 layers on multiple datasets. Models were trained on the research Hall dataset and evaluated on the test sets of all five datasets. All foundation model parameters were kept frozen during training, with only the additional classifier being fine-tuned.

Model	MLP layers	Hall	Alu	N&P	ZW	TN
DINOv2 ViT-S/14	1	**77.4**	78.3	48.4	31.7	89.9
	2	**77.4**	**79.7**	48.7	**40.6**	**90.4**
	3	75.7	79.3	**49.5**	35.2	87.8
DINOv2 ViT-g/14	1	81.7	**80.2**	52.5	33.5	**87.2**
	2	82.2	79.3	**53.2**	33.5	86.2
	3	**82.4**	79.4	51.9	**38.0**	**87.2**
CLIP ViT-B/16	1	**73.9**	**73.4**	**51.3**	65.6	**76.6**
	2	72.0	68.5	51.1	**66.6**	73.9
	3	72.0	60.6	49.2	64.8	70.2
CLIP ViT-L/14	1	**80.0**	77.5	**56.2**	25.5	**88.8**
	2	78.3	78.2	54.8	**27.4**	88.3
	3	77.2	**78.6**	52.3	35.0	81.9
ImageBind	1	80.4	82.0	54.7	55.3	91.0
	2	**80.7**	**82.8**	**57.4**	**60.3**	**92.6**
	3	77.4	80.2	54.8	56.0	89.4

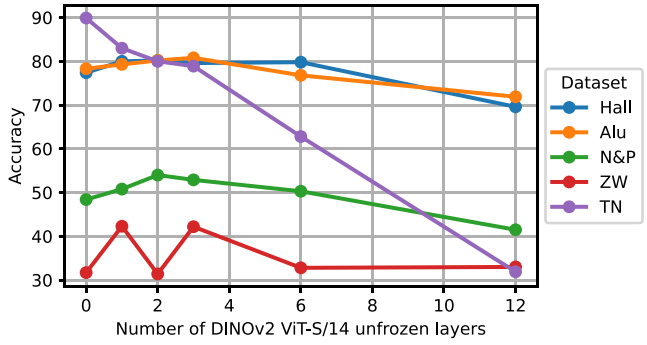

Fig. 4. Accuracy of DINOv2 ViT-S/14 with a linear classifier when fine-tuning some of the last layers of the model. Models were trained on the Research Hall dataset and evaluated on the test sets of all five datasets. Exact numerical values are presented in Supplementary Table 7.

Efficient Fine-tuning (as illustrated in Fig. 1d). We also assessed the impact of data augmentation techniques on the overall efficacy of the pipeline.

While our methodology primarily focuses on testing the generalization ability of models, this makes direct comparison with the state-of-the-art on the public TrashNet dataset challenging, as existing studies directly train their models on TrashNet. To facilitate a more direct comparison, we also perform a regular training experiment on the TrashNet dataset. Here, we extract features using a foundation model and then train a two-layer MLP classifier to map these features to the desired classes.

Foundation model. The influence of foundation model selection on feature extraction performance is still unclear. This study assessed the effectiveness of four different types of foundation models. It also explored how the pre-training dataset of the foundation model affects its performance and examined the effect of model size on performance. While smaller models are less computationally demanding and quicker to train, larger models may identify subtler features and achieve higher classification accuracy. In order to evaluate the quality of the features extracted by the foundation model, we utilized a k Nearest Neighbors (k-NN) classifier. This approach aligns with established protocols from previous studies (Caron et al., 2020, 2021; Zhou et al., 2022; Oquab et al., 2024) that have employed k-NN to assess the quality of the features computed by a model. The intuition behind the use of k-NN is that a key indicator of effective feature extraction is the ability of the resulting feature space to reflect semantic similarity. Ideally, features extracted from semantically similar images should cluster closely together, while features from dissimilar images should be well-separated. This property can be evaluated using a k-Nearest Neighbors (k-NN) classifier. We evaluated the performance of the k-NN classifier for different values of k (5, 10, 20, 40) on the research hall validation set. The k-NN model with the best performance on the validation set is then used to evaluate the model performances on all other datasets.

