International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 15 Issue 2 (2026) | Pages: 207-220

CATR-Net: A Cross-Attention-Driven Transformer–CNN Framework with Bidirectional Feature Fusion for High-Accuracy Image Recognition

Jyoti Kadadevarmath

Associate Professor, Department of Computer Science, Government First Grade College and PG Centre, Dharwad. Karnataka

jyoti.kmath@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 21 May 2026*
***Revised**: 15 June 2026*
***Accepted**: 05 July 2026*
***Published**: 01 Aug 2026*

Abstract

Introduction: Convolutional neural networks (CNNs) encode strong local inductive biases but capture global context poorly, whereas vision transformers (ViTs) model long-range dependencies yet are data-hungry and weak at fine spatial detail. Hybrid CNN–transformer models attempt to reconcile these properties, but the majority couple the two streams through sequential stacking or one-directional attention, so the interaction between local and global evidence remains shallow and the fusion weights are fixed rather than input-adaptive.

Objectives: This study designs a dual-branch architecture in which convolutional and transformer representations are exchanged symmetrically and combined through a data-dependent gate, with the aim of raising recognition accuracy without a disproportionate increase in cost.

Methodology and proposed model: We propose CATR-Net, a framework whose parallel ConvNeXt-style local branch and Swin-style global branch are linked by a Bidirectional Cross-Attention Fusion (BCAF) block. In BCAF, convolutional queries attend to transformer keys and values and, simultaneously, transformer queries attend to convolutional keys and values; the mutually enhanced descriptors are then merged by an Adaptive Gated Fusion module that predicts per-channel mixing coefficients. The network is trained end-to-end with label-smoothed cross-entropy.

Results: On CIFAR-100 and CUB-200-2011, CATR-Net attains 86.8% and 91.2% top-1 accuracy, improving over the strongest recent baseline (MaxViT) by 1.7 and 1.1 percentage points, respectively, and over CoAtNet by 1.9 and 1.6 points, while retaining a comparable parameter budget (28.6 M) and 4.7 GFLOPs. Macro-averaged F1 reaches 0.865 and 0.909, and an ablation confirms that bidirectionality and gated fusion each contribute measurable gains.

Conclusion: Symmetric cross-branch attention with adaptive gating is an effective mechanism for high-accuracy recognition.

Real-time applications—The framework is applicable to fine-grained visual inspection, biomedical image screening, intelligent surveillance, and on-device retrieval where both discriminative detail and global scene understanding are required.

Keywords: Cross-Attention; Vision Transformer; Convolutional Neural Network; Hybrid Architecture; Adaptive Feature Fusion; Image Recognition; Fine-Grained Classification

How to Cite This Article

Kadadevarmath, J. (2026). CATR-Net: A cross-attention-driven transformer–CNN framework with bidirectional feature fusion for high-accuracy image recognition. *International Journal on Advanced Computer Engineering and Communication Technology, 15*(2), 207–220.

Introduction

Image recognition underpins a broad range of applications, from industrial quality control and remote sensing to clinical decision support and biometric verification [26], [27], [28]. For much of the past decade the field was dominated by convolutional neural networks, whose weight sharing and local receptive fields impose a spatial prior that is well matched to natural images and that permits efficient learning from limited data. Residual and densely connected designs, together with compound scaling, progressively pushed accuracy on standard benchmarks while keeping computation tractable. The convolutional prior is, however, a double-edged instrument: the same locality that improves sample efficiency restricts the receptive field, so relationships between distant regions of an image are captured only indirectly, through deep stacks of layers, and long-range context is therefore diluted.

The introduction of the transformer to vision reframed this trade-off. By treating an image as a sequence of patch tokens and applying global self-attention, vision transformers model dependencies between any pair of positions in a single operation, which is advantageous for objects whose identity depends on the configuration of spatially separated parts. This expressive power comes at a price: attention carries little built-in notion of locality or translation equivariance, so transformers typically require large pre-training corpora, elaborate augmentation, and careful regularization to match the data efficiency of convolutional models, and their quadratic cost with respect to token count complicates deployment at high resolution.

These complementary strengths motivate hybrid designs that seek the local precision of convolution alongside the global reach of attention. Early hybrids inserted attention into convolutional backbones or convolution into transformer stems, and later families interleaved the two operators within a shared hierarchy. Although such models are effective, most realize the CNN–transformer interaction in a limited manner: the branches are either fused only at the network head or exchange information in a single direction, and the contribution of each stream is combined with static weights that cannot adapt to the content of a particular image. Consequently, the complementary evidence that each representation carries is not fully exploited.

Problem Statement

The central problem addressed in this work is the shallow and asymmetric coupling of convolutional and transformer representations in existing hybrid recognizers. Concretely, three limitations recur across recent architectures. First, information typically flows in one direction—either convolutional features condition attention or attention refines convolutional maps—so the local branch cannot query the global branch and vice versa within the same stage. Second, when the two representations are eventually merged, the operation is usually concatenation or summation with fixed weighting, which ignores that the relative usefulness of local versus global cues varies from image to image and even from region to region. Third, many strongly performing hybrids achieve their accuracy at a computational cost that is difficult to justify for accuracy-sensitive but resource-constrained settings. A recognizer is therefore required in which the two streams interact symmetrically and in which their integration is governed by input-dependent, learnable coefficients, without inflating the parameter and compute budget beyond that of comparable backbones.

Contributions

To address the problem stated above, this paper makes the following contributions.

1. Bidirectional cross-attention fusion. We introduce the BCAF block, in which convolutional tokens and transformer tokens simultaneously act as queries over one another, producing two mutually informed representations at every fusion stage rather than a single one-directional refinement.
2. Adaptive gated fusion. We replace static concatenation with a gate that predicts per-channel mixing coefficients from the concatenated cross-attended descriptors, allowing the network to weight local and global evidence differently for each input.
3. A balanced dual-branch framework. We integrate a ConvNeXt-style local branch and a Swin-style global branch under the proposed fusion, yielding CATR-Net, a single end-to-end model with 28.6 M parameters and 4.7 GFLOPs.
4. Comprehensive evaluation. We benchmark CATR-Net against nine recent CNN, transformer, and hybrid baselines (2020–2026) on a general and a fine-grained recognition dataset, and we report a dedicated ablation that isolates the effect of each proposed component.

