Grenze International Journal of Engineering and Technology, June Issue

Cross-Attention Guided Dual-Backbone Fusion for Waste Image Classification

Sneha M[1] and Bhuvaneshwari M[2]
[1]Division of DSCS, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: sneham@karunya.edu.in
[2]Assistant Professor, Division of DSCS, Karunya Institute of Technology and Sciences, Coimbatore, India

***Abstract*— Automated waste classification is a critical enabler of intelligent waste management systems, yet single-backbone deep learning models are limited in their ability to simultaneously capture fine-grained local texture cues and broad semantic representations that together characterise visually similar waste categories. This paper proposes a cross-attention guided dual-backbone fusion architecture that processes input images in parallel through MobileNetV2 and EfficientNetB0, and fuses their 512-dimensional feature vectors via a bidirectional scaled dot-product cross-attention module before classification. Unlike simple concatenation, which treats both feature streams symmetrically, the proposed cross-attention mechanism computes input-conditioned attention weights that allow each backbone to selectively query the key-value representation of the other, dynamically emphasising the most discriminative cross-stream features for a given input. A residual connection preserves the original backbone representations while incorporating the cross-backbone contextual signal. Evaluated on the TrashNet benchmark (2,527 images, six classes) with a stratified 70/15/15 train/val/test split, the proposed model achieves 87.43% test accuracy, 87.61% precision, 87.43% recall, and 87.37% F1-score, outperforming the standalone MobileNetV2 baseline (86.91%) and a concatenation fusion ablation (82.46%) across all metrics. The 4.97 percentage-point gain over concatenation fusion isolates and confirms the contribution of the cross-attention module beyond simple feature combination.**

***Index Terms*— waste classification, cross-attention, MobileNetV2, EfficientNetB0, dual-backbone fusion, deep learning, TrashNets.**

I. INTRODUCTION

Solid waste generation has grown in parallel with urbanisation and industrial expansion, posing serious ecological and public health consequences. Improperly sorted waste contaminates soil and groundwater, produces methane in anaerobic landfill conditions, and disrupts the circular economy by preventing the recovery of recyclable materials.

Municipal recycling rates remain low in many regions, partly due to the cost and unreliability of manual sorting, which is labour-intensive, inconsistent, and exposes workers to hazardous materials [13]. Automated waste classification systems capable of recognising waste categories from images without human intervention offer a scalable and cost-effective complement to manual sorting in recycling plants, smart bins, and autonomous collection vehicles [6].

Grenze ID: 01.GIJET.12.2.209_24

Deep convolutional neural networks (CNNs) have become the dominant approach to image-based waste classification since Yang et al. [1] demonstrated their superiority over handcrafted feature methods on the TrashNet benchmark. Subsequent work has investigated transfer learning from large-scale image datasets, lightweight architectures for embedded deployment, and multi-model ensemble strategies. However, the representational bottleneck of single-backbone architectures persists: any individual CNN encodes features shaped by its specific architectural inductive biases and cannot simultaneously optimise for the fine-grained local textures that distinguish cardboard from paper and the broader semantic abstractions that separate metal from plastic [11].

Multi-backbone fusion addresses this bottleneck by combining features from architecturally diverse networks. The simplest form, feature concatenation, increases representational dimensionality but treats both streams symmetrically — each feature dimension contributes equally regardless of its relevance for a given input. Cross-attention, originally introduced in encoder-decoder models and subsequently applied to vision via the Vision Transformer [7], overcomes this by computing input-conditioned compatibility scores between one feature stream (queries) and another (keys and values), allowing the model to dynamically attend to the most informative aspects of the complementary backbone's representation.