Classifier head. To investigate the impact of classifier size on performance, we compared the performance of three different classifier architectures: a single linear layer, a two-layer Multi-Layer Perceptron (MLP), and a three-layer MLP. For each architecture, we conducted a hyperparameter search using three learning rates (0.001, 0.0003, 0.0001) with AdamW optimizer (Loshchilov and Hutter, 2019) (weight

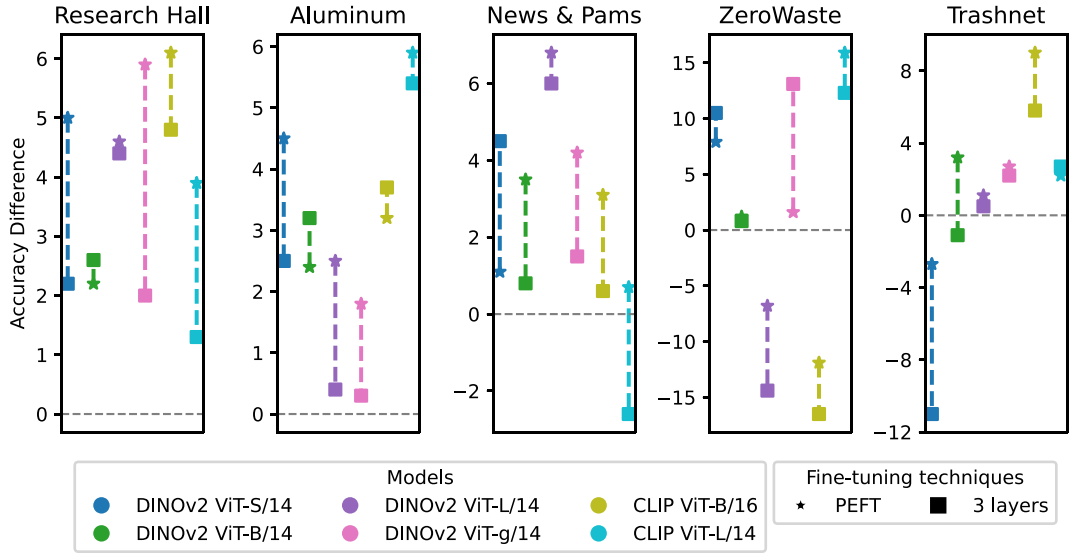

Fig. 5. Accuracy difference between fine-tuned and non-fine-tuned foundation models with an added linear layer. Models were trained on the Research Hall dataset and evaluated on the test sets of all five datasets. Fine-tuning was performed using standard methods (training the last 3 layers) or LoRA, a PEFT technique. Exact numerical values are presented in Supplementary Table 8.

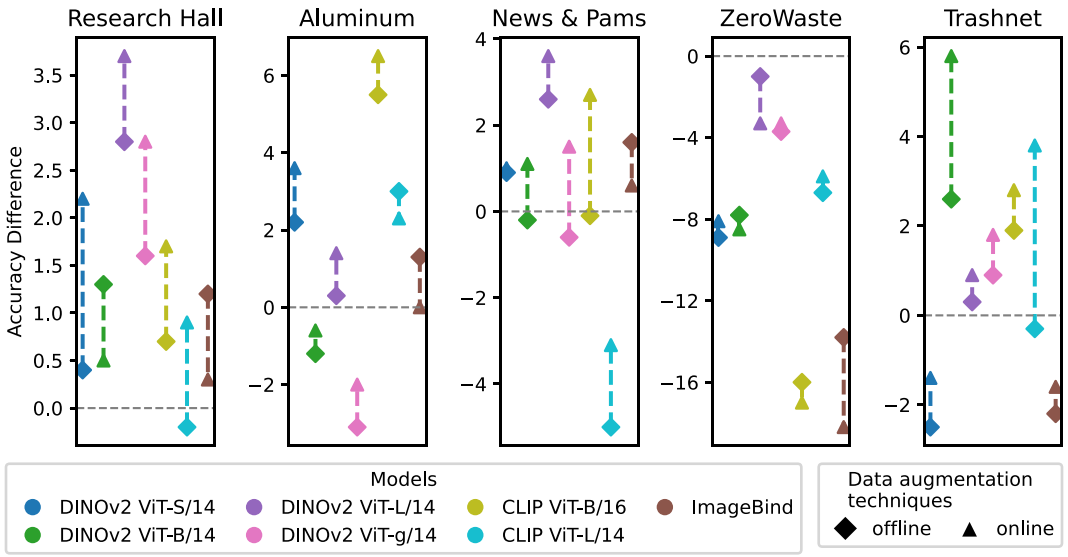

Fig. 6. Accuracy difference between foundation models trained with and without data augmentation. Models were linearly probed on the Research Hall dataset and evaluated on the test sets of all five datasets. Data augmentation was performed using offline augmentation (each image is augmented 10 times prior to training) and online augmentation (each image is randomly augmented at each epoch).

decay 0.0001, (beta1, beta2) = (0.9, 0.999)) and SGD optimizer (Sutskever et al., 2013) (weight decay 0.001, momentum 0.95). The best hyperparameters for each architecture were selected based on the performance of the linear layer on the research hall validation set. For a given foundation model the same hyperparameters were used for all classifiers.

