ConvRoute-ViT: Convolutional Token Routing for Efficient Vision Transformers

Manish Das
IIIT Allahabad
Prayagraj, India
pmm2023001@iiita.ac.in

Shiv Ram Dubey
IIIT Allahabad
Prayagraj, India
srdubey@iiita.ac.in

Satish Kumar Singh
IIIT Allahabad
Prayagraj, India
sk.singh@iiita.ac.in

arXiv:2610.38284v1 [cs.CV] 28 Oct 2026

***Abstract*—Vision Transformers (ViTs) achieve strong image recognition performance, but their self-attention increases quadratically with the number of input tokens and makes them computationally expensive. The standard implementation of ViT models treats all patch tokens with equal importance across every layer of the model. This uniform treatment introduces redundant computation and may reduce the model's focus on the most discriminative object regions. In this paper, we propose ConvRoute-ViT, a convolutional token routing mechanism for efficient vision transformers that performs token reduction through convolutional routing. Our model preserves the most relevant patch tokens according to class-token attention at selected layers, while the remaining tokens are spatially reorganized and compressed into a compact set of convolutional summary tokens. This design reduces the sequence length without completely removing the contextual information of the less informative tokens. By applying the routing process hierarchically, the proposed framework constructs compact and discriminative representations for deeper layers. Experiments on CIFAR-100 and TinyImageNet show that ConvRoute-ViT improves computational efficiency while maintaining classification accuracy compared with the base ViT.**

***Keywords*—Vision Transformer, efficient transformer, token routing, convolutional token compression, image classification**

I. Introduction

Vision Transformers (ViTs) have emerged as an effective alternative to convolutional neural networks for image recognition by modeling global relationships across diverse computer vision tasks [1]. The complexity of self-attention grows quadratically with the number of tokens, making standard ViTs computationally expensive. To improve training efficiency, DeiT [2] introduced data-efficient strategies for ViT training on moderate-scale datasets. Several efficient architectural variants incorporated inductive bias. For example, Swin Transformer [3] introduces hierarchical shifted window self-attention, PVT [4] uses a pyramid design to reduce tokens, and CVT introduces convolutional projections to capture local spatial structures [5]. Another major research direction focuses on reducing token redundancy during transformer inference. DynamicViT [6], A-ViT [7], Evo-ViT [8], SPViT [9], ATS [10], IA-RED^2 [11], Heat-ViT [12] show that a substantial portion of tokens can be reduced without severe loss of accuracy. However, directly discarded tokens may still contain important contextual information. A related line of research improves efficiency through token reorganization, merging, pooling, or learned compression. EViT [13] preserves attentive tokens but fuses all inattentive background tokens into a single representation via an attention-weighted average, losing their spatial arrangement. ToMe [14], Joint Token Pruning and Squeezing [15], TokenLearner [16], and Token Pruning and Pooling [17], DeepCompressViT [18], and TinyNext-ViT [19] further improve token efficiency through compact token modeling, while related methods further improve token interaction and fusion for efficient representation learning [20], [21]. HVCNet [22] combines local feature extraction with transformers, and HMC [23] improves the computational efficiency of Vision Transformers.

Motivated by this gap, we propose ConvRoute-ViT, a convolutional token routing framework for efficient vision transformers, as illustrated in Fig. 1. The proposed mechanism selects informative patch tokens according to CLS-attention scores and retains them for subsequent transformer processing. Rather than discarding, merging, or globally pooling the remaining less informative tokens, ConvRoute recovers their spatial layout and compresses them with a light-weight convolutional module to obtain compact ConvSummary tokens. This design reduces redundancy of tokens while preserving spatially structured contextual information to aid classification. ConvRoute-ViT learns compact, discriminative, and context-preserving token sequences by hierarchically performing the routing process at multiple depths for efficient recognition of images. The routing mechanism can be applied at multiple layers to reduce tokens while maintaining semantically meaningful and contextually important information.