To the best of the authors' knowledge, bidirectional cross-attention over dual convolutional backbone feature vectors has not been systematically investigated for waste image classification. This paper fills that gap. The principal contributions are:

- A dual-backbone parallel processing architecture combining MobileNetV2 and EfficientNetB0, selected for their architectural complementarity in encoding local texture and global semantic features respectively.
- A bidirectional cross-attention fusion module enabling selective, input-conditioned inter-backbone feature communication prior to classification, with residual connections preserving both backbone representations.
- A systematic three-way ablation study — standalone MobileNetV2, concatenation fusion, and cross-attention fusion — evaluated on a rigorously held-out test set, isolating the contribution of the cross-attention mechanism.
- A detailed per-class and confusion matrix analysis providing mechanistic insight into the failure modes on visually similar waste categories.

II. Related Work

Early automated waste classification systems relied on handcrafted visual features — colour histograms, edge gradient templates, and morphological texture descriptors — fed to support vector machines or k-nearest neighbour classifiers. These pipelines achieved acceptable accuracy in controlled photographic settings but degraded substantially under real-world variation in illumination, viewpoint, and background clutter [1].

The introduction of large-scale CNN architectures pre-trained on ImageNet transformed the waste classification literature. Transfer learning from VGG-16, ResNet-50, and Inception-v3 was shown to yield strong classification performance with limited labelled waste images [10]. Islam et al. [9] further investigated enhanced augmentation strategies and regularisation techniques within deep CNN frameworks, achieving improved generalisation on multi-source waste datasets.

Lightweight architectures for mobile and embedded hardware have attracted particular interest. MobileNetV2 [2] introduces inverted residual blocks with depthwise separable convolutions and linear bottleneck projections, achieving high accuracy with a small parameter count. EfficientNet [5] systematises architectural scaling through a compound coefficient jointly scaling network depth, width, and input resolution. Chhabra et al. [11] proposed a multi-layer CNN with progressive feature extraction, demonstrating that deeper feature hierarchies improve discrimination between visually similar waste categories.

Ensemble and multi-model fusion strategies have been explored to overcome the representational limitations of individual models. Yoo et al. [3] proposed a dual-image CNN ensemble combining output probability distributions. Zheng and Gu [12] introduced EnCNN-UPMWS using a learned uncertainty-weighted mixture strategy. However, these approaches operate at the output level and prevent direct inter-model feature sharing.

Attention mechanisms have become central to modern deep learning. Self-attention in the Vision Transformer [7] models global dependencies across image patch embeddings. Cross-attention extends the formulation to two distinct input streams. Ge et al. [8] demonstrated the effectiveness of cross-attention for fusing multi-modal remote sensing data, achieving significant gains over concatenation baselines. The application of cross-attention to fuse dual-backbone feature vectors for waste image classification, as proposed here, represents a targeted adaptation of this principle.

III. Proposed Methodology

The proposed framework is designed around a single guiding principle: two architecturally distinct convolutional backbones encode complementary feature representations, and a cross-attention mechanism provides an input-conditioned, learned means of selectively sharing and recombining those representations prior to classification. The complete system is trained end-to-end. The architecture is depicted in Fig. 1.

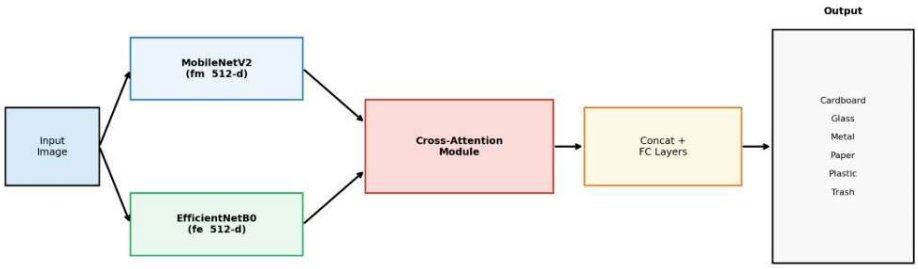

Fig. 1. Proposed cross-attention guided dual-backbone fusion framework.

