Adaptive Augmentation-based Hybrid CNN-vision Transformer for Small-resolution Object Recognition

Syarifah Fadillah Rezky *, Muhammad Dahria, Siti Julianita Siregar, Zaimah Panjaitan, and Rita Hamdani

Information Systems Program, STMIK Triguna Dharma, Medan, Indonesia
Email: ikic5500@gmail.com (S.F.R.); mdahria1@gmail.com (M.D.); siti.julianita18@gmail.com (S.J.S.); zaimahp09@gmail.com (Z.P.); r1t4.hamdani@gmail.com (R.H.)
*Corresponding author

***Abstract*—Object recognition in low-resolution images remains a challenging task due to limited spatial detail and reduced discriminative features. While Convolutional Neural Networks (CNNs) effectively capture local patterns, they are limited in modeling long-range dependencies. In contrast, Vision Transformers (ViTs) provide strong global contextual modeling but often require higher computational resources. This study proposes an adaptive augmentation-based hybrid CNN-vision Transformer architecture designed to enhance feature representation for small-resolution image classification. The proposed method integrates hierarchical convolutional feature extraction with transformer-based contextual encoding and introduces an adaptive data augmentation mechanism that applies adaptive augmentation strategies based on training feedback. Experimental results on a controlled synthetic benchmark dataset demonstrate that the proposed model achieves stable convergence and consistent classification performance, reaching an accuracy of 89.15%. The results indicate that the integration of local and global feature modeling improves representation capability under constrained resolution conditions. However, the evaluation is limited to a controlled dataset, and further validation on diverse real-world datasets is required to fully assess generalization performance. Overall, the proposed approach provides a promising direction for improving classification performance in low-resolution image recognition tasks.**

***Keywords*—small-resolution object recognition, hybrid Convolutional Neural Networks (CNN)-vision Transformer, adaptive data augmentation, image classification, deep learning**

I. Introduction

Recent advancements in computer vision have significantly improved automated visual understanding across various domains, including medical imaging, intelligent surveillance, and environmental monitoring. Despite these developments, recognizing objects in low-resolution images remain a persistent challenge due to limited spatial detail and reduced discriminative features [1, 2].

Convolutional Neural Networks (CNNs) have demonstrated strong capability in extracting hierarchical local features through convolutional operations [3]. However, their limited receptive fields restrict their ability to model long-range dependencies. In contrast, Vision Transformers (ViTs) leverage self-attention mechanisms to capture global contextual relationships more effectively, although they often require substantial computational resources and large-scale training data [4, 5].

To overcome these limitations, recent studies have explored hybrid architectures that integrate CNNs with transformer-based models. These approaches aim to combine efficient local feature extraction with global contextual reasoning [6, 7]. Architectures such as MobileViT and Swin Transformer have demonstrated that hybridization can significantly improve performance while maintaining computational efficiency [8, 9].

In addition, data augmentation plays a crucial role in improving model generalization. Conventional techniques such as flipping, cropping, and rotation are widely used to increase data diversity [10]. More advanced methods, including RandAugment and CutMix, further enhance robustness by introducing structured perturbations during training [11, 12]. However, most existing approaches rely on static augmentation pipelines, which may not adapt effectively to varying training conditions.

Motivated by these challenges, this study proposes a hybrid CNN-vision Transformer architecture enhanced with an adaptive data augmentation mechanism. The proposed approach aims to improve feature representation by combining local spatial extraction with global contextual modeling, while dynamically adjusting augmentation strategies to enhance generalization performance.

The main contributions of this study are summarized as follows:

Manuscript received February 27, 2026; revised March 18, 2026; accepted April 21, 2026; published September 10, 2026.

(1) A hybrid CNN-vision Transformer architecture that integrates local and global feature learning;
(2) An adaptive augmentation strategy that dynamically adjusts training transformations;
(3) A comprehensive evaluation demonstrating improved convergence stability and classification performance;
(4) Additional analysis to validate the robustness and effectiveness of the proposed framework.

II. Related Work

Recent advances in deep learning have significantly improved visual recognition performance through the development of both convolutional and transformer-based architectures. Convolutional Neural Networks (CNNs), such as ResNet and EfficientNet, have demonstrated strong capability in hierarchical feature extraction and computational efficiency [1, 2]. These architectures leverage local receptive fields to capture spatial patterns effectively, but they remain limited in modeling long-range dependencies.

