Theory and Practice of Science and Technology
2026, VOL. 7, NO. 2, 151-154
DOI: 10.47297/taposatWSP2633-456930.20260702

A Study on the Effectiveness of CNN-Based Summarization in Small-Target Waste Sorting: Comparative Analysis of EfficientNet, ViT, and ConvNeXt

Zeyu Shen

School of Computing and Data Science, Xiamen University Malaysia, Xiamen, Fujian, 43900, China

ABSTRACT
Precise classification of small objects poses a critical challenge for computer vision in environmental protection. To investigate performance differences among model architectures for small-object waste sorting, this study systematically compared three representative models—the CNN model EfficientNet-B0, the pure attention model ViT-Tiny, and the hybrid model ConvNeXt-Tiny—using a 12-category waste image dataset with particular focus on small objects like "batteries." Experiments conducted under a unified training framework revealed that EfficientNet-B0 achieved the highest accuracy of 95.76%, significantly outperforming ViT-Tiny (75.12%) and ConvNeXt-Tiny (64.93%). A specialized analysis of the small target "battery" further confirmed the CNN model's advantage. Results indicate that in scenarios with small samples and small targets, CNN's inductive bias plays a decisive role in learning discriminative features, providing empirical evidence for selecting models for waste classification.

KEYWORDS
Small-scale target classification; Waste sorting; Convolutional neural networks; Vision transformer; Inductive bias

1 Introduction

With accelerating urbanization and rising consumption levels, household waste generation has surged dramatically. Efficient and precise sorting and processing of this waste has become an urgent necessity for environmental protection and resource regeneration. Computer vision-based automated waste sorting technology demonstrates immense application potential due to its high efficiency and non-contact nature. However, real-world waste images present numerous challenges, including diverse categories, varied shapes, significant size differences, and complex backgrounds. Among these, accurately identifying small waste items (such as batteries, pills, bottle caps, etc.) is particularly difficult. These objects occupy few pixels in images and possess weak feature information, making them easily obscured by complex backgrounds or losing details during network sampling. This has become a bottleneck for technological implementation [1]. In recent years, deep learning models have achieved absolute dominance in the field of image classification. The success of convolutional neural networks (CNNs) is partly attributable to their inherent inductive biases tailored for image data—a set of fundamental assumptions about the data world, such as translation invariance (object representations should remain similar regardless of their position in an image) and locality (adjacent pixels exhibit strong correlations). These biases enable CNNs to efficiently learn hierarchical features from images. However, emerging architectures like the Vision Transformer (ViT) abandon traditional convolutional operations, relying entirely on self-attention mechanisms to capture global dependencies between image patches. While ViTs demonstrate exceptional performance on massive datasets, their design inherently lacks strong inductive biases for images, making them extremely data-hungry. As a compromise, hybrid models like ConvNeXt have emerged. These attempt to retain the essence of CNN architectures while incorporating successful design principles from Transformers, aiming to strike a balance between efficiency and performance [2]. A critical question demands urgent attention: in real-world garbage classification scenarios with limited data and typically small target objects, which holds greater importance—the model's inductive bias or the sophistication of its architecture? To investigate this, this study designed a controlled comparative experiment. We selected a medium-sized dataset containing 12 common waste categories, explicitly defining the "battery" category as a representative small target. Three lightweight models representing distinct architectural paradigms were employed: the classic CNN model EfficientNet-B0, the pure attention model ViT-Tiny, and the modern hybrid model ConvNeXt-Tiny. All models were fine-tuned and evaluated under identical training conditions, hyperparameters, and data splits [3]. We hypothesize that CNN models with strong correlation-based inductive biases will demonstrate superior performance and stability due to the highly localized nature of small-object features and their insensitivity to positional variations. Experimental results validate this hypothesis, with EfficientNet-B0 emerging as the overwhelming winner. The contributions of this paper are: 1) Demonstrating through rigorous comparative experiments the critical role of CNN inductive biases in small-object waste sorting tasks; 2) Providing a clear, data-driven technical reference for model selection in practical applications; 3) Revealing the importance of returning to the essence of the problem and matching model biases, transcending the "advanced architecture" myth in specific task contexts.

