Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
18/2 (2025), 261-275. DOI: http://dx.doi.org/10.21609/jiki.v18i2.1489

A Hybrid Vision Transformer Model for Efficient Waste Classification

Amir Mahmud Husein[1], Baren Baruna Harahap[2]*, Tio Fulalo Simatupang[3], Karunia Syukur Baeha[4], Bintang Keitaro Sinambela[5]

Study Program in Informatics Engineering, Faculty of Science and Technology,
Universitas Prima Indonesia, Medan, Indonesia

E-mail: [1]amirmahmud@unprimdn.ac.id, [2]*barenbarunaharahap@gmail.com, [3]fulalotio@gmail.com, [4]karuniasyukur73@gmail.com, [5]bintangkeitaro22@gmail.com

Abstract

The rapid and accurate sorting of municipal waste is essential for efficient recycling and sustainable resource recovery. Most existing AI solutions focus only on four common materials (plastic, paper, metal, and glass), overlooking many other routinely encountered waste types and losing accuracy when applied to the mixed waste compositions seen in operational environments. We introduce HR-ViT, a hybrid network that combines ResNet50 residual blocks, which capture fine-grained local cues, with Vision Transformer global self-attention. Trained on a balanced six-class benchmark of about 775 images per class (plastic, paper, organic, metal, glass, batteries), HR-ViT attains 98.27 % accuracy and a macro-averaged F1-score of 0.98, outperforming a pure ViT, VT-MLH-CNN, and Garbage FusionNet by up to five percentage points in both metrics. Gains arise from selective fine-tuning of the last ten ResNet layers, lightweight ViT hyper-parameter optimisation, and targeted data augmentation that mitigates cluttered backgrounds, uneven lighting, and object deformation. These results show that hybrid attention-residual architectures provide reliable predictions under complex imaging conditions. Future work will extend the method to multi-object scenes and domain-adaptive deployment in smart-city recycling systems.

Keywords: *Deep Learning, Fine-Tuning, Hybrid Approach, ResNet50, Vision Transformers, Waste Classification*

1. Introduction

Waste management has evolved into an increasingly urgent global crisis as population and consumption continue to rise [1], [2], [3], [4], [5], [6]. Many major cities worldwide report annual increases in waste volume, triggering various environmental problems, including soil, water, and air pollution [3], [4]. Proper waste sorting can mitigate these negative impacts, as each waste category can be treated or recycled according to its unique characteristics [5], [6]. Unfortunately, manual sorting remains time-consuming, depends on the skill of the operator, and is prone to inaccuracies [7], [8], [9]. As waste management challenges become more complex, implementing artificial intelligence (AI)-based automation is becoming increasingly necessary to develop more efficient and stable waste classification systems [10], [11].

Over the last decade, Convolutional Neural Networks (CNNs) have made a significant impact in various areas, particularly waste management. For instance, Cheema et al. [12] focused on four classes (glass, metal, plastic, and trash) using VGG16, achieving 96% accuracy. Laksono et al. [13] studied HDPE bottles, PET bottles, glass, cans, cardboard, and plastic using DenseNet-201 (95.6% accuracy), while Alrayes et al. [14] tested six classes (glass, paper, cardboard, plastic, metal, trash) with VT-MLH-CNN (95.8% accuracy). Qin et al. [15] employed SVM, and Zhou et al. [16] implemented a combination of ResNet50, YOLOv5, and CNN on the same classes, attaining 83.46% and 95.88%, respectively. Li et al. [17] used CNN & Graph LSTM for cardboard, metal, glass, paper, plastic, and organic waste, reaching 97.5%. Although these results are promising, a large fraction of existing studies still employs datasets comprising fewer than eight distinct

classes or, when larger corpora are used, they present extreme class imbalance. Larger datasets with 10 [18] or 12 [19] classes do exist, but they suffer from severe class imbalance (e.g., "clothes" comprises over 5.325 samples versus fewer than 800 for several other categories), which can hamper training stability and inflate performance metrics. This disparity destabilizes training, biases evaluation metrics toward majority categories, and increases the computational burden during re-balancing. Consequently, many CNN-based approaches that excel on small, balanced collections exhibit degraded efficiency and limited generalisation as class counts rise or as dataset imbalance worsens [20]. Moreover, CNNs have a limited capacity to capture global image context because they rely on local receptive fields, making it difficult to learn long-range relationships among objects within an image [21], [22]. Some models experience overfitting when faced with diverse backgrounds and lighting conditions commonly found in waste disposal processes [23].

Recently, the Vision Transformer (ViT) has emerged as a different approach from CNNs for image classification. By leveraging the self-attention mechanism originally introduced in Natural Language Processing [24], [25], [26], ViT represents an image as a series of patches and examines global relationships among them, proving effective on large-scale datasets such as ImageNet [24], [27]. However, pure ViT often lacks the ability to capture local features (inductive bias) and requires extensive data to achieve optimal performance [28], [29]. Various efforts have been made to enhance ViT, for instance, using the Discrete Cosine Transform [30], add-embedding [31], and Token-aware Average Pooling [32]. Although ViT performs well on standard datasets (e.g., CIFAR-10 and ImageNet), studies [33] and [34] reveal that a pure ViT remains suboptimal when data are limited or highly variable, which is typical in waste classification. For example, one implementation of ViT on five waste categories attained only 92.36% accuracy [35], indicating that ViT without additional adjustments can experience performance degradation under highly diverse backgrounds and object shapes. Thus, ViT faces a significant challenge in waste classification, where visual variation is high and data are often insufficient [34], [36].

In response to the need to harness the advantages of both architectures, several researchers have started to develop hybrid models combining CNNs and ViT [37], [38]. This approach capitalizes on the strength of CNNs in extracting local features [39] and the self-attention mechanism in ViT, which excels at understanding global information [40], [41]. For example, in [42], integrating the two architectures improved the accuracy in detecting steel surface defects. Other applications in [43], [44], and [45] have also demonstrated significant potential in the medical field. Meanwhile, for waste classification, Alrayes et al. [14] tried a hybrid approach, but still encountered efficiency challenges when the dataset size increased. Overall, these findings indicate that hybrid approaches can improve ViT in capturing local details [29], although they remain underexplored for waste classification with a larger number of classes and datasets than TrashNet [46].

This study introduces a Hybrid ResNet50-Vision Transformer (HR-ViT), a hybrid architecture that integrates the residual learning of ResNet50 with ViT's self-attention mechanism. We hypothesized that this hybrid approach could enhance ViT's ability to capture local details while drawing on ResNet50's strengths in identifying diverse visual features [24], [47]. Although newer CNN architectures such as EfficientNet and DenseNet deliver high accuracy and parameter efficiency, we selected ResNet50 as our backbone for three reasons. First, its residual block design is well validated for preventing degradation in deep networks. Second, pretrained ResNet50 weights are widely accessible, enabling stable and rapid fine-tuning in a hybrid CNN–Transformer setting [48]. Third, its moderate depth strikes an effective balance between model expressiveness and computational efficiency, making it particularly suitable for resource-constrained waste-sorting systems, where both performance and operational efficiency are key. Moreover, we emphasize fine-tuning so that HR-ViT can be optimized without excessive computational overhead [49], supporting applications in waste-sorting industries or public facilities that prioritize efficiency [46]. We also integrated two different datasets, Garbage Classification (6 Classes) [19] and Garbage Classification (12 Classes) [50], totaling 4,650 images. These datasets have been utilized in several previous studies, such as by Ahmed et al. [7], who compared various methods and reported the highest accuracy of 98.95% using ResNet50, and by Quy [51], who achieved 92% accuracy using a Vision Transformer-based approach. However, our work distinguishes itself by employing a hybrid Vision Transformer-based model and combining both datasets to enhance data diversity, going beyond conventional TrashNet-based approaches [20], [52], [53], [54], [55]. We considered six waste categories: batteries, plastic, paper, metal, organic, and glass, chosen to represent the major types of waste in management systems and provide a more comprehensive and realistic dataset. This approach is important for testing the model under realistic conditions, given

that waste issues often involve nonideal data [56]. Using a more extensive dataset, we aimed for better generalization and improved handling of noise arising from variations in texture, shape, and background [57], [58], [59].