Fine-tuning. While the extracted features are intended to be directly usable, we investigated the potential benefits of fine-tuning the foundation model for further performance improvement. We compared two fine-tuning approaches: standard fine-tuning and LoRA (Hu et al., 2022), a Parameter-Efficient Fine-Tuning (PEFT) (Lialin et al., 2023) technique, to see if fine-tuning improves feature quality and classification accuracy. Standard fine-tuning involves unfreezing the last layers of the foundation model and training a linear classifier on top of these layers, as shown in Fig. 1c. In contrast, PEFT focuses on training only a small set of parameters, either by modifying a subset of the existing parameters or by introducing new parameters, thereby significantly reducing the computational cost. LoRa is a specific PEFT technique based on the hypothesis that the modifications required during fine-tuning often have a low "intrinsic rank." LoRa exploits this by learning a rank decomposition of certain weight matrices, as depicted in Fig. 1d, allowing for minimal parameter updates. Specifically, for a projection matrix W in the self-attention module of size (d,k), it is replaced with $W + BA$, where B is of size (d,r) and A is of size (r,k), with r chosen such that $r \ll \min(d,k)$. Hence, the resulting number of trained parameters is $r(d + k)$ instead of dk. Both A and B are learned during fine-tuning, while the original W remains frozen. We employed AdamW for 100 epochs with a learning rate $lr = 0.0001$ with linear warmup from $0.1lr$ for 5 epochs and linear decay to $0.1lr$ for the last 45 epochs, $(\beta_1, \beta_2) = (0.9, 0.999)$ and a batch size of 8.

Data Augmentation. Given the robustness of foundation models, data augmentation may not be necessary for achieving optimal performance. This is particularly relevant for DINO models, which are specifically pre-trained to learn representations that are invariant to data augmentations. To investigate this hypothesis, we compared the performance of models trained with and without data augmentation, aiming to determine whether it provides any additional benefit for foundation model-based feature extraction. We employed a standard augmentation pipeline that includes scaling and cropping images to 256x256, followed by random 224x224 cropping, rotations, flips, and color jittering (using the ColorJitter class from torchvision.transforms with parameters: contrast = 0.5, saturation = 0.5, hue = 0.05). We kept the foundation model frozen and used a linear classifier. We performed both offline augmentation, where each image is augmented 10 times before training, and online augmentation, where a new augmented version of each image is generated at each epoch. For each model, we reused the optimal hyperparameters identified during the linear layer evaluation.

Table 4
Classification accuracies on the TrashNet dataset. Each model was trained on the TrashNet training set and evaluated on its test set. For foundation models, linear probing was employed, training only an additional linear classifier. State-of-the-art results are included for comparison.

Model	Accuracy	Model	Accuracy
CLIP RN50	0.841	DINOv2 ViTS14	0.912
CLIP ViTB16	0.892	DINOv2 ViTB14	0.940
CLIP RN50x16	0.884	DINOv2 ViTL14	0.940
CLIP ViTL14	0.932	DINOv2 ViTg14	0.960
CLIP RN50x64	0.912	ImageBind	0.948
(Aral et al., 2018)	0.950	(Mao et al., 2021)	**0.996**
(Bircanoğlu et al., 2018)	0.950	(Alrayes et al., 2023)	0.958

4. Results

4.1. Comparison of foundation models

Models ability to produce discriminative features is estimated through the accuracy of a k-NN classifier. Results are presented in Table 2. DINOv2 and ImageBind emerged as the top performers on three datasets, while baseline models ResNet and ViT L/16 took the lead on two datasets. Interestingly, ResNet outperformed CLIP on three datasets, demonstrating its competitive performance despite being a non-foundation model. While CLIP fell behind ResNet in terms of accuracy on these datasets, the difference remained within a margin of 0.5 points. However, on datasets where CLIP surpassed ResNet, the accuracy gap could be as significant as 16 points. Furthermore, we observed that foundation models exhibited a notable advantage over baseline models on the TrashNet dataset, which significantly differs from the training research hall dataset (manually collected waste with uniform white background, not in the context of MRF). Stable Diffusion consistently ranked last across all datasets, indicating its limitations in capturing

relevant features for image classification.

