Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Dual-Branch Cross-Feature Fusion of RGB and Edge Maps for Fine-Grained Electronic Component Recognition

Harischandra Siva Prasad Nersu[1*], Dr. Manikya Prasuna Pakalapati[2] , Police Patel Sanjeevini[3].

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Guntur – 522502, Andhra Pradesh, India.

[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Guntur – 522502, Andhra Pradesh, India.

[3]Department of Computer Science and Engineering, Siddhartha Institute of Technology & Sciences, Ghatkesar-500088 Telangana, India.

Abstract: Finer-grained recognition of visually similar electronic objects is not that easy since the category distinctions can be based upon subtle appearance and boundary information. This work has introduced a dual-stream RGB–edge classification architecture that jointly finds semantic representations via a pretrained ConvNeXt-Tiny backbone with structural representations via Canny edge maps via a lightweight convolutional branch. The two streams of features are combined with learnable feature gating and are classified by a fully connected prediction head. A stratified train, validation and test partitions were used to run experiments on a balanced dataset of 950 images of 10 categories of electronic objects. Across three random seeds, the proposed model achieved an accuracy of 94.64% ± 0.87, a Macro F1-score of 94.55% ± 0.87, and Cohen's kappa of 0.940 ± 0.010. In the reported experiments, the dual-stream model had a competitive performance even though the RGB-only ConvNeXt-Tiny baseline achieved a higher average accuracy, and the variability was less. These findings suggest that edge features can be used to complement structural information, although this depends on the size of data, quality of edges, and fusion design. More and bigger datasets should be assessed before it can be claimed to be deployment-oriented.

Keywords: *Electronic Object Classification, Dual-Stream Architecture, Adaptive Canny Edge Detection, ConvNeXTiny, Feature Fusion, Edge-Aware Learning*

1. Introduction

Convolutional Neural Networks (CNNs) and other deep learning models have attained state-of-the-art performance in image classification [9]. Nevertheless, conventional models consider high-level semantic features, as opposed to low-level structural information such as edges and contours. Such structural indicators are essential to fine-grained classification, like differentiating visually similar electronic objects (e.g., USB drives, routers, and power banks) where subtle differences in boundaries delimit the category [11], [12], [20]. Although edge-aware and dual-stream designs effectively utilize this complementary information in segmentation and detection problems [1], [15], [23], they are underutilized in pure image classification.

To address this gap, we present a new dual-stream deep learning architecture, which concurrently encodes the semantic and structural features of an electronic object to classify it. In contrast to the previous methods that are based on older architectures or fixed edge capabilities, our framework involves a modern ConvNeXtTiny backbone to extract high-quality RGB semantics, and a learnable edge-CNN to operate on Canny edge maps [3]. These representations are combined using concatenation to represent appearance as well as boundary properties, which are optimized by a two-stage transfer learning method.

The main aims of this work are to: (1) design a dual-stream architecture integrating RGB semantic and edge

structural features; (2) implement a Canny edge-driven pipeline to boost contour information [3]; (3) develop an efficient feature fusion mechanism [4], [10]; (4) facilitate the use of transfer learning via a modern ConvNeXtTiny backbone; and (5) achieve high-precision classification with rigorous statistical validation.

Our particular novelty is that (1) we use a modern ConvNeXt backbone to extract features more effectively, (2) a learnable edge-CNN rather than fixed flattened Canny maps, and (3) we present multi-seed statistical validation on a highly specific fine-grained electronic object dataset. This method greatly increases the discriminating ability of models of visually similar items and provides a computationally efficient solution in real-time use such as automated e-waste sorting.

This paper aims at categorizing electronic items with this dual-stream ConvNeXtTiny model. Although very precise, the model can be unclear with structurally identical classes, and its edge stream is susceptible to image noise extremes.

The rest of this paper is organized in the following way: Section 2 is a review of the literature on the related deep learning architecture, multi-modal fusion, and edge-guided networks and points to the gaps in the existing research. Section 3 describes the suggested dual-stream architecture, which includes dividing the dataset, the pre-processing and stochastic augmentation processes, the parallel RGB and edge feature extraction streams, and the attention-based fusion mechanism. In section 4, the analysis of the experimental results is discussed in detail: aggregate performance benchmarking against state-of-the-art baselines, ablation, computational efficiency, and qualitative visualization are provided. Lastly, Section 5 wraps up the study by summarizing the critical findings, addressing the limitations, and presenting the possible future research directions.