This study offers four main contributions. First, we propose HR-ViT, a hybrid model that integrates ResNet50's residual learning and ViT's self-attention mechanism to improve waste classification performance. Second, we tested HR-ViT on six waste classes (batteries, plastic, paper, metal, organic, and glass) comprising 4,650 images, expanding upon previous work that mostly addressed four to six categories without fully representing everyday waste. Third, we adopted a fine-tuning strategy to maintain model efficiency and minimize computational load, thus supporting real-world adoption. Fourth, we integrated HR-ViT into a cross-platform mobile application and real-time backend, enabling instant waste classification and facilitating global scalability. To support transparency and reproducibility, the source code for this work is available at https://github.com/barenbaruna/HR-ViT. Through this approach, we aim to pave new ground for developing smarter, more efficient waste classification systems, while also contributing to the broader literature on hybrid architectures.

2. Literature Review

2.1 Convolutional Neural Networks (CNN)

Waste management has become a critical global issue due to the continuous increase in waste volume, which impacts the environment and public health. Consequently, efforts to automate waste classification continue to develop greater efficiency and accuracy. One frequently used method is Convolutional Neural Networks (CNN), which excels at extracting local features from images. Several studies have highlighted the success of CNN in various waste classification scenarios, while also identifying certain limitations that motivate the development of new approaches.

Early research utilizing pure CNNs, for instance, Bobulski et al. [60], achieved 74% accuracy in distinguishing four types of plastic waste (PS, PP, PE-HD, PET), indicating limited model performance on complex backgrounds. Nnamoko et al. [61] found that image resolution and dataset size significantly affect model performance, with lower resolution achieving greater efficiency (80.88% vs. 76.19%) but still trailing behind more advanced architectures. Tatke et al. [62] reported the use of ResNet50 on the Garbage in Images (GINI) dataset with 95.93% accuracy, exceeding that of a simpler CNN at 82.19%. This result demonstrates that adding residual learning layers can mitigate vanishing gradient issues and improve training stability.

Subsequent studies adopted transfer learning using popular CNN architectures. Cheema et al. [12] achieved over 90% accuracy on TrashNet (four classes: glass, metal, plastic, and trash) using VGG16, whereas Laksono et al. [13] expanded to six classes (HDPE, PET, glass, cans, cardboard, and plastic) with DenseNet-201 (95.6% accuracy). These studies confirm that CNN-based approaches are relatively strong at extracting local features; however, when datasets become increasingly heterogeneous, models may face overfitting or require substantial computational resources to maintain accuracy.

Other studies have introduced specialized modules to complement CNNs. Li et al. [17] added Graph LSTM to CNN, raising classification accuracy for six classes of waste to 97.5%. Qin et al. [15] employed the lightweight MobileNetV2 architecture for efficiency, although its 83.46% accuracy lagged behind that of more complex CNN approaches. This underscores the trade-off between model size and performance, which is becoming increasingly important in edge device or real-time applications where resources are limited. Rayhan and Rifai [63] also found that DenseNet121 outperformed MobileNetV2 (95.2% vs. 92%) across 13 heterogeneous waste classes, yet adding more classes increased the model's susceptibility to overfitting.

Research focusing on background variation and noise further reinforces the notion that pure CNNs may not suffice in real-world conditions. Yuan and Liu [64] split the classification task into two streams (dual-stream CNN) prior to the final stage, achieving 98.5% accuracy on TrashNet. Yang et al. [65] incorporated a preprocessing step (for example, Canny edge detection) to address lighting disturbances, lifting accuracy to 96.77% on an in-house dataset and 93.72% on TrashNet. Although such techniques can improve performance, they typically rely on the local feature extraction characteristics of CNNs, making them vulnerable when objects overlap (occlusion) or when background clutter is present. While CNNs have become the backbone of waste image classification, with accuracies ranging from 74% to 98.5% across different studies, challenges emerge when data become more diverse, the number of classes increases, and systems must remain efficient. CNNs are also relatively limited in capturing the global context, especially in scenarios with complex backgrounds or overlapping classes (occlusions) [66]. These limitations have driven researchers to explore other architectures, such as the Vision Transformer

(ViT), which offers a self-attention mechanism for broader contextual understanding.

2.2 Vision Transformer (ViT)

The Vision Transformer (ViT) employs a self-attention mechanism originally prominent in Natural Language Processing. Its primary advantage is its ability to grasp the global context, addressing a key limitation of CNNs. In waste classification, ViT has been applied to a limited extent, yet the results demonstrate significant potential. Wu et al. [35] implemented Query2Label (Q2L) based on ViT-B/16 in the "Garbage In, Garbage Out" (GIGO) dataset, which includes 25,000 street images with four types of waste (bulkywaste, garbagebag, cardboard, litter). Using asymmetric loss and replacing the ResNet10 backbone with ViT-B/16 increased the accuracy by 4.75%. This improvement underlines the effectiveness of global attention in dealing with multi-label data in real-world settings. However, the study was confined to four waste categories, which do not fully capture the variety of urban waste.

Huang et al. [67] used ViT on TrashNet's six classes, achieving 96.98% accuracy. Their focus was on real-time inference on a cloud server, making ViT accessible from mobile devices. This finding highlights ViT's scalability for remote processing, while also noting that high computational power is essential. For edge-device applications, a pure ViT is often considered resource-intensive. Although ViT consistently outperforms conventional CNNs in terms of accuracy, its large data requirements and high processing costs pose significant challenges for deployment in low-power or real-time applications. Consequently, current ViT-based waste classifiers are typically limited to a few easily distinguishable categories and have yet to address more complex waste types.

Despite these advantages, certain waste categories remain particularly challenging for ViT-based classifiers. Electronic waste, such as circuit boards, batteries and cables, exhibits high intra-class variance due to pronounced differences in shape, size and material composition. Medical waste items, for example used gloves or infusion bottles, often present visual ambiguity, because stains, folds or partial occlusions can cause them to be mistaken for non-hazardous debris. Organic waste, including food scraps and vegetable peels, undergoes rapid changes in colour and morphology during decomposition, further complicating feature extraction. Moreover, the scarcity of large-scale, well-annotated datasets for these complex waste types requires robust data augmentation and transfer learning strategies to ensure reliable model generalization.

2.3 Hybrid CNN-ViT

To overcome the limitations of each method, a hybrid approach combining the benefits of local feature extraction (CNN) and global attention (ViT) has emerged. The main objective of this combination is to achieve higher accuracy while maintaining computational efficiency and robustness under various conditions. Liu et al. [68] introduced Garbage Classification Net (GCNet), integrating EfficientNetV2, ViT, and DenseNet for four main waste categories (recyclable, hazardous, kitchen waste, other garbage). GCNet achieved 97.54% accuracy, surpassing individual models such as DenseNet (96.40%), ViT (96.75%), and EfficientNetV2 (96.12%). The advantage of GCNet lies in its fusion of models and transfer learning, enabling the effective capture of local features (DenseNet, EfficientNet) and leveraging global attention (ViT). However, this approach requires more computational resources, posing a challenge for real-time or low-power devices. Additionally, the scope of waste classes remains limited to four broad categories, excluding other types of urban waste.

A similar approach was presented by Cai et al. [69] through CT-Net (CNN + Transformer), which reached 96.55% accuracy on the Huawei Cloud dataset. The authors highlighted robustness and scalability, but did not provide detailed computational overhead data for industrial environments with more classes. Alrayes et al. [14] tested VT-MLH-CNN on six classes (glass, paper, cardboard, plastic, metal, trash), achieving 95.8% accuracy. Although this represents more classes than those in Cai et al. [69] and Liu et al. [68], the dataset is relatively small and prone to overfitting.

Wang et al. [18] introduced Garbage FusionNet (GFN), combining ResNet, ViT, and additional modules like the Pyramid Pooling Module (PPM) and Convolutional Block Attention Module (CBAM). Tested on two datasets (TrashNet and Garbage Dataset), it attained accuracies of 94.21% and 96.54%, respectively, surpassing standalone ResNet and Swin Transformer. This study underscores the advantages of combining residual learning and global attention but also notes that each additional module increases the computational overhead.

From these studies, it is clear that waste classification research has focused on three main pillars: (1) local feature extraction through CNN, (2) global context awareness using ViT, and (3) hybrid architectures that blend both approaches. Through analysis of existing literature, three major

knowledge gaps have emerged. First, few studies have truly tested the scalability of hybrid models for larger numbers of classes (including e-waste, medical waste, or multi-fraction categories). Second, some studies have not emphasized optimizing models for deployment on limited-resource devices, even though real-world applications increasingly require on-device inference with low latency. Third, many studies focus on adding modules to boost accuracy yet give less attention to adaptive fine-tuning or continuous learning methods, both of which are pertinent for handling dynamic waste data.