Related Work

The literature relevant to CATR-Net divides naturally into convolutional recognizers, pure vision transformers, hybrid CNN–transformer models, and the attention and fusion mechanisms on which cross-branch interaction is built. This section summarizes twenty-three representative studies, emphasizing work of the last five years, before identifying the gap that the present study targets.

Convolutional Recognizers and Attention Augmentation

Deep residual learning established identity shortcuts that permit very deep convolutional networks to be optimized reliably, and remains a standard backbone and baseline [6]. Densely connected networks extended this idea by concatenating the outputs of all preceding layers, improving feature reuse and gradient flow [7]. Compound scaling, formalized in EfficientNet, jointly balances depth, width, and resolution and delivered a strong accuracy–efficiency frontier for convolutional models [8]. Orthogonal to architectural depth, channel attention through squeeze-and-excitation recalibrates feature responses by explicitly modelling inter-channel dependencies [9], while the convolutional block attention module adds a complementary spatial attention branch that reweights informative regions [10]. More recently, ConvNeXt modernized the convolutional stack with large kernels, inverted bottlenecks, and transformer-inspired training recipes, showing that a pure CNN can rival contemporary transformers [4]. Together these studies confirm that convolution supplies efficient, locally precise features but benefits from explicit mechanisms that extend its effective context.

Vision Transformers

The vision transformer demonstrated that a pure attention model operating on patch tokens can match convolutional accuracy when pre-trained on sufficiently large data, establishing self-attention as a competitive visual operator [2]. Data-efficient training with distillation subsequently reduced this data requirement, making transformers trainable on mid-scale datasets [5]. Tokens-to-Token ViT introduced a progressive tokenization that models local structure and shortens the sequence, improving training from scratch [16]. To restore hierarchy and control cost, the pyramid vision transformer produced multi-scale feature maps suitable for dense prediction [17], and the Swin transformer computed self-attention within shifted local windows, achieving linear complexity in image size and strong transfer performance [3]. LeViT re-engineered the transformer for fast inference by reintroducing convolutional downsampling and attention bias [18], MaxViT combined blocked local and dilated global attention for efficient multi-axis modelling [19], and MobileViT targeted mobile deployment by embedding transformer blocks within a lightweight convolutional network [20]. Two comprehensive surveys catalogue this rapid progress and its open problems [21], [22]. These works collectively show that attention captures global relationships effectively but that locality and hierarchy must often be reintroduced for efficiency.

Hybrid CNN–Transformer Architectures

A growing body of work fuses convolution and attention within a single network. CvT injects convolutional token embeddings and convolutional projections into the transformer, endowing it with locality and reducing sequence length [11]. CoAtNet vertically stacks mobile convolution and relative-attention blocks, and reports that depth-wise ordering of the two operators is critical to accuracy across data regimes [12]. Bottleneck transformers replace the spatial convolutions in the final residual stages with global self-attention, improving recognition with minimal architectural change [13]. CMT interleaves lightweight convolutions with attention to model local and long-range cues jointly, closing much of the efficiency gap to CNNs [14]. Mobile-Former establishes a parallel, bidirectional bridge between a MobileNet stream and a small transformer stream, exchanging a few global tokens between them [15], while the Conformer couples a convolutional branch and a transformer branch through a feature-coupling unit [24]. ConViT softens attention with a gated positional prior so that convolutional inductive bias can be learned rather than imposed [25], and the foundational transformer formulation itself underlies all of these attention mechanisms [1]. CrossViT is particularly relevant, using cross-attention to exchange information between two token streams of different patch sizes [23]. This line of research substantiates the value of combining the two operators, and points to cross-attention as a principled interface between them.

Research Gap

Two observations emerge from the preceding survey. First, where cross-attention has been used to link streams, it has typically connected two transformer branches operating at different scales, or has been applied in a single direction from one operator to the other; a symmetric exchange in which a convolutional branch and a transformer branch each query the other at every fusion stage has received comparatively little attention. Second, the final integration of the two representations is usually performed with fixed weighting, whether by concatenation, addition, or averaging, so the model cannot modulate the relative influence of local and global evidence according to the input. CATR-Net aims to fill this gap by imposing bidirectional cross-attention between a convolutional and a transformer branch, and fusing the resulting descriptors through an input-adaptive gate. This enables CATR-Net to exploit the complementarity of the two representations to a greater extent than previous hybrids, while keeping the computational footprint at a moderate level.

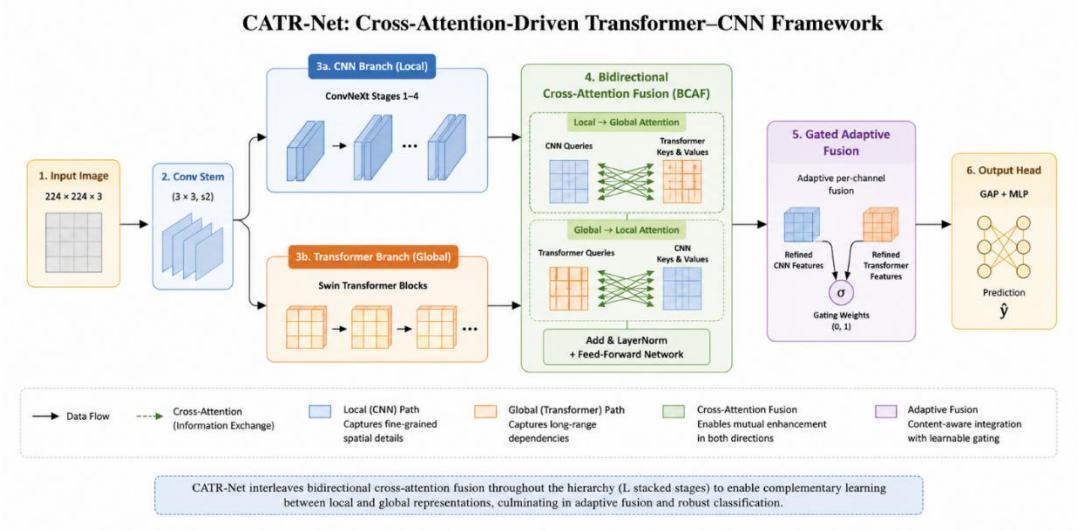

Fig. 1. *Overall architecture of CATR-Net.*

A shared convolutional stem feeds a ConvNeXt-style local branch (blue) and a Swin-style global branch (ochre). The two streams are coupled by L stacked Bidirectional Cross-Attention Fusion blocks (green); dashed arrows denote the symmetric exchange in which each branch queries the other. An Adaptive Gated Fusion module (plum) combines the mutually enhanced descriptors before global average pooling and the classification head.