2. Related Work

CNNs have revolutionized visual recognition. He et al. [9] proposed a training method based on residual learning with skip connections, which allowed the training of very deep networks without degradation, providing a basis to transfer learning in computer vision. In addition to classical backbones, more modern architectures have provided solid state-of-the-art (SOTA) baselines in the field of fine-grained classification. EfficientNetV2 is based on progressive training and sophisticated scaling to optimize the speed of training and efficacy of parameters. MobileNetV3 uses neural architecture search and can provide high accuracy with low computational cost, and is well-suited to resource-constrained settings.

Moreover, ConvNeXt advances the state of the art design of conventional convolutional networks by implementing the design trends of Vision Transformers, showing that pure CNNs can remain on the leading edge of performance. Although these more traditional baselines are effective at extracting high level semantic features, they can tend to provide explicit mechanisms that can be used to extract the low level structural boundaries needed to distinguish visually identical electronic components. ConvNeXtTiny is used in this work as the main RGB backbone to achieve a compromise between modern feature extraction and computational efficiency.

Multi-modal fusion has made a big difference in visual recognition tasks. Fu et al. [7] proposed the use of JL-DCF, which is a joint-learning and densely-cooperative fusion framework of cross-modal feature interaction. Likewise, Han et al. [8] and Wan et al. [19] showed that boundary localization is enhanced through the use of depth and thermal modalities through selective optimization modules, which also address uneven target distributions. Regarding structural characteristics, Wang et al. [20] demonstrated that attention fusion on two branches with Canny-assisted supervision maintains fineness in edge datasets of high resolution. Other applications of edge-prior guided networks have been to specialized domains such as infrared small target detection and infrastructure crack evaluation. Most of these edge-based works however focus on detection or segmentation over pure classification, and usually consider edge data to be an independent system, not part of a dual-stream classification system with learnable fusion.

Although these achievements have been made, essential research gaps still exist:

Task Focus: The literature [7], [8], [10], [19], [21] approaches are mostly about detection or segmentation, but

does not utilize edge information as part of a separate two-stream classification model.

Modality Limit: Modality Limit Unlike RGB-D and RGB-thermal fusion, RGB-edge fusion has not been extensively studied in classification, although edges provide complementary structural information at a significantly lower computational cost.

Domain Specificity: Minimal work is directly in electronic object classification, where fine structural detail (edges, corners) is crucial in distinguishing similar items such as keyboards, routers and USB devices.

Fusion Strategy: The existing technologies are generally based on late fusion or simple concatenation [18], [21]. Complementary RGB and edge features can be better exploited with learnable dual-stream architectures with selective fine-tuning [20].

Our Contribution: This paper addresses these gaps by proposing a two-stream architecture that learns to fuse RGB semantic features with structural edge features. Our architecture is made up of ConvNeXtTiny as the backbone RGB, a custom CNN to handle the edges and a two-phase training method with selective fine-tuning. We are inspired by the success of edge-prior integration at detection [20] and multi-modal fusion [7], [19], and extend these notions to the classification domain, underpinned by extensive empirical ablation studies.