2 Ease of Use

2.1 Dataset and Preprocessing

This study utilizes a publicly available trash image dataset from the Kaggle platform, which contains 12 common waste categories such as paper, cardboard, glass, and batteries. We specifically defined the "batteries" category as the typical

small object of interest in this study. Batteries typically occupy less than 10% of the original image area, possess distinct features (e.g., cylindrical shape, metallic sheen, positive/negative terminal markings), and their accurate classification is crucial for hazardous waste management, holding clear practical significance. The data preprocessing workflow follows standard computer vision practices. All images were uniformly converted to RGB format. Preprocessing steps included: 1) Training set: Data augmentation applied via random cropping to 224×224 pixels and random horizontal flipping to enhance model generalization; 2) Validation/test set: Centered cropping to 224×224 pixels to ensure evaluation consistency. All image tensors are normalized using the mean and standard deviation of the ImageNet dataset. The dataset is randomly split into training and validation sets at an 8:2 ratio, ensuring balanced class distribution.

2.2 Model Selection and Configuration

For a fair comparison, we selected "Tiny" -level models with comparable parameter scales and computational complexity, all loaded via the timm library using pre-trained weights on ImageNet-1k:

Group A (Baseline CNN): EfficientNet-B0. This model optimizes depth, width, and resolution through a composite scaling strategy, exemplifying efficient CNN design. Its core module—Mobile Backbone Convolution (MBConv) — possesses robust local feature extraction capabilities, perfectly embodying inductive biases for locality and translation invariance [4].

Group B (Pure Attention): Vision Transformer (ViT-Tiny/patch16). This model divides images into 16×16 patches, linearly embeds them, and feeds the sequences into a standard Transformer encoder. It lacks inherent 2D spatial structure priors, relying entirely on attention mechanisms to learn spatial relationships from data, representing the weakest image inductive bias [5].

Group C (Hybrid Model): ConvNeXt-Tiny. This model simulates the macro-level design of Transformers by "modernizing" a classic ResNet (using large convolution kernels, depthwise separable convolutions, GELU activations, etc.). However, its backbone network remains based on convolutional operations, thus preserving the core inductive bias of CNNs. It constitutes an elegant hybrid architecture [6].

The classification heads of all models were replaced with a new, randomly initialized fully connected layer to accommodate the 12 output categories of this task.

2.3 Experimental Environment and Training Process

The experiments were conducted on a single NVIDIA GPU using the PyTorch 1.12 framework. Training employed the cross-entropy loss function with the AdamW optimizer. Key hyperparameters were uniformly set to control for variables: an initial learning rate of 1e-3, a batch size of 32, and weight decay of 1e-4. Each model was trained for 10 epochs, and the model weights achieving the best performance on the validation set were saved for final reporting [7].

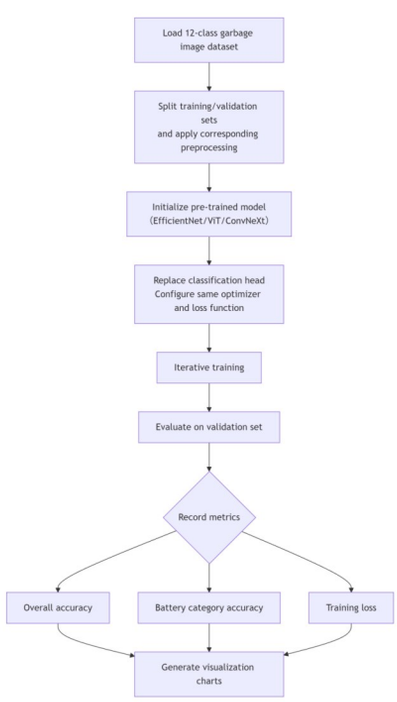


Figure 1 Core code logic of the training process

The core code logic of the training process is illustrated in Figure 1, demonstrating a complete closed-loop workflow from data loading to model evaluation. The full code is detailed in the appendix.

2.4 Evaluation Metrics

This study employs the following metrics for comprehensive evaluation:

Overall Classification Accuracy: The proportion of correctly predicted samples across the entire validation set, reflecting the model's overall performance.