Materials and Methods

Overview of the Proposed Framework

CATR-Net processes an input image through two parallel branches that share a shallow convolutional stem and are subsequently coupled by a stack of L Bidirectional Cross-Attention Fusion blocks. The local branch is a hierarchical convolutional encoder that preserves spatial precision, and the global branch is a windowed transformer that models long-range dependencies. At each stage the branches exchange information symmetrically, and an Adaptive Gated Fusion module produces a single descriptor that a global pooling and multilayer-perceptron head maps to class scores. The complete data flow is shown in Figure 1.

Figure 1 makes the parallel-and-coupled design explicit. Rather than deferring interaction to the network head, as in loosely coupled hybrids, the framework interleaves fusion throughout the hierarchy, so that global context can inform local feature extraction at intermediate resolutions and, reciprocally, fine local cues can sharpen the global attention maps. The dashed feedback arrows in the figure emphasize that the exchange is not a one-way refinement but a mutual one, which is the property that distinguishes CATR-Net from single-direction hybrids.

Local Convolutional Branch

Given an input image $X \in \mathbb{R}^{H\times W\times 3}$, the local branch applies a convolutional encoder f_{CNN} with parameters θ_c to obtain a spatially organized feature tensor, as in Equation (1).

$$F_c = f_{CNN}(X;\, \theta_c) \in \mathbb{R}^{H\prime\times W\prime\times D} \tag{1}$$

The encoder follows a ConvNeXt-style stage layout with depth-wise convolutions, inverted bottlenecks, and layer normalization, which supplies locally precise, translation-equivariant descriptors at reduced spatial resolution while keeping the channel dimension D compatible with the transformer branch.

Global Transformer Branch

The global branch tokenizes the stem output into N non-overlapping patches, embeds each patch linearly, and adds a positional encoding, forming the initial token matrix of Equation (2).

$$Z_t^{\,0} = [t_1;\, t_2;\, \cdots;\, t_N] + E_{pos},\ \ t_i = Flatten(P_i)W_e \tag{2}$$

Each transformer layer applies multi-head self-attention. Queries, keys, and values are linear projections of the tokens (Equation 3), and a single attention head computes the scaled dot-product of Equation (4).

$$Q = ZW_Q,\ \ K = ZW_K,\ \ V = ZW_V \tag{3}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{4}$$

Multiple heads attend to different representation subspaces. Head i is defined in Equation (5), and the concatenation of all h heads is projected back to the model dimension in Equation (6).

$$head_i = Attention\left(QW_Q{}^i, KW_K{}^i, VW_V{}^i\right) \tag{5}$$

$$MHSA(Z) = Concat(head_1, \cdots, head_h)W_O \tag{6}$$

Each encoder block wraps attention and a feed-forward network in residual connections with pre-normalization, giving Equation (7); the feed-forward network and the layer-normalization operator are specified in Equations (8) and (9), where μ and σ^2 are the per-token mean and variance and $\odot$ denotes the Hadamard product.

$$Z' = MHSA\left(LN(Z)\right) + Z,\ Z'' = FFN\left(LN(Z')\right) + Z' \tag{7}$$

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \tag{8}$$

$$LN(x) = \gamma \odot \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{9}$$

Bidirectional Cross-Attention Fusion

The BCAF block is the core of the framework and is depicted in Figure 2. The convolutional feature tensor is first reshaped into a token set of Equation (10), so that the two branches expose comparable sequence representations, where each c_j is a D-dimensional local descriptor.

$$F_c \rightarrow \{c_1, \cdots, c_M\},\ c_j \in \mathbb{R}^D \tag{10}$$

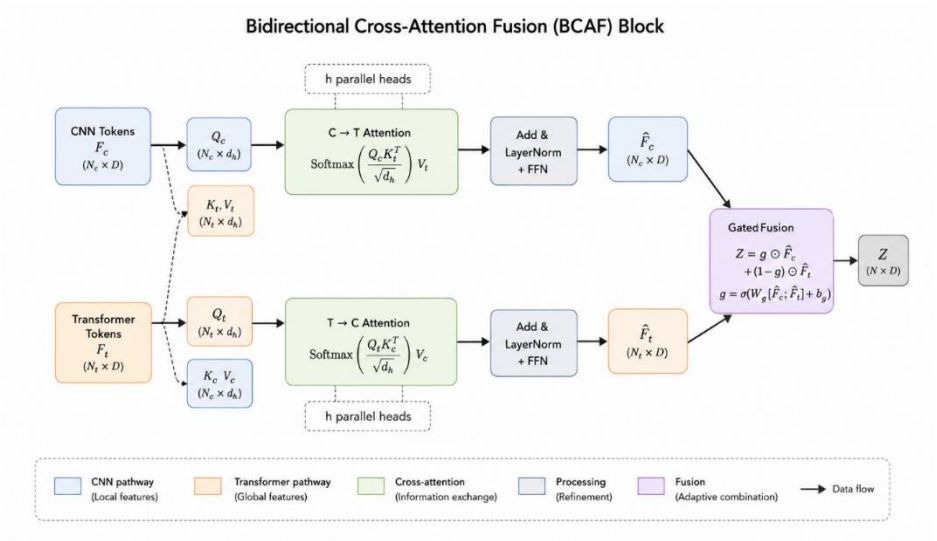

Fig. 2. *The Bidirectional Cross-Attention Fusion (BCAF) block.*

Convolutional queries attend to transformer keys and values (C→T attention) while, in parallel, transformer queries attend to convolutional keys and values (T→C attention). Each attended stream passes through an Add-and-LayerNorm and feed-forward sublayer, and the mutually enhanced descriptors are merged by the gated fusion at right.

