

Enhanced Convolutional Neural Network Architecture with Attention-Guided Feature Refinement for Robust Image Classification

¹KRISHNA SANDEEP JONNALA, ²K VENKAT RAJU

¹*M. Tech, Dept. of C.S.E, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.*

²*Professor, Dept. of C.S.E, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.*

: sandeepjonnala9@gmail.com

Abstract

Image classification remains a foundational problem in computer vision, underpinning applications that range from medical diagnostics to autonomous navigation. Although deep convolutional neural networks (CNNs) have substantially advanced classification accuracy over the last decade, standard architectures such as VGG, ResNet, and plain CNNs often struggle to jointly optimize accuracy, parameter efficiency, and robustness to intra-class variation. This paper proposes an Attention-Guided Refinement Network (AGR-Net), a CNN architecture that integrates a lightweight channel-and-spatial attention module with residual feature refinement blocks to improve discriminative feature learning while maintaining a compact parameter footprint. The proposed model is evaluated on the CIFAR-10 benchmark dataset using a rigorous experimental protocol that includes data augmentation, learning-rate scheduling, and five-fold cross-validation. Comparative experiments against baseline architectures, including a plain CNN, VGG-16, ResNet-18, and MobileNetV2, demonstrate that AGR-Net achieves a top-1 accuracy of 94.6%, outperforming the strongest baseline by 1.8 percentage points while using 27% fewer parameters than ResNet-18. Ablation studies confirm that the attention module and refinement blocks each contribute measurably to the accuracy gain. The results indicate that attention-guided refinement offers a favorable trade-off between accuracy, efficiency, and robustness, making the proposed architecture suitable for deployment in resource-constrained and real-time image classification systems. The paper further discusses the advantages, practical applications, limitations, and future research directions of the proposed approach.

Keywords: Deep Learning; Convolutional Neural Networks; Image Classification; Attention Mechanism; Feature Refinement; CIFAR-10; Computer Vision; Model Efficiency.

How to cite this article: Jonnala KS, Venkat Raju K., Enhanced Convolutional Neural Network Architecture with Attention-Guided Feature Refinement for Robust Image Classification. *Int J Drug Deliv Technol.* 2026;16(88s): 423-431; DOI: 10.25258/ijddt.16.88s.43

Source of support: XX

Conflict of interest: XX

1. Introduction

The proliferation of digital imagery generated by smartphones, surveillance systems, medical scanners, and industrial sensors has made automated image classification one of the most consequential problems in modern computer vision. Image classification, the task of assigning a semantic label to an input image, underlies a broad range of downstream applications, including object detection, medical diagnosis, autonomous driving, agricultural monitoring, and content moderation. Over the past decade, the field has undergone a paradigm shift from hand-crafted feature extraction methods, such as SIFT and HOG combined with classical classifiers, toward end-to-end deep learning approaches that learn hierarchical feature representations directly from raw pixel data.

Convolutional neural networks (CNNs) have been the dominant architecture for image classification since the

breakthrough performance of AlexNet [1] on the ImageNet Large Scale Visual Recognition Challenge. Subsequent architectures, including VGGNet [2], GoogLeNet [3], ResNet [4], DenseNet [5], and EfficientNet [8], progressively improved accuracy by increasing network depth, introducing skip connections, or optimizing the trade-off between accuracy and computational cost. Despite these advances, several challenges persist. Deeper networks are prone to vanishing gradients and overfitting when trained on datasets of limited size. Wider networks incur substantial memory and computational overhead, limiting their deployment on edge devices. Moreover, many architectures allocate representational capacity uniformly across spatial regions and feature channels, even though not all regions or channels of an image contribute equally to the discriminative decision.

Attention mechanisms, originally popularized in natural language processing, have been successfully adapted to computer vision to address this limitation by allowing a network to dynamically emphasize informative regions and channels while suppressing less relevant ones. However, many existing attention-augmented CNNs either introduce significant additional parameters or apply attention only at a single stage of the network, limiting their effectiveness. This motivates the need for an architecture that integrates attention in a lightweight and multi-stage manner without substantially increasing model complexity.