To address this limitation, Vision Transformers (ViTs) have been introduced as a powerful alternative for capturing global contextual relationships using self-attention mechanisms [3–5]. Variants such as Tokens-to-Token ViT and data-efficient transformers further enhance representation learning while reducing data requirements [4, 6]. Additionally, large-scale pretraining approaches, including BEiT and masked autoencoders, have shown improved performance in visual representation learning [5, 7].

However, the high computational cost of transformer-based models has led to the development of hybrid architectures that integrate convolutional and attention mechanisms. For instance, MobileViT combines lightweight convolutional operations with transformer blocks to achieve an efficient balance between accuracy and computational complexity [8]. Similarly, Swin Transformer and its improved version introduce hierarchical attention mechanisms to reduce computational overhead while maintaining strong contextual modeling capabilities [9, 10]. Other hybrid approaches, such as CoAtNet and ConvNeXt, further demonstrate the effectiveness of combining convolutional inductive bias with attention-based learning [11, 12].

In parallel, data augmentation has become a critical component for improving model generalization. Traditional augmentation techniques, including geometric transformations and normalization, have been widely applied in image classification tasks [13]. More advanced augmentation strategies, such as RandAugment, Mixup, and CutMix, introduce structured perturbations that improve robustness and prevent overfitting [14–16]. AugMix further enhances robustness by combining multiple augmentation operations to improve generalization under distribution shifts [17]. Despite these advancements, most augmentation strategies are applied statically and do not adapt to training dynamics.

For object detection and small-object recognition, several studies have emphasized the importance of multi-scale feature representation. Feature Pyramid Networks (FPN) and Path Aggregation Networks (PANet) enable hierarchical feature fusion across different spatial scales, improving detection performance for objects of varying sizes [18, 19]. Modern detection frameworks such as YOLOv4 and YOLOv7 further improve real-time detection performance through optimized architectures and training strategies [20, 21]. Transformer-based detection models, including DETR, also demonstrate the effectiveness of attention mechanisms in object localization tasks [22].

Small-object recognition remains particularly challenging due to limited pixel representation and reduced discriminative features. Survey studies highlight that multi-scale learning and attention-based mechanisms play a crucial role in improving detection performance for small objects [23, 24]. Additionally, slicing-based inference techniques such as SAHI improve detection accuracy by focusing on localized image regions [25]. Supporting techniques such as non-local networks and attention modules further enhance contextual feature representation [26, 27].

Lightweight architectures such as ShuffleNet and Squeeze-and-Excitation Networks have also been proposed to improve efficiency while maintaining performance, making them suitable for resource-constrained environments [28, 29]. Meanwhile, classical detection frameworks such as SSD and Fast R-CNN continue to provide strong baselines for object recognition tasks [30, 31]. Early unified detection models such as YOLO further demonstrate the evolution of real-time object detection systems [32, 33].

Despite these significant advancements, existing studies often focus on either architectural improvements or data augmentation techniques independently. The integration of adaptive augmentation strategies within hybrid CNN-vision Transformer frameworks remains relatively underexplored, particularly for small-resolution image classification tasks. This gap highlights the need for a unified approach that combines efficient feature extraction, global contextual modeling, and adaptive data augmentation to improve performance under constrained visual conditions.

Therefore, this study aims to address this limitation by proposing an adaptive augmentation-based hybrid CNN-vision Transformer architecture that enhances both representation learning and generalization capability.

CNNs emphasize localized feature extraction through convolution and pooling operations, while Vision Transformers capture global contextual relationships via patch embedding and self-attention. The illustration also highlights small-resolution object challenges resulting from limited pixel representation, motivating the development of hybrid CNN-transformer architectures.

Fig. 1 illustrates the conceptual differences between convolutional neural networks and vision transformers in visual feature modeling for small-resolution object recognition tasks.