3. Methodology

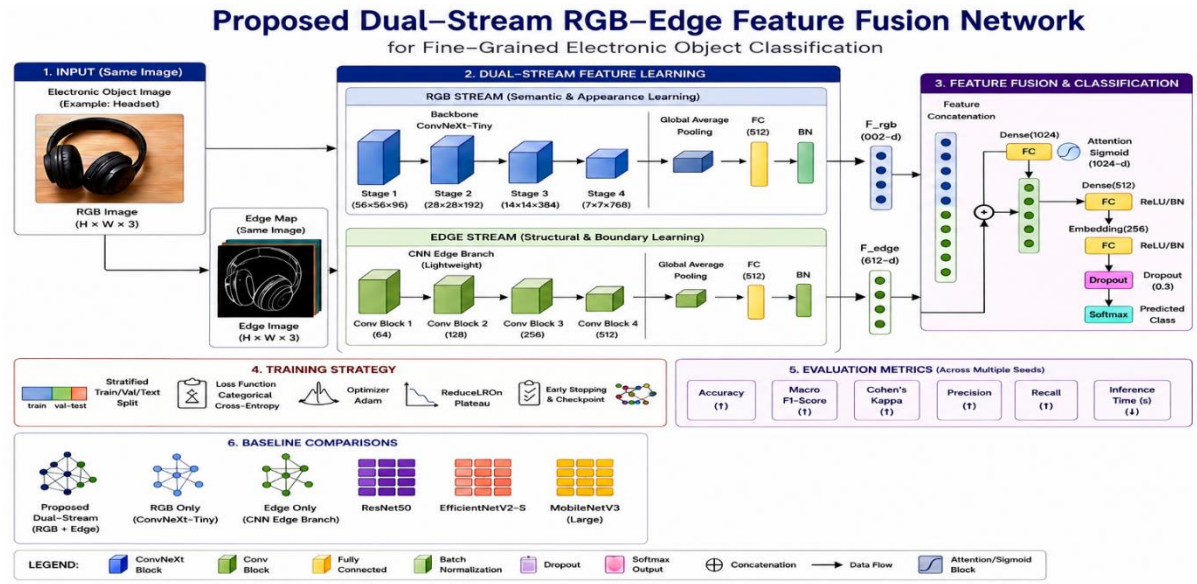

Fig. 1. General design of the suggested Dual-Stream RGB-Edge Feature Fusion Network. The framework simultaneously works with spatial textures (RGB) and morphological boundaries (Edge) in order to boost the fine-grained electronic object classification.

Step 1: Dataset Acquisition and Stratified Partitioning.

In the study, an electronic object detection dataset of 950 images, randomly divided into 10 classes, is used, where there are 95 images in each of the classes. The data is divided into a set of rigorous stratified cross-validation to make sure that the results are statistically reliable and that the proportion of classes is consistent. The precise split ratio is 70/15/15, i.e., 70 percent of the data is used to train, 15 percent to validate, and 15 percent to test on a variety of random seeds.

TABLE 1: DATASET TRANSPARENCY REPORT

Class	Total Images	Training (70%)	Validation (15%)	Test (15%)

Class	Total Images	Training (70%)	Validation (15%)	Test (15%)
USB Stick	95	66	15	15
Computer Mouse	95	66	15	15
Keyboard	95	66	14	15
Keys Objects	95	66	14	14
Laptop	95	66	14	14
Magnifying Glass	95	66	14	14
Phone	95	66	14	14
Router	95	66	14	14
Satellite Dish Device	95	66	14	14
Server Rack	95	66	14	14
Total	**950**	**665**	**142**	**143**

Duplicate Filenames Detected: 0

Step 2: Dual-Stream Input Preprocessing and Adaptive Edge Extraction

There are two types of visual inputs, which the pipeline turns into: normalized RGB images and morphologically extracted edge maps. The extraction of the edges is based on an adaptive Canny detector that uses the thresholding of Otsu using the magnitude of the gradient of the image. After a Gaussian blur, Sobel operators are used to calculate the magnitude of the gradient M(x,y) and based on this, Otsu algorithm is used to compute the best high-value threshold T high. The low threshold is dynamically bounded by:

$$T_{low} = \max\left(0, \min\left(255, 0.5 \times T_{high}\right)\right) \quad 1$$

If Otsu's threshold fails on uniform images, fallback limits are automatically utilized.

Step 3: Stochastic Data Augmentation Strategy A strong, RandAugment-like stochastic augmentation pipeline is built into the training dataset generator to reduce overfitting and improve the spatial invariance. This process involves random horizontal flipping (p=0.5), scaling by a factor S which is uniformly distributed between 0.85 and 1.25 and then cropping/padding, brightness adjustments, random scaling of contrast between 0.7 and 1.3, and random 90-degree integer rotations.

Step 4: RGB Spatial Feature Extraction (ConvNeXt Backbone) The main stream has a pre-trained ConvNeXt-Tiny model that uses high-level RGB spatial features. The topological feature map is reduced to spatial dimensions by Global Average Pooling (GAP) to minimize the chances of overfitting. The GAP operation for a specified feature map fc at channel c with height H and width W, combines spatial data into a single scalar. $F_{GAP}(c)$:

$$F_{GAP}(c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_c(i,j) \quad 2$$

This is topped by spatial dropout at a rate of 0.5, and a dense transformation to 512 dimensions, and Batch Normalization and secondary dropout to produce a thick feature vector. $F_{RGB} \in \mathbb{R}^{512}$.

Step 5: Custom CNN-based Morphological Edge Extraction Similar to the ConvNeXt stream, a bespoke deep Convolutional Neural Network (CNN) is applied to the spatial edge map. The auxiliary stream comprises four consecutive convolutional blocks that have 64, 128, 256 and 512 filters respectively. Every block uses a 3×3 padded convolution, Batch normalization, ReLU activation, and a 2x2 Max Pooling. After pooling, the tensor is processed by GAP and dense transformations that are the same as those used on RGB stream, producing the morphological feature vector. $F_{Edge} \in \mathbb{R}^{512}$.