Based on this review, a hybrid CNN–ViT approach has the potential to be a comprehensive solution for overcoming the challenges of local versus global feature understanding in waste classification. Although previous research has largely been limited to four classes, the dataset in this study covers six categories (batteries, plastic, paper, metal, organic, and glass). This broader scope requires a model that is more efficient and robust against waste diversity. Therefore, this study focuses on designing and evaluating a hybrid CNN–ViT architecture that emphasizes scalability, reliable performance under various data conditions, and computational optimization to be feasible for real-time or limited-resource scenarios. Thus, the proposed solution is expected not only to excel in accuracy, but also to be practically relevant for industries and public facilities that require modern and adaptive waste classification systems.

3. Method

3.1 HR-ViT Model

This study proposes HR-ViT, a hybrid architecture that combines the Vision Transformer (ViT) and ResNet50 to enhance image classification performance (see Figure 1).

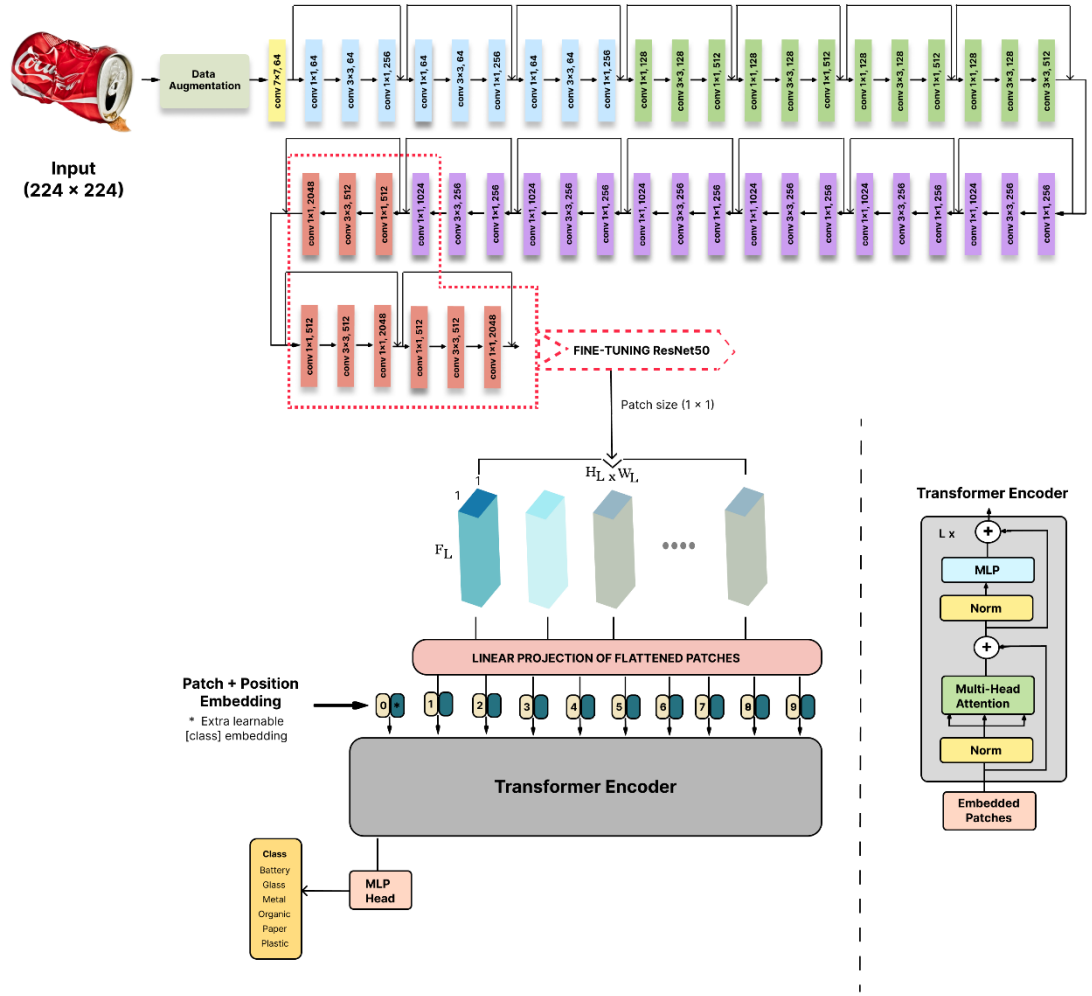


Figure 1. HR-ViT architecture.

In the ResNet50 component, we fine-tuned the last ten layers to learn domain-specific local characteristics without modifying the pretrained weights in the earlier layers. To determine the optimal depth of fine-tuning, we conducted an experiment by unfreezing 10, 15, and 20 layers of the ResNet50 backbone. After these experiments, we found that unfreezing the last ten layers offered the best balance between preserving foundational weights (capturing general features such as edges and textures), adapting waste-specific features, computational efficiency, and overfitting mitigation. This strategy preserves training stability while maximizing the adaptation of local features [70]. Meanwhile, ViT is used without further modifications to capture the global context through its self-attention mechanism, which partitions an image into smaller patches and analyzes their global relationships [24].

- **Feature Extraction with ResNet50**

The first step in HR-ViT is feature extraction using ResNet50, which processes an input image x of dimension $(H, W, 3)$ into a feature tensor $f_{ResNet}(x)$ with dimension (H', W', C). Through a series of residual blocks, local features such as edges, texture patterns, and basic shapes are extracted using convolutional strides and pooling, gradually reducing the spatial dimensions H and W to H' and W'. The output channel count, C, is typically 2048 in the final layer before global pooling. This feature tensor provides a high-level of representation of the image, making it an ideal input for the subsequent stage.

$$f_{ResNet}(x) = ResNet50(x) \in \mathbb{R}^{(H' \times W' \times C)} \quad (1)$$

- **Converting to Patches**

Next, the tensor $f_{ResNet}(x)$ from ResNet50 is transformed into a series of patches. At this stage, the spatial dimensions (H', W') are combined into $N = H' \times W'$, representing the number of patches. Each patch retains a channel dimension C, which is the output from ResNet50. This process reorganizes the spatial features produced by ResNet50 into a two-dimensional vector X_p, which is then ready for further processing by the Vision Transformer. Effectively, each spatial location in the image is mapped into a distinct feature vector.

$$X_p = Reshape\left(f_{ResNet}(x)\right) \in \mathbb{R}^{(N \times C)}, N = H' \times W' \quad (2)$$

- **Linear Projection of Patches into Embedding Space**

Each patch in X_p is linearly projected into a lower-dimensional embedding space D using a matrix $E \in \mathbb{R}^{(N \times D)}$. Matrix E is a trainable parameter that maps the ResNet features into the Transformer's embedding space. This projection reduces dimensionality while preserving essential information from the features, resulting in an embedding Z of dimension$(N \times D)$. This representation allows the Transformer to treat each patch as an individual input token.

$$Z = X_p\, E \in \mathbb{R}^{(N \times D)} \quad (3)$$

- **Adding the Class Token and Positional Embedding**

A class token $(x_{class} \in \mathbb{R}^{(1 \times D)})$ is prepended to the sequence of embeddings Z to capture global information from all patches. Additionally, a positional embedding $(E_{\text{pos}} \in \mathbb{R}^{((N+1) \times D)})$ is added to provide position information for each patch. Positional embedding is crucial because the self-attention mechanism in the Transformer is order-agnostic, necessitating an explicit spatial relationship signal among patches. This combined representation, Z', is then ready for processing by the Transformer Encoder.

$$Z' = [x_{class}; Z] + E_{\text{pos}} \in \mathbb{R}^{((N+1) \times D)} \quad (4)$$

- **Processing by the Transformer Encoder**

The Transformer Encoder processes Z' using multi-head self-attention and a feed-forward network. Self-attention enables the model to capture global relationships among patches, while the feed-forward network deepens the feature representation. The outcome is Z'', where the class token (Z_0'') now contains the global information necessary for classification. This step ensures that the model fully grasps the global context of the input image.

$$Z'' = TransformerEncoder\,(Z') \in \mathbb{R}^{((N+1) \times D)} \quad (5)$$

- **Final Classification**

The class token (Z_0'') from the Transformer Encoder output is used for classification. This class token representation is projected through a weight matrix W and activated via the softmax function on to yield a probability distribution over the classes. This distribution reflects the model's confidence in each target class, allowing for the final decision in image classification.

$$y = Softmax\,(W.Z_0'') \quad (6)$$

3.2 Model Parameters

Balancing computational efficiency with improved accuracy during training is the primary focus; therefore, the HR-ViT model parameters are carefully designed to optimize both modeling

capacity and training stability. Table 1 summarizes the key parameters used in this study, such as the embedding dimension, number of heads, and feed-forward network size. These parameter values were selected based on a series of preliminary experiments, including manual tuning strategies, considering hardware limitations, and the need for efficient inference.