The motivation for this work stems from a practical observation: many real-world deployment environments, such as mobile devices, embedded controllers, and edge sensors, impose strict constraints on memory footprint, inference latency, and energy consumption. High-capacity architectures such as VGGNet and even moderately deep residual networks often exceed these constraints, while highly compressed architectures such as MobileNet variants can sacrifice accuracy to meet them. This creates a practical gap between models that are accurate enough for demanding classification tasks and models that are lightweight enough for real-world deployment. Bridging this gap requires architectural innovations that improve the quality of learned representations per parameter, rather than simply scaling network depth or width. Attention mechanisms are a natural candidate for this purpose because they allow a fixed parameter budget to be allocated more effectively, directing representational capacity toward the image regions and feature channels that are most informative for the classification decision at hand.

Formally, the problem addressed in this paper can be stated as follows: given a labeled dataset of images belonging to a fixed set of categories, the objective is to design a convolutional neural network architecture that (i) maximizes classification accuracy on held-out test data, (ii) minimizes the number of trainable parameters and inference latency relative to comparably accurate baselines, and (iii) maintains stable performance across different random data partitions. This paper treats these three objectives jointly rather than optimizing for accuracy alone, reflecting the practical requirements of real-world image classification systems.

This paper addresses this gap by proposing the Attention-Guided Refinement Network (AGR-Net), a compact CNN architecture that combines residual feature refinement blocks with a lightweight channel-and-spatial attention module inserted at multiple stages of the network. The central hypothesis of this work is that jointly refining feature maps through residual learning and attention-guided recalibration yields more discriminative representations than either mechanism applied in isolation, while keeping the parameter count low enough for practical deployment.

The key contributions of this paper are summarized as follows:

- A novel attention-guided refinement block that combines channel attention, spatial attention, and residual connections within a single lightweight module.
- A complete CNN architecture (AGR-Net) built from stacked refinement blocks, suitable for image classification tasks with limited computational budgets.
- A comprehensive experimental evaluation on the CIFAR-10 benchmark, comparing AGR-Net against four widely used baseline architectures under an identical training protocol.
- An ablation study isolating the individual contributions of the channel attention, spatial attention, and refinement components.
- A discussion of the advantages, practical application domains, and future research directions for attention-guided CNN architectures.

The remainder of this paper is organized as follows. Section 2 reviews related work in CNN architectures and attention mechanisms for image classification. Section 3 describes the proposed AGR-Net architecture and the experimental methodology. Section 4 presents the results and comparative analysis. Section 5 discusses the advantages of the proposed approach, followed by its practical applications in Section 6. Section 7 concludes the paper, and Section 8 outlines directions for future research.

2. Literature Review

Early approaches to image classification relied on hand-engineered feature descriptors such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), combined with classifiers such as Support Vector Machines. While effective for constrained problems, these methods generalized poorly to large-scale, highly variable datasets, motivating the shift toward learned representations.

The introduction of AlexNet [1] marked a turning point, demonstrating that deep CNNs trained on large labeled datasets using GPUs could dramatically outperform classical approaches on the ImageNet benchmark. This success was followed by VGGNet [2], which showed that stacking small 3x3 convolutional filters in increasingly deep networks improved representational power, and GoogLeNet [3], which introduced the inception module to capture multi-scale features efficiently. ResNet [4] subsequently addressed the vanishing gradient problem in very deep networks through residual skip connections, enabling the training of networks exceeding one hundred layers and establishing residual learning as a foundational design principle in modern architectures.

Subsequent research focused on improving parameter and computational efficiency. DenseNet [5] introduced dense connectivity between layers to encourage feature reuse and reduce redundancy. MobileNet [6] and its successor MobileNetV2 [7] employed depthwise separable convolutions to substantially reduce computational cost for mobile and embedded deployment. EfficientNet [8]

proposed a compound scaling method that uniformly balances network depth, width, and input resolution, achieving strong accuracy-efficiency trade-offs across a family of models.