Continuing the analysis of foundation models with linear layer classifiers, similar trends can be observed on Table 2. Stable Diffusion consistently exhibited the poorest performance, falling behind all other models by a significant margin. Consequently, it was excluded from further analysis in this study. Interestingly, it can be observed that on the News and Pams datasets DINOv2 was outperformed by standard model ResNet. Similarly, on the ZeroWaste dataset, both DINOv2 and CLIP were outperformed by ResNet. However, on each dataset, the best performing model was always a foundation model. Notably, ImageBind consistently ranked among the top models, achieving the best performance on three datasets and never falling more than 1.5 accuracy points behind the best model on any dataset. Furthermore, the notable advantage of foundation models over baseline models on the TrashNet dataset is still present.

Significant variations in accuracy depending on the dataset can be observed in Table 2. These differences can be attributed to the unique characteristics of each dataset, which are of several types. Firstly, the number and granularity of classes vary. For example, the News&Pams dataset includes several types of paper, while the TrashNet and ZeroWaste datasets only have a single paper class. Additionally, the datasets corresponding to real waste streams (Aluminum, News&Pams, and ZeroWaste) exhibit class imbalance, with more instances of the main waste types in each stream. This imbalance can impact performance, as good detection of the main classes leads to better overall performance. Both the News&Pams and ZeroWaste datasets contain dense scenes, where object bounding boxes may contain other objects from different classes. This can potentially cause disturbances for the classifier and reduce performance. Furthermore, the images from the ZeroWaste dataset exhibit more blur compared to the other datasets. Additionally, it is worth noting that the TrashNet dataset is unique in that its waste items are not presented on a conveyor belt but rather isolated on a clear background.

All those differences prevent direct performance comparison between datasets. Hence direct comparisons can only be made between models for each dataset. Those comparisons reveal that foundation models shows superior performance, indicating their robustness in handling label distribution shift, cluttered background, and blur.

Inference time for the model that are used in this study are reported in Table 1, the times shown correspond to the average inference time for one image on a 16 GB NVIDIA Tesla P100. It can be noted that, with this kind of hardware, not all models are suitable for real-time use, especially on dense streams, which may involve classifying dozens of waste pieces per second. Nevertheless, there are strategies to mitigate this issue, such as reducing the size of the model. For instance, the ViT-L/14 of CLIP and DINOv2 share the same architecture, but CLIP employs a 16-bit floating point instead of a 32-bit floating point, which makes it three times faster.

Models size. To investigate the influence of model size on performance, linear classifiers were trained on the features extracted by DINO and CLIP (Fig. 3) models. Interestingly, a consistent trend can be observed across both model families: larger models generally achieved better performance. Among the DINO models, DINOv2 ViT-g/14 and ViT-L/14, the two largest models, consistently ranked among the top performers, either achieving the best performance or falling within 0.1 accuracy points of the best model across the five datasets. When computing the average rank of DINO models over all datasets, DINOv2 ViT-L/14, ViT-g/14, and DINO ViT-B/8 emerged as the top performers. Notably, there was no clear superiority of DINOv2 models over their DINO counterparts of the same size. This trend of larger models achieving better performance was also evident with CLIP models. The two largest models, ViT-L/14 and RN50x64, with comparable sizes, achieved the best performance. However, ViT-L/14 exhibited a performance drop on the ZeroWaste dataset compared to other models. Furthermore, comparing ViT-L/14 with RN50x64 and ViT-B with RN50x4 suggests that ResNet models tend to perform worse than ViT models of the same size.

Pre-training dataset. Direct evaluation of the impact of the pre-training dataset on model performance is limited due to variations in model size and architecture. Yet it can be noted that the ImageBind model, despite not having the highest number of parameters, achieved the best performance in linear evaluation. A more controlled comparison can be conducted between two models with similar architectures and parameter counts: ConvNeXt and CLIP RN50x16, both convolutional neural networks with approximately 197 and 167 million parameters, respectively. The CLIP model, pre-trained on a dataset more than 300 times larger than that of ConvNeXt, demonstrated overall superior performance.