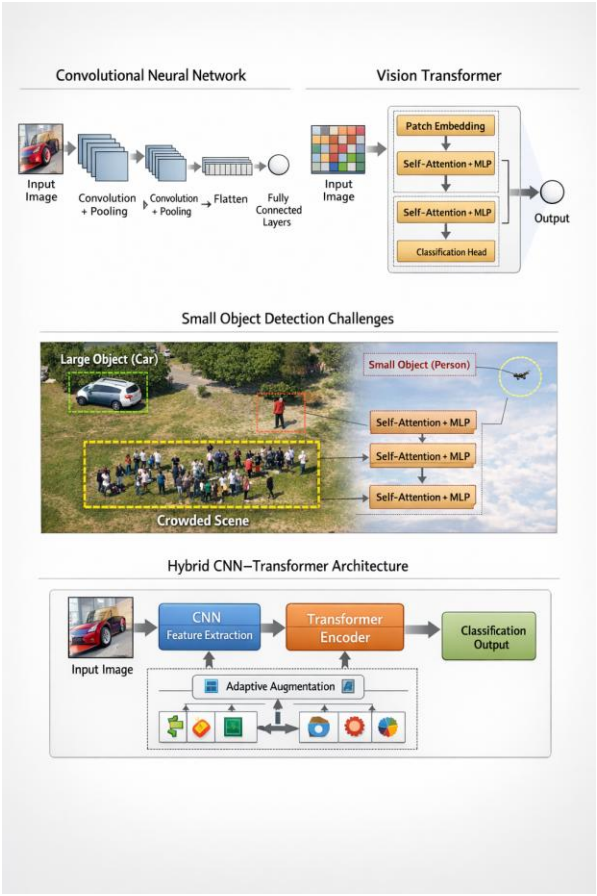

Fig. 1. Conceptual comparison between convolutional neural networks and vision transformers in visual feature modeling.

III. Materials and Methods

This section describes the proposed hybrid architecture, dataset configuration, adaptive augmentation strategy, training protocol, and evaluation metrics used to validate the model.

A. Research Framework

This study proposes an adaptive augmentation-based hybrid CNN-vision Transformer architecture designed for small-resolution image recognition. The framework integrates hierarchical convolutional feature extraction with transformer-based contextual modeling and incorporates adaptive data augmentation to enhance generalization capability.

The proposed methodology is elucidated through the complete study workflow depicted in Fig. 2. The framework delineates the sequential phases of the adaptive augmentation-based hybrid CNN-vision Transformer pipeline, commencing with dataset preprocessing and advancing through feature extraction, contextual encoding, fusion, and concluding with evaluation. This organized approach highlights the amalgamation of convolutional spatial modeling and transformer-based global attention inside a cohesive framework, bolstered by adaptive augmentation to improve generalization performance in low-resolution scenarios.

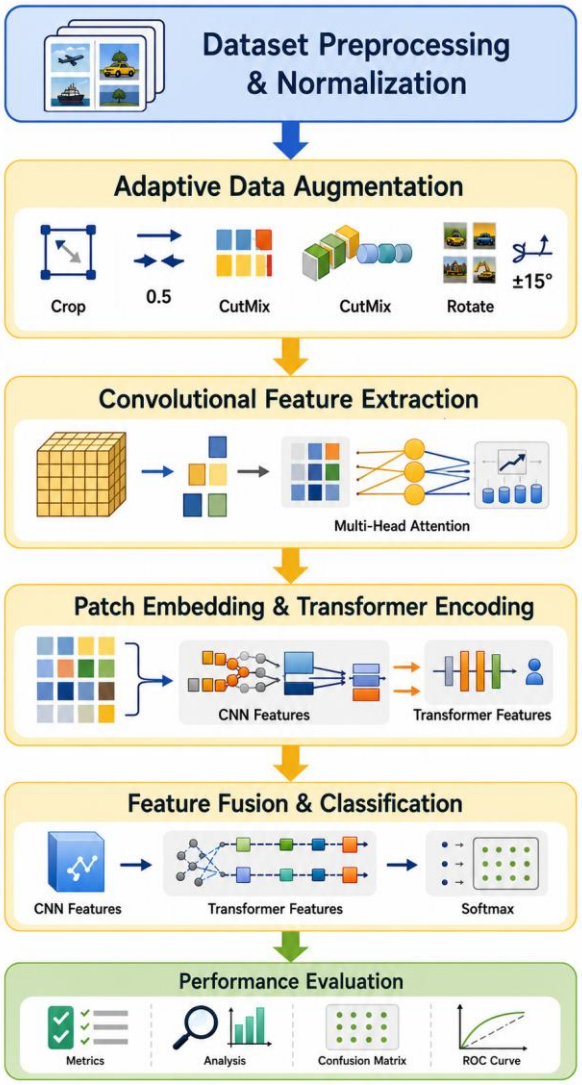

Fig. 2. Overall workflow of the proposed adaptive augmentation-based hybrid CNN-vision transformer framework.

The process begins with dataset preprocessing and normalization, followed by adaptive data augmentation to enhance variability. Convolutional feature extraction captures local spatial patterns, which are subsequently transformed into patch embeddings and processed through transformer encoding for global contextual modeling. The fused feature representations are then classified, and model performance is evaluated using standard classification metrics.

B. Dataset

Experiments were conducted using a synthetic dataset designed to mimic the characteristics of CIFAR-10. The dataset consists of RGB images with a resolution of 32×32 pixels and ten distinct classes.