Table 1. Model parameters.

Parameter	Value	Description
d_model	192	Embedding dimension
n_heads	6	Number of heads in Multi-Head Attention
d_ff	768	Hidden layer size in the Feed-Forward Network (FFN)
dropout_rate	0.15	Dropout rate in the Transformer Encoder
n_layers	4	Number of layers in the Transformer Encoder
mlp_head_size	192	Hidden layer size in the MLP for the classification head
patch_size	1	Patch size (1×1) from the ResNet backbone output

In some cases, values were also informed by configurations from prior state-of-the-art studies, particularly those involving vision transformers for similar classification tasks [71]. Additionally, the patch size was set to 1×1, representing the transformation of the ResNet50 output into individual tokens, thus providing a patch representation for each specific spatial location.

3.3 Dataset

This study used a six-class waste dataset containing plastic, paper, organic, metal, glass, and batteries, totaling 4,650 images. Each class had 775 images, ensuring a balanced distribution and minimizing model bias. Figure 2 shows the sample images for each category. The dataset was sourced from two repositories [19], [50], which were combined and standardized to offer more comprehensive class coverage than TrashNet [20], [52], [72].

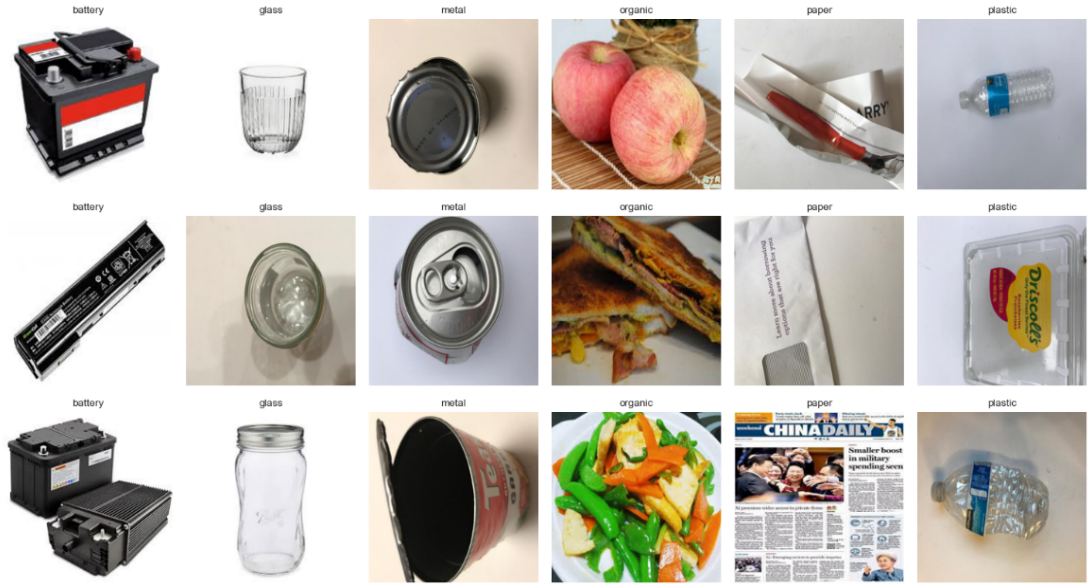


Figure 2. Samples of each class.

3.4 Data Preprocessing

All images were resized to 224×224 pixels, following the standard input dimensions of ImageNet-pretrained models, such as ResNet50. Pixel values were normalized to the range [−1,1] to stabilize the weight updates [73]. We also performed on-the-fly data augmentation (rotation, translation, shear, zoom, horizontal flip, fill mode set to "nearest") to increase the example diversity [74], [75]. Table 2 lists the augmentation parameters. This strategy is expected to improve the robustness of the model to variations in object position, orientation, and scale.

Table 2. Data augmentation parameters.

Technique	Value
Rotation	30°
Width Shift	0.3
Height Shift	0.3
Shear Transformation	0.3
Zoom	0.3
Horizontal Flip	True
Fill Mode	Nearest

After preprocessing and augmentation, the entire dataset (4,650 images) was split into training (80 %), validation (10 %), and testing (10 %) subsets using stratified sampling with random_state=42 to preserve class proportions and ensure reproducibility. The training set (3,720 images) was used for model fitting; the validation set (465 images) guided hyperparameter selection and early stopping; and the testing set (465 images) provided an unbiased estimate of final performance.

Table 3 details the per-class distribution across all subsets, confirming that each of the six waste categories remains evenly represented and that sampling bias is minimized. This approach aims to maintain the data balance while improving the generalization capability of the model prior to further training and optimization.

Table 3. Dataset split distribution.

Class	Training Set	Validation Set	Testing Set	Total
organic	620	77	78	775
metal	620	77	78	775
paper	620	77	78	775
glass	620	78	77	775
plastic	620	78	77	775
battery	620	78	77	775
Total	3.720	465	465	4.650

3.5 Model Training and Optimization

Training aims to minimize the loss function while maximizing the predictive accuracy on the validation set [76]. We chose categorical cross-entropy to handle multi-class classification [77]. For weight optimization, we used the Adam optimizer due to its stability and rapid convergence in complex architectures [78]. The initial learning rate was set to 1×10^{-5}, selected based on preliminary experiments which indicated better stability and convergence behavior at lower rates, particularly for transformer-based models.

The batch size was limited to 16 due to hardware constraints, as larger batch sizes (e.g., 32) led to memory exhaustion during training. To enhance generalization and avoid overfitting, early stopping was applied with a patience of 10 epochs, which was found effective during our internal testing to balance training duration and model performance. Furthermore, the learning rate was adaptively halved when the validation loss plateaued for 5 consecutive epochs, down to a minimum of 1×10^{-6}, allowing the model to refine weights more cautiously during the later stages of training. These values were selected based on iterative experimentation and are also commonly recommended in related transformer-based classification studies. Finally, model checkpoints were saved based on the lowest validation loss to ensure the best-performing weights were used for evaluation. Table 4 summarizes the primary training parameters used in this study.

Table 4. Training parameters.

Parameter	Value
Optimizer	Adam
Learning rate	1×10^{-5}
Loss function	Categorical Cross-Entropy
Metrics	Accuracy
Batch Size	16
Maximum Epoch	100
Early Stopping Patience	10 epoch
Learning Rate Reduction	Factor 0.5, Patience = 5, Minimum: 1×10^{-6}
Model Checkpoint Monitor	Validation Loss

3.6 Cloud-Mobile Integration

This research not only focuses on HR-ViT development but also on its large-scale deployment using a cloud–mobile architecture. Flask was chosen as the backend framework due to its flexibility in handling Python-based machine learning libraries [79], [80]. This approach facilitates the design of a REST API for server-side model inference requests.

To address the limited resources of mobile devices (computing power and storage), classification runs on an AWS-hosted server, optimizing scalability and uptime [67], [81]. To ensure cost efficiency, the backend was activated only during the study period. On the front-end, React Native supports cross-platform mobile application development, ensuring a responsive user interface compatible with various operating systems [82].

Figure 3 illustrates the system's workflow, from uploading an image via the app and invoking the Flask API for cloud-based model inference to returning the classification results to the phone. By placing computationally intensive tasks (training and inference) in the cloud, mobile devices primarily handle user interfaces and server requests. This strategy reduces user-side latency but requires a stable internet connection for real-time performance.