4.2. Evaluation of classifier heads

To further explore the factors influencing model performance, the impact of classifier head size on the performance of DINO, CLIP and ImageBind models (Table 3) was studied. For smaller models, such as DINOv2 ViT-S/14, adding a second layer to the classifier head resulted in improved performance. However, for larger models like CLIP ViT-B/16, a single-layer classifier head achieved the best results. The effect of classifier size on CLIP ViT-L/14 was less clear, with no significant difference observed between different head sizes. ImageBind consistently benefited from a two-layer classifier head across all datasets. Notably, only the largest DINOv2 model, ViT-g/14, showed improvement with a three-layer classifier head. This last finding aside, based on the observed results, it is difficult to definitively explain the relationship between model size and optimal classifier head size.

4.3. Impact of fine-tuning techniques

To assess the impact of fine-tuning foundation models on performance, we experimented with different fine-tuning strategies. Our findings in Fig. 5 revealed that both standard fine-tuning of the last 3 layers and PEFT consistently outperformed keeping all model layers frozen on almost all models and datasets. For DINOv2 ViT-S/14, Fig. 4 shows that the number of unfrozen layers had an impact on performance. Unfreezing more than 3 layers did not result in any further gains, and fine-tuning all layers even led to a significant drop in performance. For larger models (ViT-L and ViT-g), PEFT almost always outperformed fine-tuning only the last 3 layers. For other models, no clear superiority of one fine-tuning method over the other was observed.

4.4. Impact of data augmentation

To evaluate the impact of data augmentation on model performance, both offline and online augmentation techniques were experimented. Fig. 6 analysis revealed that the effect of data augmentation was limited and varied depending on the specific augmentation method and model architecture. Indeed, even if for the test set of the research hall dataset we observed that both offline and online data augmentation contributed to improved performance, when considering all other datasets used in our experiments, data augmentation techniques did not systematically produce additional enhancements in a model's ability to generalize. Notably, for both offline and online augmentation, the size of the model appears to influence its sensitivity to data augmentation only for DINOv2 models. It is noticeable on the ZeroWaste dataset, for which data augmentation has a strong negative impact, except for the two biggest DINOv2 models (Vit-L/14 and ViT-g/14).

4.5. Standard TrashNet classification

The state-of-the-art results on the TrashNet dataset hover around 95 % accuracy. Both Aral et al. (2018) and (Bircanoğlu et al., 2018) reported this figure using DenseNet121, while Alrayes et al. (2023) reported a slightly higher accuracy of 95.8 % using a Vision Transformer based on Multilayer Hybrid Convolution Neural Network. Mao et al.

(2021) achieved a reported peak accuracy of 99.6 % by employing a genetic algorithm to optimize the fully-connected-layer of DenseNet121. In contrast, our results when learning on the research hall dataset fall short of these figures, ranging from 88.8 % for CLIP ViT-L/14 to 91.0 % for Imagebind (Table 2). However, when learning directly on the TrashNet dataset, our approach demonstrates competitive performance, as evidenced in Table 4. Notably, DINOv2 ViT-g/14 achieves an accuracy of 96.0 %.

This demonstrates the capability of foundation models to achieve results comparable to the state-of-the-art with minimal adaptation and hyperparameter tuning, requiring only feature extraction and training a simple MLP, making them natural candidates for waste classification tasks, even in scenarios with minimal variations in input data distributions.

5. Discussions

Foundation models demonstrated superior generalization abilities compared to standard models, particularly on the TrashNet dataset, which significantly differed from the training data distribution, where foundation models scored at least 12 accuracy points more than standard models. These results are very promising for the application to waste classification and allow to envisage the creation of robust classifiers which could be deployed in numerous contexts eliminating the need for retraining and ensuring long-term effectiveness.

Our results confirm that foundation models are more adaptable to diverse data distributions. In particular, ImageBind emerged as the top performer across all datasets, despite not being the largest model, but the one with the largest pre-training dataset. This highlights the importance of pre-training data size and quality in achieving high generalization performance. On the contrary, Stable Diffusion encoder was found to be unsuitable as simple feature extractors. This is probably due to its inherent focus on reconstruction accuracy and thus local information preservation. This contrasts with other foundation models that prioritize capturing global and semantic content, making their features more discriminative for image classification tasks.

Standard models, ResNet and ConvNeXt, exhibited similar performance levels. Notably, ConvNeXt underperformed compared to foundation models with similar parameter numbers and architecture, such as CLIP-RN50x16. This also suggests that the pre-training dataset plays a crucial role in model performance, independently of the underlying architecture or number of parameters.