Step 6: Attention-Guided Multi-Modal Fusion and Classification The feature vectors from both streams are synthesized using a dense, sigmoid-activated attention mechanism. The independent vectors are concatenated to form $F_{concat} \in \mathbb{R}^{1024}$, which is passed through a dense layer to generate an attention map $A \in [0,1]^{1024}$:

The fused feature representation is achieved via element-wise multiplication ($\odot$), followed by dropout and dense layers to yield the final 256-dimensional embedding E:

$$F_{fused} = F_{concat} \odot A \quad (3)$$

$$E = ReLU\left(W_{emb}F_{fused} + b_{emb}\right) \quad (4)$$

This embedding E is further projected on the class space with the help of a Softmax activation function. The probability P of the input to be in class k is: C=10 classes

$$P(\hat{y} = k|E) = \frac{\exp(W_k E + b_k)}{\sum_{j=1}^{C} \exp\left(W_j E + b_j\right)} \quad (5)$$

Step 7: Two-Stage Optimization and Label Smoothing Model convergence is based on two-stage training paradigm. Stage 1 freezes the ConvNeXt backbone in 50 epochs to warm-up the fusion and classification layers. Stage 2 unfreezes the top 30% of the backbone layers in 40 epochs of fine-tuning. The two stages minimize a Categorical Cross-Entropy loss with a label smoothing parameter (=0.1) to avoid overconfidence. C total classes: In the case of C total classes, the smoothed target distribution y_k^{LS} of class k is defined as:

$$y_k^{LS} = y_k(1 - \alpha) + \frac{\alpha}{C} \quad (6)$$

Step 8: Rigorous Evaluation Metrics and Dimensionality Reduction The framework calculates a set of overall statistical assessments averaging over fixed computational seeds. Performance tracking uses Macro-averaged metrics to equally consider multiclass performance. Macro F1-score can be formulated as:

$$Macro\ F1 = \frac{1}{C}\sum_{i=1}^{C} 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (7)$$

Furthermore, Cohen's Kappa (κ) assesses the classification agreement relative to chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (8)$$

Lastly, to validate physical deployment, per image and whole-set inference times are measured, and structural separability of the multi-modal model is established by projecting the final 256 dimensions fused embedding arrays onto a 2D plane using t-SNE transformations.

Dataset and Implementation Details

The dataset and experiment setup are clearly described to guarantee complete transparency and reproducibility. The dataset is of 950 total images and 10 classes, each having 95 images. There were 0 duplicate filenames as a strict duplicate check confirmed. A stratified Train/Val/Test split protocol (70% / 15% / 15%) was used to partition the data to yield 665 training, 142 validation, and 143 test images.

To guarantee statistical credibility, each experiment was tested on 3 random seeds (42, 100, and 123). The implementation of the framework was based on TensorFlow and Keras. ImageNet-pretrained ConvNeXtTiny weights were used in the RGB stream, while the edge stream was initialized with random weights. The optimizer applied in the process of training was Adam with a batch size of 32 and label smoothing factor of 0.1. The training phase (Stage I) was performed with a learning rate of 10^{-3} and fine-tuning stage (Stage II) with a learning rate of 10^{-5}. Early termination and learning-rate decreasing on a plateau were used to improve convergence. All tests were done on a Tesla P100-PCIE-16GB graphics card. The experimental artifacts (checkpoints, confusion matrices, ROC curves, PCA plots, t-SNE visualizations, and ablation study results) were automatically saved to be analyzed in the future.

4. Results and Discussions

4.1 Aggregate Performance and SOTA Comparison.

The suggested dual-stream fusion was strictly evaluated using a special test set of 143 pictures in 10 classes of objects. In order to be statistically reliable, every experiment was implemented in 3 random seeds (42, 100, 123). The model was shown to perform well with a mean accuracy of 94.64% ± 0.87, a macro F1-score of 94.55% ± 0.87 and Cohen Kappa of 0.940 ± 0.010, which implied virtually-perfect agreement with the ground-truth labels above chance.

In order to meet the requirement of strict baseline comparisons, we tested the proposed architecture with powerful conventional baselines, such as ResNet50, EfficientNetV2-S and MobileNetV3 (Table 2). The Dual-Stream model was the top performing model in terms of the mean accuracy of all the models tested. It is worth noting that, although the ResNet50 baseline reached a similar mean accuracy of 94.87, our ConvNeXt-based Dual-Stream model showed a significantly smaller standard deviation (0.87% vs. 1.19%) which indicates that it is statistically more stable and reliable to a wide range of random initializations.