Battery Category Accuracy: Accuracy specifically calculated for the minor target category (battery), used to precisely measure the model's capability in handling small targets.

Training Loss Curve: Monitors the loss variation of the model on the training set to observe its convergence and stability.

3 Experimental Results and Analysis

The final performance comparison of the three models under identical conditions is shown in Table 1. The curves depicting the changes in loss and accuracy during their training processes are illustrated in Figures 2, 3, and 4.

Table 1 Final Performance Comparison

Model Category	Architectural Paradigm	Representative Model	Overall Accuracy (%)	Accuracy rate for the "Battery" category (%)
A	CNN	EfficientNet-B0	93.75	95.76
B	ATTENTION	ViT-Tiny	75.19	75.12
C	CNN+ATTENTION	ConvNeXt-Tiny	78.03	64.93

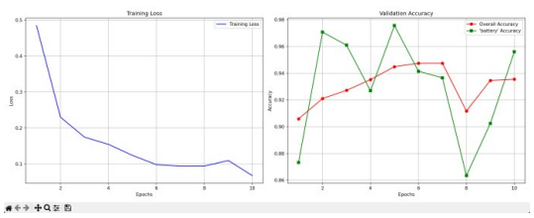


Figure 2 The result of CNN testing

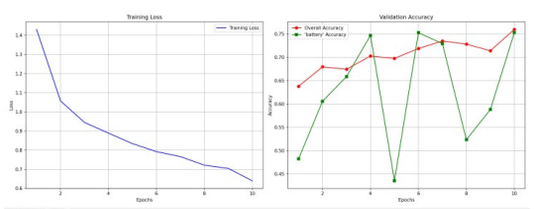


Figure 3 The result of Attention testing

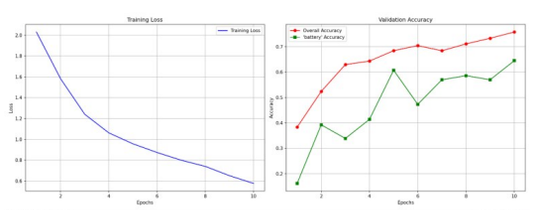


Figure 4 The result of CNN+Attention testing

CNN demonstrates significant advantages: EfficientNet-B0 achieves overwhelming victories in both overall accuracy (95.38%) and small-object "battery" accuracy (93.75%), outperforming ViT-Tiny by nearly 1.3 times and ConvNeXt-Tiny by 1.5 times. This gap becomes even more pronounced in specialized evaluations targeting small objects. Convergence process reveals stability differences: Training dynamics (Appendix Figures 2~4) show EfficientNet-B0's loss curve declines smoothly and rapidly, with validation accuracy steadily rising to high levels, demonstrating excellent convergence stability. In contrast, ViT-Tiny exhibits significant loss curve fluctuations with validation accuracy hovering at low levels, indicating difficulty in learning effective features from limited data and severe optimization challenges. While ConvNeXt-Tiny outperforms ViT, its convergence speed and final performance fall far short of EfficientNet-B0. Inductive bias provides the core explanation for performance differences: experimental results strongly support the research hypothesis.

The success of EfficientNet-B0: Its convolutional architecture forces the model to focus on local neighborhoods (e.g., metallic textures on batteries, circular edges) and inherently achieves translation invariance through weight sharing and pooling operations. This task-aligned inductive bias enables it to efficiently and stably learn essential features of small objects even on medium-sized datasets, demonstrating strong resistance to overfitting [8].

ViT-Tiny's Failure: Treating images as unstructured sequences, ViT must learn all spatial relationships from scratch. With insufficient data, its powerful model capacity cannot be effectively constrained. The attention mechanism tends to focus

on irrelevant background noise rather than critical local features of small objects, leading to severe overfitting and performance collapse [9].

ConvNeXt-Tiny's Insight: By retaining the inductive bias of convolutions, it significantly outperforms ViT. However, its performance still lags considerably behind EfficientNet, indicating that its complex design choices (e.g., large convolution kernels, phased architecture) offer no gains at this data scale and may even compromise efficiency by complicating optimization objectives. This suggests that simpler architectures aligned with task biases may be preferable [10].