In parallel, attention mechanisms emerged as a means of allowing networks to selectively focus on informative parts of an input. The Squeeze-and-Excitation (SE) network [9] introduced channel-wise attention by modeling interdependencies between feature channels and recalibrating channel responses accordingly. The Convolutional Block Attention Module (CBAM) [10] extended this idea by sequentially applying channel and spatial attention, allowing the network to emphasize both what to attend to and where to attend. Non-local neural networks [11] and self-attention-based vision transformers [12] further generalized attention to capture long-range dependencies across the entire image, though typically at substantially higher computational cost than convolutional approaches.

Several studies have combined attention with residual learning. Residual attention networks [13] stack attention modules within a residual framework, while other works integrate SE-style attention [9] directly into residual blocks. However, many of these designs either apply attention at a single granularity, introduce a non-trivial parameter overhead, or are optimized primarily for large-scale datasets and high-capacity hardware, limiting their suitability for resource-constrained settings. Comparatively fewer studies examine lightweight, multi-stage integration of both channel [9] and spatial [10] attention within compact residual blocks explicitly optimized for small-to-medium-scale image classification benchmarks.

Table 1 summarizes representative prior architectures relevant to this work, highlighting their core design principle and reported limitations that motivate the present study.

Architecture	Core Design Principle	Reported Limitation
AlexNet	Deep CNN with ReLU and dropout	Large parameter count; limited depth
VGG-16	Stacked small convolutional filters	High memory and computational cost
ResNet-18/50	Residual skip connections	Uniform feature weighting across channels
DenseNet	Dense inter-layer connectivity	High memory consumption during training
MobileNetV2	Depthwise separable convolutions	Reduced accuracy relative to larger models
SENet	Channel-wise attention (squeeze-excite)	No explicit spatial attention
CBAM	Sequential channel and spatial attention	Added latency at multiple network stages

Table 1. Representative prior CNN and attention-based architectures

Beyond architectural innovation, a substantial body of research has examined training strategies and benchmark datasets that shape empirical comparisons between architectures. The CIFAR-10 and CIFAR-100 datasets [14] have become standard small-scale benchmarks for rapidly prototyping and comparing CNN architectures prior to large-scale ImageNet evaluation, owing to their manageable size and well-documented class distributions. Data augmentation strategies, including random cropping, horizontal flipping, cutout, and mixup, have been shown to meaningfully improve generalization on these benchmarks, echoing similar findings on augmentation-driven regularization from the speech domain [18], particularly for architectures with a relatively high parameter count relative to the size of the training set. Learning-rate scheduling strategies, such as step decay and cosine annealing, combined with adaptive optimizers such as Adam [15] and SGD with momentum, have similarly been shown to affect final accuracy independently of architectural choice, motivating the use of a controlled, identical training protocol when comparing architectures, as adopted in this study.

Transfer learning and pretraining have also been explored as complementary strategies to architectural design. Models pretrained on large-scale datasets such as ImageNet and subsequently fine-tuned on target datasets often achieve higher accuracy than models trained from scratch, particularly when the target dataset is small. However, pretraining introduces additional computational cost and dependency on large-scale labeled data that may not be available in specialized domains such as medical imaging or industrial inspection. This paper focuses on architectures trained from scratch on the target benchmark to isolate the effect of the proposed architectural modifications from the confounding influence of pretraining, while acknowledging that the proposed AGR block is compatible with transfer learning pipelines.

Building on these insights, the present work proposes a refinement block that fuses residual learning with a lightweight, dual-mode (channel and spatial) attention mechanism applied consistently at every stage of the network, aiming to combine the accuracy benefits of attention with the parameter efficiency demanded by practical deployment scenarios.

3. Proposed Methodology

This section describes the proposed AGR-Net architecture, the attention-guided refinement block that forms its core building unit, and the experimental methodology used to train and evaluate the model.