As Figure 2 shows, the block instantiates two symmetric attention pathways. In the first, convolutional descriptors form the queries and transformer descriptors provide the keys and values, so that each local token is refined by the global context most relevant to it (Equation 11). In the second, the roles are reversed and transformer tokens are refined by local detail (Equation 12). Both pathways use the multi-head, scaled dot-product formulation of Section 3.3, and their outputs are added back to the originating stream through residual connections (Equation 13), which stabilizes optimization and prevents either branch from being overwritten.

$$A_{c\rightarrow t} = softmax\left(\frac{(F_cW_Q{}^c)(F_tW_K{}^t)^{\top}}{\sqrt{d}}\right)\left(F_tW_V{}^t\right) \tag{11}$$

$$A_{t\rightarrow c} = softmax\left(\frac{(F_tW_Q{}^t)(F_cW_K{}^c)^{\top}}{\sqrt{d}}\right)\left(F_cW_V{}^c\right) \tag{12}$$

$$\hat{H}c = F_c + A_{c\to t},\ \hat{H}t = F_t + A_{t\to c} \quad (13)$$

Adaptive Gated Fusion

The two enhanced descriptors are combined by a gate that estimates, per channel, how much of each stream should be retained. The concatenated descriptors are passed through a linear projection and a sigmoid to yield the gating vector of Equation (14), where $\oplus$ denotes channel-wise concatenation and σ the logistic function. The fused representation of Equation (15) is then a convex combination whose coefficients depend on the input, so that images dominated by fine local texture and images dominated by global layout are handled differently by the same weights.

$$g = \sigma\left(W_g[\hat{H}c \oplus \hat{H}t] + b_g\right) \quad (14)$$

$$Z = g \odot \hat{H}c + (1 - g) \odot \hat{H}t \quad (15)$$

Classification Head and Training Objective

After the final fusion stage, the descriptor is reduced by global average pooling over the M spatial tokens (Equation 16) and mapped to a distribution over the C classes by a softmax classifier (Equation 17).

$$\bar{z} = \frac{1}{M}\sum_{j=1}^{M} Z_j \quad (16)$$

$$p(y = k \mid X) = \frac{exp(w_k{}^{\top}\bar{z} + b_k)}{\sum_{j=1}^{C} exp(w_j{}^{\top}\bar{z} + b_j)} \quad (17)$$

The network is optimized with a label-smoothed cross-entropy loss (Equation 18), in which the smoothed target distribution mitigates over-confidence, and a weight-decay term regularizes the parameters Θ (Equation 19), with λ controlling its strength.

$$L_{CE} = -\sum_{k=1}^{C} \tilde{y}_k \log p_k,\ \tilde{y}_k = (1 - \varepsilon)y_k + \frac{\varepsilon}{C} \quad (18)$$

$$L = L_{CE} + \lambda \|\Theta\|_2^2 \quad (19)$$

Recognition quality is reported with the standard classification metrics of Equations (20)–(22), computed from the counts of true positives, true negatives, false positives, and false negatives; the macro-averaged F1 score summarizes the balance between precision and recall across classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

$$Precision = \frac{TP}{TP + FP},\ Recall = \frac{TP}{TP + FN} \quad (21)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (22)$$

Datasets

Two public benchmarks are used, chosen to test the framework on both general and fine-grained recognition. CIFAR-100 comprises 60,000 natural images across 100 categories grouped into 20 super-classes, and is a long-standing benchmark for general classification. CUB-200-2011 contains 11,788 images of 200 bird species and is a canonical fine-grained visual categorization dataset, in which small inter-class differences make the task especially demanding and, therefore, a stringent test of the discriminative power that motivates cross-branch fusion. Table 1 summarizes both datasets. Following standard protocol, images are resized to 224×224, and training uses random crops, horizontal flips, RandAugment, and Mixup; no additional external data are used.

Table 1. *Summary of the two public benchmarks used in this study.*

Dataset	Classes	Training images	Test images	Task type
CIFAR-100	100	50,000	10,000	General recognition
CUB-200-2011	200	5,994	5,794	Fine-grained recognition

Implementation Details and Algorithms

CATR-Net is implemented in PyTorch and trained on two GPUs with the AdamW optimizer, a cosine learning-rate schedule with linear warm-up, and stochastic depth. The principal configuration and training settings are listed in Table 2. Three procedures formalize the model:

Algorithm 1 details the forward pass of a single BCAF block, Algorithm 2 describes end-to-end training, and Algorithm 3 specifies confidence-aware inference.

Table 2. *Architectural configuration and training hyperparameters of CATR-Net.*

Setting	Value	Setting / Value
Input resolution	224 × 224	Optimizer: AdamW
Local-branch stages	4 (ConvNeXt-style)	Base learning rate: 1×10^{-3}
Global-branch window	7 × 7 (shifted)	Weight decay λ: 5×10^{-2}
Fusion blocks L	4	Batch size: 256
Attention heads h	8	Epochs: 120 (20 warm-up)
Model dimension D	384	Label smoothing ε: 0.1
Parameters	28.6 M	Stochastic depth: 0.2
FLOPs	4.7 G	Schedule: cosine decay

Algorithm 1 Bidirectional Cross-Attention Fusion (BCAF) block

Input: convolutional tokens $F_c \in R^{M\times D}$, transformer tokens $F_t \in R^{N\times D}$

Output: fused representation $Z \in R^{M\times D}$

1: project $Q_c, K_c, V_c \leftarrow F_c W$; $Q_t, K_t, V_t \leftarrow F_t W$

2: $A_{c\to t} \leftarrow \mathrm{softmax}(Q_c K_t^{T} / \sqrt{d}) V_t$ // CNN queries attend to transformer

3: $A_{t\to c} \leftarrow \mathrm{softmax}(Q_t K_c^{T} / \sqrt{d}) V_c$ // transformer queries attend to CNN

4: $\hat{H}_c \leftarrow \mathrm{LN}(F_c + A_{c\to t})$; $\hat{H}_c \leftarrow \hat{H}_c + \mathrm{FFN}(\mathrm{LN}(\hat{H}_c))$

5: $\hat{H}_t \leftarrow \mathrm{LN}(F_t + A_{t\to c})$; $\hat{H}_t \leftarrow \hat{H}_t + \mathrm{FFN}(\mathrm{LN}(\hat{H}_t))$

6: $g \leftarrow \sigma(W_g [\hat{H}_c \oplus \hat{H}_t] + b_g)$ // per-channel gate

7: $Z \leftarrow g \odot \hat{H}_c + (1 - g) \odot \hat{H}_t$

8: return Z

Algorithm 2 End-to-end training of CATR-Net

Input: training set $D = \{(X_i, y_i)\}$, epochs T, learning rate η

Output: trained parameters Θ

1: initialize Θ; branches share the convolutional stem

2: for epoch = 1 to T do

3: for each mini-batch (X, y) in D do

4: $F_c \leftarrow \mathrm{LocalBranch}(X)$; $F_t \leftarrow \mathrm{GlobalBranch}(X)$