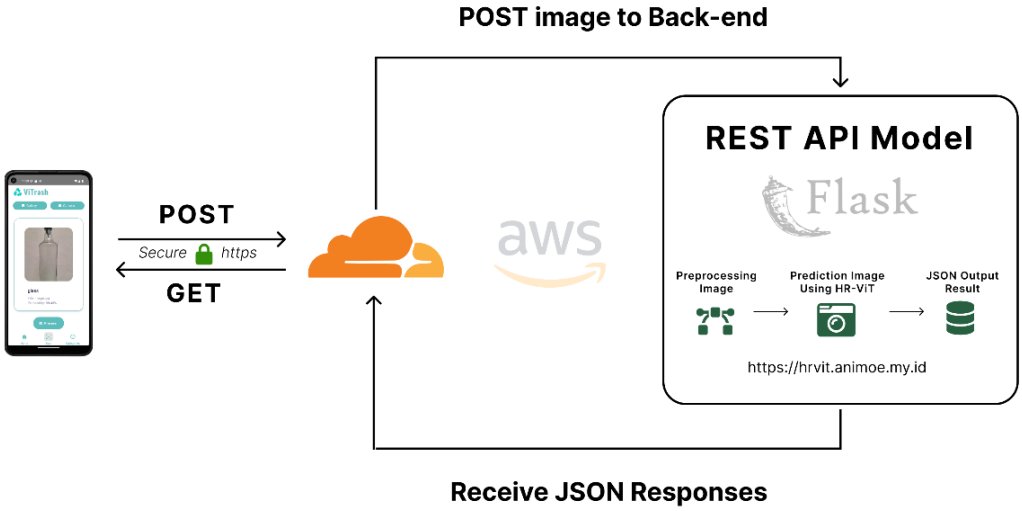


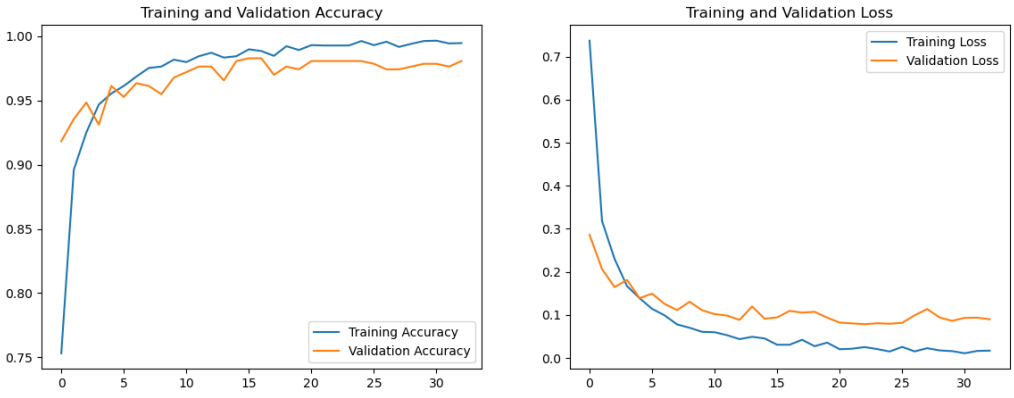
Figure 3. System architecture.

4. Result and Discussion

4.1 Model Training and Validation

This study was conducted in a Python environment under Anaconda to ensure experimental compatibility and reproducibility. The HR-ViT model was trained on an Intel Core i9-12900H with 16 GB DDR5 RAM, 512 GB NVMe SSD, and an NVIDIA GeForce RTX 3060 GPU (6 GB GDDR6). GPU acceleration is crucial because HR-ViT employs real-time data augmentation and requires intensive exploration of the network parameters [83]. TensorFlow 2.10.1 was chosen as the main framework, combining the Vision Transformer (ViT) with ResNet50 as the backbone. Keras was used modularly, and libraries such as NumPy, Pandas, and Matplotlib facilitated numerical operations, data manipulation, and result visualization.

The training exhibited rapid convergence within the first 13 epochs: at epoch 1, training and validation accuracies were 75.32 % and 91.83 % with losses of 0.737 and 0.286, respectively. By epoch 4, accuracy rose to 94.68 % (train) and 93.12 % (val) with corresponding losses of 0.167 and 0.181. Continued training reduced the training loss to 0.044 and increased accuracy to 98.71 % by epoch 13, when the minimum validation loss of 0.088 was attained (val_accuracy = 97.63 %). We applied early stopping with a patience of 10 epochs and halved the learning rate after 19 stagnant epochs. Between epochs 14–33, validation loss fluctuated modestly between 0.0785 (epoch 23) and 0.1197 (epoch 14), while validation accuracy remained within 96.56 %–98.28 %, confirming robust generalization and absence of overfitting. The loss and accuracy curves (Figure 4a–b) further substantiate the effectiveness of our training schedule.



(a) (b)

Figure 4. Training and validation accuracy (a) vs. training and validation loss (b).

4.2 Testing Results

The performance of the HR-ViT model was evaluated using 465 test samples distributed across six waste categories: battery, glass, metal, organic, paper, and plastic. As shown in Figure 5, the confusion matrix exhibits strong diagonal dominance, indicating high classification accuracy. The model achieved perfect predictions for "battery" (77/77) and "organic" (78/78). For the "glass" class, four samples were misclassified: one as "metal," one as "paper," and two as "plastic." The "metal" class included two errors, misclassified as "glass" and "plastic," while "paper" had one instance predicted as "metal." The "plastic" class showed one misclassification into "paper." These errors were isolated and did not significantly impact the overall distribution, affirming the model's ability to distinguish between classes with minimal overlap.

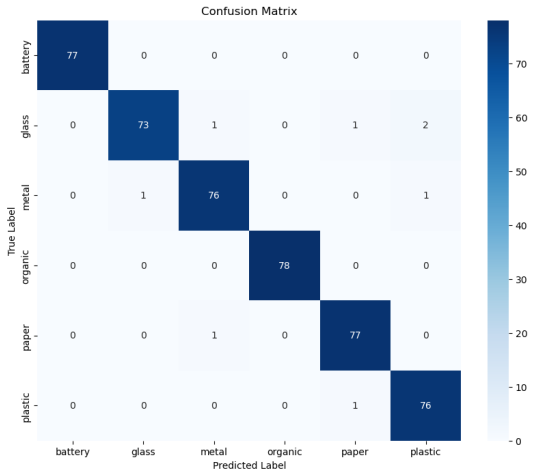


Figure 5. Confusion matrix HR-ViT.

Table 5. Classification report HR-ViT.

Class	Precision	Recall	F1-Score	Support
Battery	1.00	1.00	1.00	77
Glass	0.99	0.95	0.97	77
Metal	0.97	0.97	0.97	78
Organic	1.00	1.00	1.00	78
Paper	0.97	0.99	0.98	78
Plastic	0.96	0.99	0.97	77
Accuracy			0.98	465
Macro Avg	0.98	0.98	0.98	465
Weighted Avg	0.98	0.98	0.98	465

Quantitative evaluation results are presented in Table 5. The overall classification accuracy reached 98%, with macro-average and weighted-average precision, recall, and F1-scores of 0.98. The "battery" and "organic" classes achieved perfect scores (1.00) across all metrics. The "glass" class reported the lowest recall at 0.95 due to the aforementioned misclassifications, while "metal," "paper," and "plastic" achieved F1-scores ranging from 0.97 to 0.98. These results confirm that the proposed HR-ViT model maintains reliable and balanced classification performance across all categories, despite minor inter-class confusion.

4.3 Application Implementation Results

To test the performance of the mobile-cloud application, five real-world samples were used to ensure both diversity and model robustness. An internet connection with 50 Mbps bandwidth and a 1:1 upload-to-download ratio was maintained throughout. Users captured waste images via the in-app camera or selected them from the device gallery, then uploaded the images to a back-end server, which returned the classification results to the device. Table 6 summarizes the test data: image source, resolution, file size, and application response time.

Table 6. Characteristics of test data and application response.

No	Image Source	Image Resolution (px)	Image Size (MB)	Response Time (ms)
1	In-app camera	2448 x 3264	1,29	1500
2	In-app camera	2448 x 3264	1,21	1400
3	Gallery (reduce resolution)	1224 x 1632	411	800
4	Gallery (another source)	6120 x 8160	4,11	4000
5	Gallery (another source)	3472 x 4624	5,63	4700

The application achieved an average response time of 2,480 ms (milliseconds), with a minimum of 800 ms and a maximum of 4,700 ms. As expected, higher-resolution images incurred greater latency: the 6,120 × 8,160 px (4.11 MB) file required 4,000 ms, whereas the down-sampled 1,224 × 1,632 px (0.41 MB) image completed in 800 ms. These findings confirm that delegating inference to the cloud enables efficient classification without taxing the user device, provided network quality remains consistent. The separation of front-end UI and back-end processing further enhances scalability, although reliable connectivity remains essential to maintain low latency across deployment environments [84] [85].

By contrast, highly optimized on-device networks can perform pure inference in only a few milliseconds: for instance, EfficientFormer-L1 achieves 1.6 ms on an iPhone 12 [86], and MobileOne-S4 runs under 1 ms on the same hardware [87]. While these figures exclude network overhead, our end-to-end average of 2,480

ms aligns with reported latencies for cloud-based vision pipelines, which typically span sub-second to multi-second ranges under realistic conditions. Taken together, these results demonstrate that HR-ViT, when deployed as a mobile–cloud service, achieves a latency profile that is competitive with existing lightweight architectures while offloading all intensive computation to the cloud and thus minimizing device-side resource consumption. Figure 6 illustrates the application workflow and system responses.