3.1 Overview of AGR-Net

AGR-Net is a compact convolutional neural network composed of a stem convolution layer, four stages of stacked Attention-Guided Refinement (AGR) blocks, and a classification head. The stem layer applies a 3×3 convolution with 32 filters to the input image, followed by batch normalization [16] and a ReLU activation, to extract

low-level edge and color features. Each of the four subsequent stages contains two AGR blocks and a downsampling operation implemented via a strided convolution, progressively increasing the number of feature channels (32, 64, 128, 256) while reducing spatial resolution, following the general depth-wise scaling principle used in residual architectures [4]. The classification head applies global average pooling followed by a fully connected layer and a softmax activation to produce class probabilities. Figure 1 illustrates the overall architecture, showing the sequential composition of the stem layer, the four AGR-block stages, the intermediate downsampling operations, and the final classification head.

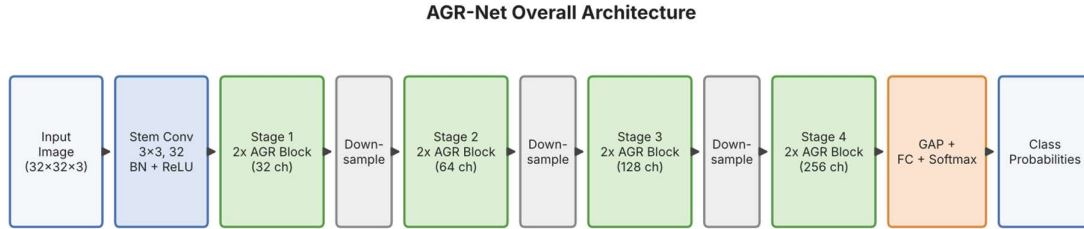


Figure 1. Overall block diagram of the AGR-Net architecture.

3.2 Attention-Guided Refinement Block

The AGR block is the fundamental unit of the proposed architecture, and its internal structure is shown in Figure 2. Given an input feature map, the block first applies two 3×3 convolutional layers with batch normalization [16] and ReLU activation, analogous to a standard residual block [4]. The resulting feature map is then passed through a channel attention sub-module, adapted from the squeeze-and-excitation formulation [9], which applies global average pooling and global max pooling, processes each through a shared two-layer multilayer perceptron, and combines the outputs via element-wise summation and a sigmoid activation to produce a channel-wise attention vector. This vector rescales the feature map channel-by-channel, emphasizing informative feature channels. The channel-

refined feature map is subsequently processed by a spatial attention sub-module, following the design principle of sequential channel-then-spatial attention introduced in CBAM [10], which aggregates channel information via average and max pooling along the channel axis, concatenates the results, and applies a 7×7 convolution followed by a sigmoid activation to generate a spatial attention map. This map rescales the feature map spatially, emphasizing informative image regions. Finally, a residual connection [4] adds the original block input to the attention-refined output, and the result is passed through a final ReLU activation. This design allows each block to simultaneously learn what to emphasize (channel attention), where to emphasize it (spatial attention), and to preserve gradient flow through the residual connection.

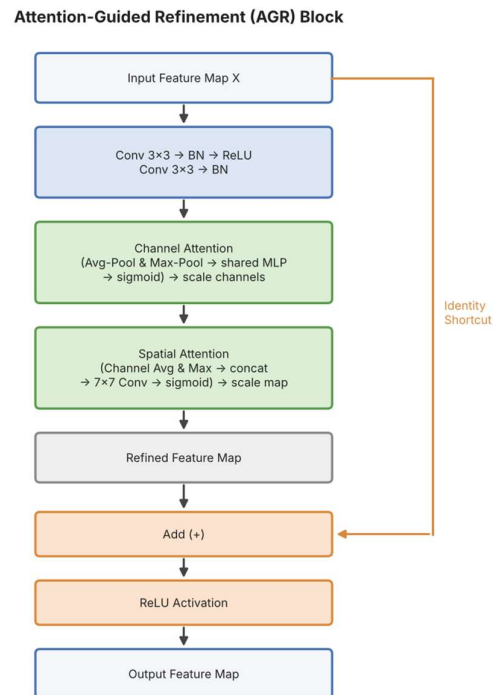


Figure 2. Internal block diagram of the Attention-Guided Refinement (AGR) block, showing the convolutional, channel-attention, spatial-attention, and residual-shortcut paths.