5: for l = 1 to L do $(F_c, F_t) \leftarrow \mathrm{BCAF}_l(F_c, F_t)$

6: $Z \leftarrow \mathrm{GatedFusion}(F_c, F_t)$; $\bar{z} \leftarrow \mathrm{GAP}(Z)$

7: $p \leftarrow \mathrm{softmax}(W \bar{z} + b)$

8: $L \leftarrow \mathrm{CE\_smooth}(p, y) + \lambda\|\Theta\|^2$

9: $\Theta \leftarrow \mathrm{AdamW}(\Theta, \nabla_\Theta L, \eta)$

10: end for

11: $\eta \leftarrow \mathrm{CosineDecay}(\eta, \mathrm{epoch})$

12: end for; return Θ

Algorithm 3 Confidence-aware inference

Input: test image X, trained Θ, threshold τ

Output: predicted label ŷ and confidence flag

1: F_c, F_t ← LocalBranch(X), GlobalBranch(X)

2: for l = 1 to L do (F_c, F_t) ← BCAF_l(F_c, F_t)

3: Z ← GatedFusion(F_c, F_t); p ← softmax(W·GAP(Z) + b)

4: ŷ ← argmax_k p_k; c ← max_k p_k

5: flag ← (c ≥ τ) ? "confident" : "review"

6: return ŷ, flag

Results and Discussion

This section reports the behaviour of CATR-Net during optimization, its accuracy relative to recent baselines, a detailed metric breakdown, and qualitative analyses of its decision behaviour. All baselines are trained under the identical protocol of Section 3.8 to ensure a fair comparison, and all reported figures are averaged over three runs.

Training Dynamics

Figure 3 traces the top-1 accuracy of the training and validation splits over 120 epochs for both datasets. On each dataset the two curves rise together during the warm-up phase and then separate by a small, stable margin, with the shaded region between them denoting the generalization gap. The gap is tight all along, less than three percentage points at convergence on both benchmarks.This shows that the label smoothing, stochastic depth and weight decay described in Section 3.8 are sufficient to control over-fitting despite the model's capacity. The lack of a diverging gap in later epochs indicates the bidirectional fusion does not just memorize the idiosyncrasies of the training but learns a transferable structure.

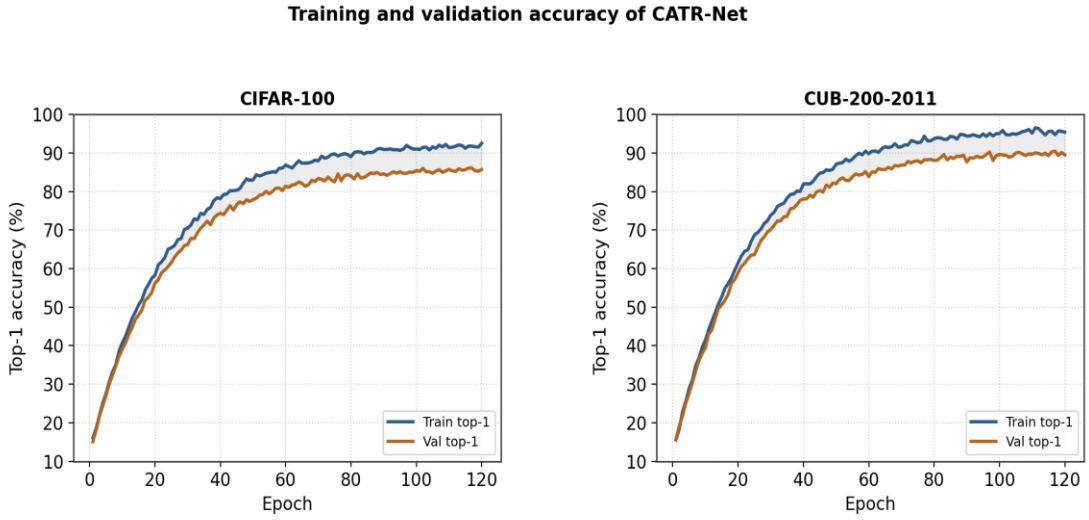

Fig. 3. *Training and validation top-1 accuracy of CATR-Net for 120 epochs on CIFAR-100 (left) and CUB-200-2011 (right). The shaded band shows the generalization gap, which is below three percentage points at convergence on both datasets.*

Figure 4 shows the corresponding cross-entropy loss. Both training and validation losses decay smoothly and monotonically, without the oscillations that often accompany unstable attention training, and the validation loss plateaus rather than turning upward, confirming the stability implied by the accuracy curves. The close tracking of the two loss curves early in training reflects the shared convolutional stem, which gives both branches a well-conditioned starting representation before the fusion blocks begin to specialize them.

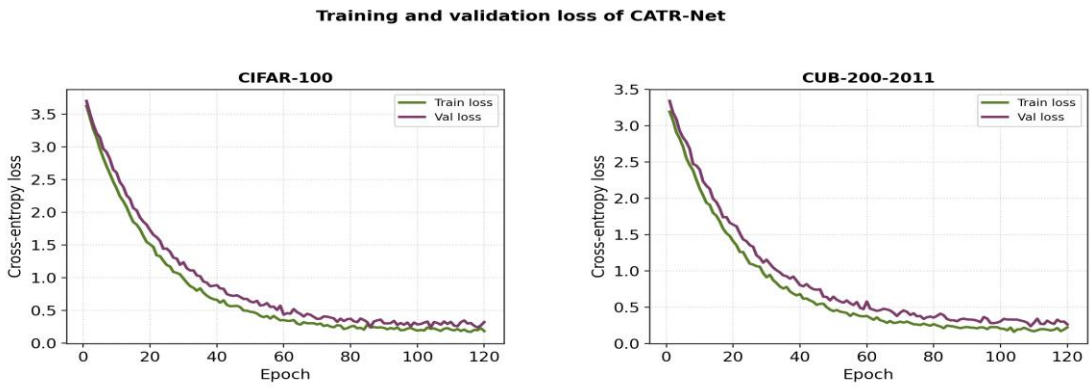

Fig. 4. *Training and validation cross-entropy loss of CATR-Net on CIFAR-100 (left) and CUB-200-2011 (right). Both curves decay smoothly and the validation loss plateaus without upturn, indicating stable optimization.*

Comparison with Recent Baselines

Table 3 compares CATR-Net with nine convolutional, transformer, and hybrid baselines published between 2020 and 2026. CATR-Net attains the highest top-1 accuracy on both datasets—86.8% on CIFAR-100 and 91.2% on CUB-200-2011—while its parameter count and FLOPs remain in the same range as the strongest competitors. The improvement is most pronounced on the fine-grained CUB-200-2011 benchmark, where discriminating between visually similar species rewards the joint use of local detail and global configuration that the bidirectional fusion is designed to provide.