The main contributions of our work:

- We introduce the **ConvRoute token routing module** that estimates the informative patch tokens using CLS-attention scores, allowing the model to focus on task-relevant and discriminative informative regions.
- We design the **ConvRoute Compressor**, which spatially reshapes and compresses them into compact *ConvSum-*

We gratefully acknowledge the Indian Institute of Information Technology Allahabad for funding and providing necessary support to pursue this research work.

This paper is accepted by the 2026 IEEE Region 10 Conference (TENCON). Copyright ©2026 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from IEEE by sending an email to pubs-permissions@ieee.org.

mary tokens, ensuring useful background structure and complementary context representations are preserved.
- We apply the routing and compression in a **hierarchical dual-stage manner** at two transformer depths, reducing the token sequence while maintaining rich discriminative representations for final classification.

II. Proposed ConvRoute-ViT Architecture

A. Preliminaries

Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, it is divided into non-overlapping patches of size $P \times P$, producing $n = \frac{H}{P} \times \frac{W}{P}$. Each patch $\mathbf{p}_i \in \mathbb{R}^{P^2C}$ is projected into a D-dimensional token as $\mathbf{x}_i = \mathbf{W}_e\mathbf{p}_i$, where $\mathbf{W}_e \in \mathbb{R}^{D\times P^2C}$. A learnable CLS token and positional embeddings form $\mathbf{Z}_0 = \mathbf{x}_{cls}, \mathbf{x}_1, \ldots, \mathbf{x}_n] + \mathbf{E}_{pos}$, where $\mathbf{Z}_0 \in \mathbb{R}^{N\times D}$ with $N = n+1$. For the r-th transformer block, attention is computed as

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right), \tag{1}$$

where $\mathbf{Q} = \mathbf{Z}_r\mathbf{W}_Q$, $\mathbf{K} = \mathbf{Z}_r\mathbf{W}_K$, and $\mathbf{V} = \mathbf{Z}_r\mathbf{W}_V$. The output is then updated and forwarded to

$$\hat{\mathbf{Z}}_r = \mathbf{Z}_r + \text{MHSA}(\mathbf{Z}_r), \qquad \mathbf{Z}_{r+1} = \hat{\mathbf{Z}}_r + \text{FFN}(\hat{\mathbf{Z}}_r). \tag{2}$$

B. ConvRoute Token Routing Module

For a transformer with L layers, ConvRoute routing is applied at two layers k and l, where $0 < k < l < L$. The k layers process the full token sequence before reduction. After layer k, the token sequence is represented as $\mathbf{Z}^{(k)} = [\mathbf{z}cls^{(k)}, \mathbf{X}^{(k)}]$, where $\mathbf{z}cls^{(k)} \in \mathbb{R}^D$ is the CLS token and $\mathbf{X}^{(k)} \in \mathbb{R}^{(N-1)\times D}$ are the patch tokens. The importance score of the i-th patch token is computed from CLS-attention as

$$s_i^{(k)} = \frac{1}{h_A}\sum_{h=1}^{h_A} A_{cls,i}^{(k,h)}, \tag{3}$$

where h_A is the number of attention heads and $A_{cls,i}^{(k,h)}$ denotes the attention from the CLS token to the i-th patch token in the h-th head. Based on the keep ratio ρ_k, the top $R_k = \lfloor \rho_k(N-1) \rfloor$ tokens are retained as $\mathbf{T}_{imp}^{(k)} \in \mathbb{R}^{R_k\times D}$, while the remaining tokens are denoted by $\mathbf{T}_{low}^{(k)}$. Instead of discarding $\mathbf{T}_{low}^{(k)}$, ConvRoute restores these tokens to their original patch-grid positions to form $\mathbf{T}_{map}^{(k)} \in \mathbb{R}^{D\times H_p\times W_p}$, where $H_p = \frac{H}{P}$ and $W_p = \frac{W}{P}$. The restored map is compressed through the ConvRoute Compressor as