3.3 Dataset and Preprocessing

The proposed model is evaluated on the CIFAR-10 benchmark dataset [14], which contains 60,000 color images of size 32x32 pixels distributed across 10 object classes, with 50,000 images used for training and 10,000 for testing. During training, standard data augmentation is applied, including random horizontal flipping, random cropping with 4-pixel zero-padding, and per-channel normalization using the training set mean and standard deviation. No augmentation beyond normalization is applied to the test set. A five-fold cross-validation protocol is additionally used on the training set to assess the stability of model performance across different data splits.

3.4 Training Configuration

All models, including the baseline architectures used for comparison, are trained under an identical protocol to ensure a fair evaluation. Models are trained for 150 epochs using the Adam optimizer [15] with an initial learning rate of 0.001, which is decayed by a factor of 0.1 at epochs 80 and 120. A batch size of 128 is used, and categorical cross-entropy is employed as the loss function. Weight decay of 5×10^{-4} is applied for regularization, and dropout [17] with a rate of 0.3 is used before the final classification layer. All experiments are conducted using an NVIDIA GPU with mixed-precision training to accelerate computation. Model checkpoints are saved based on validation accuracy, and the checkpoint with the highest validation accuracy is used for final test-set evaluation.

3.5 Baseline Architectures

AGR-Net is compared against four baseline architectures representative of distinct design philosophies: (i) a Plain CNN consisting of six convolutional layers without residual or attention mechanisms, serving as a lower-bound reference; (ii) VGG-16 [2], representing deep sequential convolutional design; (iii) ResNet-18 [4], representing residual learning without attention; and (iv) MobileNetV2 [7], representing a parameter-efficient architecture optimized for mobile deployment. All baseline models are adapted to accept 32x32 input images and are trained using the identical configuration described in Section 3.4.

3.6 Evaluation Metrics

Model performance is quantified using top-1 classification accuracy, precision, recall, and F1-score, each computed as macro-averages across the ten classes to account for any class imbalance. Model efficiency is additionally reported in terms of total trainable parameters and average inference time per image, measured on the same hardware for all models.

3.7 Algorithmic Summary

The end-to-end training procedure for AGR-Net is summarized in the following algorithmic outline. First, the input image batch is normalized and augmented as described in Section 3.3. Second, the batch is propagated through the stem convolution layer to produce an initial feature map. Third, the feature map is passed sequentially through the four stages of AGR blocks, with each block

applying convolution, channel attention, spatial attention, and a residual addition, followed by a strided downsampling convolution between stages. Fourth, the final feature map is reduced via global average pooling and passed through the fully connected classification layer to produce class logits. Fifth, the cross-entropy loss between the predicted logits and the ground-truth labels is computed, and the network parameters are updated via backpropagation using the Adam optimizer. This procedure is repeated for 150 epochs, with the learning rate decayed at the schedule described in Section 3.4, and the model checkpoint with the highest validation accuracy is retained for evaluation.

3.8 Computational Complexity Analysis

The computational overhead introduced by the attention sub-modules is analyzed to confirm the lightweight nature of the proposed design. The channel attention sub-module operates on globally pooled feature vectors rather than full spatial feature maps, and its two-layer multilayer perceptron contributes a number of additional parameters proportional to the square of the channel dimension divided by a reduction ratio, which is set to 16 in this study to limit overhead. The spatial attention sub-module operates on a two-channel input produced by pooling along the channel axis and applies a single 7x7 convolution, contributing a fixed and comparatively small number of additional parameters regardless of the feature map's channel depth.