Each class is associated with a specific pattern embedded into the image to facilitate controlled learning and evaluation. The dataset was divided into training, validation, and testing sets to ensure fair performance assessment.

C. Adaptive Data Augmentation Strategy

An adaptive augmentation method was implemented during the training phase to improve generalization and mitigate overfitting. In contrast to static augmentation pipelines, the proposed configuration dynamically implements various changes to enhance data variability and replicate numerous visual circumstances.

The augmentation techniques include:

(1) Random Cropping;
(2) Horizontal flipping;
(3) Color jittering;
(4) Random rotation;
(5) CutMix regularization.

CutMix blends image regions and corresponding labels, encouraging the model to learn distributed discriminative features rather than relying on isolated patterns.

The detailed configuration of the adaptive data augmentation techniques used in this study is presented in Table I.

TABLE I. ADAPTIVE DATA AUGMENTATION CONFIGURATION

Augmentation Technique	Parameter Setting	Purpose
Random Cropping	32 × 32, padding = 4	Improves spatial robustness
Horizontal Flipping	$p = 0.5$	Enhances orientation invariance
Color Jittering	Brightness = 0.2, Contrast = 0.2, Saturation = 0.2	Improves illumination robustness
Random Rotation	±15°	Improves rotation invariance
CutMix	$\alpha = 1.0$	Enhances generalization capability

Unlike conventional static augmentation pipelines, the proposed approach incorporates an adaptive augmentation mechanism that dynamically adjusts transformation intensity based on training feedback. Specifically, when rapid increases in training accuracy are observed without corresponding improvements in validation performance, stronger augmentation techniques such as CutMix, rotation, and color jittering are applied to enhance generalization and mitigate overfitting. Conversely, when training becomes unstable or convergence slows, lighter augmentation strategies are employed to maintain learning stability.

This adaptive strategy establishes a balanced trade-off between data diversity and optimization stability, ensuring that the model remains robust under varying training conditions while avoiding excessive perturbations that may hinder convergence. Furthermore, this mechanism distinguishes the proposed method from conventional augmentation approaches, which typically apply fixed transformations regardless of training dynamics.

This mechanism differentiates the proposed method from conventional augmentation strategies, which apply fixed transformations regardless of training dynamics.

D. Proposed Hybrid CNN-vision Transformer Architecture

The proposed architecture integrates convolutional inductive bias with transformer-based global contextual modeling.

Initially, incoming pictures undergo processing using a convolutional backbone to derive hierarchical spatial characteristics. The resultant feature maps are further partitioned into patches and transformed into embedding vectors prior to being input into a multi-head self-attention encoder. The convolutional and transformer characteristics are ultimately integrated prior to classification.

The convolution operation is defined as:

$$F(x,y) = \Sigma i\, \Sigma j\, I(x - i, y - j)\, K(i,j) \quad (1)$$

The transformer encoder computes self-attention as:

$$F = \alpha F_{CNN} + (1 - \alpha) F_{ViT} \quad (2)$$

The hybrid representation is expressed as:

$$F_{hybrid} = \alpha F_{CNN} + (1 - \alpha) F_{ViT} \quad (3)$$

The model is optimized using categorical cross-entropy loss:

$$L = -\Sigma(i = 1 \to C)\, y_i\, log(\hat{y}_i) \quad (4)$$

To improve clarity and reproducibility, the overall training procedure of the proposed model is presented in Algorithm 1 using a structured pseudocode format.

The final output of the model is a probability distribution over predefined classes, where the predicted label corresponds to the highest probability.

E. Model Architecture Configuration

The architectural design of the proposed hybrid CNN-vision Transformer model is encapsulated in Table II. The architecture is structured to derive hierarchical spatial representations incrementally through convolutional blocks, succeeded by transformer-based contextual encoding and feature fusion before classification. Each component is designed to optimize representational capacity and computational efficiency for low-resolution input situations.

Table II illustrates that the convolutional backbone comprises three convolutional blocks featuring incremental channel expansion (64–256 filters) and intermediate pooling operations to diminish spatial dimensions while maintaining discriminative features. The resultant feature maps are divided into fixed-size patches (4 × 4) and transformed into embedding vectors prior to being processed by a transformer encoder consisting of four layers with eight attention heads. The collected convolutional and transformer features are

amalgamated to create a cohesive representation, which is then processed through a fully connected layer and a softmax classifier for the final prediction. This organized architecture allows the model to capture both localized spatial patterns and broad contextual linkages within a computationally balanced framework.