Our analysis revealed a general trend of larger models achieving better performance. Interestingly, the performance improvements observed with DINOv2 over DINO models seemed to vanish when using distillation techniques. This suggests that model size is a critical parameter to benefit from larger pre-training dataset. This behaviour was described in (Kaplan et al., 2020) and (Hoffmann et al., 2022) for Large Language Models, it was found that model and dataset size should be scaled equally in order to reach compute optimally.

Furthermore, the comparison of CLIP models suggests that Vision Transformer architectures may offer better generalization capabilities compared to Convolutional Neural Networks. One possible explanation for this could be that the inductive bias of CNNs might limit their ability to learn certain types of robust discriminative features.

Our analysis revealed that the two largest models, ImageBind and ViT-g/14, benefited from using a classifier head with more than one layer. However, for other models, no clear rule emerged regarding the optimal classifier head size. This could be due to the relatively small number of classes (24) and training images (less than 5,000) in our dataset. With larger datasets and more classes, larger classifier heads might become necessary.

Fine-tuning foundation models generally led to performance improvements, suggesting that adapting the models to specific domains can be beneficial. Experiments with DINOv2 fine-tuning suggest that adaptation should be light, as fine-tuning all layers can lead to significant performance drops. PEFT appears to be particularly effective for larger models. This could be because larger models have more layers and therefore benefit from the ability to modify deeper layers sparsely. In future studies, it would be interesting to investigate whether PEFT can be further optimized by focusing on specific layers or subsets of layers, in particular this would make it possible to check whether the adaptation of the first layers of the model is really useful.

A robust classification model should ideally be insensitive to data augmentation, as it should be able to generalize well to unseen data. In fact, DINO models (Caron et al., 2021; Oquab et al., 2024) were explicitly pre-trained to produce the same features for two augmented views of the same image. This may explain why the largest DINOv2 models were less sensitive to data augmentation in our experiments. The observed random effects of data augmentation on performance suggest that simple random augmentations may not be effective for improving model robustness.

6. Conclusion

The practical application of foundation models in Material Recovery Facilities (MRFs) for waste classification highlighted their robust generalization capabilities, which are crucial for developing models that can be widely applied and remain effective over time. They therefore constitute a promising avenue to serve as a basis for such systems. Our research has demonstrated that foundation models possess superior generalization abilities compared to standard models such as ResNet and ConvNeXt. This enhanced generalization performance is primarily attributed to their large number of parameters and the extensive datasets used during their pre-training. Notably, the deployment of foundation models in practical applications does not require complex classifiers; a simple two-layer MLP suffices for feature classification. Furthermore, fine-tuning techniques, including both standard methods and PEFT, have proven beneficial, even without extensive hyperparameter optimization. PEFT, in particular, shows greater advantages over standard fine-tuning when dealing with big models (over 100 M parameters). Our study also explored the efficacy of data augmentation, which was found to impact performance inconsistently, thus rendering it less useful for our specific application. Looking ahead, further research is needed to explore the potential of foundation models in the context of object detection within waste management systems.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Veolia Secure GPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Aloïs Babé: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Rémi Cuingnet:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Mihaela Scuturici:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Serge Miguet:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wasman.2025.02.032.

Data availability

The authors do not have permission to share data.

References

Ahmad, K., Khan, K., Al-Fuqaha, A., 2020. Intelligent fusion of deep features for improved waste classification. IEEE Access 8, 96495–96504. https://doi.org/10.1109/ACCESS.2020.2995681.

Alrayes, F.S., Asiri, M.M., Maashi, M.S., Nour, M.K., Rizwanullah, M., Osman, A.E., Drar, S., Zamani, A.S., 2023. Waste classification using vision transformer based on multilayer hybrid convolution neural network. Urban Climate 49, 101483. https://doi.org/10.1016/j.uclim.2023.101483.

Anthropic, A., 2024. The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card 1.

Aral, R.A., Keskin, S.R., Kaya, M., Hacıomeroğlu, M., 2018. Classification of trashnet dataset based on deep learning models. In: 2018 IEEE International Conference on Big Data (Big Data), IEEE. pp. 2058–2062. https://doi.org/10.1109/BigData.2018.8622212.

Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Phol, K., Aljouie, A., Peng, W., 2023. Towards general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. arXiv preprint arXiv:2312.02366. https://doi.org/10.48550/arXiv.2312.02366.

Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., Ablavsky, V., Calli, B., Bargal, S.A., Saenko, K., 2022. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21147–21157. https://doi.org/10.1109/CVPR52688.2022.02047.