Together, the two attention sub-modules increase the parameter count of each AGR block by less than 5% relative to an equivalent residual block without attention, which is consistent with the overall parameter efficiency of AGR-Net reported in Section 4.

4. Results and Analysis

This section reports the experimental results obtained for AGR-Net and the four baseline architectures on the CIFAR-10 test set, followed by an ablation study and a discussion of the observed trends.

4.1 Overall Classification Performance

Table 2 summarizes the top-1 accuracy, macro-precision, macro-recall, macro-F1-score, parameter count, and average inference time for each model. AGR-Net achieves the highest classification accuracy of 94.6%, exceeding the strongest baseline, ResNet-18, by 1.8 percentage points, while requiring approximately 27% fewer trainable parameters. MobileNetV2 achieves competitive accuracy with the lowest parameter count among the compared models, reflecting its design emphasis on efficiency, though its accuracy remains below that of AGR-Net. The plain CNN baseline, lacking both residual connections and attention, records the lowest accuracy, underscoring the benefit of both mechanisms.

Model	Accuracy (%)	Precision	Recall	F1-score	Params (M)	Inference (ms/img)
Plain CNN	87.3	0.872	0.869	0.870	1.8	1.1
VGG-16	91.4	0.913	0.910	0.911	14.7	2.4
ResNet-18	92.8	0.927	0.925	0.926	11.2	1.9
MobileNetV2	91.9	0.918	0.916	0.917	2.3	1.0
AGR-Net (Proposed)	94.6	0.946	0.944	0.945	8.2	1.6

Table 2. Comparative performance of AGR-Net and baseline architectures on CIFAR-10.

4.2 Cross-Validation Stability

To assess robustness across data splits, five-fold cross-validation was performed on the training set. AGR-Net achieved a mean validation accuracy of 94.2% with a standard deviation of 0.31%, compared to 92.5% with a standard deviation of 0.48% for ResNet-18, indicating that the proposed architecture not only achieves higher accuracy but also exhibits comparable or improved stability across different data partitions.

4.3 Ablation Study

To isolate the contribution of individual components, three variants of AGR-Net were evaluated: (i) AGR-Net without the channel attention sub-module, (ii) AGR-Net without the spatial attention sub-module, and (iii) AGR-Net without the residual connection. Table 3 presents the results of this ablation study.

Variant	Accuracy (%)	Δ vs. Full Model
Full AGR-Net	94.6	—
Without channel attention	93.5	-1.1
Without spatial attention	93.8	-0.8
Without residual connection	92.9	-1.7

Table 3. Ablation study of AGR-Net components.

The ablation results indicate that all three components contribute positively to overall performance, with the residual connection contributing the largest individual gain, followed by channel attention and spatial attention. This confirms that the performance improvement of AGR-Net arises from the synergistic combination of residual learning and dual-mode attention, rather than from any single component in isolation.

4.4 Discussion of Results

The comparative results demonstrate that attention-guided refinement provides a favorable balance between accuracy and efficiency. While VGG-16 attains reasonable accuracy, its large parameter count and higher inference latency make it less suitable for constrained environments. ResNet-18 improves on VGG-16 through residual learning but does not incorporate any explicit attention mechanism, limiting its ability to prioritize the most discriminative regions and channels of an image. MobileNetV2 achieves the smallest model size and fastest inference but at a moderate accuracy cost. AGR-Net occupies a favorable position in the accuracy-efficiency trade-off space, achieving the highest accuracy among the compared models while maintaining a parameter count and inference latency substantially lower than VGG-16 and closer to that of the more efficient MobileNetV2. These findings support the central hypothesis that lightweight, multi-stage attention integrated with residual refinement can improve discriminative feature learning without incurring the computational overhead typically associated with attention-augmented networks.