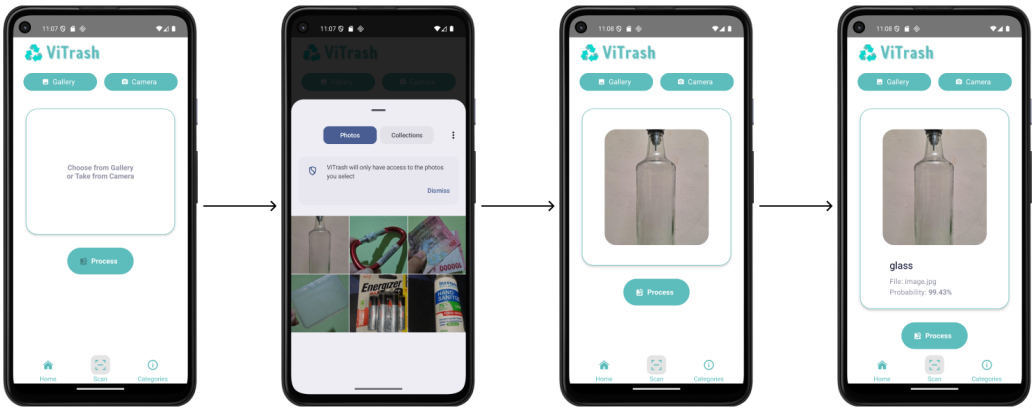


Figure 6. Implementation results.

4.4 Discussion

The HR-ViT model integrates ResNet50 and Vision Transformer (ViT), combining localized spatial feature extraction with global semantic representation. To enhance generalization, training incorporated real-time data augmentation techniques, including random rotation, scaling, flipping, and translation along with fine-tuning of the last 10 ResNet50 layers with parameters d_model = 192, n_heads = 6, d_ff = 768, and dropout_rate = 0.15. Under these settings, the final evaluation yielded a test loss of 0.0579 and a test accuracy of 0.9827, confirming the model's strong ability to generalize unseen waste images. The training curves also exhibited stable convergence with no signs of overfitting. The hybrid architecture leverages pretrained weights to minimize parameter redundancy while maximizing feature discrimination, resulting in efficient representation learning.

To assess the impact of augmentation, an ablation study was conducted using identical settings with and without it. As shown in Table 7, augmentation improved accuracy from 97% to 98%, and all key metrics (precision, recall, and F1-score) increased from 0.97 to 0.98. These gains highlight augmentation's role in mitigating overfitting and enhancing intra-class robustness, justifying its adoption as an integral part of the HR-ViT training pipeline.

Table 7. Comparative performance with and without data augmentation.

Metric	No Aug.	With Aug.	Improvement
Accuracy (%)	97	**98**	**+1.00**
Precision	0.97	**0.98**	**+0.01**
Recall	0.97	**0.98**	**+0.01**
F1-Score	0.97	**0.98**	**+0.01**

Table 8 compares the classification results reported in previous studies with those of our proposed HR-ViT, evaluated on the same datasets. Quy [51] evaluated a Vision Transformer on the Garbage Classification dataset (12 classes), reporting 92% across all major metrics. Alrayes et al. [14] tested VT-MLH-CNN on the TrashNet dataset, achieving 95.8% accuracy, though precision, recall, and F1-score were not reported. Wang et al. [18] evaluated Garbage FusionNet (GFN) on both TrashNet and a 10-class Garbage Dataset, achieving 94.21% and 96.54% accuracy, with F1-scores of 94.24% and 96.56%, respectively.

To ensure a fair and consistent comparison, we re-implemented HR-ViT using the same datasets and applied consistent preprocessing, augmentation strategies, and 10-layer fine-tuning. Our model achieved 96.04% accuracy on TrashNet, 96.56% on the Garbage Dataset, and 97.80% on the Garbage Classification dataset. Furthermore, HR-ViT yielded improvements across all key metrics (precision, recall, F1-score),

reaching 97% uniformly on the 12-class dataset, outperforming the baseline Vision Transformer by Quy [51] by 5% in each metric. All HR-ViT results were generated through controlled experimentation and reflect consistent performance gains across diverse benchmark datasets.

Although the model demonstrates strong classification performance, this study has some limitations. The dataset contains single-object images with relatively clean backgrounds, which do not fully represent complex waste disposal environments. Additionally, performance in a cloud–mobile system depends on network quality; latency may be affected under low-bandwidth conditions. Future work should address these issues through domain adaptation, synthetic data generation, and Transformer-based object detectors (e.g., DETR) to enable real-time multi-object waste detection in diverse environmental contexts.

From an application perspective, HR-ViT is well-suited for smart recycling bins, mobile environmental monitoring systems, and educational tools. Its hybrid design supports both accuracy and deployment flexibility, enabling scalable implementation across various hardware platforms. Academically, the model affirms the efficacy of combining CNN and transformer structures in visual classification tasks, paving the way for further hybrid solutions in computer vision.

Table 8. Comparison with previous studies on waste classification.

Study	Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Quy [51]	Garbage Classification (12 class)	Vision Transformer	92	92	92	92
Alrayes *et al.* [14]	TrashNet (6 class)	VT-MLH-CNN	95.8	-	-	-
Wang *et al.* [18]	TrashNet (6 class)	Garbage FusionNet (GFN)	94.21	94.31	94.21	94.24
	Garbage Dataset (10 class)		96.54	96.65	96.54	96.56
Ours	**TrashNet (6 class)**	**HR-ViT**	**96.04**	**96**	**96**	**96**
	Garbage Dataset (10 class)		**96.56**	**96**	**96**	**96**
	Garbage Classification (12 class)		**97.80**	**97**	**97**	**97**

5. Conclusion

This study presents HR-ViT (Hybrid ResNet50–ViT), effectively integrating ResNet50's residual learning with ViT's self-attention to classify six types of waste (plastic, paper, organic, metal, glass, and batteries) with 98.27% accuracy in testing, significantly outperforming previous models and achieving an average precision, recall, and F1-score of 0.98. This superior performance is attributed to fine-tuning the final ResNet layers, optimizing the ViT parameters, and incorporating diverse data augmentation to expand the training sample variety. However, this study primarily addresses single-object classification on simple backgrounds, whereas real-world waste conditions often involve complex multi-object arrangements. This limitation necessitates methods such as Transformer-based detection networks (e.g., DETR) for enhanced robustness. Additionally, mobile–cloud implementation relies heavily on stable internet connectivity, emphasizing the need for reliable infrastructure to maintain low latency. In conclusion, this study underscores the potential of hybrid approaches for waste classification and establishes a foundation for future research on computational optimization and adaptation to more complex datasets.

References

[1] F. B. Awino and S. E. Apitz, "Solid waste management in the context of the waste hierarchy and circular economy frameworks: An international critical review," Jan. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1002/ieam.4774.

[2] K. D. Sharma and S. Jain, "Municipal solid waste generation, composition, and management: the global scenario," *Social Responsibility Journal*, vol. 16, no. 6, pp. 917–948, Jul. 2020, doi: 10.1108/SRJ-06-2019-0210.

[3] K. Lipianina-Honcharenko, M. Komar, O. Osolinskyi, V. Shymanskyi, M. Havryliuk, and V. Semaniuk, "Intelligent Waste-Volume Management Method in the Smart City Concept," *Smart Cities*, vol. 7, no. 1, pp. 78–98, Feb. 2024, doi: 10.3390/smartcities7010004.

[4] D. K. Gude *et al.*, "Transforming Urban Sanitation: Enhancing Sustainability through Machine Learning-Driven Waste Processing," *Sustainability*, vol. 16, no. 17, p. 7626, Sep. 2024, doi: 10.3390/su16177626.

[5] H. Lian, D. Wang, and H. Li, "Waste sorting and its

effects on carbon emission reduction: Evidence from China," *Chinese Journal of Population Resources and Environment*, vol. 18, no. 1, pp. 26–34, Mar. 2020, doi: 10.1016/j.cjpre.2021.04.027.

[6] Z. Wang, L. Ye, F. Chen, T. Zhou, and Y. Zhao, "Multi-category sorting of plastic waste using Swin Transformer: A vision-based approach," *J Environ Manage*, vol. 370, Nov. 2024, doi: 10.1016/j.jenvman.2024.122742.

[7] M. I. B. Ahmed *et al.*, "Deep Learning Approach to Recyclable Products Classification: Towards Sustainable Waste Management," *Sustainability (Switzerland)*, vol. 15, no. 14, Jul. 2023, doi: 10.3390/su151411138.