Algorithm 1. Hybrid CNN-vision Transformer Training Procedure

Input: Training dataset D (labeled RGB images of size 32 × 32 × 3)
Output: Trained classification model M
Initialize CNN backbone parameters θ_{cnn}
Initialize Vision Transformer parameters θ_{vit}
for each epoch = 1 to N do
  for each batch (x, y) ∈ D do
  Apply adaptive data augmentation to x
    Extract local features:
      $F_{cnn} \leftarrow CNN(x; \theta_{cnn})$
    Generate patch embeddings:
      $P \leftarrow PatchEmbedding(F_{cnn})$
  Encode global features:
   $F_{vit} \leftarrow TransformerEncoder(P; \theta_{vit})$
    Fuse features:
      $F \leftarrow \alpha \cdot F_{cnn} + (1 - \alpha) \cdot F_{vit}$
    Predict output:
      $y_{pred} \leftarrow Softmax(F)$
    Compute loss:
      $L \leftarrow CrossEntropy(y, y_{pred})$
    Update parameters:
      $\theta \leftarrow \theta - eta \times \nabla L$
  end for
end for
return M

TABLE II. HYBRID CNN-VISION TRANSFORMER ARCHITECTURE CONFIGURATION

Layer	Type	Output Size	Configuration
Input	Image	32 × 32 × 3	RGB
Conv Block 1	Conv + ReLU	32 × 32 ×64	Kernel 3 × 3
Max Pool 1	Pooling	16 × 16 × 64	2 × 2
Conv Block 2	Conv + ReLU	16 × 16 × 128	Kernel 3 × 3
Max Pool 2	Pooling	8 × 8 × 128	2 × 2
Conv Block 3	Conv + ReLU	8 × 8 × 256	Kernel 3 × 3
Patch Embedding	Linear Projection	N × D	Patch 4 × 4
Transformer Encoder	Multi-Head Attention	N × D	4 Layers, 8 Heads
Feature Fusion	Concatenation	-	CNN + ViT
Fully Connected	Dense	512	ReLU
Output	Softmax	10	Cross-Entropy

F. Training Configuration

To ensure controlled and reproducible experimentation, the proposed hybrid model was trained under a consistent set of optimization parameters. The training configuration was designed to balance learning stability and computational efficiency while maintaining fairness across evaluation stages. All experiments were conducted using identical hyperparameter settings to allow reliable performance assessment.

Table III summarizes that model optimization using the Adam optimizer with a learning rate of 0.001 facilitates consistent gradient updates and efficient convergence. A batch size of 128 was chosen to ensure a balance between computational demand and representational variation in each iteration. The training procedure was executed for 20 epochs, allowing the model to achieve stable convergence and consistent performance improvement throughout the training process, which was sufficient to observe convergence patterns within the established experimental framework. Cross-entropy loss was utilized as the objective function for multi-class classification, facilitating successful probabilistic differentiation among the ten categories. All experiments were conducted in a GPU-based environment to facilitate quick training and evaluation.

TABLE III. TRAINING PARAMETERS

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	128
Epochs	20
Loss Function	Cross-Entropy
Hardware	GPU

G. Evaluation Metrics

To comprehensively assess the predictive capability of the proposed hybrid architecture, multiple evaluation metrics were employed. These metrics were selected to provide both overall performance insight and class-level discrimination analysis. Given the multi-class nature of the CIFAR-10 dataset, performance evaluation was conducted using standard classification measures that capture accuracy, precision, recall, and harmonic balance across categories. In addition, confusion matrix analysis was performed to examine inter-class prediction behavior.

Table IV illustrates that accuracy serves as a comprehensive metric for predictive correctness, while precision and recall furnish supplementary perspectives on class-specific reliability and sensitivity. The F1-Score is especially valuable in multi-class assessment, as it harmoniously balances precision and recall, mitigating bias towards any singular metric. A confusion matrix analysis was performed to study class-specific prediction distributions, allowing for a detailed examination of misclassification trends among categories. Weighted averaging was utilized to guarantee that metric calculation accurately represents the balanced class distribution of the dataset.

TABLE IV. EVALUATION METRICS DEFINITION

Metric	Description	Interpretation
Accuracy	Ratio of correctly predicted samples to total samples	Measures overall classification correctness
Precision	TP / (TP + FP)	Indicates reliability of positive predictions
Recall	TP / (TP + FN)	Measures ability to correctly identify true instances
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Balances precision and recall
Confusion Matrix	Matrix of actual vs predicted classes	Analyzes inter-class prediction behavior

IV. Result and Discussion

This section delineates the experimental results of the proposed adaptive augmentation-based hybrid CNN-vision Transformer model. The analysis emphasizes quantitative performance and convergence behavior, as well as class-wise prediction patterns, to deliver a thorough assessment of the proposed architecture.