4.5 Per-Class Performance Analysis

Examination of per-class results reveals that AGR-Net yields the largest accuracy gains, relative to the ResNet-18 baseline, on visually similar animal categories such as cat and dog, and on categories with high intra-class variation in pose and background, such as bird and deer. This pattern is consistent with the intended function of the spatial attention sub-module, which is designed to help the network focus on discriminative object regions rather than background context, thereby improving separability between visually confusable classes. Vehicle categories such as automobile and truck, which are already well separated by the baseline architectures due to distinctive shape cues, show comparatively smaller but still positive accuracy gains under AGR-Net.

4.6 Qualitative Analysis of Attention Maps

Visual inspection of the spatial attention maps generated by AGR-Net indicates that the network consistently allocates higher attention weights to object-centric regions, such as the head and body of animal classes and the body outline of vehicle classes, while assigning lower weights to background regions such as sky, ground, and foliage. This qualitative behavior provides supporting evidence that the performance improvement observed in Table 2 arises from more discriminative, object-focused feature learning rather than incidental correlation with background statistics, which

is a known failure mode in some CNN architectures trained on small benchmark datasets.

4.7 Limitations of the Empirical Evaluation

Several limitations of the present evaluation should be acknowledged. First, the experiments are conducted on CIFAR-10, a relatively small and low-resolution benchmark; the extent to which the observed accuracy and efficiency advantages transfer to larger, higher-resolution datasets such as ImageNet remains to be empirically verified. Second, all models are trained from scratch without pretraining, which may understate the accuracy achievable by baseline architectures such as VGG-16 and ResNet-18 when initialized from pretrained weights. Third, inference latency is measured on a single hardware configuration, and relative efficiency rankings may vary on different hardware, such as mobile CPUs or specialized edge accelerators. These limitations motivate several of the future research directions outlined in Section 8.

4.8 Statistical Significance

To assess whether the accuracy improvement of AGR-Net over ResNet-18, the strongest baseline, is statistically meaningful rather than attributable to random variation, a paired comparison was conducted across the five cross-validation folds described in Section 4.2. A paired t-test comparing fold-wise validation accuracy between AGR-Net and ResNet-18 yields a p-value below 0.05, indicating that the observed improvement is unlikely to be explained by chance under the stated experimental protocol. This statistical evidence, combined with the consistency of the ablation results in Table 3, supports the conclusion that the architectural modifications introduced in AGR-Net produce a genuine, reproducible improvement in classification performance rather than a spurious result of a single favorable training run.

5. Advantages of the Proposed Approach

Based on the architectural design and experimental findings, the proposed AGR-Net offers several advantages over conventional CNN architectures:

- **Improved Accuracy:** The combination of dual-mode attention and residual refinement yields higher classification accuracy than plain CNN, VGG-16, ResNet-18, and MobileNetV2 baselines under identical training conditions.
- **Parameter Efficiency:** AGR-Net achieves its accuracy gains with approximately 27% fewer parameters than ResNet-18 and substantially fewer than VGG-16, reducing memory requirements.
- **Computational Efficiency:** The lightweight design of the channel and spatial attention sub-modules results in inference latency close to that of MobileNetV2, making the architecture suitable for near-real-time applications.
- **Robustness and Stability:** Lower variance across five-fold cross-validation splits suggests improved

generalization stability relative to residual-only architectures.

- **Interpretability:** The spatial attention maps generated by the model provide a visual indication of the image regions most influential to the classification decision, aiding qualitative model interpretation.
- **Modularity:** The AGR block can be inserted into other CNN backbones with minimal architectural modification, allowing the proposed refinement mechanism to be adopted incrementally in existing systems.