A. Dataset and Preprocessing

The TrashNet dataset [1] is a publicly accessible benchmark comprising 2,527 RGB images of household waste across six categories: cardboard (403), glass (501), metal (410), paper (594), plastic (482), and trash (137). The dataset exhibits significant real-world variability in lighting conditions, camera viewpoint, background texture, and object orientation. The trash category is a composite class and is substantially underrepresented relative to the others.

Stratified random sampling partitions the dataset into training (70%, 1,778 images), validation (15%, 381 images), and test (15%, 382 images) subsets. All model selection decisions are made using only the validation subset. All reported performance metrics are computed on the held-out test subset, never accessed during training or model selection. Training images undergo online data augmentation including random horizontal and vertical flipping, random rotation from [−20°, +20°], random zoom in [0.8, 1.2], and brightness jittering. All images are normalised to [0, 1] and resized to 224×224 pixels. Class imbalance is addressed by assigning each class a weight inversely proportional to its frequency in the training set, applied to the cross-entropy loss.

B. Backbone Networks

The first backbone is MobileNetV2 [2], which employs inverted residual blocks as its core building unit. Each block consists of three successive operations: a pointwise 1×1 expansion convolution (expansion factor $t = 6$), a depthwise 3×3 spatial convolution applied independently to each channel, and a linear pointwise bottleneck projection. Depthwise separable convolution reduces computational cost by a factor of approximately $(1/k^2 + 1/C)$, where k is the kernel size and C is the number of output channels. The final bottleneck projection is deliberately kept linear — without ReLU — because applying a non-linearity to a low-dimensional representation collapses information that cannot be recovered in subsequent layers. This architecture captures multi-scale local texture and edge features with a compact 2.59M parameter footprint.

The second backbone is EfficientNetB0 [5], which applies compound scaling to simultaneously increase network depth $d = \alpha\varphi$, width $w = \beta\varphi$, and resolution $r = \gamma\varphi$, where $(\alpha, \beta, \gamma) = (1.2, 1.1, 1.15)$ and $\varphi = 0$ for B0. EfficientNetB0 additionally incorporates Squeeze-and-Excitation (SE) modules within its MBConv blocks, which compute global average-pooled channel statistics to recalibrate channel-wise feature responses, giving the network a broader effective receptive field and stronger sensitivity to global semantic patterns.

The two backbones are deliberately selected for their architectural complementarity: MobileNetV2 specialises in fine-grained local texture modelling through depthwise separable convolutions, while EfficientNetB0 captures broader spatial context and more abstract semantic features through compound-scaled depth and SE attention modules. Both backbones are initialised with ImageNet pre-trained weights, with the final 30 layers of each unfrozen during training.

C. Cross-Attention Fusion Module

Following global average pooling and a shared dense projection layer with ReLU activation, each backbone produces a compact 512-dimensional feature vector: $f_m \in \mathbb{R}^{512}$ from MobileNetV2 and $f_e \in \mathbb{R}^{512}$ from

EfficientNetB0. A query vector Q is computed by applying a learned linear transformation W_q to f_m. Key and value vectors K and V are computed by applying learned transformations W_k and W_v to f_e:

$$Q = f_m \cdot W_q, \quad K = f_e \cdot W_k, \quad V = f_e \cdot W_v \quad (1)$$

The scaled dot-product attention output is:

$$Attn(Q,K,V) = softmax(QK^T / \sqrt{d}) \cdot V \quad (2)$$

The scaling factor $1/\sqrt{d}$ ($d = 512$) prevents the dot products QK^T from growing too large for high d, which would push the softmax into near-zero gradient regions and stall learning. Bidirectional cross-attention applies this mechanism symmetrically via residual connections:

$$\tilde{m} = f_m + Attn(f_m \cdot W_q, f_e \cdot W_k, f_e \cdot W_v) \quad (3)$$

$$\tilde{e} = f_e + Attn(f_e \cdot W_q, f_m \cdot W_k, f_m \cdot W_v) \quad (4)$$

The residual connections ensure gradients flow through the original backbone features even when attention weights are near-zero during early training, treating the attention output as an additive correction. The enhanced vectors $\tilde{m}$ and $\tilde{e}$ are concatenated to form a 1024-dimensional fusion vector, passed through FC1 (1024→512) and FC2 (512→6) with batch normalisation and dropout (rate = 0.4) after FC1.