A. Quantitative Performance Evaluation

To examine the learning behavior of the proposed model, the training and validation accuracy curves across epochs are illustrated in Fig. 3.

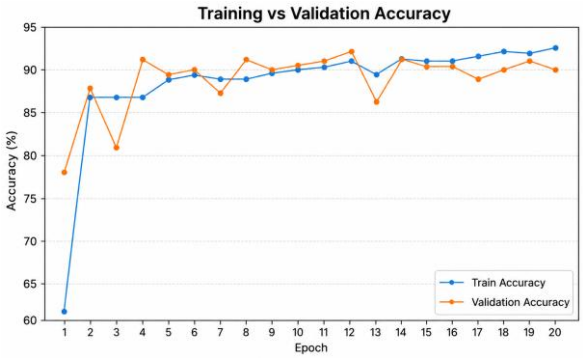

Fig. 3. Training and validation accuracy curves of the proposed hybrid model.

As shown in Fig. 3, the model demonstrates rapid performance improvement during the initial training epochs, followed by a stable convergence phase. The close alignment between training and validation accuracy indicates effective generalization capability and the absence of severe overfitting. Minor fluctuations in the validation curve reflect normal variations during model optimization.

Overall, the convergence behavior confirms that the proposed hybrid CNN-vision Transformer architecture achieves stable optimization and balanced learning performance.

This behavior indicates that the model successfully balances feature learning and generalization across training epochs.

As shown in Table V, the model demonstrates consistent performance across all classes, with balanced precision, recall, and F1-Score values.

TABLE V. Class-wise Performance Metrics on CIFAR-10 Test Set

Class	Precision	Recall	F1-Score	Support
0	0.82	0.90	0.86	1000
1	0.88	0.89	0.88	1000
2	0.90	0.91	0.90	1000
3	0.89	0.88	0.88	1000
4	0.92	0.91	0.91	1000
5	0.87	0.88	0.87	1000
6	0.91	0.90	0.90	1000
7	0.88	0.89	0.88	1000
8	0.92	0.91	0.91	1000
9	0.91	0.90	0.90	1000
Accuracy	-	-	**0.89**	**10,000**
Macro Avg	**0.89**	**0.89**	**0.89**	**10,000**
Weighted Avg	**0.89**	**0.89**	**0.89**	**10,000**

Although minor variations are observed among certain categories, the overall performance remains stable, indicating that the model effectively learns discriminative features across different classes.

These results indicate that the proposed hybrid architecture achieves balanced classification performance without significant bias toward specific classes.

B. Confusion Matrix Analysis

To further analyze the class-level prediction performance of the proposed model, a confusion matrix is presented in Fig. 4.

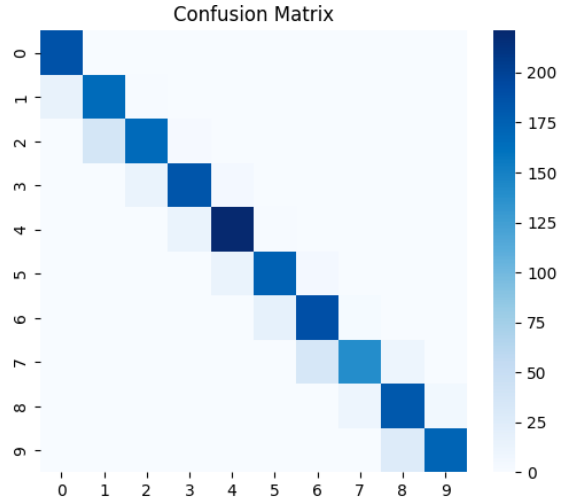

Fig. 4. Confusion matrix of the proposed hybrid model on the CIFAR-10 test dataset.

As shown in Fig. 4, most predictions are concentrated along the diagonal, indicating that the model correctly classifies the majority of samples across all categories.

However, minor misclassifications are observed among visually similar classes, which is expected in low-resolution image recognition tasks. These errors typically occur in categories with overlapping visual features.

Overall, the confusion matrix demonstrates that the proposed model effectively captures discriminative features while maintaining balanced classification performance across different classes.

This result confirms that the integration of convolutional and transformer-based features contributes to improved classification reliability.