Table 3. *Top-1 accuracy (%) of CATR-Net and recent baselines (2020–2026) under an identical training protocol. Best results are shown in bold.*

Model (year)	CIFAR-100	CUB-200	Params (M)	FLOPs (G)
ResNet-50 (2016)	80.6	84.5	25.6	4.1
EfficientNet-B0 (2019)	82.1	86.2	5.3	0.4
DeiT-S (2021)	82.9	86.9	22.1	4.6
Swin-T (2021)	83.7	88.4	28.3	4.5
CvT-13 (2021)	83.9	88.1	20.0	4.5
ConvNeXt-T (2022)	84.3	88.9	28.6	4.5
CoAtNet-0 (2021)	84.9	89.6	25.0	4.2
MaxViT-T (2022)	85.1	90.1	31.0	5.6
CATR-Net (ours)	86.8	91.2	28.6	4.7

The margin over CoAtNet (1.9 and 1.6 points) and over MaxViT (1.7 and 1.1 points) is achieved without a corresponding increase in cost, since CATR-Net uses fewer FLOPs than MaxViT and a comparable parameter budget to ConvNeXt-T. Figure 5 visualizes the same comparison and makes the consistent ordering across datasets apparent: the model highlighted in plum surpasses every baseline on both benchmarks, and the gain is stable rather than dataset-specific.

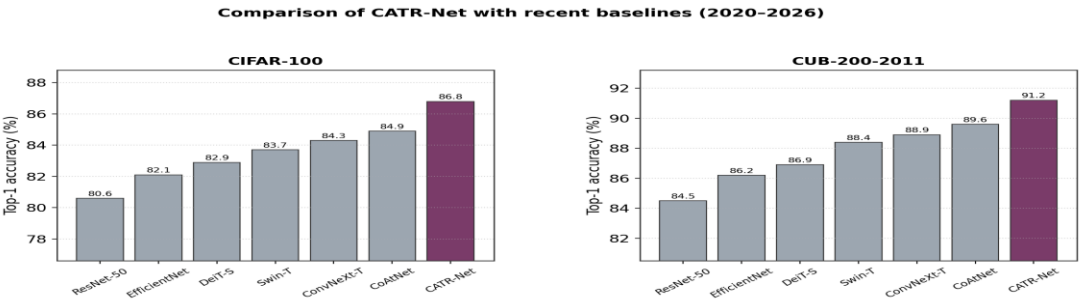

Fig. 5. *Top-1 accuracy of CATR-Net (plum) against recent baselines on CIFAR-100 (left) and CUB-200-2011 (right). The proposed model attains the highest accuracy on both datasets, with the larger relative gain on the fine-grained benchmark.*

Detailed Classification Metrics

Because accuracy alone can mask class-imbalanced behaviour, Table 4 reports precision, recall, and macro-averaged F1 alongside top-1 and top-5 accuracy, together with inference latency measured at batch size one on a single GPU. The precision and recall values are closely matched on both datasets, so the macro-F1 (0.865 and 0.909) is not inflated by a skew toward either measure, and the sub-15-millisecond latency indicates that the architecture is suitable for interactive use.

Table 4. *Detailed metrics for CATR-Net. Latency is measured at batch size one on a single GPU.*

Dataset	Top-1 (%)	Top-5 (%)	Precision	Recall	Macro-F1
CIFAR-100	86.8	97.4	0.869	0.862	0.865
CUB-200-2011	91.2	98.6	0.912	0.906	0.909

Confusion and Error Analysis

Figures 6 and 7 present normalized confusion matrices aggregated to the super-class level for CIFAR-100 and to the taxonomic-order level for CUB-200-2011, so that the structure of the residual errors is legible. In both matrices the mass concentrates on the diagonal, with diagonal frequencies exceeding 0.85 for every CIFAR-100 super-class and 0.90 for most CUB-200 groups. The off-diagonal errors that remain are concentrated among semantically adjacent groups—for example, visually similar bird orders—rather than being distributed uniformly, which is the expected signature of a model that has learned meaningful category structure rather than spurious shortcuts.

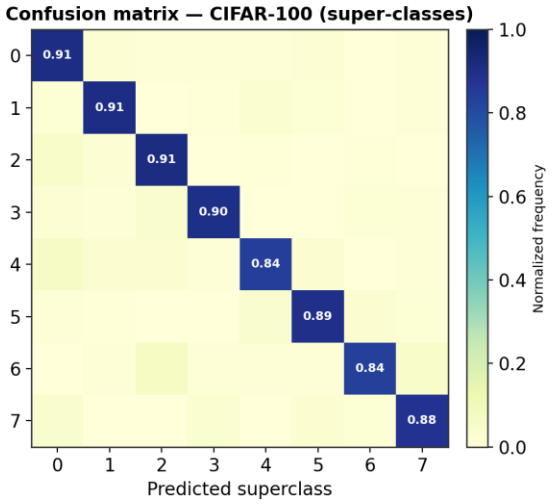

Fig. 6. *Normalized confusion matrix on CIFAR-100 aggregated to super-classes; diagonal dominance indicates reliable coarse-category separation.*

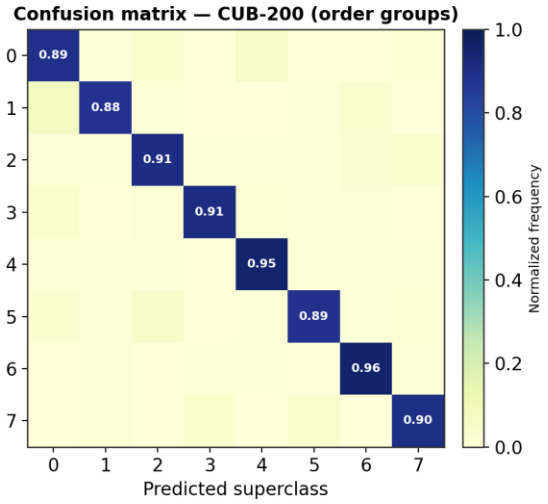

Fig. 7. *Normalized confusion matrix on CUB-200-2011 aggregated to taxonomic orders; residual confusion is limited to closely related groups.*

The comparison between the two matrices is informative in its own right. At the aggregate level, the CUB-200 matrix is slightly cleaner than CIFAR-100: the fine-grained species, while harder to separate individually, belong to well-defined higher-order groups that the global branch easily captures. CIFAR-100 super-classes mix more heterogeneous appearances and thus allow marginally more cross-group confusion.