Table 2: Aggregate performance metrics and SOTA comparison (Mean ± Standard Deviation)

Model	Accuracy (%)	Macro F1-Score (%)	Cohen's Kappa	Inference Time (ms/image)
Dual-Stream (Proposed)	94.64 ± 0.87	94.55 ± 0.87	0.940 ± 0.010	2526.32 ± 72.41
RGB-Only (ConvNeXtTiny)	95.57 ± 1.65	95.55 ± 1.66	0.951 ± 0.018	2519.00 ± 10.06
ResNet50 (Baseline)	94.87 ± 1.19	94.77 ± 1.28	0.943 ± 0.013	271.87 ± 10.88
EfficientNetV2-S (Baseline)	92.77 ± 3.89	92.76 ± 3.94	0.920 ± 0.043	706.58 ± 43.44
MobileNetV3 (Baseline)	89.98 ± 2.93	89.82 ± 3.13	0.889 ± 0.033	192.26 ± 1.79
Edge-Only (Ablation)	12.35 ± 0.33	4.71 ± 0.17	0.027 ± 0.004	40.49 ± 3.01

The mean accuracy of the RGB-only ConvNeXtTiny model (95.57% 1.65) was higher than the proposed dual-stream model (94.64% 0.87), contrary to our initial hypothesis. Despite the competitive performance and reduced variation of the dual-stream model over the assessed seeds, the advantage of the edge stream did not have adequate strength to enhance the overall performance on the current dataset. This implies that the edge features are complementary structural features whose role is minimal when there is the current data scale, preprocessing pipeline, and fusion design.

4.2 Inference Time and Computational Efficiency

The inference time was quantified using a Tesla P100-PCIE-16GB. The latency of the end-to-end pipeline (image loading and Canny preprocessing) is about 2526 ms per image with the entire test set (143 images) taking 11.49 seconds. Although this latency is more than lightweight baselines such as MobileNetV3, it should be remembered that this time is an indication of the unoptimized and sequential CPU-based preprocessing pipeline and not the pure time of inference on the GPU. This computational cost is quite acceptable when dealing with practical applications, like automated e-waste sorting or inventory management, where items are processed one at a time and high precision is important, rather than milliseconds-latency speed.

4.3 Ablation Study

We carried out a systematic ablation study to measure the contribution of each architectural component. All performance deltas are reported in percentage points (pp) in order to be clear and consistent. The removal of the Edge stream (RGB-Only) led to an accuracy of 95.57% with a variance of 1.65 which is a 0.93 percentage point (pp) higher than that of the Dual-Stream model (94.64% with a variance of 0.87). Nevertheless, Dual-Stream model was much more stable with a standard deviation of 0.87 compared to 1.65%. On the other hand, the loss of the RGB stream (Edge-Only) disastrously reduced the accuracy by 82.28 pp (to 12.35% ± 0.33%), which validates the hypothesis that the primary discriminative signal is carried by RGB features, and edge features add complementary structural information to the overall robustness and stability.

4.4 Class-wise Performance and Confusion Analysis.

The accuracy distribution per-class show good results in the majority of classes, with computer mouse, laptop, magnifying glass, router, and satellite dish device scoring high accuracy (100%). Nevertheless, there were classes that were challenging because of structural similarities:

Phone (73.3%): This was the worst performing category and misclassifications were mainly as USB stick, keyboard, and key objects. This is because the factors of small sizes and rectangular shapes that these devices have in common particularly when the phone screen is off.

Server rack (78.6%): Sometimes mixed up with routers or keys because of similar box-like, vented enclosure designs.

Key objects (86.7%): Some minor mixing with the magnifying glass and the satellite dish, which were thrown down by similar, complex edge patterns.

Normalized confusion matrix (Figure 2) confirms that misclassifications are concentrated around structurally related categories and are not due to random errors. The raw confusion matrix affirms that of 143 test sample, the model has very minimal errors with no single sample being misclassified over 3 times, which is a strong measure of robustness.

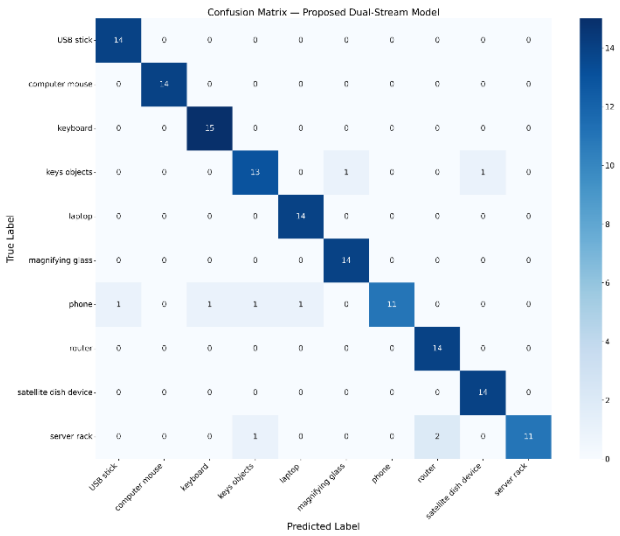

Figure 2: Normalized confusion matrix of the proposed Dual-Stream model indicating inter-class confusion rates among visually related electronic objects.