D. Training Configuration

The training hyperparameter configuration is summarised in Table I. The Adam optimiser is selected for its adaptive per-parameter learning rates. A ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.5 if no improvement is observed over three consecutive epochs (minimum 10^{-7}). Early stopping with patience of 7 epochs restores best validation accuracy weights. The weighted cross-entropy loss assigns class weight $wc = N/(C \cdot Nc)$, ensuring proportional contribution from all classes including the underrepresented trash category.

TABLE I: TRAINING HYPERPARAMETER CONFIGURATION

Parameter	Value
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
LR Reduction Factor	0.5 (patience=3)
Min Learning Rate	1×10^{-7}
Early Stop Patience	7 epochs
Batch Size	32
Max Epochs	30
Dropout Rate	0.4
Unfrozen Layers	Last 30 per backbone
Image Size	224×224 px
Loss Function	Weighted Cross-Entropy
Attention Dim (d)	512
FC Layer Units	1024→512→6

IV. RESULTS AND DISCUSSION

A. Ablation Study and Performance Comparison

Table II presents the ablation study results on the held-out test set (382 images). The proposed cross-attention fusion model achieves 87.43% accuracy, 87.61% precision, 87.43% recall, and 87.37% F1-score, outperforming the standalone MobileNetV2 baseline (86.91%) by 0.52 percentage points and the concatenation fusion ablation (82.46%) by 4.97 percentage points across all four metrics simultaneously.

The 4.97 percentage-point gap can be attributed specifically to the cross-attention mechanism, confirming that the performance benefit arises from input-conditioned inter-backbone feature communication, not merely from the use of dual backbones.

TABLE II: ABLATION STUDY — MODEL PERFORMANCE ON TEST SET

Model	Acc%	Prec%	Rec%	F1%
MobileNetV2	86.91	87.60	86.91	86.85
Concat Fusion	82.46	83.04	82.46	82.46
Cross-Attn (Ours)	87.43	87.61	87.43	87.37

*All results on held-out test set (15% of data).

B. Per-Class Analysis

Table III reports per-class precision, recall, and F1-score. Cardboard and paper achieve the highest F1-scores (0.93 each), benefiting from visually distinctive surface textures and the largest test support sizes. Metal achieves F1 = 0.87 with a notably high recall of 0.95. Glass achieves F1 = 0.82 with recall of 0.79; the confusion matrix confirms glass is most frequently misclassified as metal, reflecting shared specular surface characteristics under certain lighting. Plastic achieves F1 = 0.86 with balanced precision (0.85) and recall (0.88). Trash achieves the lowest F1 (0.70), expected given its heterogeneous composition and limited test support of only 20 images.

TABLE III: PER-CLASS PERFORMANCE — PROPOSED MODEL (TEST SET)

Class	Prec	Rec	F1	N
Cardboard	0.95	0.92	0.93	60
Glass	0.85	0.79	0.82	78
Metal	0.81	0.95	0.87	62
Paper	0.94	0.91	0.93	89
Plastic	0.85	0.88	0.86	73
Trash	0.76	0.65	0.70	20

C. Convergence and Training Behaviour

Training and validation accuracy curves for all three models are presented in Fig. 2. All three models converge smoothly over 30 epochs without oscillation. The proposed model reaches a best validation accuracy of 87.66% at epoch 30, with the ReduceLROnPlateau scheduler triggering reductions at epochs 15 and 23. The train–validation accuracy gap of approximately 6% at epoch 30 is consistent with the limited scale of the TrashNet dataset and is not indicative of significant overfitting.