$$\mathbf{F}^{(k)} = \text{DWConv}_{3\times3}\Big(\text{PWConv}_{1\times1}\big(\text{DWConv}_{3\times3}(\mathbf{T}_{map}^{(k)})\big)\Big), \tag{4}$$

$$\mathbf{T}_{conv}^{(k)} = \text{Flatten}\big(\text{AdaptiveAvgPool}_{p_h\times p_w}(\mathbf{F}^{(k)})\big), \tag{5}$$

where $\mathbf{T}_{conv}^{(k)} \in \mathbb{R}^{S_k\times D}$ and $S_k = p_h \times p_w$ is the number of ConvSummary tokens. The depthwise convolutions preserve local spatial neighborhoods, while the pointwise convolution captures channel interactions, enabling low-importance tokens to be summarized without losing contextual information.

The routed sequence after the first stage is

$$\tilde{\mathbf{Z}}^{(k)} = [\mathbf{z}_{cls}^{(k)}, \mathbf{T}_{imp}^{(k)}, \mathbf{T}_{conv}^{(k)}] \in \mathbb{R}^{M_k\times D}, \tag{6}$$

where $M_k = [1 + R_k + S_k]$ and $M_k \leq N$. The same routing process is applied again at layer l, producing $\tilde{\mathbf{Z}}^{(l)} = [\mathbf{z}_{cls}^{(l)}, \mathbf{T}_{imp}^{(l)}, \mathbf{T}_{conv}^{(l)}] \in \mathbb{R}^{M_l\times D}$, with $M_l \leq M_k \leq N$. Thus, ConvRoute progressively reduces token redundancy while preserving discriminative tokens and compact spatial-context information for final classification.

III. Experimental Settings and Results

A. Datasets and Implementation Details

Datasets: We conducted our experiments on CIFAR-100 [24] contains 60,000 color images of size 32×32 from 100 distinct classes, with 50,000 training and 10,000 test images and TinyImageNet [25] contains 200 classes, each with 500 training images and 50 validation images of size 64×64. **Experimental Settings**: Models are trained from scratch on an NVIDIA GeForce RTX 4090 GPU for 100 epochs using AdamW with an learning rate of 0.003, weight decay of 5×10^{-2}, cosine annealing, a batch size of 128 with 12 attention heads.

B. Experimental Results

Table I shows that ConvRoute-ViT consistently improves the base ViT across most routing-layer pairs. The best CIFAR-100 accuracy is obtained at layers $(2, 6)$ with $(\rho_k, \rho_l) = (0.8, 0.9)$, achieving 75.71% (+1.67), while the best TinyImageNet result is obtained at $(2, 5)$ with $(\rho_k, \rho_l) = (0.8, 0.9)$, reaching 58.04% (+1.24). Improved gains are also observed at intermediate routing positions such as $(3, 5)$ and $(3, 6)$, where the token features are sufficiently discriminative to separate important object regions from less informative context, allowing ConvRoute to remove redundancy without losing semantic information. While routing in early layers, such as at $(2, 4)$, removes useful low-level details before stable feature formation, which explains the weaker gains, since low-level visual details are reduced before the formation of feature representations. Overall, the results indicate that token reduction depends on selecting routing layers that reduce less informative tokens while preserving useful contextual information.

C. Attention Visualization Analaysis

Figure 2 presents four representative CIFAR100 [24] examples comparing the attention map visualization of the base ViT and our proposed ConvRoute-ViT. In the first example, the base ViT wrongly predicts *apple* as *orange* by focusing on the smooth, round color patterns, whereas ConvRoute-ViT attends to highly discriminative regions and predicts *apple*. In the second example, the base ViT scatters its attention to the background areas, leading to *caterpillar* instead of *beetle,* while our model concentrates over the body and preserves the fine-grained structure required for correct classification. In the third example, the base ViT predicts *orange* for *camel*, by concentrating on the bright round background regions, while ConvRoute-ViT places a coherent high response over the