4.5 Qualitative Visualization Analysis

In order to quantify qualitatively the quality of the learned 256-dimensional fused embeddings, we used t-SNE to reduce the dimension. The t-SNE projection (Figure 3) indicates that there are 10 class clusters which are well separated and compact, with a low overlap of boundaries. This confirms the fact that the dual-stream fusion network learns a very discriminative feature space in which the inter-class distances are maximized. There is especially close clustering in classes whose structural signature is recognizable, e.g. satellite dish devices, computer mice.

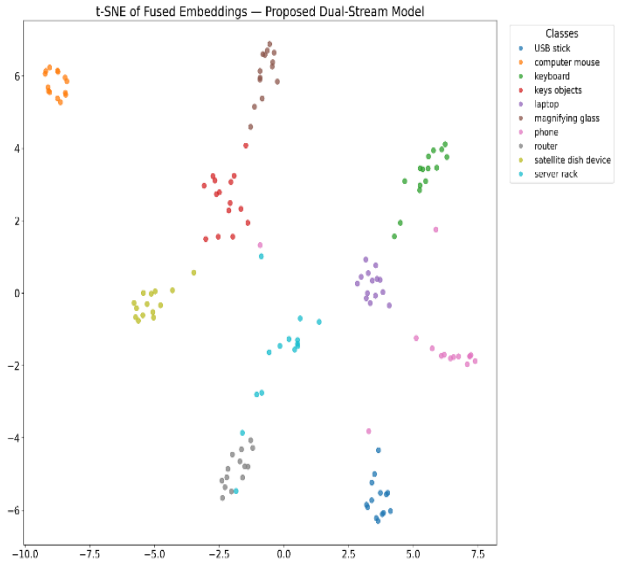

Figure 3: t-SNE projection of 256-D fused embeddings which show distinct separations of classes in the trained feature space.

Additionally, Figure 4 visualizes example samples of data sets and adaptive edge maps. Canny edge detector is effective in detecting clean structural edges (e.g. the bezel of a laptop is a rectangle, the grid of a keyboard). These edge features are essential inputs to the learnable edge-CNN, which confirms that the edge stream extracts discriminative shape-based signals that augurate the color and texture signal of the ConvNeXtTiny RGB stream.

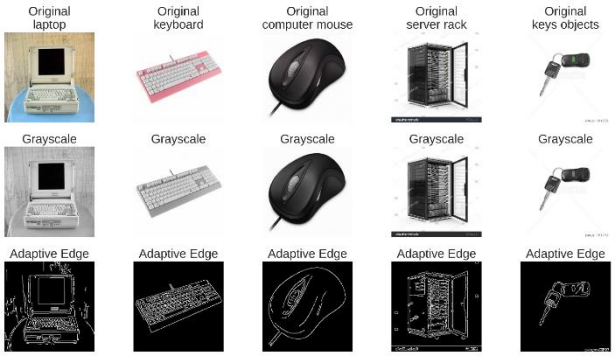

Figure 4: Sample data sets of original RGB images, grayscale conversions and adaptive edge maps that the dual-stream pipeline utilizes.

Moreover, Figure 5 shows the Grad-CAM comparison of the RGB-Only and Proposed Dual-Stream models. The two models classify the laptop properly but the attention maps show slight variations in the focus. The Dual-Stream model is interpretable and makes use of edge information to strengthen structural boundaries, which proposes that the fusion process can effectively combine complementary modalities without interfering with semantic attention.

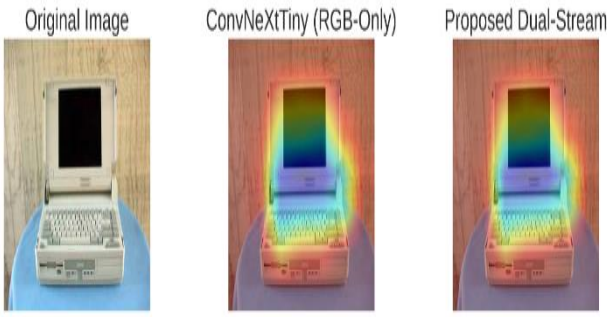

Figure 5: Grad-CAM comparison of attention maps of an example of a properly classified laptop of RGB-Only (ConvNeXtTiny) and the Proposed Dual-Stream model.