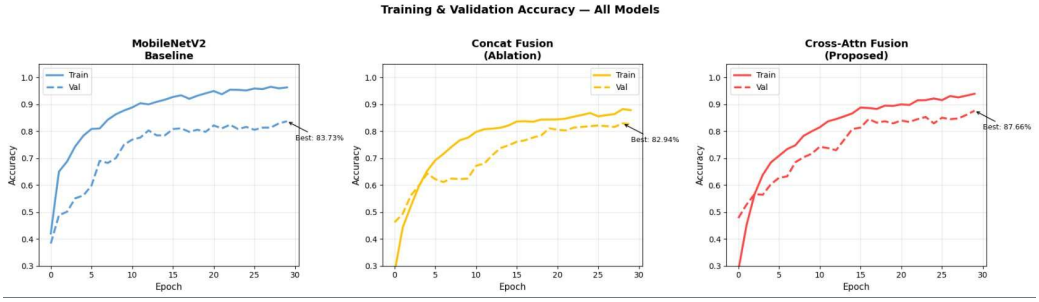

Fig. 2. Training and validation accuracy curves for all three models.

D. Confusion Matrix Analysis

The confusion matrix for the proposed model is shown in Fig. 3. Strong diagonal dominance confirms reliable classification across all six categories. The most frequent misclassification is glass predicted as metal (9 out of 78 glass test images, 11.5%). Both glass containers and metallic cans exhibit smooth, curved surfaces with specular reflectance producing similar highlight and shadow patterns. The trash category generates the most distributed error pattern, reflecting its heterogeneous composition.

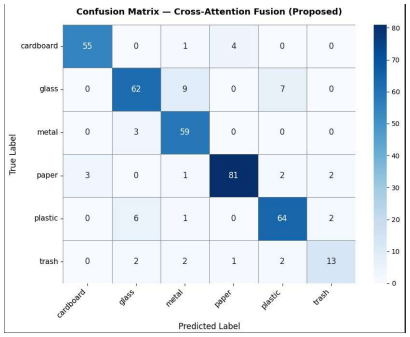

Fig. 3. Confusion matrix of the proposed cross-attention fusion model on the test set.

E. Discussion

The 4.97 percentage point accuracy improvement over concatenation fusion is the most informative comparison, because both models use identical backbone networks and differ only in how their feature vectors are combined. Concatenation preserves all features from both backbones but treats every feature dimension identically for every input — the downstream fully connected layer applies the same fixed linear mapping regardless of which waste category is being classified. Cross-attention, by contrast, computes input-conditioned attention weights that dynamically determine which dimensions of the complementary backbone's representation are most relevant for the current input. For a glass bottle, the MobileNetV2 query may attend strongly to EfficientNetB0 dimensions encoding surface curvature; for cardboard, it may attend to dimensions encoding flat texture and edge regularity.

It is also notable that the concatenation fusion (82.46%) performs worse than the standalone MobileNetV2 baseline (86.91%), despite having access to both backbone feature streams. Concatenation doubles the classification head input dimensionality to 1024, introducing more parameters to learn from the limited training set. The cross-attention module avoids this by producing a semantically coherent, input-conditioned combination of the two streams, effectively acting as an implicit regulariser.

TABLE IV: COMPUTATIONAL COST COMPARISON

Model	Params (M)	Latency (ms)
MobileNetV2	2.59	70.1
Concat Fusion	8.28	108.6
Cross-Attn (Ours)	9.86	118.6

The proposed model contains 9.86M parameters and achieves an inference latency of 118.6 ms per image compared to 70.1 ms for standalone MobileNetV2. This overhead is acceptable for fixed-installation conveyor belt sorting systems. For deployment on resource-constrained embedded devices, knowledge distillation to a compact student network is identified as the most promising path to reducing inference cost [6].