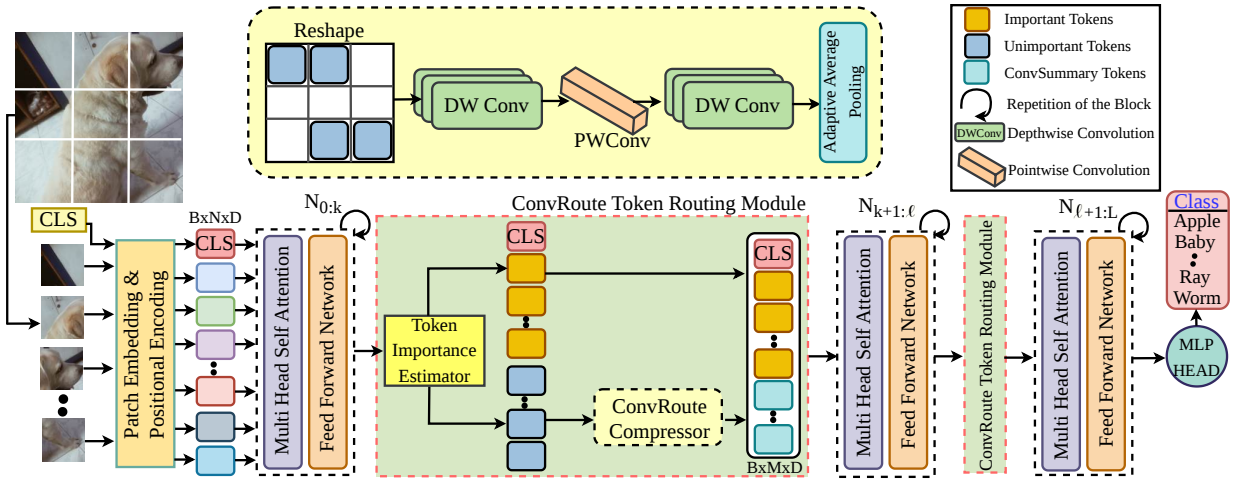


Fig. 1. Overview of the proposed ConvRoute-ViT architecture. The input image is converted into patch tokens with positional encoding and a CLS token, creating an input size of $B \times N \times D$. After initial transformer processing up to k^{th} transformer layers, the ConvRoute Token Routing Module computes the token importance score and retains the top-R important tokens. The remaining tokens are passed through the ConvRoute Compressor where they are spatially reshaped to generate compact ConvSummary tokens, resulting in $B \times M \times D$ and are processed by transformer layers from $k{+}1 \rightarrow l$. A second ConvRoute Token Routing Module is applied at layer l to further reduce the sequence and is then passed through the subsequent transformer blocks for final classification.

TABLE I

COMPARISON OF CONVROUTE-VIT UNDER DIFFERENT ROUTING-LAYER POSITIONS (k, l) AND KEEP-RATIO SETTINGS (ρ_k, ρ_l). ROWS DENOTE ROUTING-LAYER PAIRS, AND EACH COLUMN REPORTS CIFAR-100 AND TINYIMAGENET ACCURACIES FOR EACH KEEP-RATIO PAIR. THE BASE VIT ACHIEVES 74.04% ON CIFAR-100 AND 56.80% ON TINYIMAGENET. BLUE: CIFAR-100. GREEN: TINYIMAGENET.