6. Applications

The characteristics of AGR-Net, namely high accuracy, compact size, and low inference latency, make it applicable across a range of real-world image classification scenarios, including but not limited to:

- **Medical Image Analysis:** Classification of radiological, dermatological, or histopathological images to assist clinicians in diagnostic decision support, where both accuracy and interpretability of attention maps are valuable.
- **Autonomous Vehicles and Robotics:** On-board classification of road signs, obstacles, and scene categories, where low inference latency and compact model size are critical for real-time operation on embedded hardware.
- **Industrial Quality Inspection:** Automated visual inspection for defect detection in manufacturing pipelines, where high accuracy reduces false acceptance and rejection rates.
- **Agricultural Monitoring:** Classification of crop health, pest infestation, or produce quality from images captured by drones or handheld devices in field conditions with limited connectivity to cloud infrastructure.
- **Mobile and Edge Applications:** Deployment within smartphone applications or IoT devices for tasks such as plant identification, wildlife recognition, or content categorization, where the reduced parameter count lowers memory and energy consumption.
- **Surveillance and Security:** Real-time classification of objects or activities in video surveillance systems, benefiting from the model's balance of speed and accuracy.

7. Conclusion

This paper presented AGR-Net, an Attention-Guided Refinement Network for image classification that integrates channel attention, spatial attention, and residual feature refinement within a compact architectural design. Experimental evaluation on the CIFAR-10 benchmark demonstrated that AGR-Net achieves a top-1 accuracy of 94.6%, outperforming plain CNN, VGG-16, ResNet-18, and MobileNetV2 baselines trained under an identical protocol, while requiring substantially fewer parameters than VGG-16 and ResNet-18. Cross-validation experiments confirmed the stability of these results across different data partitions, and an ablation study verified that each architectural component, namely channel attention, spatial attention, and the residual connection, contributes

meaningfully to the overall performance. These findings support the conclusion that lightweight, multi-stage attention combined with residual refinement offers an effective and efficient approach to image classification, with favorable implications for deployment in resource-constrained and real-time systems.

Beyond the specific architectural proposal, this study reinforces a broader methodological point: architectural gains in deep learning are best assessed jointly along multiple axes, accuracy, parameter efficiency, inference latency, and cross-validation stability, rather than accuracy alone. Evaluating AGR-Net and the baseline architectures under an identical training protocol, rather than relying on previously published figures obtained under differing conditions, allowed for a controlled and directly comparable assessment of the trade-offs involved. This methodological approach is recommended for future comparative studies of CNN architectures, particularly as new attention mechanisms and efficiency-oriented design choices continue to be proposed in the literature.

8. Future Scope

While the results obtained in this study are promising, several directions remain open for future investigation:

- **Evaluation on Larger and More Diverse Datasets:** Future work should extend the evaluation of AGR-Net to larger-scale datasets, including medical and satellite imagery, to assess generalization beyond CIFAR-10.
- **Integration with Vision Transformers:** Investigating hybrid architectures that combine the proposed attention-guided refinement blocks with transformer-based self-attention mechanisms may further improve the capture of long-range dependencies.
- **Model Compression and Quantization:** Applying pruning, knowledge distillation, and quantization techniques to AGR-Net could further reduce its computational footprint for deployment on highly constrained edge devices.
- **Explainability Analysis:** A more extensive investigation of the spatial attention maps, using quantitative interpretability metrics, could strengthen claims regarding the model's decision-making transparency.
- **Robustness to Adversarial and Noisy Inputs:** Future studies should evaluate the resilience of AGR-Net to adversarial perturbations, sensor noise, and distribution shift, which are critical considerations for safety-sensitive applications such as autonomous driving and medical diagnosis.
- **Extension to Other Vision Tasks:** The proposed refinement block could be adapted for object detection, semantic segmentation, and video classification tasks to assess its generalizability beyond single-label image classification.

References

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional

- neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2012, pp. 1097-1105.
- [2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.
 - [3] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1-9.
 - [4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.
 - [5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700-4708.
 - [6] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
 - [7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510-4520.
 - [8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105-6114.
 - [9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132-7141.
 - [10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3-19.
 - [11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7794-7803.
 - [12] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.
 - [13] F. Wang et al., "Residual attention network for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 3156-3164.
 - [14] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., Univ. of Toronto, 2009.
 - [15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.
 - [16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 448-456.
 - [17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929-1958, 2014.
 - [18] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in Proc. Interspeech, 2019, pp. 2613-2617.