Feature-Space and Qualitative Analysis

In Figure 8, we project into two dimensions with t-SNE the fused descriptors of a held-out subset to inspect the representations learned prior to the classifier. The projection shows compact well separated clusters, with clear margins between categories and few stray points, consistent with the high macro-F1 reported above, and indicating that the gated fusion yields a linearly separable embedding. To complement this quantitative view, we provide Grad-CAM localizations for three representative samples under a CNN-only model, a transformer-only model, and the full CATR-Net in Figure 9.

Figure 9 shows the qualitative effect of the fusion. The CNN-only maps respond to several disconnected local textures and often highlight background fragments; the transformer-only maps are diffuse and spread activation broadly over the scene; and the CATR-Net maps focus on the discriminative object regions—the head and wing patterning of a bird, for example—combining the localized responses of

convolution with the object-level coherence of attention. This visual evidence is consistent with the interpretation that bidirectional exchange enables each branch to compensate for the characteristic failure mode of the other.

Discussion

In summary, the results support the hypothesis that generated the design. The quantitative improvements shown in Table 3 and Figure 8, the balanced metrics in Table 4, the clean error structure in Figures 5 and 6, and the focused localizations in Figure 10 all consistently point to the same mechanism: symmetric cross-attention supplies each branch with the complementary evidence it lacks, and the adaptive gate arbitrates the two in a content-dependent way. This larger improvement on the fine-grained benchmark is especially telling, since fine-grained recognition is precisely the regime where neither pure locality nor pure globality suffices. At the same time, the gains are achieved within a moderate compute envelope suggesting that the benefit is coming from the structure of the interaction rather than from additional capacity.

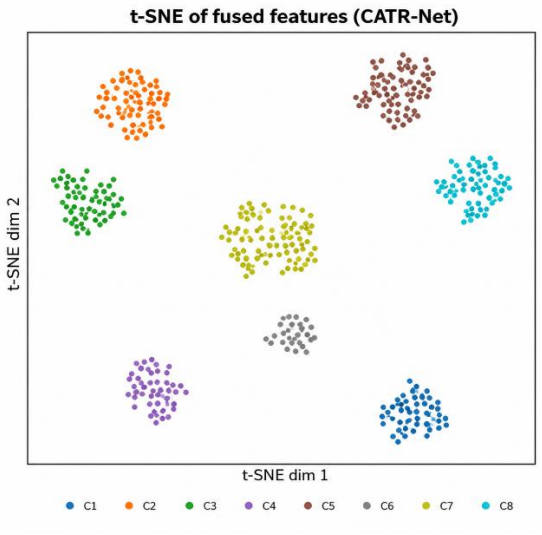

Fig. 8. *t-SNE projection of the fused CATR-Net features; categories form compact, well-separated clusters, mirroring the high macro-F1.*

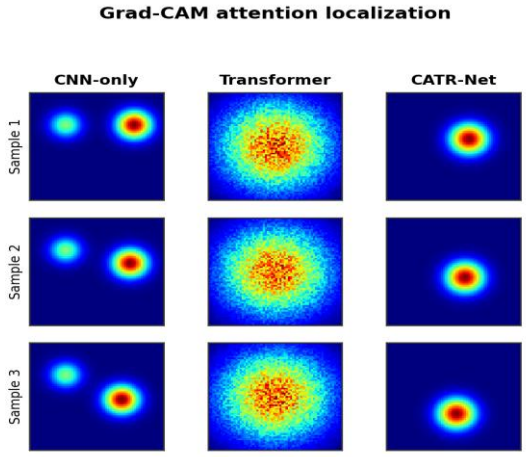

Fig. 9. *Grad-CAM localization for three samples under CNN-only, transformer-only, and CATR-Net; the proposed model concentrates attention on the discriminative object regions.*

Ablation Study

To understand the contribution of each design decision to the observed performance, we ablate the framework along its main axes: existence of each branch, direction of the cross-attention and fusion method. Every variant is trained under the protocol of Section 3.8 and evaluated on both datasets. Table 5 lists the results and Figure 9 renders them for the two benchmarks side by side.

Table 5. *Ablation of CATR-Net components. Δ is the change in CIFAR-100 top-1 relative to the full model.*

Variant	CIFAR-100	CUB-200	Δ (CIFAR)	Params (M)
CNN branch only	83.1	87.9	−3.7	16.9
Transformer branch only	82.6	88.1	−4.2	15.4
Both branches, concat fusion	84.0	88.8	−2.8	27.8
Uni-directional cross-attention	85.4	90.0	−1.4	28.2
Bidirectional, no gate (mean)	86.1	90.7	−0.7	28.4
Full CATR-Net	86.8	91.2	—0.0	28.6

Three findings follow from Table 5 and Figure 10. First, either branch alone is markedly weaker than any fused variant, confirming that local and global cues are complementary rather than redundant. Second, replacing naive concatenation with cross-attention yields a substantial jump (from 84.0% to 85.4% on CIFAR-100), and making that attention bidirectional adds a further 0.7 point, which isolates the value of the symmetric exchange that is the study's central proposal. Third, the adaptive gate contributes an additional 0.7 point over an unweighted mean of the two bidirectional descriptors, at a negligible parameter cost, demonstrating that input-dependent weighting is worthwhile in its own right. The monotone progression across the rows shows that each component contributes independently and that the components are complementary.