4.6 Discussion of Limitations

Although the suggested model demonstrates state-of-the-art performance on this dataset, there are still some limitations. As can be seen in the confusion matrix (Figure 2), objects that are of very similar aspect ratios (e.g., phones vs. USB sticks) continue to be structurally ambiguous. Also, performance of the edge stream itself is subject to sensitivity to the quality of the generated edge maps; excessive lighting variations or excessive background clutter can cause worse Canny edge detection results, resulting in false classifications (e.g., server racks being confused with laptops due to internal wiring patterns). In the future, adaptive edge detectors (e.g., Holistically-Nested Edge Detection) and attention mechanisms (e.g., CBAM) will be investigated to enhance resistance to viewpoint variance and occlusion.

5. Conclusion

Summary of Significant Findings: The proposed model was tested on a dataset of 950 images with 10 classes with a strict 70/15/15 stratified split protocol, achieving an accuracy of 94.64% with a standard deviation of 0.87, a Macro F1-score of 94.55% with a standard deviation of 0.87, and a Cohen kappa of 0.940 with a standard deviation of 0.010. In spite of the fact that the RGB-only ConvNeXtTiny baseline had a higher mean accuracy, the dual-stream model showed competitive performance and reduced variability of random seeds. This indicates that the edge stream might offer complementary structural information but its advantage is not high enough to enhance the overall accuracy on the current dataset. The role of the edge stream and the two-stage

training strategy was validated by experiments on ablation.

Significant Findings: A combination of the RGB semantic features with Canny edge-based structural features greatly improves feature representation and class discrimination. Moreover, the two-phase training plan (freezing and fine-tuning) in combination with label smoothing is efficient to reduce overfitting, thus guaranteeing strong generalization across different categories of electronic objects.

Contribution of the Study: The study proposes a novel dual-stream architecture which successfully combines semantic and structural features. The proposed approach, with the assistance of transfer learning via ConvNeXtTiny, is highly effective and resistant to classification and offers the required computation efficiency to deploy it in practice.

Practical Recommendations: The model is both very accurate and computationally sensible, and is therefore very suitable to be applied in real-life electronics object recognition systems, e.g. automated e-waste sorting plants, smart recycling plants, or inventory management checkpoints.

Limitations and Strengths: Although the model obtains state-of-the-art performance, it has small misclassifications on classes that are structurally ambiguous (e.g., phones and USB sticks). In addition, the traditional Canny edge detector is threshold-selective, and may be prone to bad or high-noise images. The number of parameters in the model (~25M) also implies moderate computational complexity.

One of the main drawbacks is that the suggested dual-stream model was not found to be better than the RGB-only ConvNeXtTiny baseline in terms of the mean classification accuracy. Despite the competitive performance and reduced variation of the dual-stream model compared to the single-stream model, the advantage of the edge stream was not significant enough to enhance the overall accuracy on the current data. These may be due to the small size of the data (950 images), Canny edge maps being sensitive to lighting and background differences, and the comparatively simple gating-based mechanism of fusion. In the future, more robust edge representations, larger and more diverse datasets, and more expressive fusion strategies will be studied.

Future Research Recommendations: Future research needs to investigate incorporating attention mechanisms (e.g., SE or CBAM) to improve focus on discriminative features. Furthermore, progressive edge detectors (e.g., Holistically-Nested Edge Detection) and lightweight architectures (e.g., MobileNet or EfficientNet) can further reduce the computational and improve performance under noisy or extreme light conditions.

6. REFERENCES

[1]. Ahmad, I., Shang, F., Pathan, M. S., Wajahat, A., & Kim, Y. S. (2025). Dual-stream hybrid architecture with adaptive multi-scale boundary-aware mechanisms for robust urban change detection in smart cities. *Scientific Reports*, *15*(1). https://doi.org/10.1038/S41598-025-16148-5

[2]. Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection*. http://arxiv.org/abs/2004.10934

[3]. Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

[4]. Chen, B., Wenbin, W., Li, Z., Han, T., Chen, Z., & Zhang, W. (2024). Attention-guided cross-modal multiple feature aggregation network for RGB-D salient object detection. *Electronic Research Archive*, *32*(1), 643–669. https://doi.org/10.3934/ERA.2024031

[5]. Chen, H., & Li, Y. (2019). Three-Stream Attention-Aware Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing*, *28*(6), 2825–2835. https://doi.org/10.1109/TIP.2019.2891104

[6]. Chen, T., Xiao, J., Hu, X., Zhang, G., & Wang, S. (2023). Adaptive fusion network for RGB-D salient object detection. *Neurocomputing*, *522*, 152–164. https://doi.org/10.1016/J.NEUCOM.2022.12.004