[8] S. Zhang, Y. Chen, Z. Yang, and H. Gong, "Computer Vision Based Two-stage Waste Recognition-Retrieval Algorithm for Waste Classification," *Resour Conserv Recycl*, vol. 169, Jun. 2021, doi: 10.1016/j.resconrec.2021.105543.

[9] M. A. Mohammed *et al.*, "Automated waste-sorting and recycling classification using artificial neural network and features fusion: a digital-enabled circular economy vision for smart cities," *Multimed Tools Appl*, vol. 82, no. 25, pp. 39617–39632, Oct. 2023, doi: 10.1007/s11042-021-11537-0.

[10] E. Z. Kuang, K. R. Bhandari, and J. Gao, "Optimizing Waste Management with Advanced Object Detection for Garbage Classification," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.09975

[11] J. Zhang, H. Yang, and X. Xu, "Research on Service Design of Garbage Classification Driven by Artificial Intelligence," *Sustainability (Switzerland)*, vol. 15, no. 23, Dec. 2023, doi: 10.3390/su152316454.

[12] S. M. Cheema, A. Hannan, and I. M. Pires, "Smart Waste Management and Classification Systems Using Cutting Edge Approach," *Sustainability*, vol. 14, no. 16, p. 10226, Aug. 2022, doi: 10.3390/su141610226.

[13] P. W. Laksono, A. Anisa, and Y. Priyandari, "Deep learning implementation using convolutional neural network in inorganic packaging waste sorting," *Franklin Open*, vol. 8, p. 100146, Sep. 2024, doi: 10.1016/j.fraope.2024.100146.

[14] F. S. Alrayes *et al.*, "Waste classification using vision transformer based on multilayer hybrid convolution neural network," *Urban Clim*, vol. 49, p. 101483, May 2023, doi: 10.1016/j.uclim.2023.101483.

[15] L. W. Qin *et al.*, "Precision Measurement for Industry 4.0 Standards towards Solid Waste Classification through Enhanced Imaging Sensors and Deep Learning Model," *Wirel Commun Mob Comput*, vol. 2021, 2021, doi: 10.1155/2021/9963999.

[16] Y. Zhou *et al.*, "Optimization of automated garbage recognition model based on ResNet-50 and weakly supervised CNN for sustainable urban development," *Alexandria Engineering Journal*, vol. 108, pp. 415–427, Dec. 2024, doi: 10.1016/j.aej.2024.07.066.

[17] N. Li and Y. Chen, "Municipal solid waste classification and real-time detection using deep learning methods," *Urban Clim*, vol. 49, May 2023, doi: 10.1016/j.uclim.2023.101462.

[18] Z. Wang, W. Zhou, and Y. Li, "GFN: A Garbage Classification Fusion Network Incorporating Multiple Attention Mechanisms," *Electronics (Basel)*, vol. 14, no. 1, p. 75, Dec. 2024, doi: 10.3390/electronics14010075.

[19] CCHANG, "Garbage Classification." [Online]. Available: https://www.kaggle.com/datasets/asdasdasasdas/garbage-classification

[20] Z. Yang, Z. Xia, G. Yang, and Y. Lv, "A Garbage Classification Method Based on a Small Convolution Neural Network," *Sustainability (Switzerland)*, vol. 14, no. 22, Nov. 2022, doi: 10.3390/su142214735.

[21] J. Spravil, S. Houben, and S. Behnke, "HyenaPixel: Global Image Context with Convolutions," 2024. doi: 10.3233/FAIA240529.

[22] R. Azad, M. Heidari, Y. Wu, and D. Merhof, "Contextual Attention Network: Transformer Meets U-Net," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01932

[23] P. Elavarthi, J. Lee, and A. Ralescu, "On the ability of CNNs to extract color invariant intensity based features for image classification," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.06500

[24] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020.

[25] G. Habib, D. Singh, I. A. Malik, and B. Lall, "Optimizing Vision Transformers with Data-Free Knowledge Transfer," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.05952

[26] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-Wise Convolutions in Vision Transformers for Efficient Training on Small Datasets," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.19394

[27] J. Ma, Y. Bai, B. Zhong, W. Zhang, T. Yao, and T. Mei, "Visualizing and Understanding Patch Interactions in Vision Transformer," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05922

[28] I. B. Akkaya, S. S. Kathiresan, E. Arani, and B. Zonooz, "Enhancing Performance of Vision Transformers on Small Datasets through Local Inductive Bias Incorporation," May 2023, [Online]. Available: http://arxiv.org/abs/2305.08551

[29] Q. Zhou, H. Zou, and H. Wu, "LGViT: A Local and Global Vision Transformer with Dynamic Contextual Position Bias Using Overlapping Windows," *Applied Sciences (Switzerland)*, vol. 13, no. 3, Feb. 2023, doi: 10.3390/app13031993.

[30] K. Su *et al.*, "DctViT: Discrete Cosine Transform meet vision transformers," *Neural Networks*, vol. 172, p. 106139, Apr. 2024, doi: 10.1016/j.neunet.2024.106139.

[31] J. Chen, P. Wu, X. Zhang, R. Xu, and J. Liang, "Add-Vit: CNN-Transformer Hybrid Architecture for Small Data Paradigm Processing," *Neural Process Lett*, vol. 56, no. 3, Jun. 2024, doi: 10.1007/s11063-024-11643-8.

[32] Y. Guo, D. Stutz, and B. Schiele, "Robustifying Token Attention for Vision Transformers," Mar. 2023.

[33] J. Chen, P. Wu, X. Zhang, R. Xu, and J. Liang, "Add-Vit: CNN-Transformer Hybrid Architecture for Small Data Paradigm Processing," *Neural Process Lett*, vol. 56, no. 3, Jun. 2024, doi: 10.1007/s11063-024-11643-8.

[34] H. Zhu, B. Chen, and C. Yang, "Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.03751

[35] R. Wu *et al.*, "An Efficient Multi-Label Classification-Based Municipal Waste Image Identification," *Processes*, vol. 12, no. 6, p. 1075, May 2024, doi: 10.3390/pr12061075.

[36] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. De Nadai, "Efficient Training of Visual Transformers with Small Datasets," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.03746

[37] Y. Zhao *et al.*, "Neural Architecture Search of Hybrid Models for NPU-CIM Heterogeneous AR/VR Devices," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.08326

[38] Z. Xin and W. U. Jialing, "AResNet-ViT: A Hybrid CNN-Transformer Network for Benign and Malignant Breast Nodule Classification in Ultrasound Images," 2024.

[39] C. Yang, X. Gan, A. Peng, and X. Yuan, "ResNet

Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments," *Sustainability (Switzerland)*, vol. 15, no. 14, Jul. 2023, doi: 10.3390/su151410762.

[40] N. Ebert, D. Stricker, and O. Wasenmüller, "PLG-ViT: Vision Transformer with Parallel Local and Global Self-Attention," *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073447.

[41] K. Zhou *et al.*, "Exploring global attention mechanism on fault detection and diagnosis for complex engineering processes," *Process Safety and Environmental Protection*, vol. 170, pp. 660–669, Feb. 2023, doi: 10.1016/j.psep.2022.12.055.

[42] M. Jeong, M. Yang, and J. Jeong, "Hybrid-DC: A Hybrid Framework Using ResNet-50 and Vision Transformer for Steel Surface Defect Classification in the Rolling Process," *Electronics (Basel)*, vol. 13, no. 22, p. 4467, Nov. 2024, doi: 10.3390/electronics13224467.

[43] J. Zhang *et al.*, "BAE-ViT: An Efficient Multimodal Vision Transformer for Bone Age Estimation," *Tomography*, vol. 10, no. 12, pp. 2058–2072, Dec. 2024, doi: 10.3390/tomography10120146.

[44] R. Yulvina *et al.*, "Hybrid Vision Transformer and Convolutional Neural Network for Multi-Class and Multi-Label Classification of Tuberculosis Anomalies on Chest X-Ray," *Computers*, vol. 13, no. 12, p. 343, Dec. 2024, doi: 10.3390/computers13120343.

[45] O. Chibuike and X. Yang, "Convolutional Neural Network–Vision Transformer Architecture with Gated Control Mechanism and Multi-Scale Fusion for Enhanced Pulmonary Disease Classification," *Diagnostics*, vol. 14, no. 24, p. 2790, Dec. 2024, doi: 10.3390/diagnostics14242790.

[46] W. Liu *et al.*, "Image Recognition for Garbage Classification Based on Transfer Learning and Model Fusion," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/4793555.

[47] J. A. Wahid, X. Mingliang, M. Ayoub, S. Husssain, L. Li, and L. Shi, "A hybrid ResNet-ViT approach to bridge the global and local features for myocardial infarction detection," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-54846-8.