(k,l)	$(\rho_k,\rho_l)=(0.6,0.7)$		$(\rho_k,\rho_l)=(0.7,0.8)$		$(\rho_k,\rho_l)=(0.8,0.9)$		$(\rho_k,\rho_l)=(0.9,0.5)$		$(\rho_k,\rho_l)=(0.9,0.7)$	
(2,4)	73.70 (-0.34)	55.56 (-1.24)	74.06 (+0.02)	55.70 (-1.10)	74.70 (+0.66)	56.62 (-0.18)	74.35 (+0.31)	56.20 (-0.60)	74.07 (+0.03)	56.06 (-0.74)
(2,5)	73.76 (-0.28)	55.74 (-1.06)	74.18 (+0.14)	56.58 (-0.22)	74.70 (+0.66)	58.04 (+1.24)	74.97 (+0.93)	56.90 (+0.10)	75.27 (+1.23)	57.18 (+0.38)
(2,6)	73.86 (-0.18)	56.08 (-0.72)	75.54 (+1.50)	56.74 (-0.06)	75.71 (+1.67)	56.74 (-0.06)	74.84 (+0.80)	57.22 (+0.42)	75.03 (+0.99)	57.70 (+0.90)
(2,7)	74.19 (+0.15)	55.98 (-0.82)	74.21 (+0.17)	56.26 (-0.54)	75.05 (+1.01)	56.96 (+0.16)	75.15 (+1.11)	56.20 (-0.60)	74.93 (+0.89)	57.26 (+0.46)
(3,5)	74.33 (+0.29)	55.78 (-1.02)	74.54 (+0.50)	56.66 (-0.14)	74.88 (+0.84)	56.58 (-0.22)	74.74 (+0.70)	57.10 (+0.30)	75.42 (+1.38)	57.28 (+0.48)
(3,6)	74.50 (+0.46)	56.34 (-0.46)	74.90 (+0.86)	56.70 (-0.10)	74.90 (+0.86)	57.28 (+0.48)	75.25 (+1.21)	57.86 (+1.06)	75.15 (+1.11)	57.76 (+0.96)
(3,7)	74.77 (+0.73)	56.58 (-0.22)	74.93 (+0.89)	56.82 (+0.02)	74.93 (+0.89)	56.86 (+0.06)	75.10 (+1.06)	57.52 (+0.72)	75.41 (+1.37)	57.04 (+0.24)
(4,6)	74.86 (+0.82)	56.16 (-0.64)	74.65 (+0.61)	56.78 (-0.02)	75.39 (+1.35)	57.18 (+0.38)	75.00 (+0.96)	56.88 (+0.08)	75.07 (+1.03)	57.70 (+0.90)
(4,7)	74.75 (+0.71)	57.20 (+0.40)	74.85 (+0.81)	57.42 (+0.62)	74.55 (+0.51)	57.28 (+0.48)	74.89 (+0.85)	57.56 (+0.76)	74.82 (+0.78)	57.02 (+0.22)

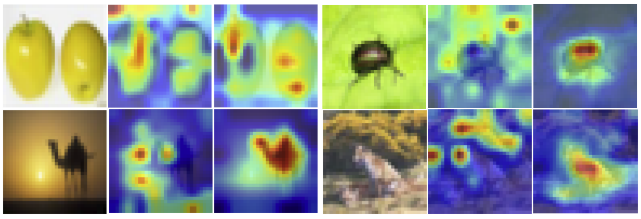


Fig. 2. Attention heatmaps for representative CIFAR-100 samples [24]. Each example consists of three panels arranged from left to right: the input image, the base ViT attention map, and the ConvRoute-ViT attention map.



Fig. 3. Comparison of prediction confidence between the proposed model and the base ViT on CIFAR-100. Input images are shown on the left, and the top-3 predictions on the right (blue: proposed; green: base ViT).

camel and correctly predicts *camel*. For the fourth example, the base ViT focuses on textured background regions and doenot focus on the main body of the fox, whereas ConvRoute-ViT produces a compact and coherent response over the fox region and gives the correct class. Overall, these visualizations show that ConvRoute-ViT improves classification by suppressing misleading background tokens and emphasizing the most informative object-centric regions.