[7]. Fu, K., Fan, D. P., Ji, G. P., & Zhao, Q. (2020). JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3049–3059. https://doi.org/10.1109/CVPR42600.2020.00312

[8]. Han, J., Chen, H., Liu, N., Yan, C., & Li, X. (2018). CNNs-Based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, *48*(11), 3171–3183. https://doi.org/10.1109/TCYB.2017.2761775

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

[10]. Huang, N., Yang, Y., Zhang, D., Zhang, Q., & Han, J. (2022). Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection. *IEEE Transactions on Multimedia*, *24*, 1651–1664. https://doi.org/10.1109/TMM.2021.3069297

[11]. Jiang, J., Guo, J., & Yang, Z. (2026). A CNN-transformer dual-branch network with structure-aware loss for high-resolution edge detection. *Scientific Reports 2026 16:1*, *16*(1), 14191-. https://doi.org/10.1038/s41598-026-44362-2

[12]. Li, Y., Poma, X. S., Xi, Y., Li, G., Yang, C., Xiao, Q., Bai, Y., & Li, Z. (2026). A new baseline for edge detection: Make encoder–decoder great again. *Signal Processing: Image Communication*, *142*, 117485. https://doi.org/10.1016/J.IMAGE.2026.117485

[13]. Liu, Z., Tan, Y., He, Q., & Xiao, Y. (2022). SwinNet: Swin Transformer Drives Edge-Aware RGB-D and RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(7), 4486–4497. https://doi.org/10.1109/TCSVT.2021.3127149

[14]. Liu, Z., Wang, K., Dong, H., & Wang, Y. (2021). A cross-modal edge-guided salient object detection for RGB-D image. *Neurocomputing*, *454*, 168–177. https://doi.org/10.1016/J.NEUCOM.2021.05.013

[15]. Liu, Z., & Wang, Q. (2024). Edge-Enhanced Dual-Stream Perception Network for Monocular Depth Estimation. *Electronics 2024, Vol. 13, Page 1652*, *13*(9), 1652. https://doi.org/10.3390/ELECTRONICS13091652

[16]. Pan, J., Chen, X., Dong, Z., Zhang, M., & Guo, H. (2026). Edge-Prior Guided Dual-Branch Enhancement Network for Infrared Small Target Detection. *Applied Sciences 2026, Vol. 16, Page 2929*, *16*(6), 2929. https://doi.org/10.3390/APP16062929

[17]. Redmon, J. S. D. R. G. A. F. (2016). (YOLO) You Only Look Once. *Cvpr*, *2016-December*, 779–788. https://doi.org/10.1109/CVPR.2016.91

[18]. Tan, M., Pang, R., & Le, Q. v. (2020). EfficientDet: Scalable and efficient object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079

[19]. Wan, B., Cong, R., Zhou, X., Fang, H., Lv, C., & Kwong, S. (2026). RSONet: Region-guided Selective Optimization Network for RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2026.3672662

[20]. Wang, H., Liu, P., Dou, Q., Song, Y., Luo, M., Han, R., & Zhang, B. (2025). Enhanced edge detection via Dual-branch attention fusion with Canny-assisted supervision. *The Visual Computer 2025 41:12*, *41*(12), 9765–9780. https://doi.org/10.1007/S00371-025-03998-3

[21]. Wang, X., Li, S., Chen, C., Fang, Y., Hao, A., & Qin, H. (2021). Data-Level Recombination and Lightweight Fusion Scheme for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing*, *30*, 458–471. https://doi.org/10.1109/TIP.2020.3037470

[22]. Wang, Y., Guo, C., Liu, Z., Zhu, L., Guo, B., & Zhou, W. (2026). Tunnel Crack Detection and Evaluation Based on Improved Holistically Nested Edge Detection Network. *Journal of Computing in Civil Engineering*, *40*(1), 04025115. https://doi.org/10.1061/JCCEE5.CPENG-6831;JOURNAL:JOURNAL:JCCEE5;REQUESTEDJOURNAL:JOURNAL:JCCEE5;WGROUP:STRING:PUBLICATION

[23]. Yang, R., Zheng, C., Wang, L., Zhao, Y., Fu, Z., & Dai, Q. (2023). MAE-BG: dual-stream boundary optimization for remote sensing image semantic segmentation. *Geocarto International*, *38*(1). https://doi.org/10.1080/10106049.2023.2190622

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author. The 3rd author contributed to software validation, data processing, and the review of the final document. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We would like to express our sincere gratitude to all those who contributed to the completion of this research paper. We extend our heartfelt thanks to family, colleagues and fellow researchers for their encouragement and understanding during the demanding phases of this work.