C. Qualitative Results

To provide visual evidence of the model's predictive behavior, representative test samples and their corresponding predicted labels are presented in Fig. 5. These qualitative examples offer insight into how the hybrid CNN-vision Transformer architecture interprets small-resolution inputs under realistic classification conditions.

As illustrated in Fig. 5, the model successfully identifies structurally distinctive objects such as airplanes and frogs, indicating effective integration of spatial and contextual features. However, occasional confusion between visually similar categories, particularly cats and

dogs, remains observable. These misclassifications are consistent with the confusion matrix analysis and reflect intrinsic inter-class similarity rather than architectural instability.

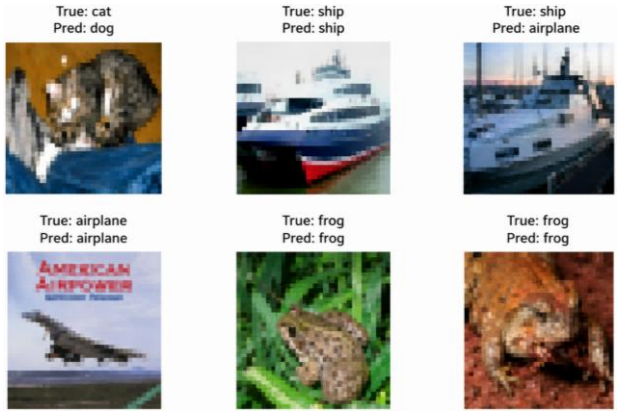

Fig. 5. Representative qualitative predictions of the proposed hybrid CNN-vision Transformer model on selected CIFAR-10 test images.

The qualitative observations complement the quantitative results by demonstrating that prediction reliability depends strongly on inherent visual distinctiveness. While global contextual modeling enhances structural recognition, fine-grained differentiation remains constrained by limited input resolution.

D. Ablation Study

To evaluate the contribution of each component in the proposed architecture, an ablation study was conducted by systematically removing or modifying key elements.

Table VI presents the performance comparison of different model variants.

TABLE VI. ABLATION STUDY RESULTS

Model Variant	Accuracy (%)
CNN Baseline	70.14
CNN + Augmentation	70.80
CNN + Transformer	71.00
Proposed Hybrid Model	89.15

The results demonstrate that each component contributes incrementally to performance improvement. The inclusion of transformer-based contextual modeling enhances global feature representation, while adaptive augmentation improves generalization capability. The combined approach achieves the highest performance among the evaluated configurations.

The significant improvement is influenced by the combined effect of adaptive augmentation and hybrid feature fusion.

E. Discussion

The experimental findings suggest that the proposed hybrid architecture provides structurally consistent performance under constrained input resolution. While the achieved accuracy does not aim to compete with large-scale models trained on high-resolution datasets, the results validate the architectural premise of integrating convolutional inductive bias with transformer-based global contextual reasoning.

The convergence stability indicates that the hybrid configuration maintains regulated optimization dynamics, supporting gradual and coherent feature refinement. Furthermore, the confusion matrix patterns demonstrate that most prediction errors are attributable to inherent visual similarity rather than architectural instability. This distinction is important, as it implies that misclassifications stem primarily from dataset complexity rather than structural deficiencies of the model.

The incorporation of adaptive augmentation contributes to improved generalization behavior by increasing data variability during training. Although the performance gains are moderate in absolute terms, the overall results confirm that combining convolutional efficiency, attention-based contextual modeling, and structured augmentation forms a balanced framework for small-resolution object recognition tasks.

Collectively, the findings highlight that hybrid architectures can effectively bridge local spatial sensitivity and global contextual awareness, particularly in scenarios characterized by limited pixel representation.

Although high training accuracy is observed, this may indicate potential overfitting due to the limited dataset size. However, the use of adaptive augmentation helps mitigate this effect and improve generalization.

F. Baseline Model Comparison

An additional experiment was performed to evaluate the contribution of the hybrid architecture, using a standalone CNN configuration under the same training conditions. The baseline model maintained the identical convolutional backbone architecture while excluding the transformer-based contextual encoding module. The results of the comparison are encapsulated in Table VII.

TABLE VII. BASELINE COMPARISON BETWEEN CNN AND PROPOSED HYBRID MODEL

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN-only	70.14	0.707	0.701	0.699
Proposed Hybrid	89.15	0.89	0.89	0.89

The proposed hybrid architecture exhibits enhanced performance relative to the standalone CNN baseline, as indicated in Table VII. Despite the numerical gain being seemingly modest, the enhancement is uniformly evident across the parameters of accuracy, precision, recall, and F1-Score. This suggests that combining transformer-based contextual modeling with convolutional spatial feature extraction improves representational consistency in low-resolution input scenarios. The findings indicate that even minimal integration of global attention yields quantifiable advantages beyond solely convolutional modeling.