D. Effect of Keep Ratio on Accuracy and Complexity

Table II shows a trade-off between classification performance and computational cost. The low complexity settings $\rho = (0.4, 0.5)$ reduce the GFLOPs to 0.1239, corresponding to a 34.44% reduction over the base model, while improving CIFAR-100 accuracy from 74.04% to 74.28%. As the keep ratio increases, the model generally preserves more informative tokens, which improves accuracy. On CIFAR-100, the best result is obtained at $\rho = (0.8, 0.9)$ with 75.39% (+1.35) with 6.72% reduction, while on TinyImageNet the best performance is achieved at $\rho = (0.8, 0.8)$ with 57.92% (+1.12), corresponding to a 9.42% reduction. Overall, moderate keep-ratio settings provide the most favorable balance between efficiency and accuracy, as they retain a sufficient number of

TABLE II

EFFECT OF KEEP-RATIO PAIR SETTINGS (ρ_k, ρ_l) AT THE 4TH AND 6TH ROUTING LAYERS. COLUMNS SHOWS DIFFERENT PAIRS, AND ROWS REPORT CIFAR-100, TINYIMAGENET, AND GFLOPS. THE BASELINE ACHIEVES 74.04% ON CIFAR-100, 56.80% ON TINYIMAGENET, AND 0.1890 GFLOPS.

(ρ_k, ρ_l)	(0.4,0.5)	(0.5,0.4)	(0.5,0.5)	(0.5,0.6)	(0.6,0.5)	(0.6,0.6)	(0.6,0.7)	(0.7,0.6)	(0.7,0.7)	(0.7,0.8)	(0.8,0.7)	(0.8,0.8)	(0.8,0.9)	(0.9,0.5)	(0.9,0.6)	(0.9,0.7)	(0.9,0.8)	(0.9,0.9)
CIFAR-100	74.28	74.34	74.55	74.25	74.70	74.55	74.86	74.75	74.87	74.65	75.05	74.40	75.39	75.00	75.01	75.07	74.01	75.06
TinyImageNet	55.48	56.36	55.34	56.56	56.82	56.96	56.16	56.90	57.12	56.78	56.32	57.92	57.18	56.88	57.20	57.70	57.22	57.26
GFLOPs	0.1239	0.1287	0.1325	0.1353	0.1395	0.1433	0.1472	0.1504	0.1553	0.1602	0.1653	0.1712	0.1763	0.1797	0.1697	0.1618	0.1736	0.1858

TABLE III

COMPARISON WITH REPRESENTATIVE CNN AND TRANSFORMER BASELINES ON CIFAR-100 AND TINYIMAGENET. RESULTS ARE TAKEN FROM [26], AND MODELS MARKED WITH † ARE IMPLEMENTED FROM SCRATCH.

Model	**GFLOPs**	**CIFAR-100**	**TinyImageNet**
ResNet-50 [27]	0.7	71.03	56.02
EfficientNet-B3 [28]	0.2	68.77	48.85
MobileNetV2-1.4 [29]	0.1	69.92	54.12
CvT-21 [30]	1.7	63.32	49.97
HSViT-C3A4 [26]	1.3	72.46	56.73
CaiT † [31]	0.6	72.44	57.59
PVT-S † [32]	0.5	61.52	49.26
ToMe † [33]	0.4	60.25	44.20
EViT † [13]	0.2	65.94	52.24
DeepCompress-ViT † [18]	–	58.39	65.02
TinyNeXt † [19]	–	49.45	43.34
Ours ConvRoute-ViT	0.1	75.42	58.56

discriminative tokens while still reducing a substantial amount of computation from less informative regions.

E. Confidence Analysis

Figure 3 compares the top-3 confidence of the predicted classes of the base ViT and our proposed ConvRoute-ViT across two distinct classes from the CIFAR100 dataset [24]. For the first example *cup*, ConvRoute-ViT assigns the highest confidence correctly to *the cup* with a score of 0.8621, whereas the base ViT spreads its confidence over the incorrect classes such as *chair* (0.1331), *cockroach* (0.0967), and *dinosaur* (0.0437). The second example, ConvRoute-ViT, correctly predicts *turtle* (0.6440), whereas the baseline favors incorrect marine categories, assigning scores to *whale* (0.2010), *shark* (0.1823), and *dolphin* (0.1470). Thus, the confidence scores show that ConvRoute-ViT produces more accurate and better concentrated class probabilities by preserving informative object-centric tokens and reducing the effect of misleading background or semantically irrelevant tokens.