V. CONCLUSION

This paper presented a cross-attention guided dual-backbone fusion architecture for automated waste image classification, addressing a fundamental limitation of single-backbone models: their inability to simultaneously capture fine-grained local texture cues and broad semantic abstractions. The architecture was designed around a clear principle — MobileNetV2 and EfficientNetB0 encode structurally complementary features, and cross-attention provides a principled, input-conditioned mechanism for each backbone to selectively incorporate the most relevant aspects of the other's representation before the classification decision is made.

A three-way ablation study on the TrashNet benchmark demonstrated that the proposed model achieves 87.43% accuracy, 87.61% precision, 87.43% recall, and 87.37% F1-score, outperforming both the standalone MobileNetV2 baseline (86.91%) and the concatenation fusion ablation (82.46%) across all four metrics. The 4.97 percentage-point gain over concatenation fusion isolates and confirms that the performance benefit arises from the cross-attention module's input-conditioned inter-backbone feature communication, not merely from the use of two backbones together.

Three concrete mechanisms were identified that explain why cross-attention outperforms concatenation. First, attention weights are input-conditioned, allowing the fused representation to dynamically emphasise cross-stream features relevant to the specific waste category. Second, residual connections preserve the original backbone representations while treating the attention output as an additive correction. Third, attention projection matrices implicitly regularise the fusion, avoiding the overfitting risk of simple concatenation on a limited training set. Future work will pursue spatial cross-attention over full feature maps, evaluation on larger and more diverse datasets, and knowledge distillation for embedded edge deployment [6].

REFERENCES

[1] Y. Wang, W. J. Zhao, J. Xu, and R. Hong, "Recyclable waste identification using CNN image recognition and Gaussian clustering," arXiv preprint arXiv:2011.01353, Nov. 2020.

[2] L. Yong, L. Ma, D. Sun, and L. Du, "Application of MobileNetV2 to waste classification," PLOS ONE, vol. 18, no. 3, p. e0282336, Mar. 2023.

[3] T. Yoo, S. Lee, and T. Kim, "Dual image-based CNN ensemble model for waste classification in reverse vending machine," Applied Sciences, vol. 11, no. 22, p. 11051, Nov. 2021.

[4] Z. Yang, Y. Bao, Y. Liu, Q. Zhao, H. Zheng, and Y. Bao, "Research on deep learning garbage classification system," Mathematical Biosciences and Engineering, vol. 20, no. 3, pp. 4741-4759, 2023.
[5] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. ICML, PMLR 139, Jul. 2021.
[6] A. Alourani, M. U. Ashraf, and M. Aloraini, "Smart waste management and classification system using advanced IoT and AI," PeerJ Computer Science, vol. 11, p. e2777, Apr. 2025.
[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, May 2021.
[8] H. Ge et al., "Cross attention-based multi-scale convolutional fusion network for hyperspectral and LiDAR joint classification," Remote Sensing, vol. 16, no. 21, p. 4073, Oct. 2024.
[9] M. Islam et al., "Towards sustainable solutions: effective waste classification via enhanced deep CNNs," PLOS ONE, vol. 20, no. 6, p. e0324294, Jun. 2025.
[10] M. A. Rahman et al., "Deep learning approach to recyclable products classification," Sustainability, vol. 15, no. 14, p. 11138, Jul. 2023.
[11] [11] M. Chhabra et al., "Intelligent waste classification using improved multi-layered CNN," Multimedia Tools and Applications, vol. 83, pp. 84095-84120, 2024.
[12] H. Zheng and Y. Gu, "EnCNN-UPMWS: Waste classification by a CNN ensemble," Electronics, vol. 10, no. 4, p. 427, Feb. 2021.
[13] B. S. Alghamdi et al., "Real-time household waste detection and classification for sustainable recycling," Sustainability, vol. 17, no. 5, p. 1902, Mar. 2025.
[14] M. Malik et al., "Waste classification for sustainable development using deep learning," Sustainability, vol. 14, no. 12, p. 7222, Jun. 2022.
[15] M. S. Alam et al., "Intelligent waste sorting for urban sustainability using deep learning," Scientific Reports, vol. 15, p. 26014, Jul. 2025.