[48] C. Zhang *et al.*, "ResNet or DenseNet? Introducing Dense Shortcuts to ResNet," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.12496

[49] A. Devoto, F. Alvetreti, J. Pomponi, P. Di Lorenzo, P. Minervini, and S. Scardapane, "Adaptive Layer Selection for Efficient Vision Transformer Fine-Tuning," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.08670

[50] Mostafa Mohamed, "Garbage Classification (12 classes)." [Online]. Available: https://www.kaggle.com/datasets/mostafaabla/garbage-classification

[51] N. T. Quy, "PERFORMANCE OF VISION TRANSFORMER ON GARBAGE IMAGE CLASSIFICATION," vol. 04, no. 01, pp. 25–36, 2026, doi: 10.61552/JEMIT.2026.01.003.

[52] W. L. Mao, W. C. Chen, C. T. Wang, and Y. H. Lin, "Recycling waste classification using optimized convolutional neural network," *Resour Conserv Recycl*, vol. 164, Jan. 2021, doi: 10.1016/j.resconrec.2020.105132.

[53] C. Shi, R. Xia, and L. Wang, "A Novel Multi-Branch Channel Expansion Network for Garbage Image Classification," *IEEE Access*, vol. 8, pp. 154436–154452, 2020, doi: 10.1109/ACCESS.2020.3016116.

[54] A. Masand, S. Chauhan, M. Jangid, R. Kumar, and S. Roy, "ScrapNet: An Efficient Approach to Trash Classification," *IEEE Access*, vol. 9, pp. 130947–130958, 2021, doi: 10.1109/ACCESS.2021.3111230.

[55] B. S. Kang and C. S. Jeong, "ARTD-Net: Anchor-Free Based Recyclable Trash Detection Net Using Edgeless Module," *Sensors*, vol. 23, no. 6, Mar. 2023, doi: 10.3390/s23062907.

[56] X. Zhang *et al.*, "Cross-Dataset Generalization in Deep Learning," 2024. [Online]. Available: https://yann.lecun.com/exdb/mnist/].

[57] Z. Lu, H. Xie, C. Liu, and Y. Zhang, "Bridging the Gap Between Vision Transformers and Convolutional Neural Networks on Small Datasets," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.05958

[58] B. Chen, T. Niu, R. Zhang, H. Zhang, Y. Lin, and B. Li, "Feature matching driven background generalization neural networks for surface defect segmentation," *Knowl Based Syst*, vol. 287, Mar. 2024, doi: 10.1016/j.knosys.2024.111451.

[59] M. Moayeri, P. Pope, Y. Balaji, and S. Feizi, "A Comprehensive Study of Image Classification Model Sensitivity to Foregrounds, Backgrounds, and Visual Attributes," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.10766

[60] J. Bobulski and M. Kubanek, "Deep Learning for Plastic Waste Classification System," *Applied Computational Intelligence and Soft Computing*, vol. 2021, 2021, doi: 10.1155/2021/6626948.

[61] N. Nnamoko *et al.*, "Solid Waste Image Classification Using Deep Convolutional Neural Network," 2022, doi: 10.3390/infrastructures.

[62] A. Tatke, M. Patil, A. Khot, P. Jadhavdr, and V. Karad', "HYBRID APPROACH OF GARBAGE CLASSIFICATION USING COMPUTER VISION AND DEEP LEARNING," 2021. [Online]. Available: http://www.ijeast.com

[63] Y. Rayhan and A. P. Rifai, "Multi-class Waste Classification Using Convolutional Neural Network," *Applied Environmental Research*, vol. 46, no. 2, Apr. 2024, doi: 10.35762/AER.2024021.

[64] Z. Yuan and J. Liu, "A Hybrid Deep Learning Model for Trash Classification Based on Deep Trasnsfer Learning," *Journal of Electrical and Computer Engineering*, vol. 2022, 2022, doi: 10.1155/2022/7608794.

[65] Z. Yang, Z. Xia, G. Yang, and Y. Lv, "A Garbage Classification Method Based on a Small Convolution Neural Network," *Sustainability (Switzerland)*, vol. 14, no. 22, Nov. 2022, doi: 10.3390/su142214735.

[66] J. Xu, Y. Li, Q. Shi, and L. He, "Occluded Scene Classification via Cascade Supervised Contrastive Learning," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 16, pp. 4565–4578, 2023, doi: 10.1109/JSTARS.2023.3274592.

[67] K. Huang, H. Lei, Z. Jiao, and Z. Zhong, "Recycling waste classification using vision transformer on portable device," *Sustainability (Switzerland)*, vol. 13, no. 21, Nov. 2021, doi: 10.3390/su132111572.

[68] W. Liu *et al.*, "Image Recognition for Garbage Classification Based on Transfer Learning and Model Fusion," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/4793555.

[69] W. Cai, M. Xie, Y. Liu, and X. Yang, "The Smart City Waste Classification Management System: Strategies and Applications Based on Computer Vision," *Journal of Organizational and End User Computing*, vol. 36, no. 1, 2024, doi: 10.4018/JOEUC.351242.

[70] C. Si, Z. Shi, S. Zhang, X. Yang, H. Pfister, and W. Shen, "Unleashing the Power of Task-Specific Directions in Parameter Efficient Fine-tuning," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.01035

[71] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling Vision Transformers," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.04560

[72] G. Thung and M. Yang, "Dataset Trashnet," p. 3, 2016, Accessed: Jan. 21, 2025. [Online]. Available:

https://github.com/garythung/trashnet

[73] E. Chai, M. Pilanci, and B. Murmann, "Separating the Effects of Batch Normalization on CNN Training Speed and Stability Using Classical Adaptive Filter Theory," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.10674

[74] Z. Zuo, J. Smith, J. Stonehouse, and B. Obara, "An Augmentation-based Model Re-adaptation Framework for Robust Image Segmentation," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.09530

[75] J. Shi, H. Ghazzai, and Y. Massoud, "Differentiable Image Data Augmentation and Its Applications: A Survey," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 2, pp. 1148–1164, Feb. 2024, doi: 10.1109/TPAMI.2023.3330862.

[76] M. Li, X. Zhang, C. Thrampoulidis, J. Chen, and S. Oymak, "AutoBalance: Optimized Loss Functions for Imbalanced Data," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.01212

[77] J. W. Shim, "Enhancing cross entropy with a linearly adaptive loss function for optimized classification performance," *Sci Rep*, vol. 14, no. 1, p. 27405, Dec. 2024, doi: 10.1038/s41598-024-78858-6.

[78] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput Appl*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.

[79] D. A. Anggoro and N. C. Aziz, "Implementation of K-Nearest Neighbors Algorithm for Predicting Heart Disease Using Python Flask," *Iraqi Journal of Science*, vol. 62, no. 9, pp. 3196–3219, Sep. 2021, doi: 10.24996/ijs.2021.62.9.33.

[80] S. B. Cleveland *et al.*, "Tapis API Development with Python: Best Practices in Scientific REST API Implementation: Experience implementing a distributed Stream API," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jul. 2020, pp. 181–187. doi: 10.1145/3311790.3396647.

[81] V. Engström, P. Johnson, R. Lagerström, E. Ringdahl, and M. Wällstedt, "Automated Security Assessments of Amazon Web Services Environments," *ACM Transactions on Privacy and Security*, vol. 26, no. 2, Mar. 2023, doi: 10.1145/3570903.

[82] Y. Liu, †⇤ X. C., P. Liu, J. Grundy, C. Chen, and L. Li, "ReuNify: A Step Towards Whole Program Analysis for React Native Android Apps." [Online]. Available: https://github.com/DannyGooo/ReuNify

[83] C. Yu, T. Chen, Z. Gan, and J. Fan, "Boost Vision Transformer with GPU-Friendly Sparsity and Quantization," May 2023, [Online]. Available: http://arxiv.org/abs/2305.10727

[84] A. Wollek, S. Hyska, B. Sabel, M. Ingrisch, and T. Lasser, "Higher Chest X-ray Resolution Improves Classification Performance," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.06051

[85] G. Nguyen, S. Dlugolinsky, V. Tran, and A. Lopez Garcia, "Deep learning for proactive network monitoring and security protection," *IEEE Access*, vol. 8, pp. 19696–19716, 2020, doi: 10.1109/ACCESS.2020.2968718.

[86] Y. Li *et al.*, "EfficientFormer: Vision Transformers at MobileNet Speed," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.01191

[87] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "MobileOne: An Improved One millisecond Mobile Backbone," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.04040