F. Analysis of the Token Reduction

Figure 4 illustrates how the proposed ConvRoute mechanism progressively reduces the token sequence across multiple routing stages. For this visualization, we set keep rates to $\rho = 0.9$ at the 3^{rd} layer and $\rho = 0.7$ at the 5^{th} layer. In each row, the first panel shows the original input image. The second panel highlights the less informative tokens routed out at the 3^{rd} layer. The third panel shows the further tokens, which are reduced at the 5^{th} layer, with both sets of removed tokens depicted as red patches. The final image presents the cumulative reduction of patch tokens after the routing layers. These visualizations show that ConvRoute-ViT reduces tokens that are mainly concentrated in less informative background



Fig. 4. Four representative CIFAR-100 examples illustrating the proposed ConvRoute mechanism. From left to right: input image, tokens removed at layer k, additional tokens removed at layer l, and the final compact token representation, which is further processed and passed for final classification.

regions, while the visually important object regions are significantly preserved. Such token reduction enables the model to reduce redundancy while retaining the most discriminative and important regions for efficient transformer processing.

G. Results Comparison

The results in Table III demonstrate that ConvRoute-ViT achieves the best overall performance among the compared CNN and Transformer baselines while maintaining only 0.1 GFLOPs. It achieves 75.42% accuracy on CIFAR-100 and 58.56% on TinyImageNet, outperforming all competing methods by margins ranging from +2.96% to +25.97% on CIFAR-100 and +0.17% to +15.22% on TinyImageNet. In particular, ConvRoute-ViT surpasses recent efficient Transformer models such as CaiT [31], PVT-S [32], ToMe [33], DeepCompress-ViT [18], and TinyNeXt [19] while requiring only 0.1 GFLOPs. These results demonstrate that ConvRoute-ViT effectively balances computational efficiency and recognition performance by reducing redundant tokens while preserving informative contextual representations. To further evaluate deployment efficiency, we evaluated the best ConvRoute-ViT configuration on CIFAR-100 against the corresponding base ViT on a single RTX 4090 GPU. ConvRoute-ViT reduced the total inference time from 0.7653 s to 0.4877 s, decreased the latency from 0.077 to 0.049 ms/image, increased the throughput from 13,066.55 to 20,506.16 images/s, and lowered the peak GPU memory usage from 399.09 MB to 357.22 MB improving both classification accuracy and deployment efficiency.

IV. CONCLUSION

In this paper, we presented ConvRoute-ViT, an efficient vision transformer framework that addresses the redundancy of uniform token processing through a convolution token routing module. By separating the token sequence into informative and compact convolutional summary tokens, our proposed mechanism reduces the sequence length in transformer layers without completely dropping the residual context information

in the less informative token patches. The overall findings show that effective token reduction should not only depend on removing tokens but should also preserve the supporting visual features that remain useful for classification. The quantitative results, along with the attention, confidence, and token reduction analyses, demonstrate that our ConvRoute-ViT learns more focused and compact representations. Overall, our proposed framework demonstrates that convolution-guided token summarization provides an effective direction for efficient vision transformer design.

Our ConvRoute-ViT preserves the discriminative CLS-attended tokens and compresses the less informative tokens into the condensed ConvSummary representations, which can improve the recognition performance by reducing the token redundancy without discarding useful contextual information. This hierarchical routing strategy enables the deeper transformer layers to process shorter but semantically richer sequences of tokens, leading to higher accuracy and computational efficiency. However, the evaluation is currently performed only on CIFAR-100 and TinyImageNet. In future work we plan to study scalability of ConvRoute-ViT on larger datasets (e.g. ImageNet-1K); conduct statistical analysis over multiple runs; perform extensive ablation studies on ConvSummary tokens; explore deployment on resource-constrained edge devices; and extend the framework to video transformers for efficient spatio-temporal token routing.