Bircanoğlu, C., Atay, M., Beşer, F., Genç, O., Kızrak, M.A., 2018. Recyclenet: Intelligent waste sorting using deep neural networks. In: 2018 Innovations in intelligent systems and applications (INISTA), Ieee. pp. 1–7. https://doi.org/10.1109/INISTA.2018.8466276.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P., 2022. On the opportunities and risks of foundation models. URL: https://arxiv.org/abs/2108.07258, https://doi.org/10.48550/arXiv.2108.07258, arXiv:2108.07258.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inform. Process. Sys. (NeurIPS) 33, 1877–1901.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inform. Process. Sys. 33, 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9650–9660. https://doi.org/10.1109/ICCV48922.2021.00951.

Chen, F., Giuffrida, M.V., Tsaftaris, S.A., 2023. Adapting vision foundation models for plant phenotyping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 604–613. https://doi.org/10.1109/ICCVW60793.2023.00067.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML), PMLR, pp. 1597–1607.

Cuingnet, R., Ladegaillerie, Y., Jossent, J., Maitrot, A., Chedal-Anglay, J., Richard, W., Bernard, M., Woolfenden, J., Birot, E., Chenu, D., 2022. Portik: A computer vision based solution for real-time automatic solid waste characterization – application to an aluminium stream. Waste Manage. 150, 267–279. URL: https://www.sciencedirect.com/science/article/pii/S0956053X2200280X, https://doi.org/10.1016/j.wasman.2022.05.021.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al., 2023. Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning (ICML), PMLR. pp. 7480–7512.

Demetriou, D., Mavromatidis, P., Robert, P.M., Papadopoulos, H., Petrou, M.F., Nicolaides, D., 2023. Real-time construction demolition waste detection using state-of-theart deep learning methods; single–stage vs two-stage detectors. Waste Manage. 167, 194–203. https://doi.org/10.1016/j.wasman.2023.05.039.

Deng, H., Ergu, D., Liu, F., Ma, B., Cai, Y., 2021. An embeddable algorithm for automatic garbage detection based on complex marine environment. Sensors 21, 6391. https://doi.org/10.3390/s21196391.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

EPA, 2013. Recycling basics and benefits. URL: https://www.epa.gov/recycle/recycling-basics-and-benefits.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inform. Process. Sys. (NeurIPS) 33, 21271–21284.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009. https://doi.org/10.1109/CVPR52688.2022.01553.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738. https://doi.org/10.1109/CVPR42600.2020.00975.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., et al., 2022. Training computeoptimal large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS), pp. 30016–30030.

Hu, E.J., Yelong, S., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR). URL: https://openreview.net/forum?id=nZeVKeeFYf9.

Huix, J.P., Ganeshan, A.R., Haslum, J.F., Söderberg, M., Matsoukas, C., Smith, K., 2024. Are natural domain foundation models useful for medical image classification?. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 7634–7643. https://doi.org/10.1109/WACV57701.2024.00746.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361.

Karbasi, H., Sanderson, A., Sharifi, A., Wilson, C., 2018. Robotic sorting of shredded e-waste: utilizing deep learning. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), the Steering Committee of the World Congress in Computer Science, pp. 119–123.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026. https://doi.org/10.1109/ICCV51070.2023.00371.

Koskinopoulou, M., Raptopoulos, F., Papadopoulos, G., Mavrakis, N., Maniadakis, M., 2021. Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste. IEEE Robotics & Auto. Magazine 28, 50–60. https://doi.org/10.1109/MRA.2021.3066040.

Lialin, V., Deshpande, V., Rumshisky, A., 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. arxiv 2023. arXiv preprint arXiv:2303.15647. https://doi.org/10.48550/arXiv.2303.15647.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 11976–11986. https://doi.org/10.1109/CVPR52688.2022.01167.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

MacKerron, C., 2012. Unfinished business: The case for extended producer responsibility for post-consumer packaging. Technical Report, As You Sow.

Malik, M., Sharma, S., Uddin, M., Chen, C.L., Wu, C.M., Soni, P., Chaudhary, S., 2022. Waste classification for sustainable development using image recognition with deep learning neural network models. Sustainability 14, 7222. https://doi.org/10.3390/su14127222.

Mao, W.L., Chen, W.C., Wang, C.T., Lin, Y.H., 2021. Recycling waste classification using optimized convolutional neural network. Resour. Conserv. Recy. 164, 105132. https://doi.org/10.1016/j.resconrec.2020.105132.