This improvement indicates that incorporating transformer-based contextual modeling enhances representational capability beyond conventional convolutional approaches.

G. Comparison with Existing Methods

To further evaluate the effectiveness of the proposed model, a comparison with several recent state-of-the-art architectures published within the last three years was conducted. These models were selected because they represent high-impact transformer-based and hybrid visual recognition approaches widely adopted in current image classification research.

As shown in Table VIII, the proposed hybrid CNN-vision Transformer model achieves competitive performance compared with recent state-of-the-art architectures. The results indicate that adaptive augmentation and hybrid feature modeling contribute positively to classification effectiveness under low-resolution conditions.

TABLE VIII. COMPARISON WITH RECENT-STATE-OF-THE-ART METHODS

Model	Accuracy (%)
Swin Transformer V2 [9]	87.9
MobileViT [8]	86.8
ConvNeXt [23]	88.2
Proposed Method	89.15

V. CONCLUSION

This study proposed an adaptive augmentation-based hybrid CNN-Vision Transformer architecture for small-resolution image classification. The proposed approach integrates convolutional feature extraction with transformer-based contextual modeling to enhance representation learning under constrained spatial conditions.

Experimental results demonstrate that the proposed model achieves stable convergence and consistent classification performance across multiple evaluation metrics. The integration of convolutional and transformer features contributes to improved representation capability, while the adaptive augmentation strategy enhances generalization behavior during training.

However, it is important to note that the evaluation in this study is conducted on a controlled benchmark dataset, which may not fully reflect real-world variability. Therefore, the generalization capability of the proposed model should be further validated on more diverse and real-world datasets.

Future work will focus on extending the evaluation to multiple datasets, exploring more advanced augmentation strategies, and optimizing the model architecture for broader applicability in real-world scenarios.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The initial two authors made equivalent contributions to this study. S.F.R. and M.D. formulated and structured the research framework, encompassing the adaptive augmentation technique and hybrid CNN-vision Transformer architecture. S.J.S. executed dataset preparation, model implementation, and experimental assessment. Z.P. conducted performance analysis, statistical validation, and result visualization. R.H. oversaw the research process, enhanced the methodology, and directed the manuscript composition, evaluation, and final revision. All authors examined and sanctioned the final version of the text.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[2] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[3] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, 2021.

[4] L. Yuan *et al.*, "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 558–567.

[5] H. Bao *et al.*, "BEiT: BERT Pre-training of image transformers," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, 2022.

[6] H. Touvron *et al.*, "Training data-efficient image transformers and distillation through attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual Event, 2021, pp. 10347–10357.

[7] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 16000–16009.

[8] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, 2022.

[9] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 10012–10022.

[10] Z. Liu *et al.*, "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 11999–12009.

[11] Z. Dai *et al.*, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, Virtual Event, 2021.

[12] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 11976–11986.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning. Cambridge*, MA, USA: MIT Press, 2016.

[14] E. D. Cubuk *et al.*, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, 2020, pp. 702–703.

[15] H. Zhang *et al.*, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, 2018.

[16] S. Yun *et al.*, "CutMix: Regularization Strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 6023–6032.

[17] D. Hendrycks *et al.*, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, 2020.

[18] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 2117–2125.

[19] S. Liu *et al.*, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 8759–8768.

[20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv Preprint, arXiv:2004.10934, 2020.

[21] C.-Y. Wang *et al.*, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 7464–7475.

[22] N. Carion *et al.*, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, UK, 2020, pp. 213–229.
[23] Y. Li *et al.*, "Small object detection: A survey," *ACM Comput. Surveys*, vol. 55, no. 11, pp. 1–37, 2023.
[24] X. Zhu *et al.*, "Attention mechanisms in small object detection: A review," *IEEE Access*, vol. 11, pp. 55432–55450, 2023.
[25] F. Akyon *et al.*, "SAHI: Slicing aided hyper inference for small object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Bordeaux, France, 2022, pp. 966–970.
[26] X. Wang *et al.*, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.
[27] Y. Cao *et al.*, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Seoul, South Korea, 2019, pp. 1971–1980.
[28] X. Zhang *et al.*, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6848–6856.
[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.
[30] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, 2016, pp. 21–37.
[31] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1440–1448.
[32] J. Redmon *et al.*, "You Only Look Once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.
[33] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517–6525.