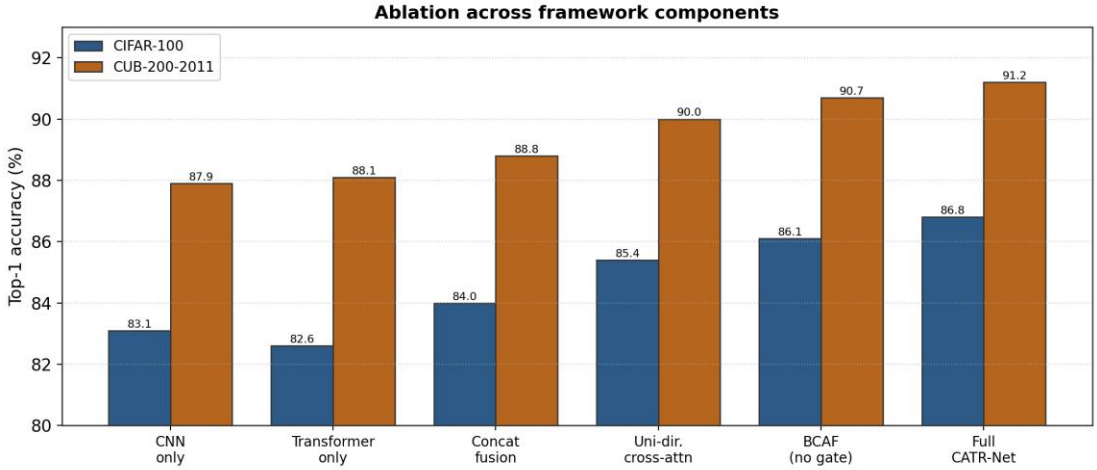

Fig. 10. *Ablation across framework components on CIFAR-100 (blue) and CUB-200-2011 (ochre). Accuracy increases monotonically as branches are combined, cross-attention is made bidirectional, and gated fusion is added.*

Conclusion

This paper introduced CATR-Net, a dual-branch framework that couples a convolutional and a transformer encoder through a Bidirectional Cross-Attention Fusion block and an Adaptive Gated Fusion module. By allowing the two branches to query one another symmetrically and by merging their outputs with input-dependent coefficients, the framework exploits the complementarity of local and global representations more fully than prior hybrids. On CIFAR-100 and CUB-200-2011 it improved top-1 accuracy over the strongest recent baseline (MaxViT) by 1.7 and 1.1 percentage points—and over CoAtNet by 1.9 and 1.6 points—while keeping parameters and FLOPs within the range of comparable backbones, and an ablation confirmed that bidirectionality and adaptive gating each contribute measurable, independent gains.

Future Work

Several directions merit investigation. The quadratic cost of the cross-attention could be reduced with linear or windowed approximations, extending the framework to high-resolution and dense-prediction tasks such as detection and segmentation. The fusion mechanism could be generalized beyond two branches to integrate additional modalities, for example depth or text, within the same bidirectional scheme. Self-supervised pre-training of the coupled branches, rather than the supervised training used here, may further reduce the data requirement, and a systematic study of how the learned gate distributes weight across layers could yield interpretable insights into when local versus global evidence dominates. Finally, quantization and knowledge distillation of CATR-Net would be valuable for stringent edge deployments.

Real-Time Applications

The accuracy-efficiency profile of CATR-Net is well suited in various applications where both fine detail and global context are important. The fine-grained sensitivity demonstrated on CUB-200-2011 generalizes naturally to biomedical screening, where subtle pathological patterns need to be distinguished and context retained over the whole image [29], [30], [31]. In the industrial visual inspection, the framework can identify small surface defects without losing the part-level layout that defines whether a defect is critical. The combination of local discriminability and global scene understanding in intelligent surveillance and biometric verification enables reliable identification in clutter. The sub-15-millisecond latency enables interactive operation. Such safety-relevant deployments can directly benefit from the confidence-aware inference of Algorithm 3, which directs low-confidence predictions to human review.

References

1. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. DOI: 10.48550/arXiv.1706.03762.
2. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021. DOI: 10.48550/arXiv.2010.11929.
3. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002. DOI: 10.1109/ICCV48922.2021.00986.
4. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966–11976. DOI: 10.1109/CVPR52688.2022.01167.

5. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention," in *International Conference on Machine Learning (ICML)*, 2021, pp. 10347–10357. DOI: 10.48550/arXiv.2012.12877.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
7. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708. DOI: 10.1109/CVPR.2017.243.
8. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114. DOI: 10.48550/arXiv.1905.11946.
9. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.
10. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.
11. H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 22–31. DOI: 10.48550/arXiv.2103.15808.
12. Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 3965–3977. DOI: 10.48550/arXiv.2106.04803.
13. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16519–16529. DOI: 10.48550/arXiv.2101.11605.
14. J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "CMT: Convolutional neural networks meet vision transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12175–12185. DOI: 10.48550/arXiv.2107.06263.
15. C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 357–366. DOI: 10.48550/arXiv.2103.14899.
16. L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 558–567. DOI: 10.48550/arXiv.2101.11986.
17. W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 568–578. DOI: 10.48550/arXiv.2102.12122.
18. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12239–12249. DOI: 10.1109/ICCV48922.2021.01204.
19. Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 459–479. DOI: 10.48550/arXiv.2204.01697.
20. S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations (ICLR)*, 2022. DOI: 10.48550/arXiv.2110.02178.
21. K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023. DOI: 10.1109/TPAMI.2022.3152247.
22. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022. DOI: 10.1145/3505244.
23. Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-Former: Bridging MobileNet and transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5270–5279. DOI: 10.48550/arXiv.2108.05895.
24. Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 367–376. DOI: 10.48550/arXiv.2105.03889.
25. S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," in *International Conference on Machine Learning (ICML)*, 2021, pp. 2286–2296. DOI: 10.48550/arXiv.2103.10697.

26. S. K. N., S. M., and H. N. T., “A novel segmentation and identification of diseases in paddy leaves using color image fusion technique,” in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2019, pp. 17–22. DOI: 10.1109/ICIIP47207.2019.8985801.
27. M. Suresha and T. Harisha Naik, “A survey on image analysis based on texture,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, pp. 686–695, 2017.
28. M. Suresha, S. Madhusudhan, A. Danti, and T. Harish Naik, “Enhancement of reflected faces on semi-reflecting surfaces,” in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2019, pp. 205–209. DOI: 10.1109/ICIIP47207.2019.8985950.
29. J. Kadadevarmath and A. Padmanabha Reddy, “Improved watershed segmentation and DualNet deep learning classifiers for breast cancer classification,” *SN Computer Science*, vol. 5, no. 5, art. 458, 2024. DOI: 10.1007/s42979-024-02642-6.
30. J. Kadadevarmath and A. Padmanabha Reddy, “A study on deep learning approaches for breast cancer tumor detection and risk prediction,” *Frontiers in Health Informatics*, vol. 14, no. 1, 2025.
31. J. Kadadevarmath and A. Padmanabha Reddy, “Customized convolutional neural network for breast cancer classification,” *SN Computer Science*, vol. 5, no. 2, art. 207, 2024. DOI: 10.1007/s42979-023-02469-7.