Mayilvahanan, P., Wiedemer, T., Rusak, E., Bethge, M., Brendel, W., 2024. Does CLIP's generalization performance mainly stem from high train-test similarity?. In: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=tnBaiidobu.

OpenAI, 2023.Gpt-4 technical report. arXiv preprint arXiv:2303.08774 https://doi.org/10.48550/arXiv.2303.08774.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR). URL: https://openreview.net/forum?id=a68SUt6zFt.

Pnvr, K., Singh, B., Ghosh, P., Siddiquie, B., Jacobs, D., 2023. Ld-znet: A latent diffusion approach for text-based image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4157–4168. https://doi.org/10.1109/ICCV51070.2023.00384.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International conference on machine learning (ICML), PMLR. pp. 8748–8763.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical Textconditional Image Generation with Clip Latents. Arxiv Preprint arXiv: 2204.06125 1, 3. https://doi.org/10.48550/arXiv.2204.06125.

Richards, M., Kirichenko, P., Bouchacourt, D., Ibrahim, M., 2024. Does progress on object recognition benchmarks improve generalization on crowdsourced, global data?. In: The Twelfth International Conference on Learning Representations (ICLR). URL: https://openreview.net/forum?id=rhaQbS3K3R.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comp. Vision 115, 211–252. https://doi.org/10.1007/s11263015-0816-y.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al., 2022. Photorealistic text-toimage diffusion models with deep language understanding. Adv. Neural Inform. Process. Sys. (NeurIPS) 35, 36479–36494.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inform. Process. Sys. (NeurIPS) 35, 25278–25294.

Singh, R.S., Pathapati, S.V.S.H., Free, M.L., Sarswat, P.K., 2024. Identification and Separation of E-Waste Components Using Modified Image Recognition Model Based on Advanced Deep Learning Tools. In: Technology Innovation for the Circular Economy; John Wiley & Sons, Ltd, 2024; pp. 115–127. https://doi.org/10.1002/9781394214297.ch10.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: International Conference on Machine Learning (ICML), PMLR, pp. 1139–1147.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML), PMLR, pp. 6105–6114.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. https://doi.org/10.48550/arXiv.2312.11805.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971.

Tu, W., Deng, W., Gedeon, T., 2024. A closer look at the robustness of contrastive language-image pre-training (clip). In: Advances in Neural Information Processing Systems (NeurIPS), p. 36.

Vogt-Lowell, K., Lee, N., Tsiligkaridis, T., Vaillant, M., 2023. Robust finetuning of vision-language models for domain generalization. In: 2023 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, pp. 1–7. https://doi.org/10.1109/HPEC58863.2023.10363450.

Wang, C., Jia, R., Liu, X., Song, D., 2024. Benchmarking zero-shot robustness of multimodal foundation models: A pilot study. arXiv preprint arXiv:2403.10499. https://doi.org/10.48550/arXiv.2403.10499.

Wu, T.W., Zhang, H., Peng, W., Lu, F., He, P.J., 2023. Applications of convolutional neural networks for intelligent waste identification and recycling: A review. Resour. Conserv. Recy. 190, 106813. https://doi.org/10.1016/j.resconrec.2022.106813.

Yang, M., Thung, G., 2016. trashnet. https://github.com/garythung/trashnet.

Yudin, D., Zakharenko, N., Smetanin, A., Filonov, R., Kichik, M., Kuznetsov, V., Larichev, D., Gudov, E., Budennyy, S., Panov, A., 2024. Hierarchical waste detection with weakly supervised segmentation in images from recycling plants. Eng. Appl. Artif. Intel. 128, 107542. https://doi.org/10.1016/j.engappai.2023.107542.

Zhang, C., Zhang, X., Tu, D., Wang, Y., 2021. Intelligent garbage detection system based on neural networks. In: International Conference on Image Processing and Intelligent Control (IPIC 2021), SPIE. pp. 252–257. https://doi.org/10.1117/12.2611334.

Zhang, J., Fang, I., Wu, H., Kaushik, A., Rodriguez, A., Zhao, H., Zhang, J., Zheng, Z., Iovita, R., Feng, C., 2024. Luwa dataset: Learning lithic use-wear analysis on microscopic images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22563–22573. https://doi.org/10.1109/CVPR52733.2024.02129.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2022. ibot: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=ydopy-e6Dg.