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] H. T. et al., "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021.

[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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[4] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[5] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[6] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, 2021.

[7] H. Yin *et al.*, "A-vit: Adaptive tokens for efficient vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun, "Evo-vit: Slow-fast token evolution for dynamic vision transformer," in *Proceedings of the AAAI conference on artificial intelligence*, 2022.

[9] Z. Kong, P. Dong, X. Ma, X. Meng, W. Niu, M. Sun, X. Shen, G. Yuan, B. Ren, H. Tang *et al.*, "Spvit: Enabling faster vision transformers via latency-aware soft token pruning," in *European conference on computer vision*, 2022.

[10] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsiavash, and J. Gall, "Adaptive token sampling for efficient vision transformers," in *European conference on computer vision*, 2022.

[11] B. Pan, R. Panda, Y. Jiang, Z. Wang, R. Feris, and A. Oliva, "Ia-red^2: Interpretability-aware redundancy reduction for vision transformers," *Advances in neural information processing systems*, 2021.

[12] P. Dong, M. Sun, A. Lu, Y. Xie, K. Liu, Z. Kong, X. Meng, Z. Li, X. Lin, Z. Fang *et al.*, "Heatvit: Hardware-efficient adaptive token pruning for vision transformers," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2023.

[13] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, and P. Xie, "Not all patches are what you need: Expediting vision transformers via token reorganizations," *arXiv preprint arXiv:2202.07800*, 2022.

[14] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," in *The Eleventh International Conference on Learning Representations*, 2023.

[15] S. Wei, T. Ye, S. Zhang, Y. Tang, and J. Liang, "Joint token pruning and squeezing towards more aggressive compression of vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[16] M. S. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: What can 8 learned tokens do for images and videos?" *arXiv preprint arXiv:2106.11297*, 2021.

[17] X. Wu, F. Zeng, X. Wang, and X. Chen, "Ppt: Token pruning and pooling for efficient vision transformers," *arXiv preprint arXiv:2310.01812*, 2023.

[18] S. Ahmed, A. Al Arafat, D. Najafi, A. Mahmood, M. N. Rizve, M. Al Nahian, R. Zhou, S. Angizi, and A. S. Rakin, "Deepcompress-vit: Rethinking model compression to enhance efficiency of vision transformers at the edge."

[19] F. Zeng, H. Li, J. Guan, R. Fan, T. Wu, X. Wang, and R. Lai, "An efficient hybrid vision transformer for tinyml applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 19 914–19 924.

[20] X. Xu, S. Wang, Y. Chen, Y. Zheng, Z. Wei, and J. Liu, "Gtp-vit: Efficient vision transformers via graph-based token propagation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[21] S. Lee, J. Choi, and H. J. Kim, "Multi-criteria token fusion with one-step-ahead attention for efficient vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[22] R. Shao, X.-J. Bi, and Z. Chen, "Hybrid vit-cnn network for fine-grained image classification," *IEEE Signal Processing Letters*, 2024.

[23] Y. Xie, A. Liu, X. Lu, and D. Chong, "Hybrid multi-class token vision transformer convolutional network for doa estimation," *IEEE Signal Processing Letters*, 2025.

[24] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Tech. Rep., 2009.

[25] J. Wu, Q. Zhang, and G. Xu, "Tiny ImageNet Challenge," Stanford University, Tech. Rep., 2017.

[26] C. Xu, C.-T. Li, C. P. Lim, and D. Creighton, "Hsvit: Horizontally scalable vision transformer," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[28] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[30] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[31] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[32] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[33] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," in *The Eleventh International Conference on Learning Representations*, 2023.