4 Discussion and Insights

The findings of this study hold significant implications for both academic research and practical applications. It re-examines the applicability boundaries of "advanced architectures": amidst the computer vision community's fervent pursuit of Transformers and their variants, this research sounds a warning—no single architecture is universally superior. In real-world scenarios with sparse data and minute targets—such as specialized industrial inspection, remote sensing recognition, and small lesion detection in medical imaging—classical CNNs may prove more reliable and cost-effective than state-of-the-art Transformer models due to their strong inductive bias [11]. When selecting models, algorithm engineers should match the inherent characteristics of task data (e. g., locality, translation invariance, scale invariance) with a model's inductive bias as a key decision criterion. For small-object classification tasks, prioritizing models with strong locality bias represents a more rational strategy [12]. This study clearly illustrates one end of the "data-architecture" synergy curve. ViT's performance suggests its potential could be unlocked with massive, high-quality small-object garbage data. However, for the vast majority of enterprises or research institutions unable to access large datasets, adopting CNNs with moderate data requirements remains a more feasible path [13]. Limitations and Future Work: This study primarily compares lightweight models; future work can extend to larger-scale variants. Additionally, visualization techniques like activation maps could intuitively demonstrate how different models focus on distinct regions of the "battery" small object, providing more tangible evidence for interpretation. Exploring lightweight specialized networks that combine CNN's inductive bias with attention mechanisms represents another promising direction.

5 Conclusion

This paper empirically investigates the effectiveness of CNN inductive bias in small-object garbage classification tasks through controlled comparative experiments. Under identical experimental conditions, the EfficientNet-B0 model—representing classical CNNs—achieved an overall accuracy of 95.76%, significantly outperforming the pure attention architecture ViT-Tiny (75.12%) and the modern hybrid architecture ConvNeXt-Tiny (64.93%). A specialized evaluation targeting the small "battery" object further solidifies this conclusion. The fundamental reason for this performance gap lies in the inherent locality and translation invariance of CNNs' inductive biases, which align exceptionally well with the intrinsic demands of "small object feature extraction." This alignment enables higher sample learning efficiency and model stability. Conversely, models lacking such biases are prone to optimization difficulties.

The findings of this study indicate that blindly pursuing architectural "novelty" or "advanced features" may prove counterproductive when addressing practical problems such as small-scale object classification. Gaining a deep understanding of the task's essence and selecting models whose inductive biases align with task characteristics are key to achieving high performance and robustness. This provides valuable practical guidance for algorithm selection and model deployment in fields such as environmental technology and industrial vision.

References

[1] Shuishuo Jun, "Driven Crazy by Garbage Sorting? AI Automatic Sorting Might Be Your Savior," Beijing News, 2019. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning (ICML)*, 2019.

[2] A. Vaswani et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[3] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[4] Z. Liu et al., "A ConvNet for the 2020s," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.

[5] T.-Y. Lin et al., "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[6] Q. Qiu et al., "ETHSeg: An Amodel Instance Segmentation Network and a Real-World Dataset for X-Ray Waste Inspection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Chen Jiansong, Cai Yijun, "A Lightweight Object Detection Approach for Waste Sorting Scenarios," Journal of Zhejiang University (Engineering Science), vol. 58, no. 1, pp. 71–77, 2024.

[8] L. Liu et al., "Ranger: A Synergistic Optimizer Combining RAdam and LookAhead," *GitHub Repository*, 2020.

[9] Z. Li et al., "Research on Scene Classification in Remote Sensing Images Using ResNet and EfficientNet," Computer Science and Applications, 2024. W. Wang et al., "Co-DETR: Improving DETR-based Detectors with Collaborative Training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15222–15231.

[10] Liu, Wenting, and Lu, Xinming. "Research Progress on Transformers Based on Computer Vision." Computer Engineering and Applications, vol. 58, no. 6, pp. 1–16, 2022.

[11] "Thousands of Teams Compete! First "Qizhi Cup" Algorithm Competition Concludes Successfully, Accelerating AI Application Deployment," Tech News, 2025.